# COLOUR STIMULUS AND COLOUR SENSATION

by P. J. BOUMA and A. A. KRUITHOF.

535.65:535.733

The difference between the two conceptions "colour sensation" and "colour stimulus" is explained. Colour sensation, besides being dependent upon the spectral distribution of the light striking our eye from a coloured object, is strongly influenced by various kinds of secondary circumstances, such as the colour of the surroundings of the object observed and the condition of our eyes. Characteristic features of a colour sensation are its hue, saturation and brightness. A colour stimulus is the subject of trichromatic measurements, by methods which, thanks to standardisation, practically exclude such secondary circumstances. The results of the measurement of a colour stimulus may be expressed by its dominating wavelength, colorimetric purity and luminosity. Two colour stimuli are compared by examining them side by side (simultaneous comparison), whereas two colour sensations must be compared successively, for it is only then that the influence of the surroundings upon the state of chromatic adaptation of the eye becomes fully manifest. Owing to chromatic adaptation the variations in colour sensation accompanying the changing over from one kind of light to another are much smaller than the differences in colour stimuli brought about by the two kinds of light. This phenomenon greatly affects the impressions we get from our surroundings.

## Introduction

It is a common experience that a white tablecloth gives a "white" impression not only in daylight but also at night under the light of incandescent lamps. If this is considered more closely it is found to be very remarkable. In fact if an experiment is made where one part of the tablecloth is lighted by lamplight and another part by daylight — an experiment which anyone can improvise in the living room — the artificially lighted part will not appear to be white at all, but distinctly yellowish compared with that part under daylight. The spectral composition of the light striking the eye from the lamplighted part of the cloth (when it evokes a yellowish sensation) is, however, the same as that of the light striking the eye from the same cloth under the same lamp at night (when it is seen as white).

If the same experiment is made with coloured objects they also show unexpected differences in colour. A violet dress, for instance, will appear to be pinkish where it is under the lamplight, as compared with the part that is under daylight. In itself this observation is not very surprising, for the dress reflects light of different spectral distributions, corresponding to different colours according to the spectral composition of the two kinds of light falling upon it. The remarkable fact, however, is that if the room is lighted exclusively with electric light the colour of the dress is not pink at all, but violet.

These differences in colour only occur in everyday life when conditions are such that we have in the room daylight and electric light of about equal intensities but striking the objects from different sides. One then speaks of "false light". We then observe much greater differences in colour (*i.a.* coloured shadows) than the differences in shade which we are accustomed to observe when daylight is replaced by artificial light. That is why these differences in colour are often unexpected and disagreeable.

In order to explain the contrast between our observations in the experiments previously mentioned and our experience in everyday life, where no great differences in colour are noticed, we have to distinguish between the conceptions "colour stimulus" and "colour sensation". A colour stimulus is the subject of technical colour measurements. It is entirely different from the colour sensation, which is sometimes indicated in the following by the word colour, as applying to the subjective sensation of the observer. The fact that a given colour stimulus may give rise to different colour sensations is ascribed to a change in the retina under the influence of the light. This phenomenon is termed "chromatic adaptation" and it greatly affects the impressions we obtain from our surroundings, as is apparent from the foregoing. The object of the present article is to formulate clearly the difference between the conceptions "colour stimulus" and "colour sensation", preparatory to dealing with chromatic adaptation in a following article.

## Difference between colour sensation and colour stimulus

In the following we shall explain the difference between the conceptions "colour sensation" and "colour stimulus". To that end the characteristic features of the two conceptions will be enumerated in succession, each one first for the sensation and then for the stimulus.

*Colour sensation:* The colour sensation we get from an object in our surroundings depends on the following three groups of causes:

a) The spectral composition of the light that the coloured object throws upon the eye.

b) The "normal" laws of additive colour mixing of the eye, *i.e.* the laws governing the results of additive mixing of coloured light for the normal eye undcr standardised conditions.

c) All sorts of incidental circumstances affecting the state of our organs of sight at the moment.

*Colour stimulus:* The result of a technical measurement of colour is called colour stimulus, which only depends upon the factors a) and b). Methods of technical colorimetry have already been discussed in this periodical [1]).

---

*Colour sensation:* The circumstances c) affecting the colour of a beam of light of given physical properties come under the following headings:

1) Characteristics of the eye of the individual observer.

2) Properties of the objects viewed which evoke psychical influences (*e.g.* memory).

3) The state of the retina, which is affected by other light impinging upon other parts of the retina while the beam from the object is under view, or by such other light as may have just previously reached it.

*Colour stimulus:* In colorimetry the effect of these circumstances is precluded by various specific measures, such as:

1) The observer must have normal colour vision and the measurements must be performed with sufficiently high brightness [2]).

2) Essentially the measurement consists in the judging of the possibility to distinguish the two halves of a circular spot of light, one half having the colour to be measured and the other half a reference colour.

3) The rest of the visual field is dark.

---

*Colour sensation:* It might be considered ideal to possess a complete set of specifications for predicting the nature of a colour sensation from the various physical conditions. As, however, a colour sensation is difficult to express in numerical terms and, moreover, depends upon so many circumstances, some of which are of a non-physical nature, such an ideal can never be fully realised.

*Colour stimulus:* Thanks to the simple normal laws of additive colour mixing of the eye it has been possible to draw up complete specifications for the measuring of colour stimuli, so that nowadays there are tables enabling one to calculate a colour stimulus from the results of purely physical measurements, without any recourse to visual judgement. Colour stimuli are defined by means of the colour coordinates [3]) X, Y and Z (C.I.E. 1931).

---

*Colour sensation:* The characteristic features of a colour sensation are:

1) **Hue**: the property of colour sensation causing us to give the colour a name, such as red, green and blue.

2) **Saturation**: the extent to which a colour sensation differs from "white", or to which the sensation is "coloured"; the property that causes us to speak of faded colours or of vivid colours.

3) **The impression of brightness**: the property that causes us to speak of light and dark colours.

*Colour stimulus:* Instead of using the coordinates $x$, $y$ and $z$ a colour stimulus may be given by:

1) The dominating wavelength $\lambda_d$.

2) The colorimetric purity $p$: a colour stimulus is reproduced by a mixture of the spectral colour $\lambda_d$ and "white", the ratio of brightness being $p : (1-p)$.

3) **The brightness**, arising from the sensation of brightness by standardisation of the measuring conditions.

[1]) J. P. Bouma: The Perception of Colour, Ph. Tech. Rev. 1, 283, 1936.

[2]) Below a certain level of brightness the characteristics of the normal eye depend upon the brightness (Purkinje effect). See, *e.g.*, P. J. Bouma: The definitions of brightness and their importance in road lighting and photometry, Ph. Tech. Rev. 1, 142, 1936.

[3]) See *e.g.* P. J. Bouma: The representation of colour sensations in a colour space diagram of colour triangle, Ph. Tech. Rev. 2, 39, 1937.

This procedure is similar to the manner in which the conception of colour stimulus arises from the colour sensation.

---

*Colour sensation:* When judging a colour sensation one usually considers each colour separately, and if comparison is needed another colour is considered afterwards, so that the comparison is successive.

*Colour stimulus:* In visual colorimetry, which is the foundation upon which the aforementioned specifications are drawn up for determining the colour coordinates pertaining to a given colour stimulus, always two coloured patches of light are examined simultaneously: the stimulus to be measured and the reference colour stimulus, so that in this case a simultaneous comparison is made.

**Colour sensations and colour stimuli due to different kinds of light**

Bearing in mind the differences enumerated above, it is easy to understand the difference between the determination of colour sensations and the measurement of colour stimuli. The phenomena occurring when white or coloured objects are illuminated with different kinds of light also become clear. We will now proceed, in the same way as before, to compare in more detail the behaviour of colour sensations and colour stimuli due to different kinds of light.

*Colour sensation:* The question in how far the same colour impressions are obtained from a given coloured object under two kinds of light of different spectral composition is of importance when the whole of our surroundings is illuminated first with one of the sources of light and then with the other.

*Colour stimulus:* The question in how far the same colour stimulus is obtained from an object in our surroundings under two different kinds of light is particularly of importance when one desires to use both sources of light simultaneously.

---

*Colour sensation:* Changes occurring as a consequence of the transition from one kind of light to another may consist, *i.a.*, in a change in hue of some objects, or in a colour becoming more prominent owing to stronger sensations of brilliancy or of saturation. Such changes are quite marked when for instance incandescent light is replaced by mercury light.

*Colour stimulus:* Differences between two kinds of light present at the same time manifest themselves, *i.a.*, in the occurrence of the unpleasant phenomena of false light and coloured shadows referred to in the introduction. These phenomena are observed when an electric lamp is switched on in a room already illuminated by daylight.

---

*Colour sensation:* The equivalence of two kinds of light in respect to the colour sensations they evoke can as yet only be judged experimentally.

*Colour stimulus:* For the examination of the equivalence of two kinds of light as regards the colour stimuli they evoke for a given object there are, among others, two methods previously described in this periodical [4]); the first is more of an experimental nature and the second more theoretical.

---

*Colour sensation:* In essence the experiment consists in the examination of a large number of coloured cards against a white background first under one kind of light and then under the other. The comparison of the colour sensation obtained, as far as hue is concerned, is relatively simple because in this respect our impression can be adequately expressed in words ("I call this colour impression yellowish-green", etc). On the other hand it is much more difficult to compare the other two characteristics of the colour sensations (brightness and saturation). Therefore, the experiments to be described in a subsequent article will only have reference to hue denomination.

It is to be noted that here we have a successive comparison, where, after the change from one source of light to the other, the surroundings also throw a different light upon the eye and fully exercise their influence. This fits with the fact that we are dealing with colour sensations.

[4]) P. J. Bouma: Colour reproduction in the use of different sources of "white" light, Ph. Techn. Rev. 2, 1, 1937.

*Colour stimulus:*

In the experimental method a number of coloured cards are held against a black background and one half of each is exposed to one of the kinds of light while at the same time the other half is exposed to the other light. The difference in colour stimulus (also with regard to luminosity and colorimetric purity) showing itself in the simultaneous examination of the two halves is a measure for the difference between the two sources of light.

Here we have a simultaneous comparison and the conditions of measurement are approximately those of colorimetry. This fits with the fact that we are dealing with differences in colour stimuli.

*Colour sensation:*

A more theoretical method of comparison based on the spectral composition curves of the sources of light can only be developed if the influence of the surroundings appears to be subject to definite laws. Then the influence can be accounted for and predictions can be given as to the changes that will take place in the colour sensations evoked by a coloured object when changing over from one kind of light to another. This, too, will be gone into more fully in a subsequent article.

*Colour stimulus:*

The second, more theoretical, method of comparison already mentioned [5]) is based on the spectral composition curves of the sources of light. The spectrum is judiciously divided into 8 sections. It may then be said that the two sources of light are to be regarded as being practically equivalent if the relative contributions of each of them to the luminous flux in every section is the same within certain tolerances. Where there is a noticeable difference between the sources of light the degree of that difference can be judged by calculating the differences in trichromatic specifications that will occur for the coloured object when changing over from one of the kinds of light to the other.

The examples quoted in the introduction have already given some idea how great the differences may be under two different kinds of light, and this can be further illustrated in the following way. The colour points of the reflected light for 24 cards of a "circle" of an Ostwald colour atlas are calculated when exposed successively to daylight and to the light from an incandescent lamp [6]). We selected for this purpose a circle of rather saturated colours, the *nc* circle, consisting of 100 cards in all, of which the saturation impression is fairly constant. The hue of the cards numbered 0 to 99 extends from yellowish-green to yellow, orange, red, purple, blue, green and back to yellowish green [7]). The results of the calculations made are given in *fig. 1*.

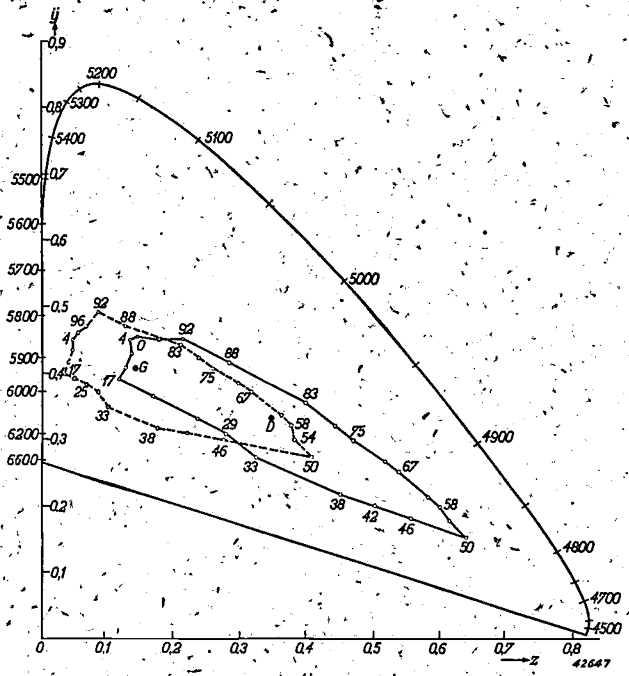


Fig. 1. Position of the colour points for the Ostwald cards *nc* O, 4, 8, 13, . . . . in the international colour triangle. Broken line: under the light from an electric incandescent lamp. Full line: under daylight. *D* is the colour point for daylight, *C* for incandescent lamplight, *y* and *z* are colour coordinates of the C.I.E. The numbers from 4500 to 6600 indicate the wavelength in Ångstrom units on the curvc of the spectral colours.

*Colour sensation:*

When, for instance, card 83 is first examined under daylight and then under electric light, the colour appears to be practically the same: under both kinds of light this card gives a purely green sensation.

*Colour stimulus:*

Fig. 1 shows, however, that the two colour points 83 are far apart: the colour stimuli are different. This difference in colour stimulus is obvious when one half of the card is exposed to daylight and the other

[5]) See also P. M. van Alphen: A photometer for the investigation of the colour rendering reproduction of various light sources, Ph. Techn. Rev. 4, 66, 1939.

[6]) For the method of calculations see Philips Techn. Rev. 2, 45, 1937.

[7]) In the new editions of the Ostwald atlas the circle has only the 24 cards used here.

half to electric light and both are examined simultaneously.

---

*Colour sensation:* If one examines in succession card 47 under electric light and card 31 under daylight a large difference is noticed, the former showing a violet hue and the latter crimson.

*Colour stimulus:* Fig. 1 shows that the two colour stimuli, 47 under electric light and 31 under daylight, are practically the same. This is confirmed by a test similar to that with the two halves, holding side by side card 47 under electric light and card 31 under daylight (simultaneous comparison).

The results of these experiments are to be explained by the fact that when comparing colour sensations we find the results are influenced by the activity of an additional factor, *viz.* a factor classified under c) in the beginning of this article. The experiments show that this factor may contribute towards making two equal colour stimuli appear to be different in the sensations they evoke (the last example), but that on the other hand, under certain circumstances, it may neutralise a difference in colour stimuli in so far as the colour sensation is concerned (first example). In these cases this factor is mainly to be sought in the surroundings.

*Colour sensation:* The surroundings usually have a considerable influence upon colours, as is evident from the last mentioned example where in different surroundings one and the same colour stimulus gives an entirely different sensation. This influence lies mainly in the fact that the retina gets a different sensitivity for the various colours. This phenomenon is called chromatic adaptation of the eye.

*Colour stimulus:* The surroundings do not affect the colour stimuli. Though the specifications for colour measuring prescribe entirely dark surroundings, any non-dark surroundings have practically no influence upon the result, because the two coloured patches to be compared are always shown simultaneously and immediately adjacent to each other; consequently the two parts of the retina on which the images of the coloured patches are formed are always adapted in approximately the same way.

The great influence of the surroundings upon colours can be further demonstrated in the following way (see *fig. 2*). A transparent window 15 × 15 cm can be illuminated at the back with a number of differently coloured lamps. Around the window is a field of 80 × 80 cm which may emit incandescent lamplight or artificial daylight as desired, without affecting the light passing through the window.

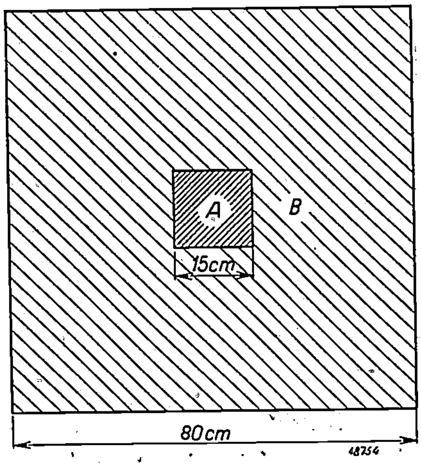


Fig. 2. The chromatic adaptation of the eye is very well demonstrated by illuminating a transparent window *A* with a light of a certain colour and the surrounding field *B* first with artificial daylight, for instance, and then with incandescent lamplight. Shortly after changing over from daylight to lamplight the eye directed upon *A* sees a change of colour in the window.

If the whole set-up is viewed from a few yards away and the light in the surrounding field is changed from daylight to incandescent lamplight then in most cases after a while a decided change of colour seems to take place in the small lighted window.

*Colour sensation:* The colour sensation may change from say orange-yellow to yellowish-green, or from crimson to bluish-purple. In a few cases, however, the colour does not change at all (in the cases of blue and orange).

*Colour stimulus:* The colour stimulus that can be measured at the window does not, of course, change because the lamps are still the same. Thus we see here a change in colour sensation due to the influence of the surroundings in its unqualified form [8]).

In the example of card 83 we saw that it may happen that under certain circumstances a change in colour stimulus does not bring about a change in colour sensation: the process of chromatic adaptation

[8]) The fact that the change in colour is not noticed until some time after the surrounding light is changed indicates that the eye requires some time to adapt itself to the changed surroundings.

of the eye compensates a difference in stimulus arising in the beginning. This phenomenon occurs in practice (consider the examples in the introduction) so frequently that it has been formulated in a "law", which for a number of years already has been widely adopted by psychologists and physiologists, *viz*:

*Colour sensation:* "The colour sensations created by the coloured objects in our surroundings are practically independent of the kind of light with which the whole scene is illuminated".

*Colour stimulus:* At the same time, however, the colour stimuli may differ appreciably, as illustrated in fig. 1.

The aforementioned rule of the unchangeability of colour sensations, however, holds only for kinds of light having a spectral composition differing little from that of black body radiators, and even so it is only an approximative rule. Small deviations can easily be observed experimentally. For instance, pigments having a different spectral remission curve may have the same colour when exposed to one certain kind of light but they are certainly different after the light has been changed.

We therefore prefer to formulate the rule as follows: *With most kinds of light commonly used the difference between the colour stimuli going out from a given pigment under two different kinds of light is usually much greater than the difference between the colour sensations evoked by those colour stimuli.*

For example, when comparing the light from an incandescent lamp with daylight:

*Colour sensation:* Everyone knows from experience that there is very little difference in the colours of the same object when illuminated by these two sources of light.

*Colour stimulus:* From fig. 1, or when applying the "8-sections method" or "the comparative method of two halves" (see the foregoing), the conclusion to be drawn is that the difference in colour stimuli for a given card under different light may in fact be surprisingly great.

Here we may conclude the comparison of the two conceptions, colour stimulus and colour sensation. It has already been pointed out that much less is known about colour-sensations than about colour stimuli. In particular very few investigations have as yet been made in regard to the determination of colour sensations produced under different circumstances. It is hoped that the experiments and observations to be dealt with in the next article will contribute towards our knowledge of this subject.

# A REMARKABLE PHENOMENON WITH STEREOPHONIC SOUND REPRODUCTION

by K. de BOER.

To a practised listener the sound image heard with stereophonic reproduction generally appears to lie above (and at times below) a line between the loudspeakers, in other words it has elevation. This remarkable phenomenon can be explained in every detail, even quantitatively, by the familiar hypothesis that the impression of elevation of sound above the horizontal plane of its source is due to slight movements of the head. For stereophonic reproduction this phenomenon is of no consequence in practice.

When a system for stereophonic sound reproduction is installed, such as already described in this journal, [1]) two loudspeakers are set up on either side of a platform or screen, each with its own channel (amplifier, etc.) from a separate microphone set up in the recording room (*fig. 1*). For cinema reproduction the sound striking each of the two microphones in the film studio is first recorded on a sound track on the film and each of the loudspeakers reproduces the sound from the corresponding sound track. This case, too, can be represented diagrammatically by fig. 1.

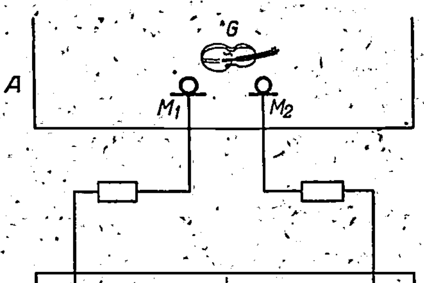


Fig. 1. Schematic representation of a set-up for stereophonic reproduction. Two loudspeakers $L_1$ and $L_2$ set up in the reproducing space $B$ are each connected to a corresponding microphone $M_1$ and $M_2$ in the recording space $A$. Here, for the sake of clearness, the microphones are shown separated, but in reality they are usually contained in a dummy head. The listener $W$ observes a virtual picture $G^1$ of the source of sound $G$. The recording space may be a film studio and the reproduction space the auditorium of a cinema, in which case the "channel" between each microphone and its corresponding loudspeaker comprises, *inter alia*, a sound track in which the proportions of sound are preserved.

Let us suppose that a listener is standing in the middle of a hall where the loudspeakers are set up. Particularly when the spoken word is being reproduced, a "sound image" can very well be localised, that is to say the sound is heard to come from a certain spot between the two loudspeakers, depending on the position of the source of the sound in the recording room. For the present we will confine our attention to the case where the sound image appears to lie just halfway between the loudspeakers.

With the help of a number of trained listeners the astonishing fact has been discovered that as a rule the sound image is not situated on the line running direct from one loudspeaker to the other, as one would expect, but a distance above it [2]): it has a certain elevation above the horizontal plane. Upon walking down the middle of the hall towards the loudspeakers the listener hears the sound coming from a higher level, rising gradually at first but more quickly as he gets nearer, until when he has reached the centre of the line between the loudspeakers the sound is practically perpendicular above him.

Still more astonishing is the sensation when the listener raises his head to try to "see" whence the sound is coming, as one does when there is some source of sound with elevation (*e.g.* an aeroplane). The sound image seems to climb higher and higher, so that one cannot catch it in the eye, as it were.

Some observers find that when they concentrate their minds and keep looking straight ahead the sound does not come from above but rather from below the horizontal plane, while as they move closer forward it goes deeper and deeper, finally disappearing through the floor, vertically underneath. To some this downward effect comes more readily than the upward movement.

In practice, with stereophonic reproduction in a concert hall or cinema such an effect has no

[1]) K. de Boer, Stereophonic Sound Reproduction, Philips Techn. Rev. 5, 107, 1940.

[2]) K. de Boer, A remarkable phenomenon in direction of hearing, Ned. T. Natuurk. 11, 75, 1944.

adverse consequences worth mentioning. From most seats in the auditorium the distance from the loudspeakers is so great that the elevation of the sound can only be small; something further will be said about this in the following pages. Moreover, on the one hand most of the seats are not in the middle of the hall, while on the other hand the sound image lies (or the images lie) closer either to the right-hand loudspeaker or to the left one, both tending to reduce the elevation observed. Furthermore, this is by no means an effect that strikes everyone. In fact it takes some practice to notice it at all, as is understandable considering that in general it is much more difficult to observe and determine the elevation of a sound than the azimuthal angle. As a consequence the observation of the elevation of a sound image is more susceptible to suggestion. In the case, for example, of stereophonic reproduction of the sound recorded of a passing aeroplane, all listeners in different places seemed to hear the sound directly overhead. It is just by reason of this susceptibility to suggestion that where stereophonic reproduction of sound accompanies picture projcction (in a cinema) the sound image is always attracted by the visual picture towards the horizontal plane, even if a listener is sufficiently trained to observe an elevation effect. A similar suggestive influence is present in the case of reproduction of a musical performance, for then the listeners picture to themselves the places usually occupied by the various members of the orchestra.

Although this effect has no practical consequences, it is instructive to discuss it and to show how the phenomenon can be explained both qualitatively and quantitatively. It is to be remarked that it can be observed at home, for instance when in order to improve the quality of one's radio set ("smoothing out" the sound) one has two loudspeakers set up some distance apart [3]).

### Description of the experiments and results of measurements

In order to investigate the effect described and in particular to measure the apparent elevation observed, an arrangement was set up as shown in *fig. 2*, with two loudspeakers at the level of the listener's ears and halfway between them a vertical measuring rod. The listener is placed in the plane of symmetry with the set-up and told to look straight ahead of him, at a mark on the measuring rod. When there are small angles of elevation of the sound image he can read their position directly on the rod, without lifting his head (this he must not do, because then the image moves higher up). Because large angles of elevation cannot be read directly from the vertical measuring rod, another was set up horizontally, along which an accessory loudspeaker could be moved (fig. 2). When this loudspeaker is in a suitable position the sound from the apparent source to be localised seems to come from the same direction as a momentary signal emitted at intervals from this loudspeaker. In this way the direction of the sound image is determined by a "zero method"; the observer has only to judge whether there is any difference in direction in respect to a comparativc signal.

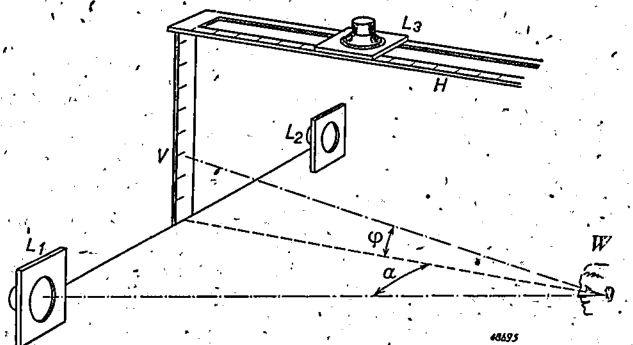


Fig. 2. Set-up for measuring the elevation effect observed. $V$ is a vertical measuring rod on which the angle $\varphi$ for small elevations can be read directly by the listener $W$. $H$ is a horizontal rod with auxiliary loudspeaker $L_3$ that can be moved along it for determining larger elevations.

The results of the measurements are reproduced in *fig. 3*. They were obtained for the greater part with little trained listeners. With the difficulty already mentioned in this kind of localisation the observations were consequently not very consistent and showed considerable variations. Nevertheless, the diagram shows clearly the qualitative details of the phenomenon. The angle of elevation $\varphi$ of the sound image observed is plotted for different distances from the listener to the pair of loudspeakers, measured through the angle $a$ between the plane of symmetry and the line from the loudspeaker to the listener (figs. 1 and 2). It is seen that as the angle $a$ becomes larger and thus the listener gets nearer the sound image rises first slowly and then more quickly. At $a = 90°$, that is to say when the observer is just between the loudspeakers, the

[3]) When two loudspeakers are used, either connected to the same radio set or placed for stereophonic reproduction, in order to get a natural impression with a well defined sound image it is necessary that the two loudspeakers should be in phase with each other, that is to say the diaphragms must both move together towards and away from the listener. One loudspeaker is put in phase or counter-phase with the other simply by turning it or changing the polarity of its connection.

sound does not come from exactly overhead but from a short distance behind ($\varphi \approx 100°$).

## Explanation of the phenomenon

Since the phenomenon described consists of an elevation of an apparent source of sound, obviously its explanation must be related to the theory accounting for the perception of the elevation of an actual source of sound.

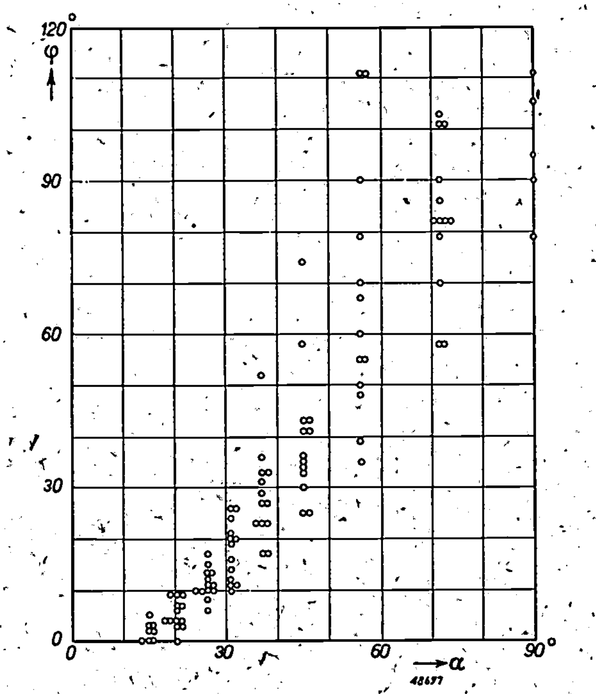


Fig. 3. Measured values of the angle of elevation $\varphi$ observed by a number of (little trained) listeners, as a function of the distance between the listener and the pair of loudspeakers (measured through the angle $a$, see figs. 1 and 2).

This theory has already been dealt with *in extenso* in this journal [4]). Briefly it comes to this, that a perception of direction is due to a difference in intensity between the two ears [5]). Where there is no such difference, *i.e.* where the intensity ratio $v = 1$, this corresponds to the perception of direction "straight ahead". Equally so, however, $v=1$ happens with a source of sound "immediately overhead" or "immediately behind", or in general with any elevation of the source of sound in the plane of symmetry with the observer (*fig. 4*). A criterion for distinguishing these cases and "determining" the angle of elevation is obtained by the listener making small more or less arbitrary turns of his head around a verticale axis (as shaking the head for "no"). When giving the head a turn $\mathrm{d}a$ the ratio $v$ of the intensity between the ears will change by an amount $\mathrm{d}v$ approximately proportional to $\mathrm{d}a$. In particular, for instance when the source of sound is "straight ahead", when the plane through that source of sound and the aural axis is horizontal, $\mathrm{d}v = 3\%$ (0.14 db) for $\mathrm{d}a = 1°$. If, however, the source of sound has an elevation $\varphi$, then for a given turn $\mathrm{d}a$ the effective turn $\mathrm{d}a'$ in the plane through the source of sound and the aural axis (fig. 4) is smaller, *viz.*

$$\mathrm{d}a' = \mathrm{d}a \cdot \cos\varphi .$$

Also the change $\mathrm{d}v'$ in the intensity ratio is proportionately smaller for a turn $\mathrm{d}a$. Now the fact that for a given $\mathrm{d}a$ a too small $\mathrm{d}v'$ occurs is interpreted by the sense of hearing (due to its experience) as an elevation of the source of the sound, and it does this quantitatively, according to the equation [6]):

$$\cos\varphi = \frac{\mathrm{d}a'}{\mathrm{d}a} = \frac{\mathrm{d}v'}{\mathrm{d}v} \quad . \quad . \quad . \quad . \quad . \quad (1)$$

The elevation thus "measured" by the ear is in reality not the angle in respect to the horizontal plane but one in respect to a fixed plane of reference connected with the head, which with the head in the normal position is horizontal. This is of importance in the event that the head is lifted, for the

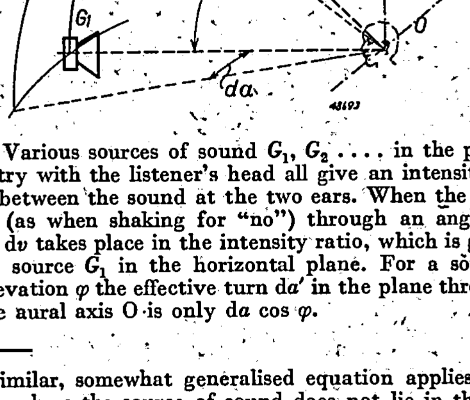


Fig. 4. Various sources of sound $G_1$, $G_2$ . . . . in the plane of symmetry with the listener's head all give an intensity ratio $v = 1$ between the sound at the two ears. When the head is turned (as when shaking for "no") through an angle $\mathrm{d}a$ a change $\mathrm{d}v$ takes place in the intensity ratio, which is greatest for the source $G_1$ in the horizontal plane. For a source $G_2$ with elevation $\varphi$ the effective turn $\mathrm{d}a'$ in the plane through $G_2$ and the aural axis O is only $\mathrm{d}a \cos\varphi$.

[4]) K. de Boer and A. Th. van Urk: Some facts about the direction of hearing, Philips Techn. Rev. 6, 363, 1941.

[5]) Accompanying differences in time also contribute to a perception of direction (*cf.* the article quoted in footnote [1])), but for very small angles $a$ they are of such little influence that we may here confine our considerations to the differences in intensity.

[6]) A similar, somewhat generalised equation applies to the case where the source of sound does not lie in the plane of symmetry of the observer. when the intensity ratio $v$ differs from unity already in the state of rest. For the sake of simplicity, however, we confine ourselves to sources of sound in the plane of symmetry.

plane of reference moves with it. When the listener gets the source of sound in line with the eye it is then again in the plane of reference and he therefore expects his sense of hearing to register the elevation zero.

So much for the explanation of the perception of the elevation of an actual source of sound. We will now go back to our set-up for stereophonic reproduction, fig. 1. There we have an apparent source of sound observed straight ahead as the result of the joint action of two loudspeakers, one on the right and one on the left. We can again put the question what change will take place in the intensity ratio at the ears when the listener turns his head an angle $d\alpha$ from the vertical axis. *It is found that this change is in fact smaller than what would be expected with an actual source of sound straight ahead:*

The fact that the $dv'$ will be smaller than for an actual source of sound may be seen from the following:

We will use $I_1$ and $I_2$ to denote the sound intensities from one loudspeaker at the ear closest to it and at the other ear respectively, when the loudspeaker is placed at an angle $\alpha$ from the plane of symmetry (*fig.* 5*a*). When the listener looks straight ahead each ear receives the total intensity $I_1 + I_2$. Upon the head being turned an angle $d\alpha$ to the right the angle of deviation of the loudspeaker $L$ is increased to $\alpha + d\alpha$ while that of the loudspeaker $R$ is reduced to $\alpha - d\alpha$. The left ear then receives (fig. 5*b*) the quantities of sound

$$I_1 + \frac{dI_1}{d\alpha}\,d\alpha \text{ from } L, \text{ and } I_2 - \frac{dI_2}{d\alpha}\,d\alpha \text{ from } R.$$

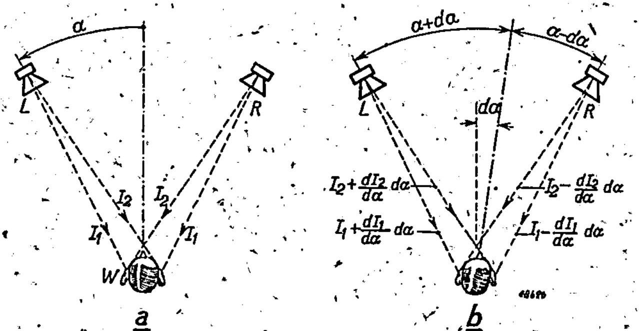
The same applies for the right ear. The intensity ratio at both ears, which was originally unity, is now:

$$1 + dv' = \frac{I_1 + \dfrac{dI_1}{d\alpha}\,d\alpha + I_2 - \dfrac{dI_2}{d\alpha}\,d\alpha}{I_1 - \dfrac{dI_1}{d\alpha}\,d\alpha + I_2 + \dfrac{dI_2}{d\alpha}\,d\alpha}$$

Thus $1 + dv'$

$$\approx 1 + \frac{2}{I_1 + I_2} \cdot \frac{d}{d\alpha}(I_1 - I_2)\,d\alpha.$$

$$\text{or } dv' = \frac{2}{I_1 + I_2} \cdot \frac{d\,(I_1 - I_2)}{d\alpha}\,d\alpha \quad . \; . \; . \quad (2)$$



a
L
R
$I_2$
$I_2$
$I_1$
$I_1$
W
a
α+dα
α−dα
dα
L
R
$I_2+\frac{dI_2}{d\alpha}d\alpha$
$I_2-\frac{dI_2}{d\alpha}d\alpha$
$I_1+\frac{dI_1}{d\alpha}d\alpha$
$I_1-\frac{dI_1}{d\alpha}d\alpha$
48498
b

Fig. 5. *a*) The loudspeaker $L$, placed at an angle $\alpha$ from the plane of symmetry of the head of the listener $W$, produces the sound proportions $I_1$ and $I_2$ respectively at the ear closest to it and at the other ear. The same is the case with the loudspeaker $R$ (the sound image is presumed to lie in the plane of symmetry). *b*) After a small turn of the head $d\alpha$ the left-hand loudspeaker is at an angle $\alpha + d\alpha$ from the plane of symmetry of the listener, while the right-hand one is at an angle $\alpha - d\alpha$. The respective sound proportions at each ear have been altered accordingly, to the values indicated in the diagram.

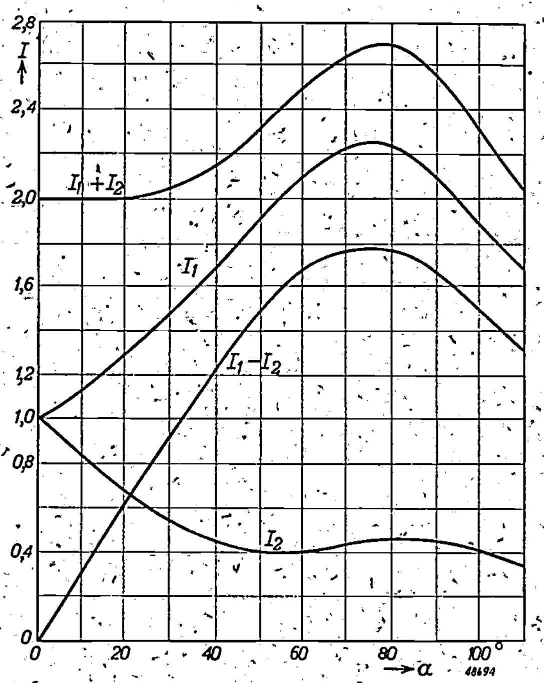


Fig. 6. The sound proportions from a loudspeaker at the closest ear ($I_1$) and at the other ($I_2$) as functions of the angle $\alpha$ through which the loudspeaker is turned away from the plane of symmetry of the listener (see the arrow at $\alpha$ in fig. 5*a*). The curves have been plotted from measurements taken by Sivian and White [7]). There are also plotted the quantities $I_1 + I_2$ and $I_1 - I_2$ used for calculating the elevation observed.

The functions $I_1(\alpha)$ and $I_2(\alpha)$ indicate simply the sound intensities from one single loudspeaker, say $L$ in fig. 5*a*, at the ear closest to it and at the other ear respectively when that loudspeaker is moved in respect to the listener's head in the manner indicated by the arrow in fig. 5*a*. This intensity as a function of $\alpha$ has been dertermined by Sivian and White [7]), whose results are reproduced in *fig.* 6 by the curves $I_1$ and $I_2$. This figure also gives the curves $I_1 + I_2$ and $I_1 - I_2$ as a function of $\alpha$. By this means it is possible to calculate the effective change $dv'$ in the intensity ratio according to (2) for any

[7]) J. Sivian and S. B. White, J. Acoust. Soc. Amer. 4, 288, 1933.

angle $a$ (corresponding to a certain distance between the listener and the pair of loudspeakers) and then to predict the "elevation" to be expected according to (1) by comparison with the "normal" change already given ($dv \approx 3\%$ for $da = 1$).

The fact that $dv' < dv$, thus that there will actually be a real angle of elevation ($\cos\varphi < 1$), may be deduced from fig. 6, considering that in the limit case $a = 0$, when the listener is removed far from the loudspeakers, the two loudspeakers together function as one actual source of sound straight ahead. The change $dv'$ calculated from (2) for this case will therefore be the same as that to which the ear is accustomed and used as criterion ($dv$) in normal hearing in such a situation as this. From fig. 6 it will now be seen that the curve $I_1$—$I_2$ (denominator of the fraction in (2)) lies lower for $a = 0$ than for any other value of $a$, whilst the slope of the curve $I_1$—$I_2$ (numerator of (2)) is greater for $a = 0$ than for any other value of $a$. The change observed ($dv'$) is thus certainly smaller for any angle $a$ than that for $a = 0$, *i.e.* the angle normally expected.

### Further proof of the explanation

The calculation as described of the elevation to be expected with the aid of the curves in fig. 6 and the equations (2) and (1) gives as a result the curve plotted in *fig.* 7, which shows also the elevations determined experimentally, *viz.* the mean values of the series of measurements reproduced in fig. 3. It will be seen that there is a fairly good agreement. In the range of the small angles the calculation is very inaccurate, due to the not inconsiderable influence of the limited accuracy of Sivian and White's measurements (which, moreover, depend somewhat upon the frequency spectrum of the sound). For large angles $a$ the calculation is more reliable and even leads in fact to a prediction of the peculiarity, already mentioned, that as the listener reaches the line connecting the two loudspeakers the sound comes from immediately behind him instead of directly overhead. It is certainly remarkable that the simple theory developed here should produce also this detail.

The fact that some listeners may observe a negative elevation instead of a positive one is not contradictory to the theory, because since a too small $dv'$ occurs just as well with a negative elevation as with a positive one the theory makes no differentiation between these two cases. Why, however, in some cases there happens to be a preference for the perception of the rather unusual negative elevation is difficult to account for.

There remains to be considered the impossibility of getting the apparent source of sound in the line of sight by lifting the head. This, too, is not difficult to comprehend. If, when listening to the real source of sound with an elevation, one gradually raises the head, the change $dv'$ in the intensity ratio at the ears becomes greater and greater when the head is moved to either side, owing to the elevation of the source above the plane of reference decreasing.

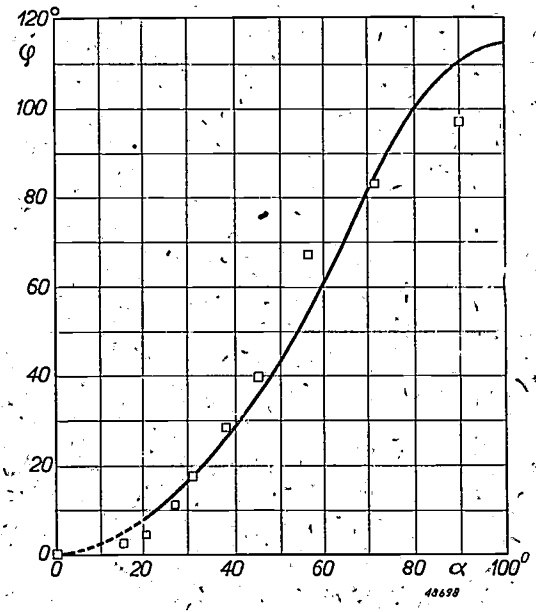


Fig. 7. Calculated elevation $\varphi$ of an apparent source of sound straight ahead, as function of the distance to the two loudspeakers (angle $a$). For very great distances (small angle $a$) the calculation is rather unreliable; the curve for this range is drawn in a broken line. The squares are the mean values of the series of measurements given in fig. 3.

When the plane of a reference passes through the source, $dv'$ becomes equal to the value that the sense of hearing regards as normal, thus no longer "too small", and the impression of an elevation disappears. In our experiments with an apparent source of sound the "too small" value of $dv'$ was caused by the joint action of the two loudspeakers. Upon the head being raised, while still looking straight ahead, the proportions of sound at the two ears, including their changes when shaking the head, remain approximately the same, so that $dv'$ continues to be "too small" and the impression of elevation remains. Consequently the sound image rises at the same time.

The more complicated phenomena occurring when the apparent sound image does not lie halfway between the loudspeakers or the listener is not in the plane of symmetry can also be explained in the manner described. We shall not enter into these cases here. We would only recall the fact already mentioned that in such cases the elevation is always smaller than in the case we have considered above. By this means and with the aid of fig. 7 we can give an indication of what significance this effect might have in stereophonic reproduction in practice, for instance in a cinema. Any perceptible elevation of about 20° will easily be corrected by the suggestion of the visual picture, so that no elevation will be noticed in the seats farther away from the screen than the place corresponding to the angle $a = 30°$ (see fig. 7). If the distance between the loudspeakers is say 8 meters the elevation effect will therefore only be noticeable for a trained listener at places closer than about 7 meters to the screen, and the number of such seats — which also in other respects are to be regarded as unfavourable — is only a small percentage of the total number of seats in a normal cinema.

# A NEW ELECTRICAL METHOD FOR DETERMINING MOISTURE CONTENT

by J. BOEKE.

543.812: 621.317.39

A description is given of a new method for determining the water content of liquid or solid substances, such as grain, wood, textiles, butter, etc. The water is extracted from the material with acetone in which oxalic acid is dissolved. The increase in the conductivity of the extraction liquid serves as a measure of the amount of water taken up. Compared with those already existing this method has the advantage that the apparatus required is inexpensive and simple in operation, while at the same time it is a very reasonably quick method and not dependent on the form in which the substance to be tested occurs.

The determination of the water content in gases, liquids and solids is one of the most important analyses regularly applied in commerce and industry. The moisture content of grain, of tobacco, of butter, to name only a few examples, is also paid for by the purchaser, and large profits or losses may be involved when a kilogram of the product contains a few grams more or less of water than were estimated. It is therefore understandable that numerous attempts have been made to develop methods of measuring moisture content quickly and accurately.

If we confine ourselves to solids and liquids, the only universal method used until now consisted in drying a sample and determining the amount of evaporated water by weighing. This, however, is often a lengthy and far from easy procedure. More rapid work is made possible by a number of electrical methods where measurements are taken directly on a sample without any preliminary treatment. Three main methods of measurement have been developed on this principle. In the first the difference in dielectric constant is determined between the moist and the dry substance; in the second the dielectric losses are measured; in the third the conductivity. In all these methods, however, the results depend very much upon the form of the substance to be examined (lumps, grains, powder, liquid), for this form affects the filling factor when the sample being measured is introduced into a measuring vessel. The result is that only relative values of moisture are measured, and if absolute quantities of water are desired a separate calibration curve is needed for each substance in its given form.

In order to eliminate the effect of the form of the substance (which means that the absolute water content is determined with only one calibration curve for all substances), another group of methods have been developed which are based on the extraction of the water from the moist substance. The most important methods of this type are tabulated and briefly characterized below. Except for the drying already mentioned, which actually also belongs to this group and is therefore included in the table, the methods indicated here are hardly less rapid than the direct electrical measurements, but they all have one drawback in that they require a fairly expensive apparatus which cannot be operated by a layman [1]).

We have now found that a method of measurement on the basis of extraction can be developed which while still being reasonably quick offers the advantages of very simple operation and an inexpensive apparatus.

The basic idea is the following. Water is extracted from the substance to be tested with a hygroscopic

[1]) A survey of the different methods of moisture determination will be found in: E. Eckert and P. Wulff, Z.angew. Chemie, **53**, 403-405, 1940.

mentioned that in such cases the elevation is always smaller than in the case we have considered above. By this means and with the aid of fig. 7 we can give an indication of what significance this effect might have in stereophonic reproduction in practice, for instance in a cinema. Any perceptible elevation of about 20° will easily be corrected by the suggestion of the visual picture, so that no elevation will be noticed in the seats farther away from the screen than the place corresponding to the angle $a = 30°$ (see fig. 7). If the distance between the loudspeakers is say 8 meters the elevation effect will therefore only be noticeable for a trained listener at places closer than about 7 meters to the screen, and the number of such seats — which also in other respects are to be regarded as unfavourable — is only a small percentage of the total number of seats in a normal cinema.

# A NEW ELECTRICAL METHOD FOR DETERMINING MOISTURE CONTENT

by J. BOEKE.

543.812: 621.317.39

A description is given of a new method for determining the water content of liquid or solid substances, such as grain, wood, textiles, butter, etc. The water is extracted from the material with acetone in which oxalic acid is dissolved. The increase in the conductivity of the extraction liquid serves as a measure of the amount of water taken up. Compared with those already existing this method has the advantage that the apparatus required is inexpensive and simple in operation, while at the same time it is a very reasonably quick method and not dependent on the form in which the substance to be tested occurs.

The determination of the water content in gases, liquids and solids is one of the most important analyses regularly applied in commerce and industry. The moisture content of grain, of tobacco, of butter, to name only a few examples, is also paid for by the purchaser, and large profits or losses may be involved when a kilogram of the product contains a few grams more or less of water than were estimated. It is therefore understandable that numerous attempts have been made to develop methods of measuring moisture content quickly and accurately.

If we confine ourselves to solids and liquids, the only universal method used until now consisted in drying a sample and determining the amount of evaporated water by weighing. This, however, is often a lengthy and far from easy procedure. More rapid work is made possible by a number of electrical methods where measurements are taken directly on a sample without any preliminary treatment. Three main methods of measurement have been developed on this principle. In the first the difference in dielectric constant is determined between the moist and the dry substance; in the second the dielectric losses are measured; in the third the conductivity. In all these methods, however, the results depend very much upon the form of the substance to be examined (lumps, grains, powder, liquid), for this form affects the filling factor when the sample being measured is introduced into a measuring vessel. The result is that only relative values of moisture are measured, and if absolute quantities of water are desired a separate calibration curve is needed for each substance in its given form.

In order to eliminate the effect of the form of the substance (which means that the absolute water content is determined with only one calibration curve for all substances), another group of methods have been developed which are based on the extraction of the water from the moist substance. The most important methods of this type are tabulated and briefly characterized below. Except for the drying already mentioned, which actually also belongs to this group and is therefore included in the table, the methods indicated here are hardly less rapid than the direct electrical measurements, but they all have one drawback in that they require a fairly expensive apparatus which cannot be operated by a layman [1]).

We have now found that a method of measurement on the basis of extraction can be developed which while still being reasonably quick offers the advantages of very simple operation and an inexpensive apparatus.

The basic idea is the following. Water is extracted from the substance to be tested with a hygroscopic

[1]) A survey of the different methods of moisture determination will be found in: E. Eckert and P. Wulff, Z.angew. Chemie, **53**, 403-405, 1940.

| Process | Extraction medium | Procedure | Determination of the water extracted by: |
|---|---|---|---|
| Drying | air<br>(vacuum) | circulation, or heating | weight |
| Distillation of mixture | toluene | distillation | volume |
| "Exluan" process | dioxane | mixing, or grinding together | dielectric constant |
| Titration according to Fischer | methanol | mixing, or grinding together | potentiometric titration with Fischer's reagent |

liquid which in itself has a low conductivity and does not readily dissociate electrolytes. An electrolyte is dissolved in the extraction liquid, thereby being only very slightly dissociated in it, so that the solution possesses only a low conductivity. The absorption of water by the extraction liquid increases its power to dissociate, the electrolyte is thus more dissociated and the result is a considerable increase in the conductivity of the liquid, which can be measured by simple means.

As extraction liquids possessing the desired properties methyl and ethyl alcohol can be used, but acetone is still more suitable. Oxalic acid may be used as electrolyte, since it dissolves very readily in acetone [2]). When a solution of 10% oxalic acid in acetone is used the conductivity of the solution as a function of the water content increases very rapidly, as may be seen from the curve in *fig. 1*. Variation in the content of oxalic acid makes relatively little difference, as shown by the dotted-line curves; the effect of changes in temperature is also slight.

The possibility that the test substance contains common salt or some other salts must be kept in mind. Common salt is practically insoluble in pure acetone and has scarcely any effect on the low conductivity of acetone. A slight addition of water, however, causes the salt to dissolve more readily in the acetone, the dissolved salt is dissociated and the conductivity increased. This effect, which is based upon the increased dissolving power of the acetone upon absorption of water, is even stronger than the effect of the increased dissociating power, of which use is made in our method of measuring, and it may therefore obviously be asked why the whole method should not be based rather upon the first effect. That would mean that an excess of solid sodium chloride would have to be added instead of oxalic acid, but then there is the objection that sodium chloride dissolves very slowly in acetone containing little water (say $< 5\%$), and it may take days to establish the equilibrium, which would be very objectionable for practical measurements. For our method, based on the dissociation of oxalic acid, on the other hand, the slow solution of sodium chloride has just the advantage that little difficulty is experienced from the salt content of the substance to be examined. It is only necessary that the extraction should be completed within two hours and that the final acetone-oxalic acid-water mixture does not contain more than a few percent of water. Provided these conditions are fulfilled, large deviations from the calibration curve of fig. 1 are only observed in the case of substances with an extremely high content, such as salted fish. The apparent amount of water in such a case may be double the real amount determined by drying. With wood, paper, textiles and suchlike, however, the figures found for the water content, taken absolutely, differ by only a few tenths percent at most from those found by drying. The accuracy of the method, like that of the other methods based upon extraction, is for a large part limited by the familiar phenomenon that the test object very stubbornly retains the last traces of water and the state of equilibrium, where only a very minute quantity of water remains definitively in the sub-

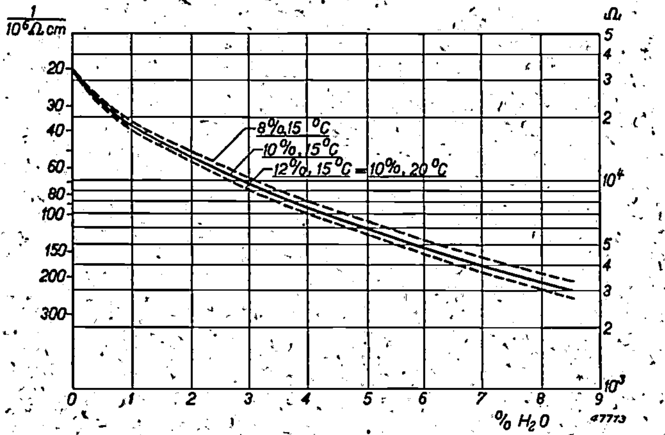


Fig. 1. Conductivity and resistance, measured in the measuring cell, of acetone with a certain percentage of dissolved oxalic acid, as a function of the amount of water absorbed. The three curves apply for different concentrations of oxalic acid and temperatures as indicated.

[2]) It would be simpler if the extraction liquid itself could function as electrolyte; concentrated sulphuric acid for instance is hygroscopic and at the same time upon taking up water its dissociation and consequently its conductivity is very much increased. It has not been possible, however, to find a substance which combines all the desired properties. Sulphuric acid, for example, is too aggressive chemically.

stance, is only slowly attained. Substances containing albumins are especially difficult in this respect. Grains of wheat, for example, which had lain in a vacuum of 20 mm Hg for three da s, while for 10 hours the temperature had been maintained at 65 C°, again lost 1.6% by weight of water during the next two days in a vacuum. This phenomenon makes it very difficult, whatever the method, to determine the absolute moisture content of such substances. Nevertheless, the extraction with acetone has at least the advantage that the above-mentioned equilibrium is established much more quickly than when drying in air or a vacuum.

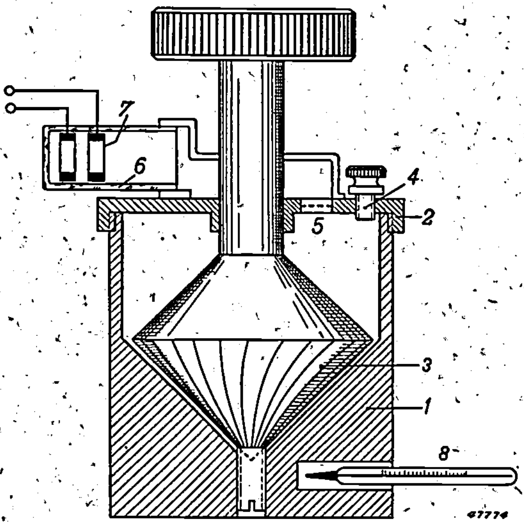


Fig. 2. Apparatus for determining the moisture content by the extraction of water and the measurement of conductivity. After the vessel *1* is filled with a weighed sample of the substance to be investigated, the cover *2* with the grinding cone *3*, which can be rotated and moved up and down, is screwed on. Through the opeuing *4* a known amount of extraction liquid is poured in. By rotating and pressing at the same time on the cone, which is provided with grinding grooves, the sample and the liquid are mixed. After the extraction is complete, the apparatus is tipped and the liquid runs through the sieve *5* into the measuring cell *6* which is attached with an air-tight joint and contains the platinum electrodes *7*. The resistance between these electrodes can be measured. *8* is a thermometer.

In the practical application of the method care must be taken, *i.a.*, that during the extraction and the subsequent measurement of the conductivity no acetone can evaporate, which would increase the concentration of the oxalic acid. It is therefore desirable to carry out the measurement in a completely closed vessel, as shown in *fig. 2*. The apparatus consists of a kind of mill with a measuring vessel attached to it. A weighed quantity of material and a measured volume of acetone with 10% oxalic acid are placed in the mill, which is then closed tight. After the substance and the extraction liquid have been thoroughly mixed together and left to stand for one to two hours, the mill is tipped so that the extract, filtered through a sieve, runs into the measuring vessel. This is a cell with two platinum electrodes. Since the configuration of the elcctrodes and the liquid between them is fixed, from a measurement of the electrical resistance of the cell the conductivity of the liquid and, with the help of the calibration curve in fig. 1, its water content can immediately be derived. The resistance may be measured by means of a measuring bridge, for example the "Philoscope"[3]), which is specially adapted for such simple measurements. In the choice of the calibration curve the temperature of the extract, which can be read off on a thermometer attached to the mill, must be taken into account.

Since oxalic acid attacks metals, all the surfaces of the mill coming into contact with the extraction liquid, except the platinum electrodes, must be made for instance of ceramic material, plastics or glass.

[3]) See Philips Techn. Rev. **2**, 270, 1937.

# VIBRATION-FREE MOUNTINGS WITH AUXILIARY MASS

by J. A. HARINGX.

621.752

The use of sensitive instruments such as balances, galvanometers and microscopes is often made difficult or even impossible by vibrations in the surroundings. In order to reduce the amplitudes of these forced vibrations the instruments can be placed upon sufficiently weak springs. Then, however, a damping must be introduced in order to stop the free vibrations of the system after a slight impulse or an initial displacement. The most obvious manner of applying this damping, namely between the apparatus and the foundation upon which it is placed, is indeed favourable for the rapid decay of the free vibrations, but it promotes the forced vibrations. A better method consists in introducing the damping between the apparatus and an auxiliary mass attached to it by means of springs. The features of this system and the choice of the different parameters (masses, rigidities of the springs, damping) are discussed in this article for the one-dimensional case.

## Various systems of vibration-free mountings

When sensitive instruments such as balances, galvanometers, microscopes, and dial gauges are used difficulties are often experienced due to vibrations transmitted to the apparatus through the floors, walls and tables. In such a case an arrangement will be needed in which the transmission of the vibrations to the instrument in question is avoided. Very good results can be obtained by using a spring-mounting, but it is quite impossible to construct in this way a support which completely prevents the transmission of vibrations. It is only possible to limit the amplitudes of the forced vibrations of the measuring instrument to such a degree that these vibrations no longer present any difficulty, so to that extent one may then say that the instrument is supported „vibration-free".

When a resilient layer is introduced between the instrument and its foundation, for instance a rubber cushion or a set of helical springs, a system is obtained like that shown diagrammatically in *fig. 1a*: a mass $m$ joined by a spring with the rigidity $c$ (*i.e.* the force per unit of elongation of the spring) to a foundation which vibrates in a vertical direction with an angular frequency $\omega$ and an amplitude $a_0$.

This system behaves as follows. The mass $m$ vibrates at the same frequency [1]) $\omega$ with the foundation, but with an amplitude $a$ which depends very much on the frequency. From the differential equation for the motion of the mass it may be derived that

$$\frac{a}{a_0} = \left| \frac{c}{c - m\omega^2} \right|.$$

This "frequency characteristic" is shown in

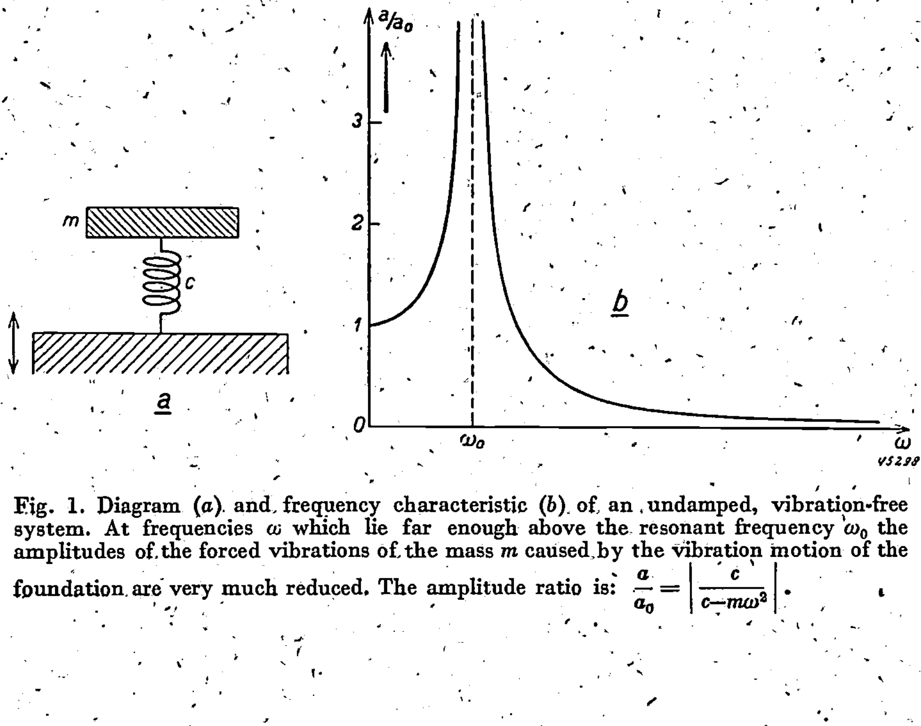


Fig. 1. Diagram (*a*) and frequency characteristic (*b*) of an undamped, vibration-free system. At frequencies $\omega$ which lie far enough above the resonant frequency $\omega_0$ the amplitudes of the forced vibrations of the mass $m$ caused by the vibration motion of the foundation are very much reduced. The amplitude ratio is: $\dfrac{a}{a_0} = \left| \dfrac{c}{c - m\omega^2} \right|$.

[1]) For the sake of brevity we use the term "frequency" here meaning (unless otherwise stated) the angular frequency $\omega = 2\pi$ times the frequency.

fig. 1*b*. In the neighbourhood of the frequency

$$\omega_0 = \sqrt{\frac{c}{m}},$$

the so-called resonant frequency of the system, the mass takes on very large amplitudes which are much larger than $a_0$. On the other hand at frequencies far enough above $\omega$ the amplitude becomes smaller than $a_0$. At very high frequencies it even gradually approaches the zero, changing in inverse proportion to the square of the frequency:

$$\frac{a}{a_0} \approx \frac{c}{m\omega^2} = \left(\frac{\omega_0}{\omega}\right)^2 \quad . \ . \ . \ . \ . \ (1)$$

If the spring ($c$) is sufficiently weak and the mass ($m$) large enough to make the resonant frequency $\omega$ at least 3 to 5 times as low as the lowest frequency occurring in the interfering vibrations of the foundation, the instrument will take up these vibrations only to a very slight extent.

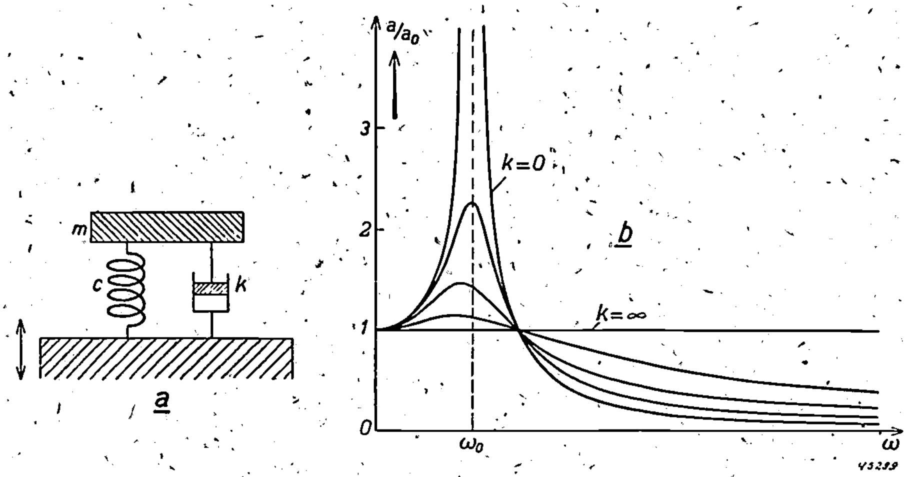


Fig. 2. Diagram (*a*) and frequency characteristic (*b*) of a vibration-free system with „relative" damping. With increasing damping ($k$) the resonance peak is gradually lowered and at the same time the free vibrations die out more quickly, but at high frequencies the amplitudes of the forced vibrations become larger. The amplitude ratio is given by the expression [2]):

$$\left(\frac{a}{a_0}\right)^2 = \frac{1+q^2\bar{\omega}^2}{(\bar{\omega}^2-1)^2+q^2\bar{\omega}^2}, \text{ where } \bar{\omega} = \frac{\omega}{\omega_0},\ q = \frac{k}{m\omega_0},\ \omega_0^2 = \frac{c}{m}.$$

Besides the limitation of the amplitudes of the forced vibrations in the frequency region of the permanently occurring vibrations, it is also desired that after an impulse or initial displacement, caused either accidentally or by the operation of the instrument, the free vibrations of the system should come to rest as quickly as possible. Once it has taken up vibration energy the system according to fig. 1*a* continues to move and, theoretically, it takes an infinitely long time before the free vibration dies out. In practical systems, however, the vibration energy will gradually disappear, because there is always a certain damping even if it is only the internal damping of the material of the springs or the damping due to air resistance. It is clear that the stronger this damping the sooner the system will come to rest. This naturally suggests the introduction of an extra damping in the manner shown in *fig. 2a*. The damping force is assumed to be proportional to the relative velocity of the mass with respect to the foundation (viscous damping) and the proportionality factor is called $k$. As the free vibration dies out the amplitude then decreases as a function of time $t$ proportional to $e^{-kt/2m}$. Such an arrangement, however, behaves entirely different from the first one, not only as far as the decay of the free vibrations is concerned but also as regards the forced vibrations. This is clearly shown by the frequency characteristics drawn in fig. 2*b* for different values of the damping $k$. The resonance peak is more or less reduced, but at the same time, at high frequencies, the decrease is much more gradual than in fig. 1*b*. It is found that in this case the amplitude ratio at high frequencies is inversely proportional to the frequency:

$$\frac{a}{a_0} \approx \frac{k}{m\omega}.$$

Thus the larger the damping $k$ the more the resonance peak is cut down and the quicker the system comes to rest after an impulse, but at high frequencies the amplitudes are less effectively reduced;

[2]) See for example E. Lehr, Schwingungstechnik, Vol. II, pp 171 et seq. (J. Springer, Berlin 1934).

thus in order to limit suffciently the amplitudes of the forced vibrations the resonant frequency will have to be placed farther below the interfering frequency region.

Since, however, in practice, the resonant frequency cannot be made arbitrarily low (if the mass should be limited the mounting would become too "weak"), an arrangement according to the principle of fig. 2$a$ does not usually give satisfactory results: with low damping it takes too long for the free vibrations to die out, whereas with large damping amplitudes of the forced vibrations are not sufficiently limited.

The situation would be quite different if the damping were not introduced between the mass $m$ and the vibrating foundation as in fig. 2$a$ (relative damping), but between the mass and a fixed point in space, so-called absolute damping, see *fig. 3a*. The frequency characteristics of this system, drawn in fig. 3$b$ for different values of the damping, show that in this case increased damping is advantageous in every respect. The resonance peak is very much flattened, while even at high frequencies, *i.e.* in the interfering frequency region of the foundation, the amplitude is always less than in the arrangement with no damping (fig. 1) and, as in that case, at sufficiently high frequencies its trend is again according to equation (1). The difference between the effect of this absolute damping and the relative damping indicated in fig. 2 is easily understood when one considers that the damping introduced between two points constitutes a hindrance to their relative movement and an infinitely large damping even has the effect of a rigid coupling. Thus, if in fig. 2$a$ the damping increases, the mass $m$ will finally be rigidly joined with the foundation and will therefore follow its movement completely ($a/a_0 = 1$). If, on the other hand, in fig. 3$a$ the damping is increased, the mass is finally rigidly bound to a stationary point in space and thus remains completely at rest ($a/a_0 = 0$).

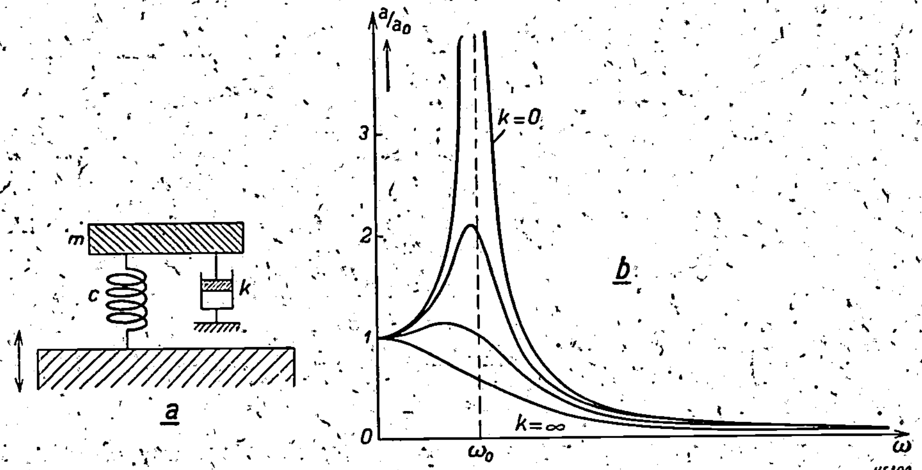


Fig. 3. Diagram ($a$) and frequency characteristic ($b$) of a vibration-free system with "absolute" damping. With increasing damping ($k$) not only is the resonance peak lowerd and the duration of the free vibrations shortened, but also the amplitudes of the forced vibrations are reduced. The amplitude ratio is given by the expression [3]):

$$\left(\frac{a}{a_0}\right)^2 = \frac{1}{(\overline{\omega}^2 - 1)^2 + q^2\overline{\omega}^2}, \text{ where } \overline{\omega} = \frac{\omega}{\omega_0},\ q = \frac{k}{m\omega_0},\ \omega_0{}^2 = \frac{c}{m}.$$

The case of absolute damping is of course only of theoretical interest. If it were actually possible to have a fixed point in space it would be advisable to make the instrument vibration-free by attaching it rigidly to the point in question directly, instead of mounting it with springs on the vibrating foundation.

Nevertheless, even without having a fixed point at our disposal, it is possible to imitate absolute damping to a certain extent by introducing the damping element ($k$) between the main mass $m$ and an auxiliary mass $M$ attached to the main mass by a spring (rigidity $C$). This arrangement is represented schematically in *fig. 4a*. At high frequencies the auxiliary mass $M$, due to its inertia, will in general have a tendency to remain at rest and in that way will furnish an almost fixed — or at least a slightly vibrating — point of contact for the damping force. Such a conception is indeed confirmed theoretically and the arrangement is found to possess very favourable properties not only as regards the amplitudes of the forced vibrations but also in [illegible]

[3]) See for example E. Lehr, Schwingungstechnik, Vol. II, pp. 135 et seq. (J. Springer, Berlin 1934).

respect to the decay of the free vibrations. Several vibration-free mountings have already been constructed on this principle in the Philips factories.

We will now look more closely into the behaviour of a system according to fig. 4*a* and consider the choice of the various parameters ($m$, $c$, $M$, $C$, $k$), confirming our attention to the one-dimensional case.

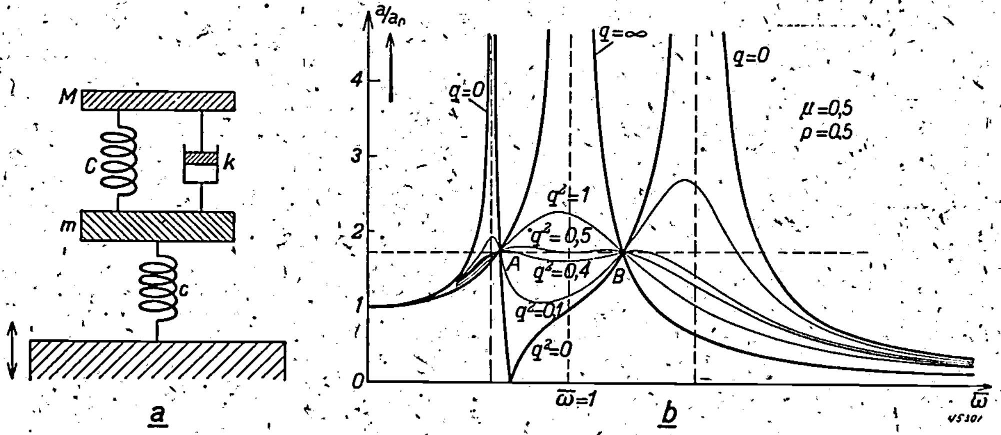


Fig. 4. Diagram ($a$) and frequency characteristic ($b$) of a vibration-free system with auxiliary mass $M$. The damping ($k$) is introduced between the main mass ($m$) and the auxiliary mass. The amplitude ratio is given by formula (3) on page 20. The figure here is drawn for the special case $\mu = 0.5$ and $p = 0.5$, where the points of intersection $A$ and $B$ of all the frequency characteristics lie at the same height.

## The behaviour of the system with auxiliary mass in relation to the parameters

### *The frequency characteristic*

In fig. 4*b* we have the frequency characteristic for a given case at different values of damping. We will begin to analyse this figure by taking the extreme case where the damping $k$ is infinitely large. The auxiliary mass $M$ is then rigidly bound to the main mass $m$; we then have in fact a system like that of fig. 1, with spring rigidity $c$ and mass $m + M$. The frequency characteristic of this system is the fully drawn curve in *fig. 5* (identical with fig. 1*b*), having the resonance frequency

$$\omega_0 = \sqrt{c/(m+M)}.$$

Since $\overline{\omega} = \omega/\omega_0$ is taken as abscissa, resonance occurs at $\overline{\omega} = 1$.

If, on the other hand, $k = 0$, we have the familiar case of two coupled, undamped oscillators. Such a system with two degrees of freedom is in resonance at two different frequencies and therefore has a frequency characteristic like the dotted curve in fig. 5. The exact position of the resonant frequencies on either side of $\omega_0$ ($\overline{\omega}=1$), as well as of the intermediate zero point $P$ of the curve, depends upon the parameters of the system (masses and rigidities of the springs).

Finally, in the general case where the damping $k$ has a value between 0 and $\infty$, the frequency characteristic always passes through the two points of intersection $A$ and $B$ of the two extreme curves first considered and for the rest lies entirely between those two curves, in general either with one maximum between $A$ and $B$ or with two maxima one on either side of $A$ and $B$. See the family of curves in fig. 4*b*.

It will perhaps be interesting and useful to explain this in somewhat more detail with formulae, the derivation of which may be briefly outlined [4]). When considering the forces acting on the main and auxiliary masses, we get for the motion of the two masses two coupled linear differential equations of the second order, from which by elimination of the coordinate of the auxiliary mass a linear differential equation of the fourth order is obtained for the coordinate $x$ of the main mass. If now the foundation vibrates according to $a_0 \sin \omega t$, the main mass will vibrate at the same frequency but generally with a different amplitude and phase: $x = a \sin (\omega t + \varphi)$. By substituting this in the differential equation, an equation can be derived for the frequency characteristic, namely for the ratio $a/a_0$ as a function of $\omega$ and of the parameters $m$, $c$, $M$, $C$, $k$. These quantities, however, are

[4]) Cf. J. P. den Hartog, Vibrations et mouvements vibratoires, p. 99 (Ed. Dunod, Paris 1936); E. Hahnkamm, Die Dämpfung von Fundamentschwingungen bei veränderlicher Erregerfrequenz, Ing. Arch. 4, 192, 1933; L. Geislinger, Theorie des Resonanzschwingungsdämpfers, Ing. Arch. 5, 146, 1934.

found to occur in the result only in certain combinations, so that the result can be written much more simply and be made more comprehensible by introducing the following (dimensionless) quantities:

$$\left.\begin{aligned} \bar\omega &= \frac{\omega}{\omega_0} = \omega\sqrt{\frac{m+M}{c}},\\ p &= \frac{C}{c}\,\frac{m+M}{M},\\ q &= \frac{k}{M}\sqrt{\frac{m+M}{c}},\\ \mu &= \frac{m}{m+M},\end{aligned}\right\}\quad \ldots \quad (2)$$

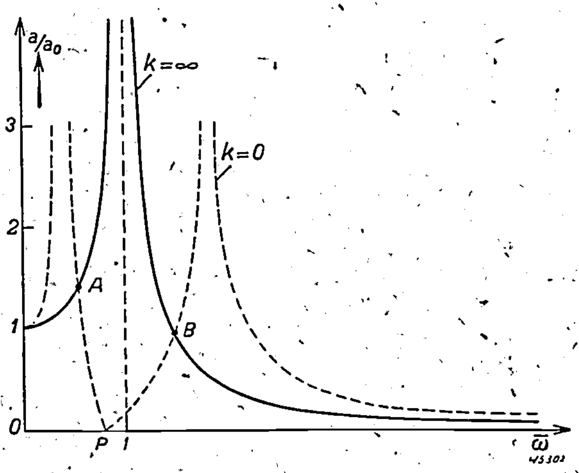


Fig. 5. Frequency characteristic of the system with auxiliary mass (fig. 4*a*) in the two extreme cases of damping $k = \infty$ and $k = 0$. In the first case main mass and auxiliary mass are rigidly connected and we have a system like that in fig. 1*a* (fully drawn curve), while in the second case we have a combination of two coupled (undamped) vibration systems. There are then two resonant frequencies (dotted curves) corresponding to the two degrees of freedom; $\bar\omega = \omega/\omega_0$ is the reduced frequency.

The quantity $p$ will be called the rigidity parameter, $q$ the damping parameter and $\mu$ the mass parameter. The formula for the frequency characteristic now becomes

$$\left(\frac{a}{a_0}\right)^2 = \frac{(\bar\omega^2-p)^2+q^2\bar\omega^2}{[\mu\bar\omega^4-(1+p)\,\bar\omega^2+p]^2+q^2\bar\omega^2(\bar\omega^2-1)^2}. \quad (3)$$

In the case of infinitely large damping, thus $q = \infty$, the formula becomes

$$\left(\frac{a}{a_0}\right)_{q=\infty} = \left|\frac{1}{\bar\omega^2-1}\right| \quad \ldots \quad (4)$$

This is the equation for the fully drawn curve in fig. 5. In the case without damping, thus $q = 0$, the formula becomes

$$\left(\frac{a}{a_0}\right)_{q=0} = \left|\frac{\bar\omega^2-p}{\mu\bar\omega^4-(1+p)\,\bar\omega^2+p}\right|,$$

which is the equation for the dotted curve in fig. 5.

If for the sake of simplicity we now call the functions of $\omega$ occurring in the four terms of numerator and denominator of (3) $\alpha$, $\beta$, $\gamma$, $\delta$, we may then write for (3):

$$\left(\frac{a}{a_0}\right)^2 = \frac{\alpha+q^2\beta}{\gamma+q^2\delta} = \frac{\alpha\,(1+q^2\cdot\beta/\alpha)}{\gamma\cdot(1+q^2\cdot\delta/\gamma)}$$

and it is immediately clear that with the condition that

$$\frac{\beta}{\alpha} = \frac{\delta}{\gamma} \quad \ldots \ldots \quad (5)$$

the quotient $(a/a_0)^2$ becomes equal to $\alpha/\gamma$ and thus independent of $q$. At those values of $\omega$ for which (5) is satisfied, therefore, the curves have the same ordinate for all values of $q$, including $q = 0$ and $q = \infty$, from which it follows, as already stated, that all the curves pass through the two points of intersection $A$ and $B$ of the fully drawn and the dotted curves in fig. 5.

For the abscissae $\bar\omega_A$ and $\bar\omega_B$ of these points of intersection, which we shall presently need, we can derive the following equation by substituting the four functions $\alpha \ldots \delta$ in (5):

$$\bar\omega^4 - 2\,\frac{1+p}{1+\mu}\,\bar\omega^2 + 2\,\frac{p}{1+\mu} = 0,$$

thus:

$$(\bar\omega^2)_{A,B} = \frac{1}{1+\mu}\,(1+p \pm \sqrt{p^2-2p\,\mu+1}), \quad (6)$$

where

$$\bar\omega_A < 1 \text{ and } \bar\omega_B > 1.$$

For the ordinates of the points of intersection one then finds according to (4)

$$\left(\frac{a}{a_0}\right)_A = \frac{1}{1-\bar\omega_A^2} \text{ and } \left(\frac{a}{a_0}\right)_B = \frac{1}{\bar\omega_B^2-1}. \quad (7)$$

*Behaviour at high frequencies*

From the general shape of the curves in fig. 4*b* it may be seen that also for this type of system the resonance region must in any case lie far below the interfering frequency region of the vibrations of the foundation. The degree to which the amplitudes of the forced vibrations of the main mass are then restricted can easily be deduced from equation (3). For sufficiently high frequencies it becomes

$$\left|\frac{a}{a_0}\right| \approx \frac{1}{\mu\bar\omega^2} = \frac{c}{m\omega^2}. \quad \ldots \quad (8)$$

When we compare (8) with (1), we see that with the same values of $c$ and $m$ we obtain the same favourable behaviour as in the elementary arrangement entirely without damping (fig. 1) or as in that with absolute damping (fig. 3). The presence of the auxiliary mass $M$ therefore apparently plays no part here.

Such behaviour has already been presumed quantitatively from the tendency of the auxiliary mass $M$ to remain stationary at high frequencies.

Actually, of course, $M$ does move slightly, and it may be deduced that the amplitude $a$ of this movement at high frequencies, is determined by $|a'/a| \approx kc/mM\omega^3$. The ratio $a'/a$ thus decreases more rapidly with increasing $\omega$ than the amplitude ratio for the main mass according to formula (8).

### *The optimum choice of parameters*

In the practical realisation of a vibration-free mounting according to fig. 4 the question will naturally arise as to how stiff the springs must be made, and how heavy the masses and how strong the damping will have to be. The behaviour at high frequencies, equation (8), gives the indication already known, that the resonance region of the system must lie at the lowest frequencies possible. As to the choice of $C$, $M$ and $k$ this does not help us at all, since according to equation (8) these parameters do not affect the amplitudes of the forced vibrations at high frequencies. The choice of these parameters will, however, be decisive for the behaviour at lower frequencies and for the decay of the free vibrations of the system after an accidental impulse or initial displacement.

During the free vibrations the motion of the system is in general composed of two vibrations with the two "resonant frequencies", the amplitude of each of these vibrations decreasing exponentially with the time: $e^{-a_1 t}$ and $e^{-a_2 t}$ respectively, while the relation between the initial amplitudes of the two vibrations depends upon the initial conditions (initial displacement or impulse). The rate of decay of the free vibration will thus depend not only on $a_1$ and $a_2$ but also on the accidental initial conditions. It is therefore impossible to speak of *the* rate of decay of the free vibrations, and even when the initial conditions are given the influence of the parameters of the system on the decay of the vibrations is still very difficult to ascertain. When, however, we assume that the limitation of the duration of the free vibrations runs more or less parallel with the decrease in the amplitudes of the system in the resonance region, thus in a manner similar to the behaviour of the systems with one mass (figs. 2 and 3), we only need to study the frequency characteristic at the lower frequencies. We might then state the condition that the highest peak in the frequency characteristic should be as low as possible, and from that requirement derive the optimum choice of the parameters.

Among the curves with different values of the damping parameter $q$ there will be one which has its highest peak just at the highest of the two points $A$ and $B$. When the position of $A$ and $B$ is known this is evidently the most favourable possibility, because then the ordinate of $A$ or $B$ is never exceeded. The value of $q$ corresponding to this curve is, it is true, still unknown, but that does not affect our argument. We now study the position of $A$ and $B$ with reference to fig. 5. The fully drawn curve ($q = \infty$) to which equation (4) applies is entirely fixed if the resonant frequency $\omega_0$ is taken as given. The points $A$ and $B$ will then always lie on this curve, and from formula (6) it can be proved that when we vary the parameter $p$ both points are displaced to the right or to the left. From fig. 5 it may be seen that one point therefore always rises as the other falls, and *vice versa*. The highest peak of the frequency characteristic can then also be considerably lowered with respect to fig. 5 by giving $p$ a value such that the points $A$ and $B$ lie at the same height as is the case in fig. 4*b*. Although the optimum frequency characteristic then exhibits two maxima lying at the same height but coinciding neither with $A$ nor with $B$, the differences are so extremely small that in this way a good approximation is attained. From (7) it follows that $A$ and $B$ lie at the same height when the following condition is satisfied:

$$\overline{\omega}_A^2 + \overline{\omega}_B^2 = 2.$$

If we substitute here expression (6) for $\overline{\omega}_A$ and $\overline{\omega}_B$ we obtain as "optimum" value for the rigidity parameter

$$p_{\text{opt}} = \mu.$$

For the corresponding maximum amplitude ratio we find

$$\left(\frac{a}{a_0}\right)^2_{\text{opt}} \approx \left(\frac{a}{a_0}\right)^2_{A,B} = \frac{1+\mu}{1-\mu} \quad . \quad . \quad . \quad (9)$$

On the basis of the calculations of Collatz[5]) it can further be shown that the optimum value of the damping parameter can be taken with a very close approximation to be:

$$q_{\text{opt}} \approx \sqrt{1{,}5\,\mu\,(1-\mu)}.$$

[5]) L. Collatz, Über den günstigsten Wert der Kopplungskonstanten bei reibungsgekoppelten Systemen, Ing. Arch. **10**, 269, 1939.

*Important practical case: mass parameter $\mu = 0.5$*

After the foregoing the problem of the choice of the parameters is reduced to the choice of the mass parameter $\mu = m/(m + M)$. Since we try to keep the frequency characteristic as low as possible we must try to make the largest ordinate now occurring as small as possible. According to equation (9) the smallest possible value of $\mu$ is desired, *i.e.* in our vibration-free mounting the auxiliary mass $M$ would have to be large compared with the main mass $m$. In practice, however, one is not likely to make $M$ larger than the main mass. If we therefore assume that the two masses are chosen of equal size, thus $\mu = 0.5$, we obtain:

$$p_{\iota\mathrm{pt}} = \mu = 0.5,$$

$$q_{\mathrm{opt}} \approx \sqrt{1.5\ \mu\ (1 - \mu)} = 0.612.$$

Having regard to equation (2), it also appears that the following relations must be satisfied:

$$C/c = \mu(1-\mu) = 0.25,$$

$$\frac{k}{\sqrt{mc}} \approx (1-\mu)\sqrt{1.5\ (1-\mu)} = 0.433.$$

Finally, the maximum amplitude ratio according to equation (9) amounts to only

$$\left(\frac{a}{a_0}\right)_{\mathrm{opt}} \approx \sqrt{\frac{1+\mu}{1-\mu}} = 1.728.$$

When $p = \mu = 0.5$ we find exactly [6])

$$q_{\mathrm{opt}} = 0.624 \text{ and } (a/a_0)_{\mathrm{opt}} = 1.746.$$

The approximations given are thus found to agree very well with the exact values. As a matter of fact even if we make the damping parameter much too large or too small, $(a/a_0)$ still varies only slightly; between $q = 0.4$ and $q = 0.95$ the increase of the amplitude ratio with respect to its optimum value (1.75) amounts at the most to 25%. The same is true for the choice of $p$. If we keep to $\mu = 0.5$ and take in each case the best value of $q$, a variation of $p$ between 0.3 and 0.7 results in a maximum rise of 25% in $(a/a_0)_{\mathrm{max}}$.

When we compare these optimum results of our mounting with auxiliary mass with the vibration-free mounting without auxiliary mass, we assume the total mass $m + M$ or $m$ and the rigidity of the spring $c$ to be given, since in the practical construction a certain weight will in general be available and it is, moreover, required that the resilient attachment of the main mass should possess a certain degree of rigidity. The resonant frequencies

$$\omega_0 = \sqrt{c/(m+M)} \text{ resp. } \omega_0 = \sqrt{c/m}$$

are then automatically equal, so that for each $\omega$ the reduced frequencies $\overline{\omega} = \omega/\omega_0$ have the same value in both cases. In *fig. 6* the following curves are given for the sake of comparison.

1. The frequency characteristic of the undamped system according to fig. 1*a*;
2. that of the system with auxiliary mass according to fig. 4*a* with the same rigidity $c$ and the same total mass divided into two equal parts: thus $\mu$ is 0.5 and further $p = 0.5$; $q = 0.62$;
3. that of the system according to fig. 2*a* with relative damping, with the same rigidity and mass and a damping ($q = k/m\omega_0 = 0.74$) such that the maximum of the frequency characteristic lies just as low as that of curve *2*.

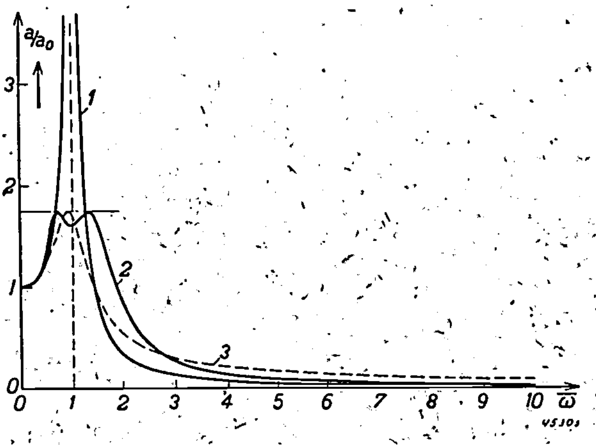


Fig. 6. Frequency characteristics of an undamped vibration system (curve *1*), a system with auxiliary mass and "optimum" parameters chosen for $\mu = 0.5$ (curve 2), and a system with relative damping (curve 3). It is assumed that the total mass and also the rigidity of the spring $c$ is the same in all three systems. Further for curve *3* the damping was so chosen that the maximum amplitude ratio is the same as in case 2.

It is clearly seen how in case (3) the trend of the frequency characteristic is much less favourable at high frequencies. [6]) Moreover, the last mentioned system has the undesirable property that an accidental increase in the damping coefficient $k$ (for instance with viscous damping due to a fall in temperature) causes the mass to vibrate with a proportionally larger amplitude, with the result that the system, intended as a vibration-free mounting, becomes much less effective.

On the other hand, if an auxiliary mass is applied the damping does not — or at least not perceptibly — affect the amplitudes of the forced vibrations at high frequencies, while, as we have seen, the maximum amplitude ratio in the resonance region reacts to a change in damping only to a slight extent.

[6]) Because $a/a_0 \approx km/\omega = q/\overline{\omega}$ instead of $a/a_0 = 1/\overline{\omega}^2$ and $1/\mu\overline{\omega}^2$ respectively.

### *Decay of the free vibrations*

In order to avoid the difficulties encountered in studying the effect of the parameters on the behaviour of the system with auxiliary mass with respect to the decay of the free vibrations after an impulse or an initial displacement, we have adopted the simple point of view that the duration of the free vibrations would be restricted parallel with the reduction of the amplitudes at resonance, and consequently we aimed at the lowest possible maximum of the frequency characteristic. However, once all the parameters have been chosen, the rate of decay of the free vibrations can be determined exactly. For the case where $m = M$ with the corresponding "optimum" values $p = 0.5$ and $a = 0.62$ we find that after a certain initial displacement the free vibration of our main mass diminishes in the manner shown in *fig. 7*.

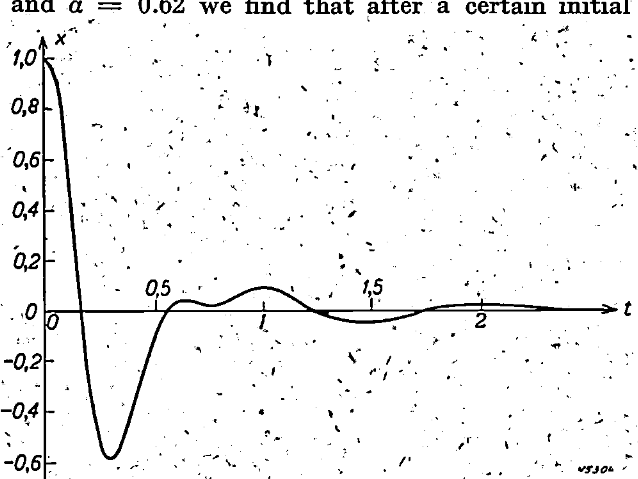


Fig. 7. The decay of the free vibration after a given initial displacement for a system according to fig. 4*a*, with $\mu = 0.5$ and the "optimum" parameters $p = 0.5$ and $q = 0.62$.

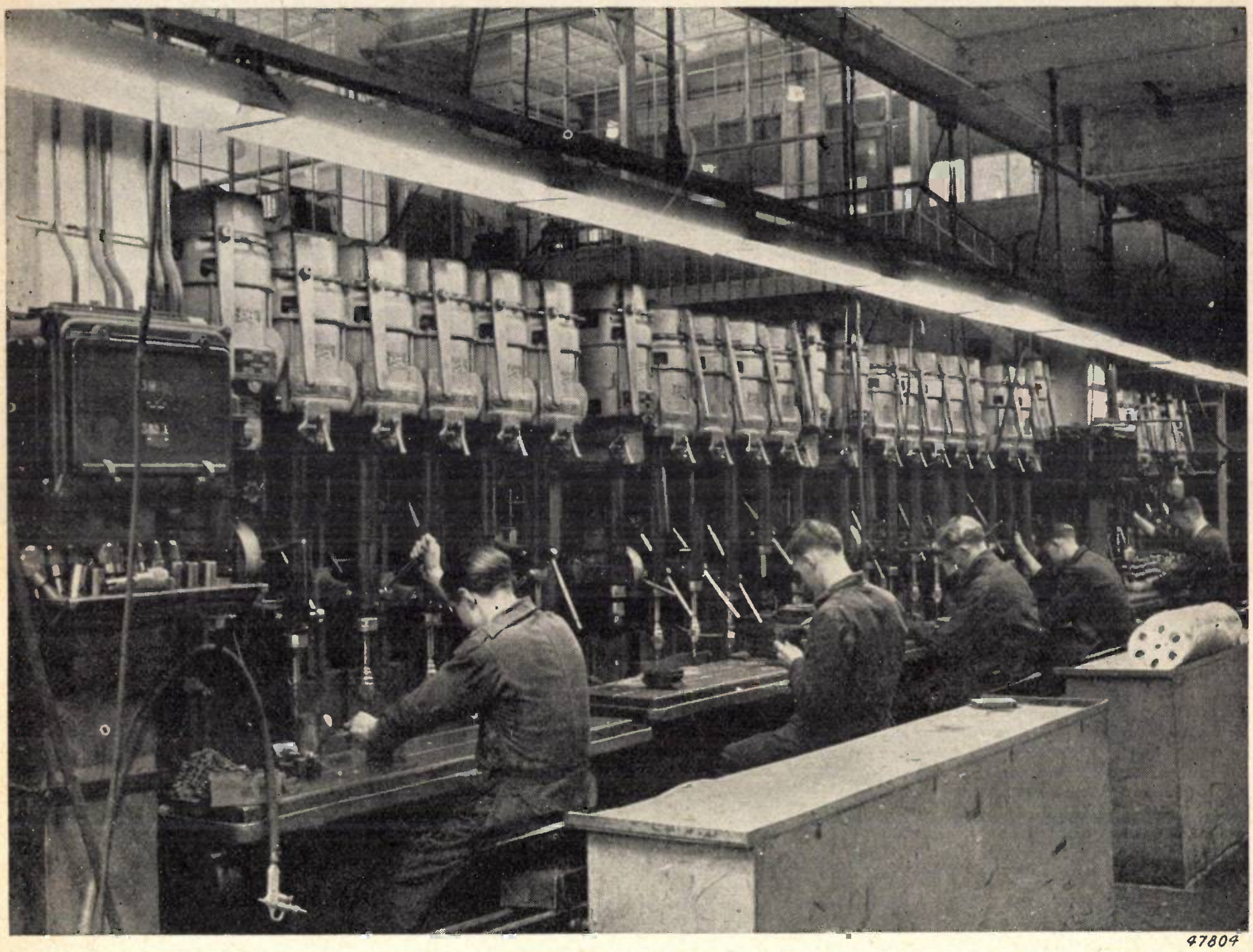
After the first period already the deviation has fallen to one-tenth of the initial displacement, while after two periods only $1^1/_2\%$ of the initial amplitude remains. This very favourable behaviour shows that our choice of parameters was a good one, although of course we may not assume that we have found the most favourable conditions with respect to the rate of decay of the free vibrations [7]).

If, under the same initial conditions, we likewise investigate the free vibration of the elementary system with relative damping and with the same maximum in the frequency characteristic (case 3), we find that in this case the system comes to rest about one and a half times as quickly.

When it comes to putting the theory here developed into practical application for the construction of vibration-free mountings, one is faced immediately with the fact that as a rule the foundation of an apparatus may vibrate in different directions. We hope to discuss this more general case in a following article in this periodical, where at the same time we shall deal with the practical construction in more detail.

[7]) Closer consideration shows, for example, that in general at $p = \mu$ and $q \approx \sqrt{4\mu\,(1\text{-}\mu)}$ the system comes to rest somewhat more quickly than with our "optimum" values of the parameters $p = \mu$ and $q = \sqrt{1.5\,\mu\,(1\text{-}\mu)}$. These considerations, which would take us too far afield here, will be published elesewhere.

# FACTORY LICHTING WITH GAS-DISCHARGE LAMPS



47804

Several articles have appeared in this journal from time to time dealing with the development, properties and applications of gas-discharge lamps.

The photograph reproduced here shows the lighting of 5 groups each of 5 drilling machines with gas-discharge lamps, type TL 100, mounted in metallic reflectors suspended $1^1/_2$ meters (abt. 4′10″) above the working plane. The intensity of light on the working plane is 100-150 lux, which is amply sufficient. This replaces the lighting system with movable reflectors fitted with incandescent lamps, and it is much more satisfactory.

# STABILISED AMPLIFIERS

by J. J. ZAALBERG van ZELST.

621.394.645.3

At a given frequency the degree of amplification of an amplifier depends upon the properties of the valves used (particularly their slope) and the impedances of the other elements in the circuit. Assuming that the latter are fairly constant (given the right choice of material and proper construction, it is then a matter of designing the circuits in such a way as to be dependent as little as possible upon the properties of the valves, which may vary according to the amplitude of the anode voltages, the temperature, the contact potentials, etc., or when a valve is replaced. This is of great importance, for instance, when taking measurements.

In this article two groups of circuits are discussed (each of which may be divided into two sub-groups) which tend to provide for a high degree of constancy in the amplification. Some of these circuiting schemes need very few extra parts. Sometimes there is the additional advantage of reduced distortion.

In many cases combinations of the various methods are possible, resulting in an exceptionally constant amplification.

## Introduction

In many cases, for instance for measuring purposes, it is desired to have an amplifier that does not change with time, even though variations may occur in such factors as anode voltages, contact potentials, temperature, etc. which may affect the amplification directly or indirectly. This demand should not be taken too literally; if the factors referred to remain within certain reasonable limits — limits which in practice are exceeded only in exceptional cases — it is sufficient if the amplification is also kept constant within certain limits narrow enough for the deviations from the nominal value to be negligible for the purpose in view. Here in this article we will deal with the principles that count in the designing of such an amplifier as this, confining ourselves to those cases where the amplifying action is obtained, through the alternating voltage to be amplified between two electrodes of a valve (*e.g.* cathode and control grid) resulting in a current of the same frequency to another electrode of the valve (*e.g.* the anode). We will consider only the influence that the variations of the valve properties have upon the amplification, because the changes taking place in the other elements of a circuit can be kept within certain narrow limits by suitable construction or choice of material.

In the arrangement as found in an amplifier there is for each valve or each set of valves a certain relation between the alternating current generated and the alternating voltage applied, a relation which depends somewhat upon the amplitude but with the usual order of magnitude only to a small extent. For the sake of brevity this relation will here be termed the slope (symbolised by $S$), generalising somewhat the usual meaning of the word. The actual amplification, *i.e.* the ratio of the output voltage to the input voltage, is proportional to that slope.

The methods for improving stability of the amplification may be divided into two groups. In the first group the valves retain their slope as given by circumstances, the improved stability being obtained by adding a compensating quantity to the input or output signal; this can be done outside the amplifier, so that nothing need be altered in the amplifier itself.

This group comprises:

Ia) feedback: part of the output signal is fed back to the input circuit and amplified with it;

Ib) the input signal is equated with a part of the output signal and the difference, after being amplified in an extra amplifier, is added to the output signal.

In the second group of methods the operation of the valves is arranged in such a way that the slope remains constant, using a control voltage derived from an auxiliary alternating voltage of a frequency that does not cause any interference. There are two possibilities, according to the origin of the auxiliary voltage:

IIa) where it is extraneous to the amplifier;

IIb) where it is generated in the amplifier itself.

Each of these methods will now be considered in turn.

### Ia) Adding a compensating quantity to the input signal (feedback)

Much has already been written, also in this periodical [1]), about feedback and in particular

[1]) Philips Techn. Review 1, 268, 1936 and 2, 289. 1937,

about the special form of feedback termed **negative feedback**, which we have mainly in mind here. Nevertheless, it is well to recall briefly the principle of this system.

The output current $I_a$ of an amplifier $A$ (*fig. 1*) or the output voltage as the case may be, or a part of it, is conducted to a network $T$ called the feedback circuit. The output voltage $V_t$ of this circuit, together with the signal $V_i$ to be amplified, forms the input voltage $V_g$ of the amplifier. One speaks of **negative feedback** when the phase difference between $V_i$ and $V_t$ is such that the amplitude of $V_g$ is less than that of $V_i$.

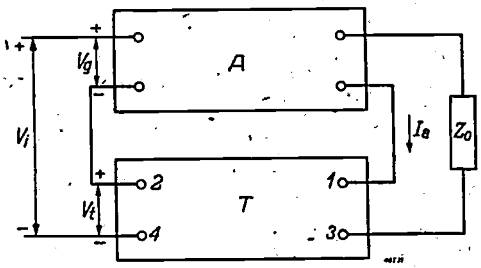


Fig. 1. The output a.c. current $I_a$ from an amplifier $A$ is conducted to the input terminals *1-3* of a feedback system $T$. At the output terminals *2-4* there consequently arises a voltage $V_t$ which together with the input signal $V_i$ forms the voltage $V_g$ supplied to $A$. The + and — signs at the terminals indicate in what direction the a.c. voltages $V_i$, $V_t$ and $V_g$ are positive. $Z_o$ is an external impedance.

The amplifier $A$ may be characterised by the slope $S$:

$$I_a = SV_g, \quad \dots \dots \quad (1)$$

the feedback circuit $T$ by a **transfer impedance** $Z$:

$$V_t = ZI_a = (R + jX)\, I_a \quad \dots \quad (2)$$

The terminals *1* and *2* of the circuit $T$ may be coincident; equally so the terminals *3* and *4*. In that event $Z$ is simply the impedance connected between the two points *1-2* and *3-4*. In more complicated cases the network as shown in fig. 1 has **four poles** and $Z$ then indicates the ratio of the output voltage to the input current of this four-polar system; one then speaks of a **transfer impedance**.

Finally, with the positive direction of the voltages indicated in fig. 1 we have the relation

$$V_g = V_i - V_t \quad \dots \dots \quad (3)$$

In the absence of feedback ($Z = 0$; thus $V_t = 0$) the slope $S'$ of the whole system is identical with that of the amplifier itself:

$$S' = \frac{I_a}{V_i} = \frac{I_a}{V_g} = S.$$

Where feedback is applied we find however:

$$S' = \frac{I_a}{V_i} = \frac{SV_g}{V_i} =$$

$$= S \cdot \frac{V_i - V_t}{V_i} = S\left(1 - \frac{ZI_a}{V_i}\right) = S\,(1 - ZS')$$

or

$$S' = \frac{1}{Z + \dfrac{1}{S}} = \frac{1}{\dfrac{1}{S} + R + jX},$$

hence

$$|S'| = S_{eff} = \frac{1}{\sqrt{\left(\dfrac{1}{S} + R\right)^2 + X^2}}, \quad \dots \quad (4)$$

where $|S'|$ is denoted by $S_{eff}$.

The question is now how $R$ and $X$ can best be chosen so that certain variations of $S$ have the minimum influence upon $S_{eff}$ without the latter becoming appreciably smaller than the original slope $S$. It is to be borne in mind that the components $R$ and $X$ of the transfer impedance $Z$ may be negative, whilst the four-polar system $T$ need not necessarily have any negative resistance, self-inductances or capacities (an example of such a case will be given presently). Consequently the aim will be to give $R$ such a negative value as will just compensate the mean value of $1/S$. Supposing that through some cause or other $S$ fluctuates between the limits $S_{min}$ and $S_{max}$, one will then choose

$$R = -{}^1\!/_2 \left(\frac{1}{S_{\max}} + \frac{1}{S_{\min}}\right) \quad \dots \quad (5)$$

By this means the first term in the denominator of equation (4) is caused to disappear as far as possible, with the result that 1) variations of $S$ — which factor occurs only in this term — have the least possible influence and 2) for a given value of $X$ the denominator of (4) is reduced to the lowest possible value and consequently the amplification becomes as high as possible.

By substituting (5) in (4) one finds that the effective slope will lie between the limits

$$S_{eff\,\max} = \frac{1}{|X|}$$

and

$$S_{eff\,\min} = \frac{1}{\sqrt{\dfrac{1}{4}\left(\dfrac{1}{S_{\min}} - \dfrac{1}{S_{\max}}\right)^2 + X^2}} \quad (6)$$

Such a variation can be made relatively as small as one desires by choosing $|X|$ only just large enough; as $|X|$ is increased, however, so the amplification is reduced, and one must therefore find the compromise best suitable for each particular case. Given a sufficiently low value of $|X|$, $S_{eff}$ becomes greater than the original $S$. According as $S_{eff}$ is greater or smaller than $S$, so one gets positive or negative feedback respectively. As may be calculated with the aid of eq. (6), when applying negative feedback the sacrifice in amplification is accompanied by a large gain in stability; with a weak positive feedback there is much less gain and a strong positive feedback even results in a loss of stability. This will be illustrated by a concrete example.

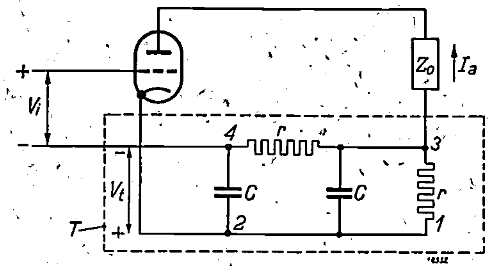


Fig. 2. Example (see footnote [2]) of an amplifier with feedback system $T$ consisting of two resistances $r$ and two capacitors $C$. The numbering *1-4* of the terminals of $T$ corresponds to that of fig. 1.

*Fig. 2* shows the circuit diagram [2]) of an amplifier where the feedback circuit consists of a combination of two resistances and two capacitors. For the sake of simplicity it is assumed that both resistances have the value $r$ and both the capacitors the value $C$, though this is by no means essential.

As a simple calculation will show, for an angular frequency $\omega$ we get for this circuit

$$R = \frac{1-p^2}{1+7p^2+p^4}\,r,\quad X = \frac{-3p}{1+7p^2+p^4}\,r, \qquad (7)$$

where $p = \omega Cr$.

If, for example, an average slope of 5 mA/V is subject to variations of + and — 10%, we find from eq. (5) for $R$ the value —202 ohms. From (7) it appears that to get negative values of $R$ it is necessary that $p$ should be greater than unity. If one chooses for instance $p = 2$ and the frequency of the signal to be amplified is say 1000 c/sec, it then follows from (7) that $r = 3030$ ohms, $C = 0.105$ μF, $X = -404$ ohms. $S_{eff}$ is then 2.5 mA/V ± 0.064%, so that by sacrificing only a factor of 2 in slope an enormous gain is obtained in stability.

It would also have been possible to leave the slope unaltered, by choosing $X = 1/S$, thus $X = -200$ ohms. From eq. (7) it follows that in that case we have to take $r = 4000$ ohms and $C = 0.131$ μF. Eq. (6) shows that the result is then $S_{eff} = 5$ mA/V ± 0.25%, which for many purposes is still quite satisfactory.

Any gain in amplification can only be realised at the cost of gain in stability. In the example just given, for instance, the average slope could be increased say by a factor of 5, thus to 25 mA/V, by choosing $X = -38.2$ ohms, taking $r = 54\,500$ ohms and $C = 0.047$ μF (again for 1000 c/sec), but with the assumed 10% variation in $S$ there would still be a fluctuation of 6.3% in $S_{eff}$, so that in this respect there is no improvement worth mentioning. If a still stronger positive feedback were to be applied $S_{eff}$ would in fact fluctuate much more than $S$.

In the designing of the feedback circuit care is to be taken to avoid oscillation, which would be undesirable. If there is a freqeuncy for which the value $Z$ assumes the proportions of $1/S$ then the amplifier will start oscillating in that frequency.

This can be explained as follows: Suppose that the output of the feedback system $T$ and the input of the amplifier $A$ (*fig. 3*) are disconnected from each other for a moment while there is no signal to be amplified. Upon applying a voltage $V_{t1}$ to the amplifier an anode current $I_a = SV_{t1}$ is generated, which in turn supplies a voltage $V_{t2} = ZI_a = ZSV_{t1}$ to the output of the system. The condition for $V_{t2} = V_{t1}$ is therefore

$$ZS = -1 \qquad (8)$$

If that condition is satisfied and $V_{t1}$ were of that frequency, then upon $T$ and $A$ being connected again the situation would remain as it was; in other words the amplifier would oscillate.

With the feedback system of fig. 2 there is no risk of oscillation, for the limit of $ZS = -1$ for oscillation can only be reached when $Z$ is the real value (assuming that $S$ is a real value, which is usually the case), thus when $X = 0$, which according to eq. (7) is only the case for $\omega = 0$ and $\omega = \infty$ (moreover with $\omega = 0$ according to eq. (7) $R$ is

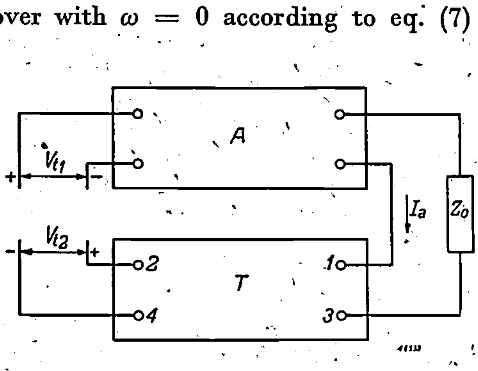


Fig. 3. When for a certain frequency the transfer impedance $Z = V_{t2}/I_a$ of the feedback system $T$ assumes the value $-1/S$ (eq. (8)) then the voltage supplied by $T$ is just identical with the voltage required at the input of the amplifier $A$ to produce $V_{t2}$. When the output of $T$ is connected to the input of $A$ the amplifier begins to oscillate in that frequency.

[2]) For the sake of clarity the sources of grid and anode voltage have been omitted in figs. 2 and 8.

positive, *viz.* equal to $r$, so that with this solution equation (8) $ZS = -1$ is not satisfied.

Finally it is to be remembered that when applying negative feedback the non-linear distortion may be reduced [3]).

Summarizing, we come to the conclusion that this method of feedback answers the purpose and is simple, with little or no sacrifice of amplification, and that it may lead to reduced distortion. It is to be borne in mind, however, that both the amplification and the improvement of stability are functions of the frequency, so that as a rule this method is less suitable if the signal to be amplified covers a wide frequency band. Furthermore, unless the feedback system is properly circuited there is a risk of troublesome oscillation.

### Ib) Addition of a compensating quantity to the output signal

Briefly the principle of this method lies in the output voltage of an amplifier being reduced, by means of, for instance, a potentiometer or a transformer, in a ratio equal to the amplification required (which we shall call $A_o$), the fraction of the output voltage thus obtained being compared with the input voltage. If the amplification were exactly $A_o$ then that fraction would be equal to the input voltage. Any difference is conducted to a separate amplifier having an amplification factor $A_o$. Combination of the output voltage of the two amplifiers then produces a voltage exactly $A_o$ times the input voltage.

When, instead of the output voltage, the anode current is taken as the output signal, then one proceeds as follows: with the help of a transfer impedance of the order of $1/S_o$ one derives from the anode current $I_{a1}$ of the main amplifier with the desired slope $S_c$ a voltage that is equated with the input voltage. Any difference between these two voltages results in an anode current $I_{a2}$ in the anode circuit of the auxiliary amplifier (slope $S_o$), and the sum of the two anode currents is exactly $S_o$ times the input voltage.

Schematically this could be represented as follows: let $S_o + \varDelta S$ be the actual slope of the main amplifier, then

1) the input voltage $V_i$, from which we start, supplies

2) in the main amplifier an anode current

$$I_{a1} = (S_0 \pm \varDelta S) \cdot V_i.$$

3) With the aid of a transfer impedance $1/S_c$ we derive from this a voltage

$$\frac{S_0 \pm \varDelta S}{S_0} V_i = \left(1 \pm \frac{\varDelta S}{S_0}\right) V_i$$

4) and compare this with the input voltage. The difference is

$$\mp \frac{\varDelta S}{S_0} V_i.$$

5) This difference is conducted to the auxiliary amplifier with slope $S_o$, which therefore supplies the anode current

$$I_{a2} = \mp V_i \varDelta S,$$

6) which, added to the anode current $I_{a1}$ of the main amplifier, produces just the desired output current $S_0 V_i$.

It may be thought that this is only transferring the difficulty to the auxiliary channel, the amplification of which was assumed to be $A_o$ and the slope $S_o$ but which may, of course, likewise be subject to variations. It must not be forgotten, however, that the auxiliary amplifier only supplies a compensating current, so that any fluctuations that may occur in this will have very little effect upon the ultimate result. This will be made quite obvious from a numerical example.

Suppose that a signal of 1V is required to yield a current of 100 mA, but that the slope of the main channel happens to be only 95 mA/V. The output current of 95 mA is conducted through a transfer impedance of 10 ohms, so that at the output terminals of that impedance we get a voltage of 0.95 V. The difference between this and the original signal, 0.05 V, is conducted to the auxiliary amplifier. If the latter has a slope of exactly 100 mA/V it will produce just the current of 5 mA lacking at the output of the main amplifier, but if for instance that slope should be 90 or 110 mA/V the final result would be 95 + (90 or 110) · 0.05 = 99.5 or 100.5 mA. Consequently a deviation of 10% in the slope of the auxiliary amplifier results in an error of only 0.5% in the total amplification.

This can be achieved both for a narrow and for a broad frequency band, according to the dimensions of the coupling elements (in contrast to the first method, which does not lend itself so well for a broad bandwidth).

It is easily realised that with this second method it is also possible to counteract the non-linear distortion. For instance a peak cut off in the main amplifier through over-loading is supplemented from the auxiliary channel.

[3]) See the articles referred to in footnote [1]).

It will be equally obvious that if desired a second or third auxiliary channel can be employed to reach still greater accuracy. *Fig. 4* gives an example of a system with one auxiliary channel (for details see the text below the diagram), but all sorts of variations are possible, which it is not necessary to enter into here.

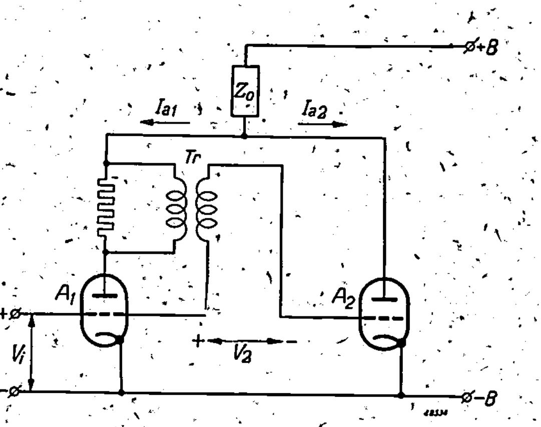


Fig. 4. The secondary voltage $V_2$ from the transformer *Tr* is just equal to the input voltage $V_i$ when the slope of the main amplifier valve $A_1$ has the nominal value. Any deviation from that nominal slope causes a difference to arise between $V_i$ and $V_2$, which is then conducted to the auxiliary amplifier valve $A_2$. The slope of $A_2$ is such that the anode current $I_{a2}$ just compensates the surplus of deficit in the anode current $I_{a1}$, so that a constant current flows through the external impedance $Z_a$. $+B$ = positive, $-B$ = negative pole of the anode voltage source.

Summarizing briefly, the pros and cons of this method balance out as follows:

Advantages: no amplification loss in the main channel; greatly reduced distortion; no risk of oscillation; suitable for a broad frequency band.

Disadvantages: requires some additional circuiting elements, including at least one amplifying valve (the auxiliary channel, however, only need be dimensioned for a much smaller power than the main amplifier).

### IIa) Controlling the slope with a separately generated auxiliary voltage

In this group of circuiting systems an auxiliary signal of a non-interfering frequency is supplied to the amplifier together with the main signal. On the output side the auxiliary signal is filtered out, rectified and smoothed, and the resultant d.c. voltage, after deduction of a fixed amount, is utilised as control voltage for the amplifying valve or valves.

*Fig. 5* gives a diagram of an amplifier stage for high frequency. $HF_i$ and $HF_o$ are respectively the input and output terminals. An auxiliary voltage with low frequency is supplied to the transformer $Tr_1$. The voltage across the winding *sec 1* when rectified yields the d.c. voltage $V_1$. The low frequency a.c. voltage of the winding *sec 2* is amplified together with the h.f. signal and produces — *via* the tuned transformer $Tr_2$ and rectified by the diode $D_2$ — the d.c. voltage $V_2$.

The d.c. voltages $V_1$ and $V_2$ are both taken up in the control grid circuit of valve *A*, in such a way that $V_1$ works in a positive sense and $V_2$ in a negative sense. The ratio of the a.c. voltages from *sec 1* and *sec 2* and the transforming ratio of $Tr_2$ are of such dimensions that if the slope of *A* is of the right value $V_1$ and $V_2$ just compensate each other. If, however, the slope of *A* should decrease then $V_2$ drops, making the control grid voltage of *A* less negative, so that the valve operates in a part of the response curve where the slope is larger; with an increasing slope the reverse takes place. In this way any deviation from the normal slope is automatically corrected, at least in part.

Any variations in the amplitude of the l.f. input signal affect $V_1$ and $V_2$ to the same degree and are therefore of no consequence.

Obviously one would think that the l.f. auxiliary signal could be drawn from the a.c. mains supplying the amplifier, but this is undesirable for the following reasons:

Owing to the very low frequency of the mains large capacitors are necessary to smooth out the voltages $V_1$ and $V_2$ sufficiently, and this causes

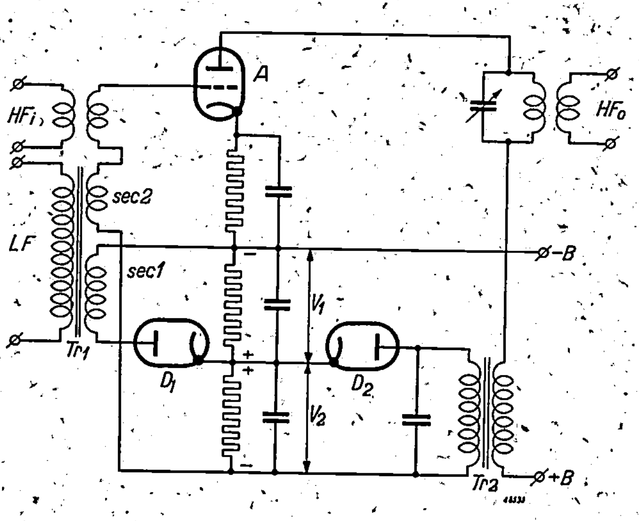


Fig. 5. Amplifier for high frequency (input $HF_i$, output $HF_o$) to which a l.f. auxiliary signal *LF* is applied together with the signal to be amplified. The winding *sec 1* on the transformer $Tr_1$ supplies a voltage $V_1$ which, after being rectified by the diode *D*, acts positively upon the control grid of the amplifying valve *A*. The voltage from winding *sec 2* is amplified together with the main signal and *via* the transformer $Tr_2$ (tuned to the low frequency) and the diode $D_2$ gives a rectified voltage $V_2$, which just neutralizes $V_1$ when the amplification is of the nominal value. In case of any deviation from the nominal value then the difference between $V_1$ and $V_2$ acts as a compensating control voltage. $+B$ and $-B$ are poles of the anode voltage source.

such a lag in the working that the control cannot respond quickly enough to sudden fluctuations in the mains voltage, with the result that there may be temporarily considerable deviations from the nominal amplification. For that reason it is preferable to employ an auxiliary signal with a higher frequency, say of the order of 1000 c/sec. The drawback of having to generate this signal separately is overcome by the following method, where the amplifier itself produces the auxiliary voltage.

### IIb) Controlling the slope with an auxiliary voltage generated in the amplifier

From the circuiting system of fig. 5 it is only a short step to the much better solution of causing the amplifier itself to oscillate in the desired non-interfering auxiliary frequency. For this purpose a signal has to be drawn from the output side and fed back to the input *via* a transfer impedance, in the manner described under Ia) except that whereas with that method the signal fed back has the frequency of the signal to be amplified and it may on no account be allowed to oscillate, in this case the amplifier does oscillate, preferably with a frequency differing from that of the signal to be amplified. In this state of oscillation we have

$$SZ = -1, \quad \ldots \ldots \ldots \quad (8)$$

so that the slope $S$ is fixed, since it must equal a given admittance $-1/Z$. If, therefore, just this admittance has been chosen equal to the desired value of the slope, the occurrence of oscillation proves that the slope is indeed of that value. Deviations from that value are corrected by a control voltage.

Whereas with the method IIa) this control voltage had to be taken from the amplified auxiliary voltage by equation with a fixed voltage, here the rectified oscillation voltage itself is used as control voltage. Thanks to this, smaller auxiliary voltages can be employed, so that there will be much less modulation of the auxiliary signal upon the h.f. signal, which in the circuiting of fig. 5 always occurs more or less. The circuiting, too, can be much simpler, as will be evident presently.

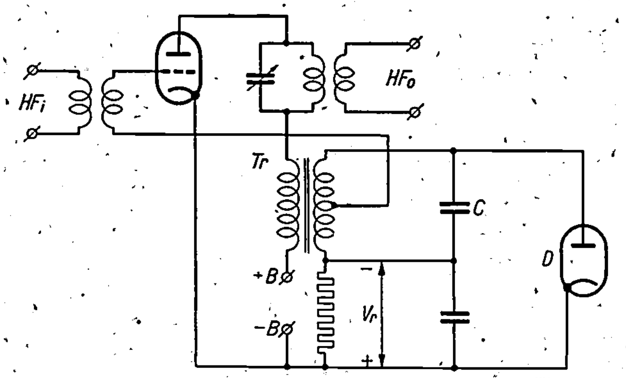


Fig. 6. H.f. amplifier (input $HF_i$, output $HF_o$) analogous to that of fig. 5 but with the l.f. auxiliary signal generated in the amplifier itself (oscillating circuit with feedback, consisting of transformer $Tr$ and capacitor $C$). The l.f. voltage is rectified by a diode $D$ and forms directly the control voltage $V_2$, used for compensating deviations in slope. $+B$ and $-B$ are poles of the anode voltage source.

The question may now arise whether an exactly constant amplification is really attained in this way. This would indeed be the case if $S$ were not dependent upon the signal amplitude, or, in other words, if the valve characteristic were perfectly linear in the working zone. Actually the valve characteristic is more or less curved; the quantity $S$ in eq. (8) is a sort of average slope of the part of the characteristic traversed, and therefore more or less dependent upon the amplitude. To this extent the ampli-

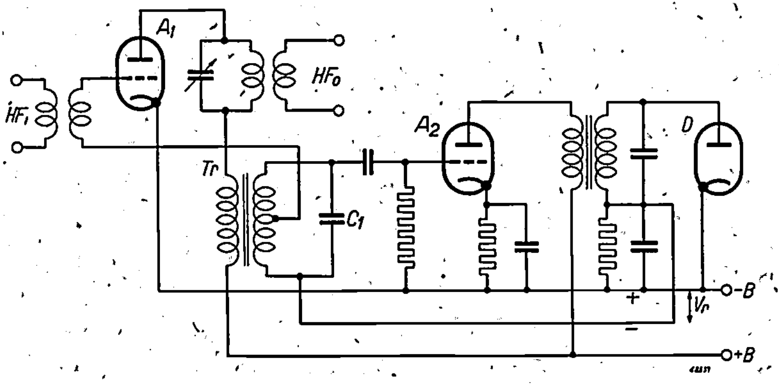


Fig. 7. A variation of the system of fig. 6. The l.f. auxiliary signal generated in the oscillating circuit $Tr$-$C_1$ is amplified by the auxiliary valve $A_2$ before being rectified by the diode $D$ into the control voltage $V_r$. The advantage here is that the amplitude of the voltage across the circuit $Tr$-$C_1$ and also the fraction of it fed to the control grid of the main valve $A_1$ can be kept very low, with a correspondingly reduced chance of modulation of the l.f. signal upon the h.f. one. $+B$ and $-B$ are poles of the anode voltage source.

fication is not absolutely constant, but with small amplitude the deviations will be only very small.

In the case where the characteristic is absolutely or practically straight the control voltage on the control grid has of course little or no effect, for the slope is (practically) independent of the bias on the control grid. The slope can be influenced, however, by varying a d.c. voltage on the third grid (suppressor grid) of the valve, which must then be a pentode. In such a case, therefore, the control voltage should be applied to the suppressor grid.

We will now give some examples of systems where this method is applied. In *fig. 6* $HF_i$ and $HF_0$ again represent respectively the input and output terminals of the h.f. signal. The primary coil of the transformer *Tr* is taken up in the anode circuit of an amplifying valve, while the secondary coil forms a l.f. oscillating circuit with the capacitor *C*; part of the a.c. voltage across this circuit is fed back to the control grid, bringing about the required feedback. The l.f. voltage is rectified by the diode *D* to the control d.c. voltage $V_r$, which, acting negatively upon the control grid, exercises a correcting influence. A comparison of fig. 6 with fig. 5 shows that the former is much simpler.

If it is desired to keep the auxiliary a.c. voltage in the amplifying valve exceptionally low, in order to reduce still further its modulation on the h.f. signal, then the oscillation voltage can be amplified with a separate valve before drawing the control voltage from it. *Fig. 7* gives an example of this, explained in the text underneath.

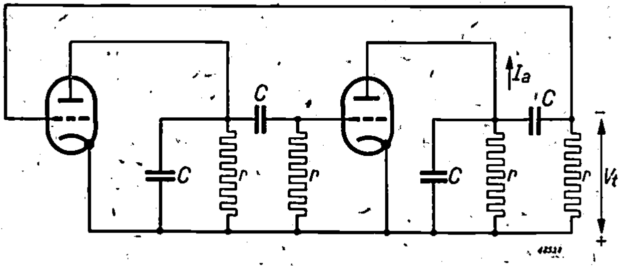


Fig. 8. Example of a system (see footnote [2]) without oscillating circuit, which can still oscillate (so-called *RC* generator).

As already mentioned earlier on, to cause a circuit to oscillate in a certain frequency it is necessary to feed back to the input a voltage with the right phase and amplitude. It is not at all necessary to do this *via* oscillation circuits, for any suitably chosen feedback system consisting only equally well. Especially with low frequencies, self-inductions (*RC* or *RL* systems) can serve of resistances and capacitors or of resistances and

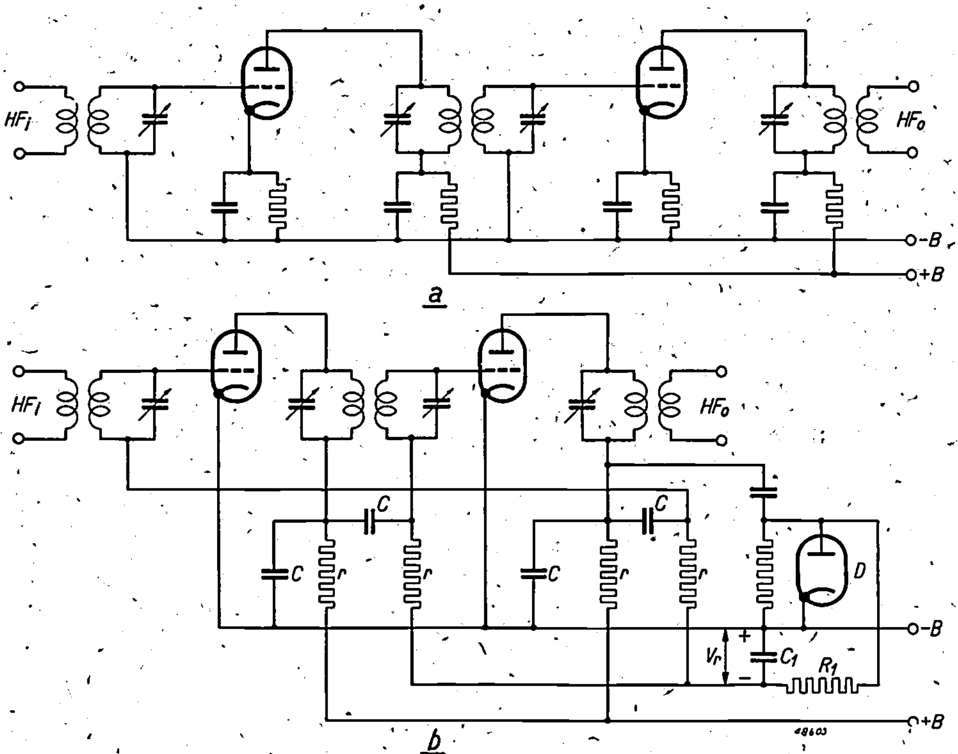


Fig. 9. H.f. amplifier in two stages, *a*) without and *b*) with stabilized amplification. In the latter case the amplifier acts as an *RC* generator on the principle of fig. 8 and thus produces itself the l.f. auxiliary voltage required. By means of the diode *D* and the smoothing circuit $R_1$-$C_1$ a negative grid bias $V_2$ is derived which not only acts as control voltage but also has the function of amplitude limiter. $HF_i$ = input, $HF_o$ = output of the h.f. signal. $+B$ and $-B$ are poles of the anode voltage source. The resistances and the capacitors denoted by *r* and *C* form *RC* systems analogous to that of fig. 8.

where oscillating circuits would involve large and expensive coils, an $RC$ system may be much more economical. Some of these circuiting systems are already familiar as generators of voltages with rectangular or other non-sinusoidal curves, such as Abraham and Bloch's multivibrator. It is less known, however, that if in these circuits only the amplitude is limited in the right way the oscillations remain practically sinusoidal. They are then very usuful for stabilizing the slope of an amplifier in the manner just described.

Of the many circuiting systems that might usefully be employed *fig. 8* gives a simple example (see footnote [2]). The only difference from fig. 2 is that the resistance and capacitor farthest to the left in fig. 2 are changed round and two stages are in cascade connection (with only one stage the voltage $V_t$ would have to be displaced 180° in phase to allow of oscillation). In fig. 8 we get for the transfer impedance $Z = V_t/I_a$:

$$Z = \frac{-3\,p^2 + jp\,(1-p^2)}{1 + 7\,p^2 + p^4} \cdot r, \quad . \; . \; . \quad (9)$$

where $p$ again represents $\omega Cr$. As will be seen, $Z$ can now be made a real value for $p = 1$, for which it assumes the value $-r/3$. Oscillation will therefore occur if $S = 3/r$, and then with the frequency $f = 1/2\pi Cr$.

Inversely, when the circuit does actually oscillate one can conclude that the slope is $S = 3/r$ and is therefore fixed by the resistance value $r$.

Limitation of the amplitude can be effected in a simple manner by rectifying and smoothing the a.c. voltage generated and using the resultant d.c. voltage (or a part of it) as grid bias for the oscillating amplifier valve. One finds this applied in the two-stage amplifier for high frequency shown in *fig. 9b* (on the principle of fig. 8). A comparison with fig. 9*a* of the same amplifier without stabilized amplification shows that only a few, inexpensive components are required for stabilizing, and there is no loss of amplifying power.

In conclusion it is to be observed that two or more of the methods described here can be combined for raising the degree of stability in amplification to an exceptionally high level.

VOL. 9 No. 2, p. 33-64 1947

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 ELECTRON MICROSCOPE WITH CONTINUOUSLY VARIABLE MAGNIFICATION

by J. B. le POOLE.

621.385.833

Several articles concerning electronic microscopes will be published in this periodical. In this first article of the series some characteristics of ordinary optical microscopes are reviewed. The wave nature of light sets a limit to the resolving power. The smallest distance between two distinguishable details is 1000 Ångström units with the best optical microscopes. A significant improvement in resolving power is obtained by using electron beams instead of light. The principle of the electron microscope is first explained with particular reference to the focusing of electron beams by magnetic lenses. A description is then given of a new electron microscope now in use at the Institute for Electron Microscopy at Delft. The advantages of the new construction over previous models are explained. The resolving power amounts to about 25 Å and the magnification is continuously variable from 1000 to 80 000 times. With this instrument it is also possible to get an electron diffraction pattern of a part of the specimen which has first been studied electron-optically, which offers the advantage of an easier identification of the materials which are being investigated. In conclusion several applications of electronic microscopes are mentioned.

In recent years investigations have been carried out in many countries with microscopes where use is made of electron beams. This has also been the case in the Netherlands, especially at the Technical University at Delft and in the Philips Laboratory at Eindhoven. It is our intention to devote several articles in this periodical to that subject. In this first article a description is given of the electron microscope that was constructed by the author for the Institute for Electron Microscopy at Delft, which is under his direction. This instrument was completed in 1944 and was then used for one month. It was then taken apart and the parts were hidden to prevent their being carried off by the enemy. Immediately after the liberation the instrument was assembled and taken into use again.

As an introduction to the description of the new microscope we shall first review several characteristics of the optical microscope for purposes of comparison, and then deal briefly with the general principles upon which the functioning of the electron microscope is based.

## The optical microscope

It is a matter of general knowledge that a light microscope contains a condenser lens which concentrates a beam of light on the object, further an objective which forms an enlarged intermediate image of the object, and an ocular with which this image is observed. The total magnification is found by multiplying the enlargement of the objective by that of the ocular.

It is easy to see that the wave nature of light sets a limit to the resolving power, *i.e.* to the smallest distance between two details which can just be distinguished from each other. When a beam of light rays strikes an object they will be stopped by the non-transparent parts. The transparent parts transmit the rays, which, when a lens is placed in their path, converge again in the plane of the image. If, however, the opening between two non-transparent parts is of the order of magnitude of the wavelength of the light used, the light is strongly diffracted and scattered at the other side of the object over a wide angle. Because only a small part of this beam reaches the lens the part of the image corresponding to the opening will still remain dark. In this way it can be understood that the resolving power of an optical microscope is of the same order of magnitude as the wavelength of the light used. This wavelength of visible light lies between 0.8 $\mu$ and 0.4 $\mu$.

This subject is dealt with more precisely in the theory of Abbe for the optical microscope. From this theory it follows that the resolving power

$$d_{\min} = \lambda/2A,$$

where $\lambda$ is the wavelength of the light used and $A$ is the numerical aperture of the objective.

This last quantity is equal to $n \sin u$ where $n$ is the index of refraction of the medium in which the object lies for the light used and $u$ is half the angle aperture of the objective.

In order to prove this we first examine how an image is formed by an objective of a grating $S$ which is struck perpendicularly by a parallel beam of light (*fig. 1*). Diffraction beams will be diffracted by the object in a number of directions.

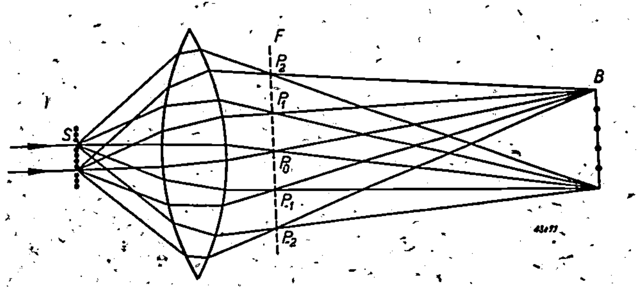


Fig. 1. Formation of the image of a grating $S$ by an objective lens. The deflected rays converge on the focal plane $F$ of the objective to give primary diffraction images $P_0$, $P_1$, $P_{-1}$, $P_2$, $P_{-2}$, etc. The image $B$ is formed by interference of the beams coming from the primary images.

Part of these beams are captured by the objective and lead to the formation of a number of diffraction images (when white light is used: diffraction spectra) in the focal plane of the objective. Abbe calls these the primary images. According to the wave theory these primary images $P$ must be regarded as new sources of light from which issue the waves which form an image of interference in the image plane. Tbis interference pattern is the image which is observed with thc ocular of the microscope.

It is possible to calculate which diffraction beams form the primary images. Let us call the angle which a beam issuing from the grating makes with the main axis $\varphi$. According to the wave theory a large part of the light is now diffracted in very definitc directions $\varphi_k (k = 1, 2, 3 \ldots)$, which are determined by the relation

$$\sin \varphi_k = \frac{k\lambda}{ny} \quad (k = 1, 2, 3, \ldots\ldots),$$

where $y$ is the linear spacing of the grating (the so-called grating constant). The first diffraction beams on either side of the main axis thus have the direction $\varphi_1$, for which is valid $\sin \varphi_1 = \lambda/ny$.

Abbe has proved that an absolutely true image of the object is formed only when all the diffraction beams issuing from the object combine to form the image. This cannot be realized in a microscope. But a very good image is also obtained when a smaller number of beams combine.

With perpendicular illumination, in connection with the symmetrical arrangement, not only the central beam but also at least two others, thus a total of at least three beams, must combine in the formation of the image (*fig. 2a*). Thus the half angle aperture $u$ must be at least equal to the angle $\varphi_1$. This means that $\sin u \geq \lambda/ny$ or $y \geq \lambda(n \sin u) = \lambda/A$.

This argument shows why, with perpendicular illumination, only those structures can be resolved by a microscope objective with the numerical aperture $A$ whose mutual spacing $y$ is at least $\lambda/A$.

Nevertheless, more can be achieved with the same objective. An impression of the grating structure can also be obtained when only two neighbouring beams cooperate, thus for example the "central" one and one of the first two diffraction beams. This can be realized by allowing the light to fall obliquely on the object in such a way that it enters the tube of the microscope just at the edge of the objective (fig. 2*b*). The angle $\varphi_1$ may now be twice as large as in the case just referred to. That means that in this case a grating with twice the fineness of structure can still be observed. With oblique illumination therefore the limit of the resolving power in one definite direction is given by $y = \lambda/2A$.

When a condenser is used not only a unilaterally oblique illumination is obtained, but also a universally oblique illumination. Provided the aperture of the condenser is not smaller than that of the objective, the following is then generally valid:

$$d_{\min} = \lambda/2A.$$

For an optical microscope with oil immersion $A$ is a maximum of 1.5 for $u = 90°$, so that under the most favourable circumstances one finds $d_{\min} = \lambda/3$. For such a microscope, therefore, the resolving power is about 0.2 μ.

The magnification which must be used to see an object of these minimum dimensions clearly is connected with the resolving power of the human eye. If the latter is 0.1 mm, as is usually assumed, a magnification of 500 times would be sufficient. It will be preferable, however, to use a stronger

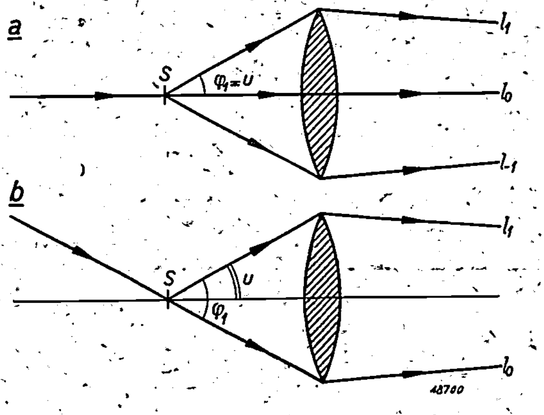


Fig. 2. The formation of primary images with a lens.
*a*) Upon perpendicular illumination the aperture $2\mu$ of the objective must be at least equal to $2\varphi_1$ in order to separate the structure of the grating $S$.
*b*) With oblique illumination an image can still be formed when the aperture is equal to $\varphi_1$.
Here $\sin \varphi_1 = \lambda/ny$, where $\lambda$ is the wavelength of the light used, $n$ the index of refraction of the medium and $y$ the grating constant. The central ray which forms the primary image $P_0$ (fig. 1) is indicated by $l_0$; $l_1$ and $l_{-1}$ are the rays which form the primary images $P_1$ and $P_{-1}$.

magnification than strictly necessary in order not to tire the eye too much and to make every detail easily visible. A magnification of 1000 × or 1500 × is usually used. Greater magnifications are in general of little use with an optical microscope.

The only way of improving the resolving power when the numerical aperture is a maximum is to use light of a shorter wavelength. The use of microscopes for ultra-violet light is based on this fact. Since glass does not transmit ultra-violet rays these instruments must be equipped with quartz lenses. In this way the resolving power is increased by a factor 2.

For the distance between two details which can just be seen separately a limit of 0.1 $\mu$ or 1000 Ängström is then found. In many modern investigations, however, there is need of an instrument with a still higher resolving power.

## Principle of the electron microscope

An important improvement in the resolving power is obtained by using electron beams instead of light.

In 1924 Louis de Broglie announced that a wave nature must be assigned to moving electrons. Experimentally this was confirmed shortly afterwards by interference experiments carried out by the American investigators Davisson and Germer. The wavelength of the electron depends upon its velocity. The greater the velocity the shorter the wavelength. According to de Broglie the wave nature of an electron is characterized by

$$\lambda = \frac{h}{mv},$$

where $h$ is Planck's constant, $m$ the mass and $v$ the velocity of the electron.

When an electron with the charge $e$ passes through a potential difference $V$ its kinetic energy is

$$^1/_2\, mv^2 = eV,$$

from which it follows that

$$v = \sqrt{\frac{2eV}{m}} \text{ and } \lambda = \sqrt{\frac{h^2}{2meV}}.$$

When we substitute the known values in this: $h = 6.6 \times 10^{-27}$; $m = 9.1 \times 10^{-28}$; $e = 4.8 \times 10^{-10}$, e.s.u., expressing $V$ in volts, we obtain

$$\lambda = \sqrt{\frac{150}{V}} \cdot 10^{-8}\,\text{cm} = \frac{12.3}{\sqrt{V}}\,\text{Å} \quad . \quad . \quad . \quad (1)$$

Corresponding to electrons of an energy of 150 kV, we therefore have waves of matter of about 0.03 Å $= 3 \times 10^{-10}$ cm, thus a wavelength of the order of magnitude of hard X-rays.

Since with electron beams wavelengths are thus reached which are, for example, 100 000 times as small as those of the light used for ultra microscopy, theoretically an increase in the resolving power by the same factor is possible when the image of an object is formed with electron rays. This does not mean that this is a practical possibility. In the first place it is necessary that it should actually be possible to form an image with electron rays, and, in the second place, if that can be done, the resolving power attained still depends upon the quality of the lenses to be used for the image formation. In the case of X-rays, which also possess the very short wavelengths mentioned and therefore promise theoretically a high resolving power, the first condition is not satisfied; no medium is known which has a refractive index for X-rays appreciably different from unity. Therefore it is impossible to make "X-ray lenses". Electron rays on the other hand can be refracted and focused by means of magnetic or electrostatic fields, and thus an image can be obtained. The terminology of the optical microscope has been taken over and one speaks of magnetic and electrostatic lenses. The unavoidable errors of these lenses, as a closer theoretical consideration shows, make it impossible to obtain anywhere the limit of the resolving power that would be expected on the basis of the wavelength. With an accelerating voltage of 150 kV it will not be possible to go farther than 5 Å. In practice, at this voltage, a resolving power of 15 to 30 Å has already been reached. A magnetic electron lens is a short coil which thus gives a non-homogeneous magnetic field with rotational symmetry. An electrostatic lens is usually some combination or other of electrodes which gives an electrostatic non-homogeneous field, likewise with rotational symmetry. In both cases the field often has, moreover, a plane of symmetry perpendicular to the axis of symmetry. Concepts such as object distance $a$ and image distance $b$, focal distance $f$ and power $1/f$ of a lens are also defined in electron-optics in the same way as in ordinary optics. The relation between $a$, $b$ and $f$ here is also given by

$$\frac{1}{a} + \frac{1}{b} = \frac{1}{f} \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

One also speaks of thin lenses, meaning lenses so constructed that the thickness of the region within which the influence on the path of a moving electron cannot be disregarded is small compared with the object and image distance. For the power of a thin lens in the electrostatic case the following holds:

$$\frac{1}{f}=\frac{1}{8\sqrt{V_0}}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{V^3}}\left(\frac{\mathrm{d}V}{\mathrm{d}x}\right)^2\mathrm{d}x \quad . \; . \; . \quad (3)$$

and in the magnetic case:

$$\frac{1}{f}=\frac{e}{8mV_0}\int_{-\infty}^{+\infty}H^2\,\mathrm{d}x, \quad . \; . \; . \; . \; . \quad (4)$$

where the $x$-axis is the axis of symmetry, $V_0$ is the accelerating potential, $V$ the electrostatic potential on the axis and $H$ the magnetic field on the axis. The magnification is given in both cases by $b/a$.

The fact that it is possible to focus electron rays is by no means self-evident, especially in the magnetic case where the force acting on the electron depends upon the velocity of the latter. We shall prove this below [1]), for the magnetic lens in a special case, namely for electrons which start from a point of the axis and follow trajectories in the neighbourhood of the axis (paraxial rays). We wish to show that all the electrons issuing from a point $P_1$ on the axis which lies in front of the coil, and whose trajectories make a small angle $\alpha$ with the axis, converge again at the other side of the lens at a point $P_2$ on the axis. In *fig. 3* the trajectory of one of these electrons through the field of the lens is represented diagrammatically. We call its distance from the axis $r$ and its velocity $v$. The small radial velocity of the electron then amounts initially to $v\alpha$. The distances from $P_1$ and $P_2$ to the middle of the lens are respectively $a$ and $b$. From the fact that we confine ourselves to paraxial rays it also follows that the maximum distance $r_0$ to the axis is so small that at that distance the axial component $H$ of the magnetic field is practically the same as on the axis.

As soon as the electron enters the magnetic field it is deflected laterally. It will then travel along a helix around the axis with a gradually changing radius. The change in the radius, which amounts to a movement of the electron towards the axis, is caused by a force which is the result of the axial magnetic field and the angular velocity with respect to the axis. This angular velocity $\dot{\varphi}$ can be calculated in the following way. The change per unit of time in the impulse moment of the electron with respect to the axis is equal to the moment of the Lorentz force with respect to the axis. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}m\,r^2\dot{\varphi}=r\,(e\,v_r\,H-e\,v\,H_r), \quad . \; . \; . \; . \; . \quad (5)$$

where $v_r$ is the radial velocity and $H_r$ the radial magnetic field (calculated as positive in the direction of the axis). Here $v\cos\alpha$ is replaced by $v$.

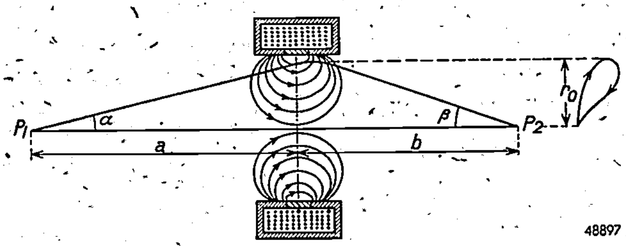


Fig. 3. Diagram of the path of an electron in a thin magnetic lens: projection on a plane through the axis of symmetry (left) and projection on a plane perpendicular to this axis (right). The lines of force drawn characterize the magnetic field of the lens. The electron leaves the axis at $P_1$ and reaches it again at $P_2$. The distances from $P_1$ and $P_2$ to the middle plane of the lens are $a$ and $b$, respectively; the angles which the ray make with the axis at $P_1$ and $P_2$ are $\alpha$ and $\beta$ respectively. The maximum distance of the electron from the axis is $r_0$.

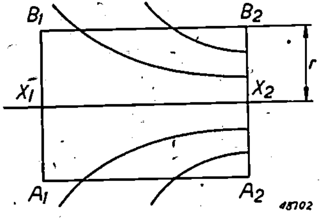


Fig. 4. The magnetic lines of force in a small cylinder whose axis coincides with the axis of symmetry of the magnetic lens. The cylinder with the radius $r$ is shown in cross section. $A_1B_1$ and $A_2B_2$ represent two circular cross sections of the cylinder perpendicular to its axis at the points $x_1$ and $x_2$. The axial components of the field strength at $x_1$ and $x_2$ are $H_1$ and $H_2$ respectively.

On the basis of *fig. 4* we shall now calculate this $H_r$. In the figure $A_1B_1$ and $A_2B_2$ represent two circular cross sections with the radius $r$ taken perpendicular to the axis of the coil at the points $x_1$ and $x_2$. The axial components of the field strength at $x_1$ and $x_2$ we shall call $H_1$ and $H_2$ respectively. The number of magnetic lines of force passing through the cross section $A_2B_2$ but not through $A_1B_1$ is $\pi r^2(H_2-H_1)$. This is the number of lines of force cutting the cylindrical surface with the area $2\pi r(x_2-x_1)$. The average value of the radial magnetic force is therefore $H_r = r(H_r-H_1)/2(x_2-x_1)$, from which, upon passing over to an infinitesimal distance, one finds that

$$H_r=-\frac{r}{2}\cdot\frac{\mathrm{d}H}{\mathrm{d}x} \quad . \; . \; . \; . \; . \; . \; . \; . \quad (6)$$

The above equation of motion (5) thus becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}r^2\dot{\varphi}=\frac{e}{m}\left(r\frac{\mathrm{d}r}{\mathrm{d}t}H+\frac{r^2}{2}\frac{\mathrm{d}H}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}t}\right).$$

The expression between parentheses in the second member is nothing else than the differential quotient of the product $^1/_2\,r^2H$ with respect to time, so that we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}r^2\dot{\varphi}=\frac{e}{m}\frac{\mathrm{d}}{\mathrm{d}t}\,{}^1/_2\,r^2H.$$

From this it follows that

$$\dot{\varphi}=\frac{e}{2m}H \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad (7)$$

We see, therefore, that the angular velocity of the electron is proportional to the axial magnetic field. It becomes zero as soon as the electron has passed through the field of the coil. The electron cannot pass the axis. Behind the coil it moves in a plane which passes through the axis and then, when the field of the coil is strong enough, it must strike the axis due to its inwardly directed velocity.

In order to find the inwardly directed acceleration of the electron it must be noted that a centripetal force $m\dot{\varphi}r^2$ would be necessary to keep the electron at a constant distance $r$ from the axis. But the axial magnetic field and the lateral velocity together cause an inwardly directed Lorentz force:

$$-K_r=e\,\dot{\varphi}r\,H.$$

[1]) A similar proof may be found also in A. Bouwers, Physica 4, 200, 1937.

The resulting acceleration is therefore

$$\ddot{r} = K_r/m + \dot{\varphi}^2 r,$$

or, according to (7)

$$\ddot{r} = -\frac{e^2 r}{4m^2} H^2 \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad (8)$$

For the case of a thin lens, thus where the field acts only over a short distance, $r$ may be considered a constant $r_0$. Equation (8) can then easily be integrated. By setting $\mathrm{d}t = \mathrm{d}x/v$ — which is permissible, because we have confined ourselves to rays in the neighbourhood of the axis — we find that

$$\dot{r} = \frac{v r_0}{a} - \frac{e^2 r_0}{4m^2 v} \int_{-\infty}^{+\infty} H^2 \,\mathrm{d}x \quad . \; . \; . \; . \; . \quad (9)$$

Due to this velocity the electron will reach the axis after the time $\tau = r_0/\dot{r}$. Since it is found from formula (9) that this time is the same for all electrons issuing from a point $P_1$ on the axis, they are indeed all focused at the same point $P_2$, which concludes the proof.

For a thin lens it is now easy to prove formulae (4) and (2). Just as the electron has the radial velocity $\dot{r} = v r_0/a$ before entering the magnetic field, thus between $P_1$ and the lens, after leaving the magnetic field, thus between the lens and $P_2$, it will have the radial velocity

$$\dot{r} = -\frac{v r_0}{b},$$

$b$ being the image distance, *i.e.* the distance between $P_2$ and the lens. When this is substituted in (9) we find that

$$\frac{1}{a} + \frac{1}{b} = \frac{e^2}{4\,m^2 v^2} \int_{-\infty}^{+\infty} H^2 \,\mathrm{d}x,$$

which corresponds to (2) when the expression (4) is substituted for $1/f$ and $^1/_2 mv^2 = eV_0$ is taken into account.

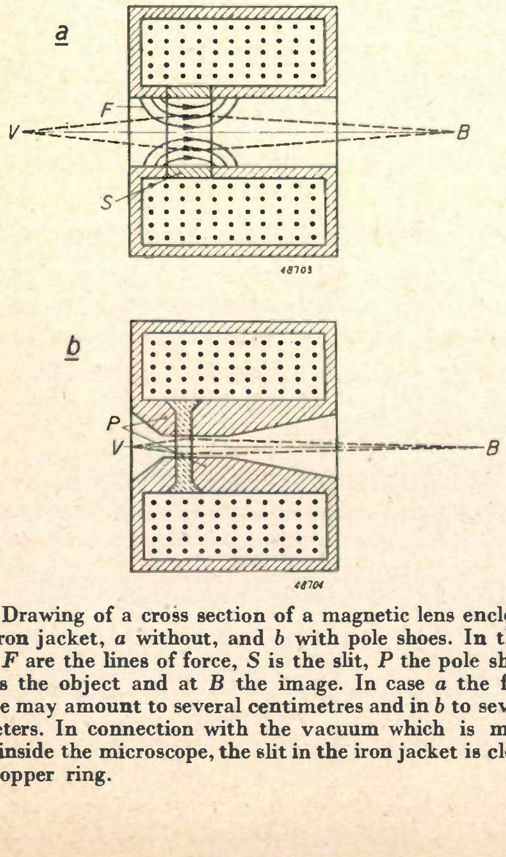


Fig. 5. Drawing of a cross section of a magnetic lens enclosed in an iron jacket, $a$ without, and $b$ with pole shoes. In these figures $F$ are the lines of force, $S$ is the slit, $P$ the pole shoes. At $V$ is the object and at $B$ the image. In case $a$ the focal distance may amount to several centimetres and in $b$ to several millimeters. In connection with the vacuum which is maintained inside the microscope, the slit in the iron jacket is closed by a copper ring.

In order to obtain strong magnetic lenses the coil was encased in an iron jacket, just leaving a narrow annular slit on the inside. *Fig. 5a* is a diagram of this construction. The whole magnetic field is then concentrated on the immediate vicinity of the slit. At the acceleration voltages customary in electron microscopy such iron-clad coils can be used as magnetic lenses with a focal distance of several centimeters. In order to increase the power of such a lens still more pole shoes are introduced at the slit (fig. 5*b*). In this way the focal distance is reduced to not more than a few millimeters. It is easy to understand that with the construction sketched in fig. 5 the total number of ampere turns is a measure of the strength of the field between the pole shoes. By varying the current the strength of the magnetic field is changed and thereby also the focal distance of the lens. A magnetic lens of modern construction is shown in *fig. 6*.

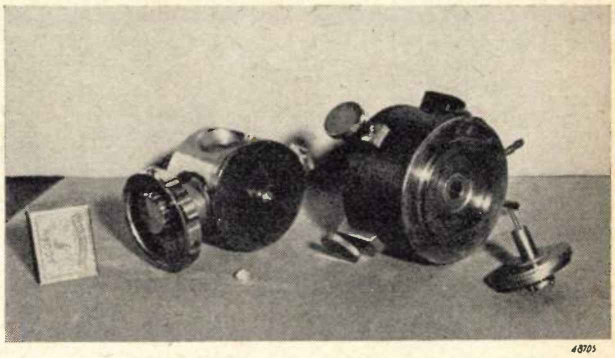


Fig. 6. Parts of a modern electron microscope. On the left a sample holder with air lock. The substance is placed on the point at the middle of the front surface of the holder. On the right an objective, and lying in front of it the holder for the objective diaphragm.

## The arrangement of an electron microscope

The first electron microscopes were constructed in Germany around 1934. Later on this work was continued in other countries. *Fig.* 7 gives a diagrammatic representation of the arrangement of such an instrument with magnetic lenses.

The source of electrons is usually so arranged that the electrons emerging from the cathode get an energy of 50 to 100 kV. They pass through the circular opening in the anode and leave the accelerating tube as a beam with a constant velocity. The magnetic field of a condenser lens concentrates the electrons more or less, as desired, on the object situated close above the objective. In order to obtain a sufficiently narrow beam — the marginal rays blur the image — a physical aperture is usually placed in the condenser.

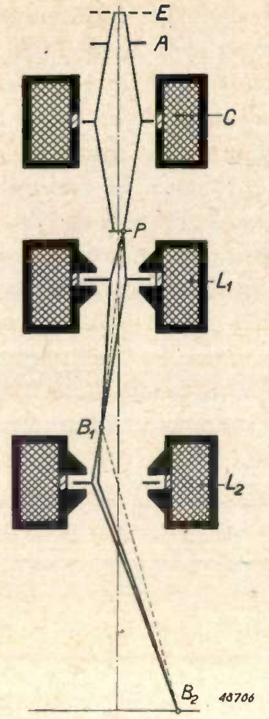


Fig. 7. The path of the rays of a magnetic electron microscope. $E$ is the source of electrons, $A$ the anode, $C$ the condenser, $L_1$ the objective, $L_2$ the projector, $P$ the object, $B_1$ the intermediate image, $B_2$ the final image.

The electrons, which converge on the object within a small area, are scattered by the different parts of the object in varying degrees. In most instruments the rays which are scattered too much in a part of the object strike a diaphragm and thus do not contribute to the formation of the image. At the place in the image corresponding to that part of the object a dark spot is then left. The other electrons which upon passing through the specimen are scattered only over small angles then pass through the objective and are focused into an image by the magnetic field of that lens. It can be shown that the solid angle over which the electrons are scattered is proportional to the mass through which the rays have passed. The contrast effect in the image is thus based upon a difference in mass in the parts of the specimen.

In most microscopes this intermediate image formed by the objective is made visible on a fluorescent screen situated at $B_1$ in fig. 7. There is then a side opening through which it can be observed visually. For this first stage a magnification of 100 to 125× is usually employed.

The above-mentioned screen is situated immediately above a second microscope lens (disregarding the condenser), which is called the projector, and has a hole in the centre. The part of the image lying above this hole passes through the projector and is further magnified on the fluorescent screen at the bottom of the instrument. This final image can be observed through a side tube likewise provided with a glass window. The magnification of the projector, which is only slightly variable in each instrument, lies between 50 and 300× as needed. It is obvious that if desired the last mentioned fluorescent screen can be replaced by a photographic plate.

The whole tube in which the electrons move must be evacuated as completely as possible, since otherwise the electrons would be very much disturbed in their motion. During the observation the pressure must remain below 0.001 mm of mercury. With the help of a so-called air lock a new object can be put into position without seriously disturbing the vacuum.

It is desirable to take measures for the stabilisation of the high tension. Since the focal distance depends upon the acceleration voltage (*cf.* formula (4)), the image would otherwise be blurred owing to fluctuations of the voltage.

From the same formula it follows that the strength of a magnetic lens also depends upon the field strength. From that it follows that it is essential to keep the current in the lenses highly constant, because otherwise their focal distance would vary. This can be achieved by taking care that the lens current is supplied by a source with constant voltage, while water-cooling prevents the temperature and thus the resistance of the coils from changing too much.

Some investigators have preferred electrostatic lenses. The choice between the two systems is determined by the requirements made of the instrument. When a high magnification is desired magnetic lenses offer some advantage over electrostatic ones. Magnetic lenses have been succesfully constructed with a very short focal distance, whereas until now this has proved very difficult with electrostatic lenses. With magnetic lenses the spherical aberration is also less. We shall return to this later.

Further, the strength of magnetic lenses, in contrast to that of electrostatic lenses, can easily be varied from the outside, which is an advantage for adjusting magnification and focusing.

On the other hand, an advantage of electrostatic lenses is that the supply voltage of the accelerating tube and the lenses need not be so rigorously stabilized [2]).

[2]) From the above formula (3) it would even follow that if $V_0$ and $V$ come from the same source (mains voltage) no stabilization is necessary at all. Secondary effects, among which a correction to be introduced on the basis of the theory of relativity, make it necessary, however, that there should still be stabilization (to 0.5%).

The instrument decribed in this article is equipped with magnetic lenses.

### The use of an intermediate lens

In the discussion of the new electron microscope we shall give special attention to the properties which distinguish it from the electron microscopes so far described in literature.

An important improvement is the introduction of a third lens between objective and projector which makes it possible to vary the magnification continuously between wide limits.

In order to make this clear we shall first explain that in a two-stage microscope it is impossible to alter the magnification over a wide range. This is connected with the aberrations occurring under certain circumstances and with the construction of the microscope.

In electron microscopes aberrations occur which are analogous to those in optical lenses. One of the most important of these is the spherical aberration, resulting from the fact that the edge zones of a lens are stronger than the centre. By increasing the field strength of a lens, its focal distance becomes smaller, and with the same angular aperture a part of the magnetic field lying closer to the axis is used. The influence of spherical aberration then becomes less. It is therefore desirable to increase the number of ampere turns as much as possible. A limit is set to this by the magnetic saturation taking place in the pole pieces. Now in order to avoid this, all the dimensions of the lens could be increased, but this raises the difficulties that the lenses then consume too much energy and the microscope assumes too large proportions. A compromise must therefore be sought. A satisfactory solution can be found, at the voltage of 150 kV chosen for our instrument, by using lenses with about 3000 ampere turns. With such lenses a focal distance for the objective of $f = 4.5$ mm has been obtained. For the other lenses a focal distance of 3.5 mm can be attained, because in those cases there are no difficulties connected with the introduction of the specimen.

From calculations of Glaser and Dosse it is found that only about 1/4 to 1/3 of the inner diameter of a magnetic lens may be used if it is desired to prevent distortion, an image aberration closely connected with spherical aberration, from being larger than 5 to 10 percent, a limit which is considered permissible also in the optical microscope. As a result of this limitation the magnification of a magnetic lens in a microscope can be regulated within rather narrow limits, as we shall prove below.

In the case of the projector of the microscope described here the diameter of the bore amounts to 3.5 mm, so that the diameter of the useful field, *i.e.* of the cross section of the beam permissible within the lens, is 1 mm. The diameter of the final screen amounts to 180 mm. If this is to be used to its full extent, therefore, a magnification of 180× is necessary. The projector has therefore been constructed for this normal maximum magnification; with a further increase in power the strength of the projector increases only slightly. It is of course possible to obtain an image with a lower magnification, but then only part of the final screen is filled, so that the possibilities of the instrument are not fully utilised.

The magnification by the objective is entirely fixed. In this case the position of the object is fixed, while the image distance is determined by the position of the object plane of the projector lying immediately above that lens.

In a two-stage microscope the magnification can therefore only be varied within narrow limits.

When it is required of a microscope that the magnification shall be completely controllable, special measures must be taken. It is advisable to find a solution where the projector, which in any case can only be changed to a slight extent, is maintained at a perfectly constant strength. When this is done it is possible to adapt this lens entirely to the magnification it is required to give, *i.e.* to construct it in such a way that with respect to distortion it is corrected to a high degree, while the largest possible field angle is used.

A simple method of varying the magnification is now obtained by adding an extra lens between objective and projector. When the projector is kept constant the aperture of the projector may be considered as the projection screen of the "two-stage microscope" consisting of the objective and the new lens. Because the projector has a high magnification this ocular aperture will be small, even when a reasonably large final screen is provided. We have already stated that in our case the diameter of the bore of the projector amounts to 3.5 mm and consequently that of the ocular aperture amounts to about 1 mm.

The extra lens — which we shall call the intermediate lens — need only fill this small field free of distortion. In the case of the two-stage microscope discussed above (formed by objective and projector) the possibility of regulation was only slight because of the requirement made that the final screen of 180 mm diameter should always be filled. Without that requirement variation of the magnification would also have been possible with that microscope. With the two-stage microscope formed by objective and intermediate lens the requirement is now made that it shall merely fill a "screen" of 1 mm diameter. It will be clear that this can easily be satisfied also with a great variation of magnification.

The introduction of an intermediate lens not only makes it possible to vary the magnification, but also the maximum magnification now considerably exceeds that which is attainable with two lenses. In the case of the microscope described here, objective and projector together give a magnification of 6000 times. By changing the powering of the intermediate lens the magnification can be varied from 6000 to 80 000 times. This total magnification is of course the product of the magnifications of the three separate lenses.

The trajectories when three lenses are used are shown diagrammatically in *fig. 8a.*

Is it now also possible by means of the intermediate lens to vary the magnification from 6000 times to smaller values? Theoretically it certainly is. When the power supplied to the intermediate lens is reduced its magnification decreases. The point of intersection of the rays, seen in fig. 8*a*, between the intermediate lens and the projector then shifts towards the projector. When the power of the intermediate lens is further reduced there comes a moment when the above-mentioned point of intersection of the rays lies in the object plane of the projector. This position is very important, because, as will appear in the continuation of this article, a diffraction image is then obtained. The magnification is then zero. As the power is further reduced the point of intersection falls still lower. Upon passing through the magnification zero the image is inverted. The magnification then increases again until finally no power current at all flows through the intermediate lens and we have once more the two-stage microscope with which we began. In practice, however, there is a serious objection to this method. When the point of convergence of the rays is shifted to a lower level, an increasingly large part of the bore of the intermediate lens has to be used. (The final screen must in any case be filled.) Therefore the distortion increases. At low magnification, therefore, this image aberration becomes much too large. This could be remedied by making the bore of the pole pieces of the intermediate lens larger. Then, however, its minimum focal distance also becomes larger, so that the maximum magnification of this lens is reduced.

A better solution is to use another intermediate lens with a large bore for the range of low magnifications. For these reasons a fourth lens was introduced into the microscope. Since this is also used for making diffraction photographs (this will be discussed later) it is called the diffraction lens. It is situated between the objective and the intermediate lens. When this fourth lens is used the intermediate lens is switched off. By varying the current in the diffraction lens the magnification can be made variable for all values between 6000 and 1000 times. The magnification of the instrument is thus as a whole continuously variable from 1000 to 80 000 times. The lower limit gives the desired transition to the magnification of the optical microscope.

In fig. 8*b* the path of the rays is shown diagrammatically for the case where the diffraction lens is used. This lens is given such a power that the point of convergence of the rays mentioned above lies below the object plane of the projector. The diffraction lens is weaker than the intermediate lens. It does not give a real intermediate image. It refracts more or less the rays coming from the objective and in this way reduces the

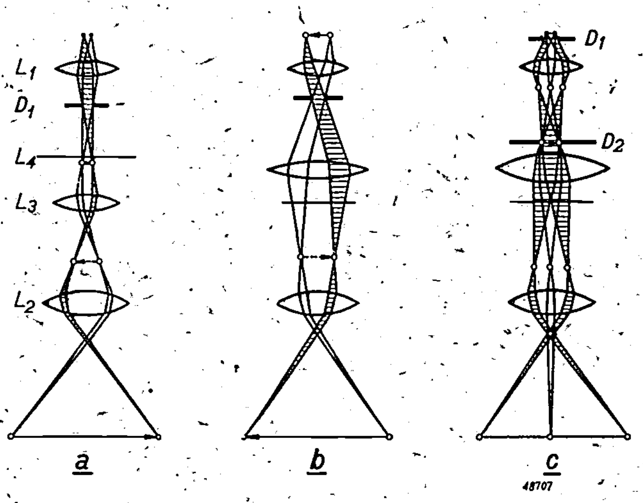


Fig. 8. Diagram of the path of the rays in the electron microscope described in this article.
$L_1$ is the objective, $L_2$ the projector, $L_0$ the intermediate lens, $L_4$ the diffraction lens, $D_1$ the objective diaphragm and $D_2$ the diffraction diaphragm.

*a*) The intermediate lens is in action. This lens converges the rays so strongly that they meet above the projector and then form a new intermediate image of the image formed by the objective lens. The intermediate lens makes it possible to vary the magnification continuously between 6000× and 80 000×.

*b*) The diffraction lens deflects the rays which have passed through the objective, before they have formed an image, and thereby reduces the size of the image. The magnification is variable from 6000× to 1000×.

*c*) The microscope is used for studying a diffraction pattern. As for image formation, the setting is for the magnification 0. The diffraction image, which is formed as "primary image" in the focal plane of the objective, is projected in magnified form on the final screen by the diffraction lens and the projector. The objective diaphragm is now pushed up to allow the passage of the diffraction beams.

magnification as compared with that of the original two-stage microscope.

We shall now first discuss the second function of the diffraction lens.

### The recording of diffraction patterns

In 1927 G. P. Thomson demonstrated in a perfect manner the correspondence between moving electrons and short electromagnetic waves (X-rays), when he showed that it is also possible to obtain the well-known Debije-Scherrer rings when beams of fast electrons are sent through thin foils.

The significance of these diffraction diagrams for the identification of substances and the study of their crystalline state has already been discussed in this periodical [3]).

When a sample of a substance is investigated with an electron microscope it is in many cases desirable to have a diffraction pattern of the substance at one's disposal at the same time.

When a diffraction diagram is to be made with the electronic microscopes so far described in literature the lenses are switched off. But the small diameter of the pole pieces of the projector makes it impossible to record a diffraction pattern immediately, because the diffraction rings are then cut off. Thus either these pole pieces must be removed, which takes some time, or the sample must be placed in an air lock anew under the projector, or an extra plate camera has to be introduced above the projector, which makes the construction much more complicated. Only in the last case is it still possible to find the diffraction image of a part of the specimen of which the electron-optically magnified image has first been observed.

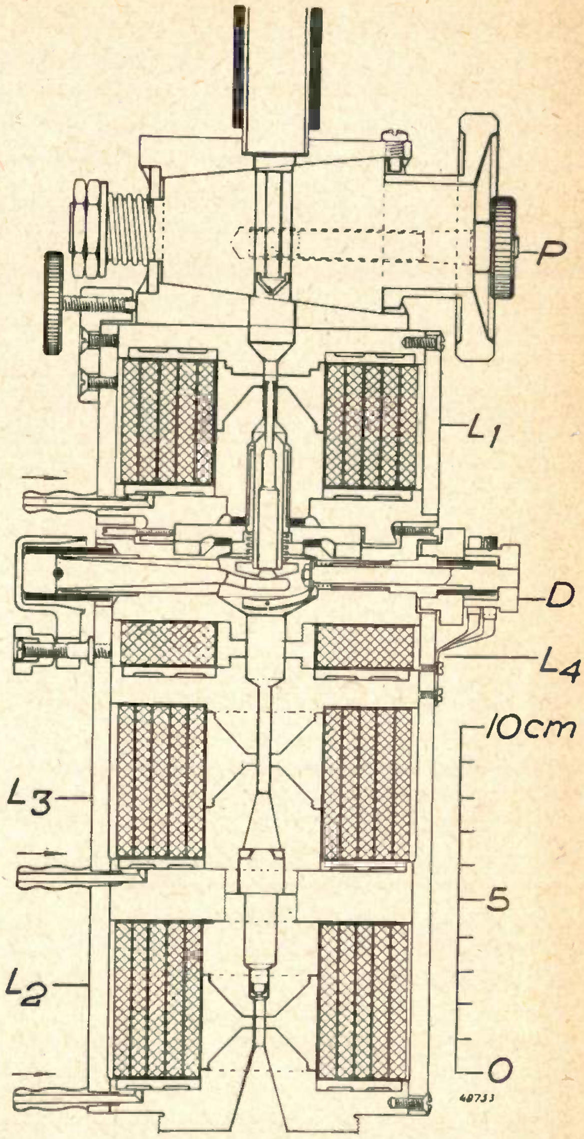


Fig. 10. Drawing of a cross section of the microscope. The magnetic lenses may be seen whose action is described in the text. *P* is the lock for the object, $L_1$ the objective, $L_2$ the projector, $L_3$ the intermediate lens, $L_4$ the diffraction lens, *D* the diffraction diaphragm. Since the magnification is easily variable with the help of intermediate and diffraction lenses, no use is made of the intermediate image formed between objective and projector for seeking the detail of the object to be magnified. This intermediate image is not, therefore, made visible on a fluorescent screen, as is customary in most other constructions.

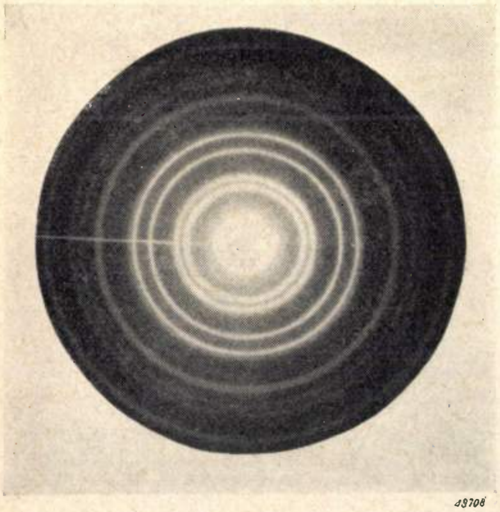


Fig. 9. Diffraction diagram of a gold sol taken with the instrument described in this article.

The introduction of an extra lens as described above makes it possible to simplify considerably the method of taking diffraction diagrams. We then make use of the weak lens called the diffraction lens. With its help the focal plane of the objective is projected on to the object plane of the projector (fig. 8*c*). In this focal plane, as follows from the theory of Abbe, lies the primary image and, with sufficiently wide aperture, therefore also the

[3]) W. G. Burgers, X-rays and electron rays as aids in chemical and metallographic investigation. Philips Techn. Rev. 5, 161, 1940.

diffraction image (*cf.* fig. 1). This diffraction image is very small owing to the small focal distance of the objective, but the projector enlarges it to the size customary for such patterns. *Fig. 9* gives an example of a diffraction pattern obtained in this way.

In order to obtain good image formation with sufficient contrast when using the instrument in a normal way, it is necessary to screen off the rays that are too strongly scattered by the specimen. For that purpose a diaphragm with an aperture of about 70 μ is introduced in the objective, which can be accurately adjusted from the outside. But when diffraction is employed it is just the strongly scattered rays which are needed. For that purpose the objective diaphragm is raised and thus put out of action, the diaphragm in the diffraction lens then being brought into play. In this lens there are two diffraction apertures of different size, which can be used at discretion. When the small aperture is used it is possible to make diffraction diaphragms of very small regions, for when the object is sharply focussed on this diaphragm only those electrons can pass through which come from the corresponding part of the object. In the case of the microscope described here the magnification on the diaphragm is 13 ×. When the aperture of 40 μ is used it is possible to scan the object and to cut out regions with a diameter of 3 μ.

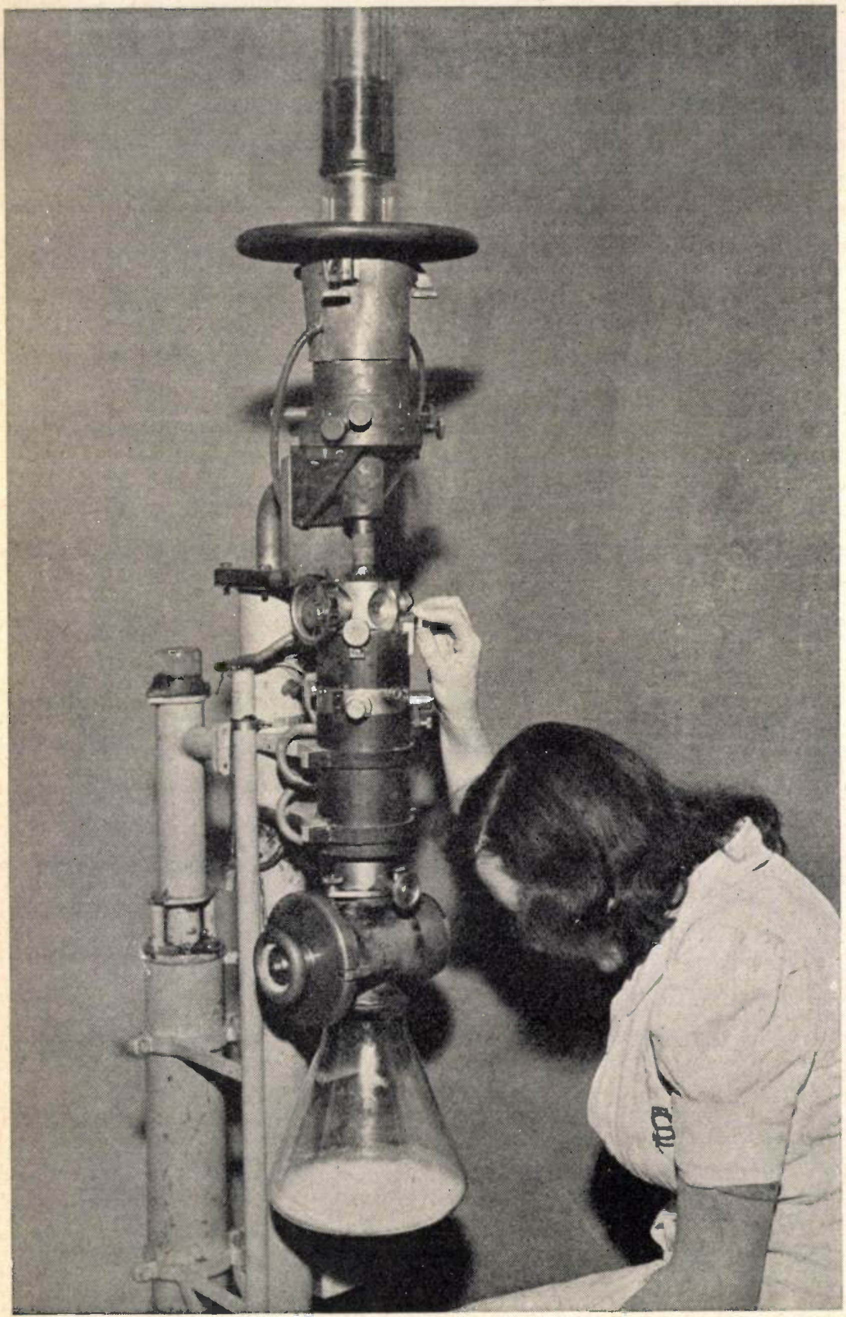


Fig. 11. The electronic microscope for 150 kV in the Institute for Electronic Microscopy at Delft.

The great advantage of this construction is that it is easy to switch over from the electron image to the diffraction image and that one can then observe a diffraction pattern of that part of the object whose image has first been studied microscopically. This furnishes a good method for the identification of the crystals in the specimen being investigated.

### Some further details of the construction

*Fig. 10* shows a cross section of the microscope and *fig. 11* is a photograph of the instrument.

For the maximum accelerating voltage 150 kV was chosen because this gives the electrons a reasonably large power of penetration and the necessary high-tension installation had already been developed by the Philips Physical Laboratory [4]).

The current for the lenses is supplied by an accumulator battery of 35 volts with a sufficiently

[4]) This installation, which furnishes a voltage constant to within $0.2\,^{0}/_{00}$, will be described by A. C. van Dorsten in one of the following numbers of this periodical.

constant voltage. The maximum current per lens is about 1.4 ampere. Each of the lenses is cooled separately. In order to obtain a rapid dissipation of heat, layers of copper are introduced between the windings at regular intervals. The conduction of heat towards the cooled side walls is thus increased, so that a current density of 6.5 A/mm$^2$ became permissible.

The introduction of a third lens between objective and projector also gives the advantage that the length of the microscope tube could be considerably decreased. The total distance from the object to the final image in the new microscope amounts to 600 mm, which is short considering that the diameter of the final image is 180 mm.

The electrons thus have short trajectories, and that also has several advantages. In the first place there is less chance of collision with residual gas molecules in the tube, and in the second place the influence of disturbing fields is less. The path of the rays between the objective and the next lens is most sensitive to these fields. Its length in our microscope amounts to only 10 cm, compared with 30 cm in the model which until now was the best in that respect. Moreover, owing to the compact structure the path of the rays is doubly shielded by the tube itself and by the coil jacket. As a third advantage of a short microscope tube the small volume may be mentioned, which makes it possible to reach a high vacuum more quickly.

The coils of the various lenses are separated by the iron side walls of the jackets. Only between the objective and the diffraction lens is a small space left free in which several essential components are housed: in the first place the three adjusting screws by means of which the objective can be aligned with respect to the other lenses; further a similar arrangement for adjusting the objective diaphragm — here is also located the fork with which the last mentioned diaphragm can be moved up for making diffraction photographs — and finally in this space the diaphragm holder is housed with the two already mentioned apertures of different size for the electron diffraction. The larger or the smaller of these can be used at will, or the passage can be made entirely free. All these arrangements can be operated while the microscope is in use.

In electron microscopy the visual image is usually only used to obtain an impression of the specimen and for sharp focusing. The most important observations, however, are made photographically. A photographic plate or film must then be brought into the path of the rays.

It is desirable to obtain sharp photographs of the whole image. As already stated, the diameter of the final image is 180 mm. It would be difficult to introduce a photographic plate of that size into the vacuum close to the final screen. We therefore use a 35 mm film, which, in order to be able to cover the whole image, is introduced at a spot where the cross section of the beam is still sufficiently small. Because of the small apertures customary in electron microscopy, the depth of focus is more than sufficient to obtain a sharp image at that spot without a new adjustment. In this way the exposures are made with 1/5 of the total magnification.

This must not, however, by any means be considered as a disadvantage. The resolving power of the film is much better than that of the eye. It is therefore possible subsequently to enlarge the small photographs more than five times.

The fact that the camera is placed close to the projector has also the advantage that because of the high current density of the beam impinging on the film a short exposure is sufficient.

It is obvious that in visual observations the camera must be removed from the beam of the rays. It was found possible to introduce a simple arrangement with which the film holder can be tipped

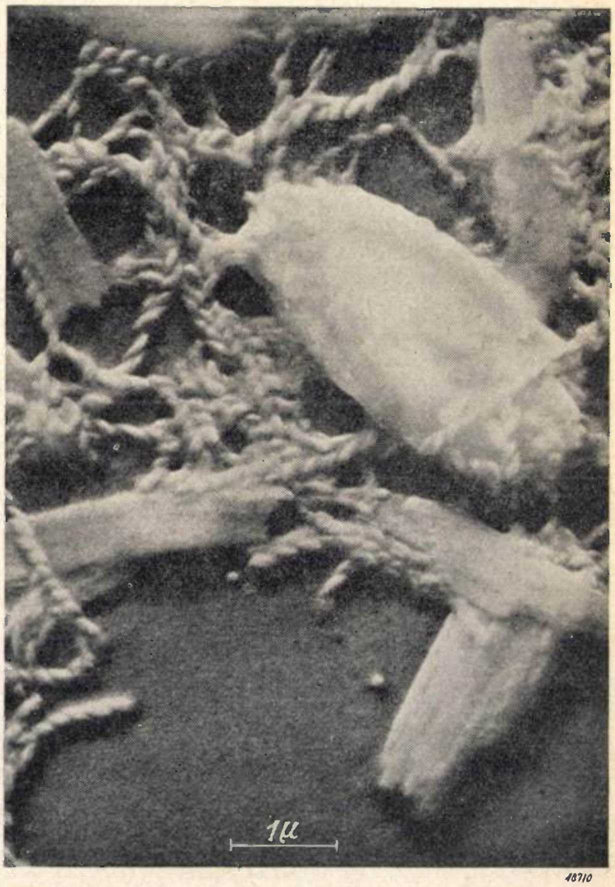


Fig. 12. A sample of fat magnified 20 000 times electron-optically with an accelerating voltage of 90 kV.

upwards from the outside. In order to prevent double exposure the film is automatically shifted when the camera is tipped up. When 25 photographs have been made a new film must be placed in the vacuum. Then the whole microscope has to be exhausted anew. Although in this process the gases which the new film brings with it have also to be removed, this operation does not take longer than about 10 min.

One of the difficulties attending the use of high magnifications is the low brightness of the final image. This makes the focusing very difficult. A special focusing arrangement has therefore been introduced.

By means of the electric field between two sets of parallel plates situated between the condenser and the objective lens, the incident ray is caused to oscillate back and forth with a frequency of 50 c/sec. When the microscope is not exactly focused this oscillation will blur the image. The power current of the objective is now varied until, in spite of the oscillation of the incident ray, the image remains sharp, whereupon the voltage on the deflection plates is switched off again.

The focusing arrangement is so constructed that the incident beam oscillates over an angle of 1/100 radian. Very good results are obtained with this method. Since in this way rays which have an aperture 20 times as large as the beam used for image formation are focused as well as possible, the adjustment can be improved by a factor of 20.

In conclusion it may be mentioned that the resolving power of the microscope lies between 25 and 30 Ångström. It will be possible to approach the theoretical limit of 5 Å, mentioned at the beginning of this article, more closely by improving the rotational symmetry of the lenses.

In *figs. 12* and *13* two photographs are reproduced which were taken with this instrument.

Summarizing, the most important advantages of the electron microscope described here as compared with previous types are the following:

1) The magnification can easily be continuously varied between 1000× and 80 000×.
2) The length is shorter and therefore the effect of disturbing fields is also smaller.
3) The image field is much larger.
4) With the 35 mm camera it is possible to make a large number of exposures in a short time.
5) A special arrangement provides for accurate focusing.
6) It is possible to pass over immediately from electron to diffraction pattern and to make a diffraction pattern exposure of a selected region of 3 μ diameter in the sample whose image has first been studied microscopically.

## Applications of the electron microscope

The electron microscope has opened up entirely new fields of investigation for various branches of science. Its applications are widely varied and we can only mention a few of them.

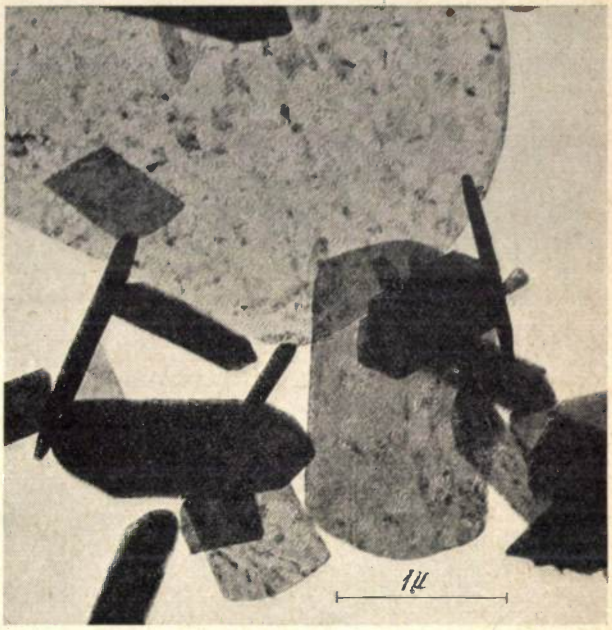


Fig. 13. Molybdenum oxide magnified 30 000 times electron-optically with an accelerating voltage of 82 kV.

Although the idea of examining specimens with such a high magnification is very attractive, some investigators at first had objections to this new instrument. They feared that it would make a difference having to observe specimens in a vacuum. In order to settle this, various kinds of specimens were also investigated with an optical microscope in a vacuum. It was found that in most cases the vacuum presents no difficulty. Primitive organisms such as bacteria remain alive in spite of a certain drying out. It was also asked whether, owing to the continuous electron bombardment, the specimens might not be fundamentally changed. It is possible to determine whether changes occur during an observation. When sufficiently thin preparations are examined and the radiation intensity is limited, there are usually no serious objections. The bacteria will indeed be killed by the electron bombardment, but with the optical microscope the bacteria are usually dyed and this likewise kills them.

In medical science it was long a difficulty that various diseases are caused by organisms or substances which are invisible under an ordinary microscope. Such a virus can as a rule be observed with an electron microscope.

With this instrument it is also possible to observe structural elements only a few mμ in size in tuberculus bacillae and other bacteria, which provides the possibility of distinguishing morphologically different types of disease producers. Further, medical investigators and biologists have made use of the instrument to study the structure of muscles and nerve fibrilla as well as the structure of protoplasm and the problems of chromosome structure and the arrangement of the genes. It was also very important that various not yet completely explained processes, such as the occurrence of coagulation in blood, could be studied under very high magnification.

Many plant diseases are also caused by viruses. The virus of the tobacco mosaic disease and that of a very destructive potato disease have, among others, been made visible with the electron microscope. This has led to great advances being made in phytopathology in recent years. The viruses often prove to be rodlike in shape and of the order of magnitude of molecules.

It has also been possible with an electron microscope to make various bacteriophages visible. These are kinds of viruses which destroy certain bacteria. This process — called bacteriophage lysis — has also been investigated with the electron microscope.

Bacteria had already long been observed with the optical microscope, but the cilia of these organisms, which have a thickness of from 0.02 to 0.02 μ, and all kinds of details and complicated structures of the bacterial cell were only made clearly visible with electron rays.

For mineralogy and technology the new microscope is also an important aid. In the problems there encountered it is often desired also to be able to observe a diffraction diagram in order to identify the crystals with the help of the Debije-Scherrer rings. It is of great advantage when one can easily switch over from the electron to the diffraction image.

In the field of silicates, by the application of the new method discoveries have been made which are of great importance in ceramics.

It is important to note that here there has been a transition from qualitative to quantitative methods. The adsorption power of the surface of particles of clay has often to be investigated. One then counts and measures under an electron microscope, for instance, 1000 crystals of a substance and makes a distribution curve of their sizes. In making such distribution curves it is very desirable to be able to change over easily to a lower magnification. The study of chemical processes at a magnification of 20 000 to 30 000× is also very instructive. As an example we may mention an excellent investigation of the manner in which an image is formed on a photographic plate, which was carried out by Von Ardenne before 1940.

It may be assumed that within a reasonable time every large hospital and every laboratory for research in the field of microbiology, mineralogy and technology will be employing electron microscopes.

At the Institute for Electron Microscopy at Delft the microscope described here has long been in daily use for investigations in the interest of industry and science.

# ELECTRONIC CONDUCTIVITY OF NON-METALLIC MATERIALS

by E. J. W. VERWEY.

537.311.32

Generally speaking, non-metallic materials are electrical insulators. In certain cases, however, conduction — caused by electrons — may take place also in these substances. This article deals with the relation frequently existing in oxides and halogenides between electronic conductivity, photo-electrical phenomena, light absorption and deviations from the stoichiometric composition. From the simultaneous occurrence of all these phenomena some insight is obtained into the manner in which they are brought about. The picture that can be formed is especially worked out for the case of the crystal of potassium chloride, for which many observations and theoretical data are available.

Electrical conductivity in metals is brought about by the movement of free electrons, which are always present. In non-metallic solids, on the other hand, conductivity may be due to two causes: in the first place — as in the case of metals — to the movement of free electrons (electronic conductivity) and in the second place to the movement of ions (electrolytic conductivity). Often both movements take place simultaneously. In this article, however, we will confine our considerations to those non-metallic substances in which the electronic conductivity plays the main part.

The interest in electrotechnics in the phenomenon of electronic conductivity in non-metallic substances is due in the first place to two applications of these substances.

The first relates to the use of all kinds of substances as insulators, thus as media that are required to have the least possible conductivity for an electric current. As a rule the insulating medium is applied for two purposes, either as a dielectric in condensers or as insulating material. In the first case the highest possible capacity is required and, therefore, often a high dielectric constant; as a dielectric titanium oxide (rutile: $TiO_2$) may for instance be used, the dielectric constant of which in the sintered condition is over 100. In the second case, however, a low dielectric constant is often required and, for instance, organic substances like the well-known polystyrene are used as insulating material. When inorganic substances are indicated because greater chemical and thermal stability are required, materials like porcelain or many of its various modern varieties are used, sometimes in the porous state.

A general answer to the question which material is an insulator and which properties are connected with a low conductivity is therefore of direct technical interest.

In the second application it is rather a high conductivity that is required; in recent years more and more use is being made in electrotechnics of more or less well conducting materials of a non-metallic nature as resistors, advantage being taken, among other things, of the fact that these substances possess a much higher specific resistance than metals. According to the application envisaged, materials are required with the lowest possible temperature coefficient, sometimes a positive one or in other cases rather a very strongly negative one. The conductivity of these insulators must be purely electronic, because electrolytic conductivity always gives rise to chemical changes at the contacts of the resistor or to polarisation phenomena. In this case it is also often required that the resistance mass must be able to withstand high temperatures, and in such cases inorganic substances are indicated.

The materials referred to above, having a very much lower conductivity than metals but a better conductivity than insulators, are often called semi-conductors. Rather arbitrarily the following distinction is usually made:

Metals: specific resistance $< 1\ \Omega$ cm (mostly $10^{-5}$—$10^{-6}$, sometimes higher).

Semi-conductors: specific resistance between 1 and $10^{10}\ \Omega$ cm.

Insulators: specific resistance $> 10^{10}\ \Omega$ cm.

All these limiting values for the specific resistance are those applying at room temperature.

Conductivity and non-conductivity appear to be distributed among the elements in such a way that the permanent non-conductors occur in the top right-hand corner of the periodic system. Typical examples of non-conductors are sulphur and diamond. Carbon, however, is on the border line, for its stable modification graphite, is a conductor, at least in certain directions of the crystal. There are also a few intermediate cases occurring in the boundary region, such as selenium, which is a semi-conductor.

These intermediate cases are also found among the compounds. For instance in a series $Ag_2O$, $Ag_2S$, $Ag_2Se$, $Ag_2Te$ we find the conductivity gradually increasing parallel with the metallic character, which manifests itself, for instance, in increased reflectivity. The problem of electronic conductivity may be identified here with that of the metallic state. Electronic conductivity, however, also frequently occurs in compounds which in other respects are typically non-metallic. In such cases the occurrence of conductivity appears to coincide with a number of other phenomena which throw considerable light upon the process of electronic conductivity and which will be discussed further in this article.

### Relation between electronic conductivity, deviation from stoichiometry and light absorption

It appears that in general we have to make a distinction between stoichiometric and non-stoichiometric compounds, Formerly the belief was held that the typical feature of compounds was that they contain the various elements in a simple ratio. In the case of many compounds this is so, but it is not always true for compounds in the solid state. In elementary instruction in chemistry one often uses the familiar example of 56 parts Fe to 32 parts S, forming the solid FeS. It is just for a compound like FeS that it has now been found that it always contains an excess of sulphur, and equally so FeO always contains an excess of oxygen.

These are compounds which cannot exist at all in the stoichiometric proportion, or, to be exact, can only exist in the metastable state. There are, however, also many compounds that may occur both in the exactly stoichiometric proportion and with a different composition. Typical examples are: ZnO, with or without an excess of Zn; $TiO_2$, with or without an excess of Ti; $Cu_2O$ and NiO on the other hand sometimes contain an excess of O. In such cases there is a discoloration of the material towards black: ZnO is white, whilst with an excess of Zn it becomes grey; $TiO_2$ is pale yellow and by chemical reduction becomes dark blue or even black; $Cu_2O$, in itself red, through oxidation becomes black; NiO, itself pale green, with a small excess of oxygen likewise turns black.

All these substances have the common property that in the stoichiometric form they are good and sometimes even excellent insulators, but that if there is any deviation they get a more or less strong electronic conductivity. $TiO_2$, in itself possessing a specific resistance of at least $10^{10}$ Ω cm, upon being heated in hydrogen is reduced to $TiO_{1.75}$ with a specific resistance of $10^{-2}$ Ω cm. As a rule electronic conductivity increases with the deviation from the stoichiometric proportions: the specific resistance of $TiO_{1.9995}$ is 10 Ω cm; and that of $TiO_{1.995}$ is 1.2 Ω cm.

Compounds are also known in which deviation from the stoichiometric proportion does not immediately result in a blackening of the material but rather in an absorption band in the visible spectrum. BaO (white) can incorporate about one per cent Ba in excess and thereby turns red. KCl heated in K vapour incorporates a small excess of K (at most a few hundredths percent) and becomes violet. Similary NaCl may be given an excess of Na, the crystals thereby turning yellow. In these cases the electronic conductivity resulting from the deviation from the stoichiometric composition is not noticeable at room temperature; slightly higher temperatures are required to be able to observe a decided difference compared with the conductivity of the pure compound. For the purpose of our considerations, however, this makes no essential difference.

We therefore observe the simultaneous occurrence of three phenomena: deviation from stoichiometry, absorption of visible light and conductivity caused by electrons.

In how far these phenomena are related to each other can be illustrated by the following experiment (*fig. 1*), which was first described by Stasiw, one of Pohl's co-workers, in whose laboratory an intensive study was made of the phenomena occurring with this kind of discoloured crystals.

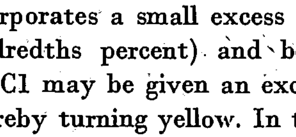


Fig. 1. When a violet-coloured KCl crystal is placed between two electrodes at a temperature of about 500° C the colouring moves out of the crystal. The arrow indicates the direction in which the boundary between the violet (shaded) part and the colourless part of the crystal moves.

Two electrodes are applied to the two opposite faces of a KCl crystal which has been made violet by the incorporation of additional potassium atoms and which is kept at a temperature of 500° C (KCl was chosen by preference for this experiment because the violet colour can be clearly observed). In the first place it is found that the conductivity of the coloured crystal is much greater than that of the non-coloured crystal at the same temperature. If the contact between the crystal and the negative electrode is rather faulty, then, on the current being passed through, a non-coloured zone is seen to form near the negative electrode, which zone gradually extends over the whole crystal. The discoloration of the crystal, which at this temperature has become deep blue, is seen to move through the crystal and after some time disappears into the electrode; this is accompanied by a gradual reduction in conductivity. After completion of the experiment the crystal again becomes stoichiometric, colourless KCl, and the extra conductivity has disappeared.

This remarkable experiment, the interpretation of which will be reverted to later, demonstrates very clearly that the conductivity created by the excess of potassium and the violet colour are closely related to each other.

### Phenomena connected with irradiation

There is a second group of phenomena which throws some light upon the mechanism of the conductivity of these substances, *viz.* the photo-conductivity and the photo-chemical changes taking place under the influence of rays of different wave-lengths.

The phenomenon of photo-conductivity occurs in all sorts of substances: it implies that when a substance is irradiated with light of certain wave-lengths an originally non-conductive substance becomes conductive, or rather that the conductivity of the substance is increased by the irradiation. It occurs for instance with the KCl crystals just mentioned (which by heating in K vapour have turned violet) when the crystals are irradiated with light that is absorbed by the violet coloured crystal. The violet colour of this KCl is caused by an absorption band in the yellow part of the spectrum with a maximum at about 5600 Å. The phenomenon of photo-conductivity therefore occurs under irradiation with yellow light of about this wavelength.

From this behaviour it is to be concluded that under the influence of light of this wavelength electrons which before the irradiation were bound to certain places inside the crystal are released. The energy with which these electrons are bound to those places and which, therefore, has to be supplied to release an electron follows directly from the wavelength of the active light. This wavelength $\lambda$ (frequency $\nu$) corresponds to the energy $eV$ supplied to the electron by the absorption according to the formula

$$eV = h\nu; \quad \text{or:} \quad V(\text{volt}) = \frac{hc}{e\lambda} = \frac{12390}{\lambda(\text{Å})},$$

in which $V$ is the voltage through which a free electron has to pass in order to accumulate the energy required ($h$, $c$ and $e$ indicate respectively Planck's constant, the velocity of light and the charge of the electron). From this we calculate for the band at 5600 Å in coloured KCl that the electrons are bound with an energy of 2.2 $eV$ (the electron energy is expressed in "electron volts" $eV$ by indicating the voltage $V$ that corresponds to the energy $eV$).

With the excess of potassium we have thus introduced additional electrons in the crystal, which may be in two states; the state in which they are still bound and a state with 2.2 $eV$ higher energy in which they have freedom of movement and may cause conduction. Further we have discovered two different means of raising the electrons from the lowest state of energy to a state of higher energy, namely by supplying thermal energy (raising the temperature to about 500° C suffices to make the conductivity perceptible) and by supplying radiation energy [1]).

The occurrence of conductivity may also be observed in crystals of potassium chloride coloured violet by photo-chemical means. The same violet colour that can be obtained by heating in potassium vapour is also obtained when KCl is irradiated, for instance, with X-rays or with ultra-violet light. KCl crystals coloured in this way, however, behave in a slightly different way than do the crystals with an excess of potassium, but, nevertheless, the electrons bound with an energy of 2.2 eV can still be released, *i.e.* on heating or by irradiation with yellow light conductivity is observed. The photo-chemically coloured crystals, however, differ from the others in one respect: upon irradiation or heating the violet colour very soon fades away and the crystal returns completely to its original state. Apparently under this treatment the electrons find an opportunity to return to their original positions. Photo-chemically coloured crystals are therefore in a metastable state; when by the addition of energy a return to the stable state is made possible, this takes place by transport of electrons inside the crystal lattice.

The discoloration with X-rays is not really a purely photo-chemical process but is caused mainly by the secondary electrons released in the substance. As a matter of fact the same result can be obtained by the irradiation of a KCl crystal with electrons. A recent example of this phenomenon is found in the cathode ray tubes where the screen consists

[1]) For the sake of clarity, on a few minor points we take the liberty of representing the position in a somewhat too simple manner. One of these points is that the energy to be supplied by the thermal movement in order to release an electron from the bound state is quantitatively not equal to the radiation energy required for the same purpose, but in general about half of that. Another simplification is that in reality the above-mentioned transition to the state with a higher energy probably takes place *via* an intermediate state; the energy of this intermediate state, however, is only very little less than that of the final state.

of a thin layer of vaporised alkali-halogenide; in this kind of tube advantage is taken of the fact that this layer has the property of discoloring more or less permanently at the places where it is struck by the electron beam.

### Further considerations about conductivity and light absorption

At room temperature the mean energy of the thermal movement of an ion or a free electron amounts to 0.04 eV. In coloured KCl the transition from a weakly bound electron into a free electron involves many times this amount, so that the chance of this transition being brought about by thermal movement is practically nil at room temperature. When the temperature is raised, however, this chance increases very rapidly, with the result that at a temperature of about 500° C a noticeable number of electrons are released and conduction can be observed in the manner just described.

The fact that non-coloured KCl is an insulator and at about 500 °C still does not yet show any noticeable conductivity for electrons may be related in the same way to the absorption in the ultra-violet part of the spectrum. As already remarked, this absorption brings about a discoloration of KCl, *i.e.* the release of electrons in non-coloured KCl, so that the electrons released in this way may at least partly be caught in the same positions which are also taken up by the additional electrons of KCl heated in K vapour. The energy required to release an electron inside the crystal of non-coloured KCl is thus much greater than that required in the case of coloured KCl; the corresponding maximum in the absorption band lies at 1310 Å, from which we can calculate, with the aid of the conversion formula previously used, a correponding energy of 9.4 eV. The probability of this large amount of energy being supplied by the thermal movement is again very much less than in the case of coloured KCl, and consequently even at a higher temperature the electronic conductivity in non-coloured KCl is practically negligible.

In the case of the oxides previously mentioned, where there is a deviation from the stoichiometric composition the colour is mostly black, absorption thus occurring throughout the whole of the visible spectrum; moreover, this absorption extends as a rule far into the infra-red. This means that these substances contain electrons which are bound to certain places with very low energies. This is in accordance with the fact that already at room temperature a noticeable number of electrons are released and electronic conductivity takes place in these substances.

### Interpretation in connection with the crystal structure

Potassium chloride forms an ion lattice with positive metal ions and negative halogen ions. The occurrence of ions is related to the tendency of all atoms to surround themselves with completed shells of electrons. Potassium, which has one electron in the N shell, loses this, whilst chlorine, with 7 electrons in the M shell, supplements it to eight electrons.

To take one electron from a potassium atom an energy of only 4.1 eV is required; to remove a second electron would cost 44 eV, because it would then have to be taken out of the completely filled M shell. The ionisation energy of the chlorine atom is 13.1 eV, which is much greater than that of the potassium atom. Apparently there is little tendency on the part of the chlorine atom to give off an electron out of its M shell. Rather it would be inclined to supplement this M shell, which is already almost full; in taking up the one missing electron an energy of 3.6 eV — called the electron affinity — is gained, whereas in the case of potassium no gain in energy can be attained by the taking up of an electron.

Although the ionisation energy of potassium is still 0.5 eV greater than the electron affinity of chlorine, yet the formation of an ion lattice is advantageous from the point of view of energy. As the positively and negatively charged ions approach each other a large amount of potential energy is released which makes the ion lattice more favourable from the point of view of energy in comparison with an atom lattice. This energy is usually written in the form

$$E = M \cdot \frac{e^2}{r},$$

in which $r$ is the distance between adjacent ions and $M$ is a numerical factor depending upon the type of lattice (the so-called Madelung constant). For the lattice type of KCl $M = 1.75$. Further $e = 4.8 \cdot 10^{-10}$ e.s.u. and $r = 3.14 \cdot 10^{-8}$ cm, so that

$$E = 12.8 \cdot 10^{-12} \text{ erg} = 8{,}0 \text{ eV}.$$

As a consequence of the formation of the lattice 8.0 eV is thus gained per pair of ions, and since the formation of the ions out of the atoms costs only 0.5 eV the ionised state in the crystal is still more favourable by 7.5 eV than that of the free neutral atoms.

It is therefore readily seen that an extra electron added to this lattice built up of ions will be able to move through the lattice quickly and that, therefore, the same will be the case with an electron originating from the lattice itself, once it has been released from the ions making up the lattice. Let us suppose that at a given moment this extra electron is at the position of a K$^+$ ion and forms with it a potassium atom. Owing to the large radius of the potassium atom (or in other words owing to the expansion of the trajectory of this electron), the electron will not merely stay in the cavity of the K$^+$ ion but will penetrate into the neighbouring chlorine ions. *Fig. 2,* in which the dimensions of the ions and of the potassium atom are represented in their true proportions, shows that in the course of its normal movement the electron will frequently reach a point half-way between two adjacent K$^+$ ions, so that it will easily pass on from one ion to the other. A small electrical field strength is sufficient to cause a directed movement of the electrons, which manifests itself in the form of electrical conductance.

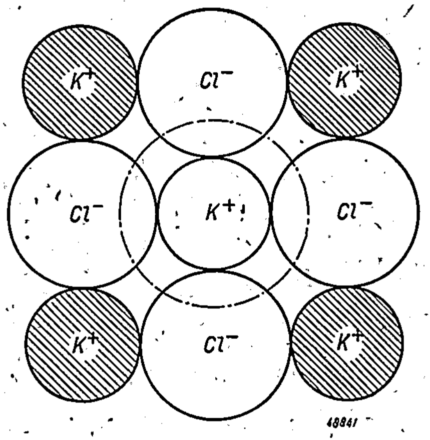


Fig. 2. Structure of a KCl crystal with potassium ions and chlorine ions in their true position and size. The dot-dash circle indicates the space which a potassium atom would occupy in the position of a K+ ion.

It is, however, likewise readily to be understood that in a stoichiometric KCl lattice it will be very difficult to release an electron from one of the ions of which the lattice is composed. For this only a Cl$^-$ ion can be considered, since the ionisation of the K$^+$ion costs 44 eV. It still requires rather a lot of energy to remove an electron from a Cl$^-$ ion, owing to the fact that, as we have seen, the electrostatic binding forces stabilise the state where the binding is brought about by ions. Consequently we have to supply not only the energy required to release an electron from Cl$^-$ (3.6 eV; see above) but also that required to overcome the electrostatic forces originating in the surrounding lattice, whereby the attraction from the adjacent positive K$^+$ions predominates. The potential energy of an electron in the position of a Cl$^-$ion as a result of the electrostatic lattice forces is in fact $-M{\cdot}e^2/r$ or $-8.0$ eV. In stoichiometric KCl, where the electrons derived from the potassium atoms are bound by just as many chlorine atoms, and in general in ion lattices of stoichiometric composition, the energy required to release electrons from the negative ions is therefore considerable, with the result that these substances are not coloured and show no electronic conductiviy. We need not, it is true, supply the full $3.6 + 8.0 = 11.6$ eV[2]) to remove the electron from the Cl ion out of the field of the surrounding ions and to let it move freely through the lattice.

Only 9.4 eV suffices, corresponding to the aforementioned absorption at 1310 Å. Nevertheless, this is still a considerable amount of energy.

The consideration given here regarding the various states of energy in which the electrons may exist in a KCl crystal might also be cast in a somewhat different form, seeking a closer connection with the quantum-mechanical theory of the solid substance. According to quantum-mechanics the energy of the electron in a solid cannot *a priori* have all possible values. Certain energy zones are "forbidden" for electrons while others are "allowed", the latter being called "energy bands". The situation is diagrammatically represented in *fig. 3a*, where the different shading of the three energy bands indicates that the two lower ones are "occupied" by electrons, whereas the top one is "unoccupied". This way of expressing the picture calls for some explanation. One must imagine an energy band as being formed by a number of energy levels lying very close together, According to quantum-mechanics any two electrons forming part of the same system (*e.g.* a piece of solid substance) can never have exactly equal energy. In our case, expressed somewhat differently, this means that each level in an energy band can only be occupied by at most one electron. When all levels of an energy band are occupied by electrons the band is said to be "occupied" or "full". If no level in a band is occupied by electrons one speaks of an "unoccupied" or "empty" band. Of course an energy band may also be partly occupied (see fig. 3*b*). The conductivity of a substance depends upon the manner in which the energy bands are occupied by electrons. If all the energy bands are either fully occupied or fully unoccupied then as a rule the substance is an insulator (fig. 3*a*); if on the other hand there is an energy band only partly occupied then we have a conductor (fig. 3*b*). This is closely related to the fact that in the former case a

[2]) For instance owing to the fact that the surroundings of the remaining "positive hole" adjust themselves to the change, by which process a certain polarisation energy is gained. Further, in the estimation it is presumed that in the conductive state the electrons are bound with a binding energy equal to zero because on their way through the lattice they are alternately located near K+ and Cl$^-$ ions. In reality these electrons still have a certain interaction with the lattice. The whole estimation given above is only approximative.

finite. in some cases large energy increase is necessary to pass an electron to a higher unoccupied level, whereas in the latter case a practically infinitely small energy increase is needed. The position and width of the energy bands depend upon the chemical composition and crystal structure of the substances [3]).

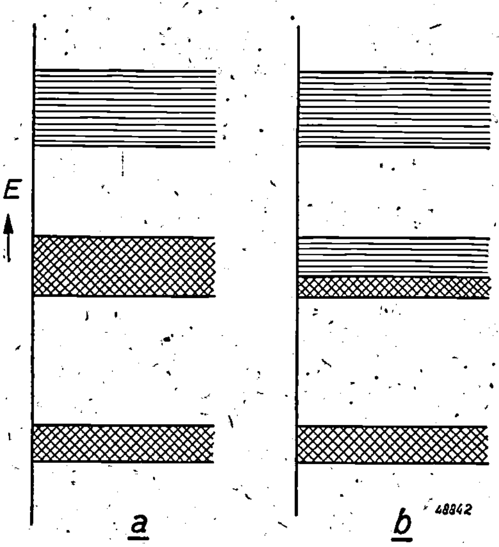


Fig. 3. Diagrammatic representation of the energy bands of electrons in a solid. The values of the electron energy $E$ lying between the values corresponding to the bands are "prohibited". All energy bands situated below the three drawn here are fully occupied by electrons and all energy bands above those drawn are unoccupied. In case *a*) there are no partly occupied energy bands; the two lower bands are entirely occupied (cross-hatched), the top band is unoccupied (shaded); the material is an insulator. In case *b*) there is one partly occupied energy band (the middle one): the material is a conductor (*e.g.* a metal).

In this article our considerations have been based on the simplified picture that the KCl lattice is an ion lattice. In this picture the electrons that can be released at the lowest possible cost of energy belong to the $Cl^-$ ions. Also according to quantum-mechanics such a picture is very near the truth for a strongly polarised substance such as KCl. The energy state of the electrons bound in the $Cl^-$ ions therefore corresponds to that of the highest occupied band of KCl. The state of electrons which by the addition of an amount of energy of at least 9.4 eV have been released from these $Cl^-$ ions and are more or less free to pass through the lattice corresponds to the state of the lowest unoccupied band. The position of the bands in stoichiometric KCl is diagrammatically represented in *fig. 4a*. At absolute zero temperature all electrons are in their fundamental state, and since all $Cl^-$ ions are complete the band $A$ is fully occupied. Consequently there are no electrons in the excited state and the energy band $B$ is completely "empty". The distance between these two bands is so great that even at room temperature only very few electrons pass from band $A$ to band $B$.

In potassium chloride containing an excess of potassium the extra electrons derived from the excess potassium atoms have to find room in other positions where they are naturally not so strongly bound as in the $Cl^-$ ions. Regarding the nature of this state of binding, which must exist both in KCl discoloured through heating in potassium vapour as well as in photo-chemically discoloured KCl, the following picture has recently been formed (J. H. de Boer, Mott).

If an excess of potassium is incorporated in a KCl crystal this is probably not brought about by the $K^+$ ions formed in the lattice finding their place between the normal lattice ions. Schottky and others have made it plausible that also a stoichiometric KCl crystal contains a number of lattice defects in the form of unoccupied places; $K^+$ ions and $Cl^-$ ions are then missing in equal numbers (*fig. 5a*). In non-stoichiometric KCl with an excess of potassium there are more $Cl^-$ ions missing than $K^+$ ions, the excess of $K^+$ ions being located in normal lattice position; the extra electrons from the atoms having been introduced into the crystal will preferably occupy the position of the missing $Cl^-$ ions (fig. 5*b*). The above mentioned energy of 2.2 eV is the energy binding the extra electrons of these "lattice defects".

When a stoichiometric KCl crystal is irradiated with X-rays or ultraviolet light the result is that an electron is taken away from some of the $Cl^-$ ions; these released electrons will again preferably occupy the places of the missing $Cl^-$ ions, so that

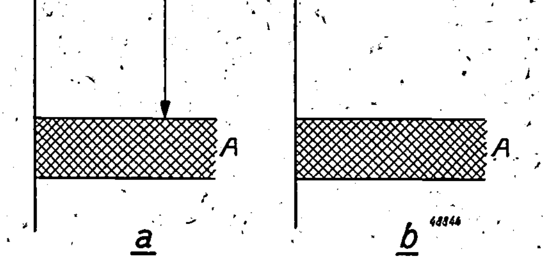


Fig. 4. Simplified diagram of energy bands:
*a*) in the case of stoichiometric KCl with a crystal lattice without defects and therefore not coloured (band $A$ is entirely occupied, band $B$ is empty);
*b*) in the case of KCl with lattice defects, *i.e.* holes (*cf.* fig. 5). In this case there is a new and very narrow energy band $C$, drawn here as a sharp level; if the KCl is non-stoichiometric or has been irradiated with X-rays then band $C$ is partly occupied and the crystal has a violet colour. Through exposure to yellow light or through heating, electrons may then pass over from $C$ to $B$, the latter then becoming partly occupied, thus making coloured KCl a conductor.

[3]) Considerations about energy bands have also been applied in this periodical for explaining the behaviour of blocking-layer photocells (see Philips Techn. R. 8, 56, 1946) and of luminescent materials (see Philips Techn. R. 6, 349, 1941).

the binding energy of these electrons is again 2.2 eV (fig. 5*c*).

The difference between KCl coloured by X-rays and KCl coloured by heating in potassium vapour consists in the fact that the former contains in addition to ions and extra electrons also neutral chlorine atoms, namely in the places where a chlorine ion of the lattice has given off an electron under the influence of the irradiation. For this reason only the colour of KCl that has been coloured by X-rays will fade again when heated or irradiated with light from its absorption band of 5600 Å. It is only in this case that the electrons that have again been released have an opportunity to move into the free places and again form chlorine ions.

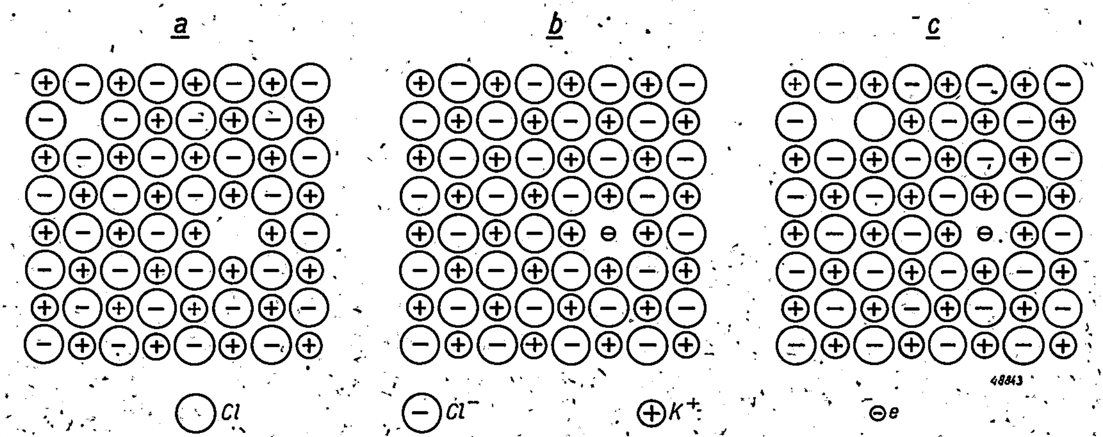


Fig. 5. Diagrammatic representation of the structure of KCl:
*a*) stoichiometric KCl with lattice faults, *i.e.* unoccupied positions (in the figure one K+ hole and one $Cl^-$ hole have been drawn).
*b*) KCl with an excess of potassium; the atom introduced into the lattice separates into a K+ ion, so that the K+ hole becomes occupied, and an electron (*e*) which is situated in the position of the $Cl^-$ hole.
*c*) Stoichiometric KCl irradiated with X-rays or ultra-violet rays; the electron thereby separated from a $Cl^-$ ion (at the K+ hole) comes to lies in the $Cl^-$ hole.
In the process of discoloration (due to excess of potassium or to irradiation) as a rule the holes shift to different positions. For the sake of clarity this has not been taken into account in the figure.

From the estimations made in the foregoing it follows that the positions where $Cl^-$ions are missing are indeed well suited to hold an electron. As a matter of fact for the potential energy of an electron in the centre of the Cl-cavity we again find —1.75 · $e^2/r$ = —8.0 eV. Actually the potential energy in the cavity is very much less negative, owing to the polarisation of the immediate surroundings (*cf.* note [2]). Moreover, the electron retains some degree of movement, so that it does not always remain in the centre of the cavity, and possesses a certain kinetic energy. Consequently the resulting energy lies approximately 2.2 eV lower than that of the conducting state.

One may look for some connection with the quantum-mechanical considerations introduced above concerning the energy bands in solids also of non-stoichiometric KCl. A diagram of the energy bands of KCl crystal coloured through an excess of potassium is given in fig. 4*b*. At absolute zero temperature all the extra electrons in such a crystal occur in the energy band *C*. This energy band may also be present in stoichiometric KCl but there is quite unoccupied (unless electrons pass into it from band *A* through irradiation with X-rays). The only condition for the presence of band *C* is the presence of unoccupied positions in the crystal lattice of KCl.

Since the extra electrons in coloured KCl are rather great distances apart and thus influence each other very little, they will all have approximately the same energy, in other words band *C* is very narrow, almost a sharp level. The photoconductivity caused by irradiation with yellow light or the conductivity arising from the heating of the coloured KCl to about 500° C is interpreted in this system of bands as a consequence of the fact that a number of electrons are passed from band *C* into the unoccupied "conductivity band" *B*.

The estimations given here for the energy of the various states of the electrons in the lattice of the alkali halogenides could be replaced by more accurate calculations, so that it may be said that our knowlegde and understanding of the conductivity phenomena in the alkali halogenides forms a more or less complete picture. Regarding other cases, such as the technically so much more important oxidic semi-conductors, the theory has not been so far developed, because in these cases quantitative calculations are much more difficult. Nevertheless, the phenomena are very similar.

We can now look somewhat more closely into the experiments described in the foregoing, where the colouring of non-stoichiometric KCl disappeared from the crystal under high temperature.

At the temperature of the experiment, owing to thermal movement electrons are continuously changed from the weakly bound state (in "$Cl^-$ holes") into the conductive state. Under the influence of the electric field they will in course of time assume on the average a directed movement towards the positive electrode. Sometimes they will thereby again be caught in empty "$Cl^-$ holes", and after a certain time again be released, and so on. Upon reaching the positive electrode they will ultimately disappear into it. If the contact with the negative electrode is a very good one then electrons will again be supplied there through transition from the electrode to the crystal. The passage of the current can then only be registered by the current intensity observed. The crystal remains coloured.

The phenomena described in the experiment occur when the contact at the negative electrode is incomplete, thereby making it difficult for electrons to pass from the electrode to the crystal. A thin layer of colourless KCl is then formed between the electrode and the coloured crystal. This layer has a much smaller conductivity than a coloured KCl, though it is not entirely zero owing to the fact that at the temperature of the experiment KCl has a small electrolytical conductivity due to the ions or, more probably, the "ion holes" already possessing a certain freedom of movement. The current, through gradually decreasing with increasing thickness of the non-coloured part of the crystal, is thus maintained by a stream of ions, the K+ ions collecting on the surface on the crystal at the negative electrode, where they are discharged and form free potassium. Since the whole crystal remains electrically neutral everywhere, the quantity of separated potassium is equivalent to the quantity of electrons passing into the other electrode. The whole phenomenon therefore resolves itself into the disappearance of the excess of potassium out of the crystal owing to electrons being given off to one electrode and the equivalent quantity of K+ ions to the other electrode. It is thereby observed that the extra electrons and the violet colour are also locally connected one with the other. It may therefore be said that this experiment makes the movement of the conductivity electrons visible.

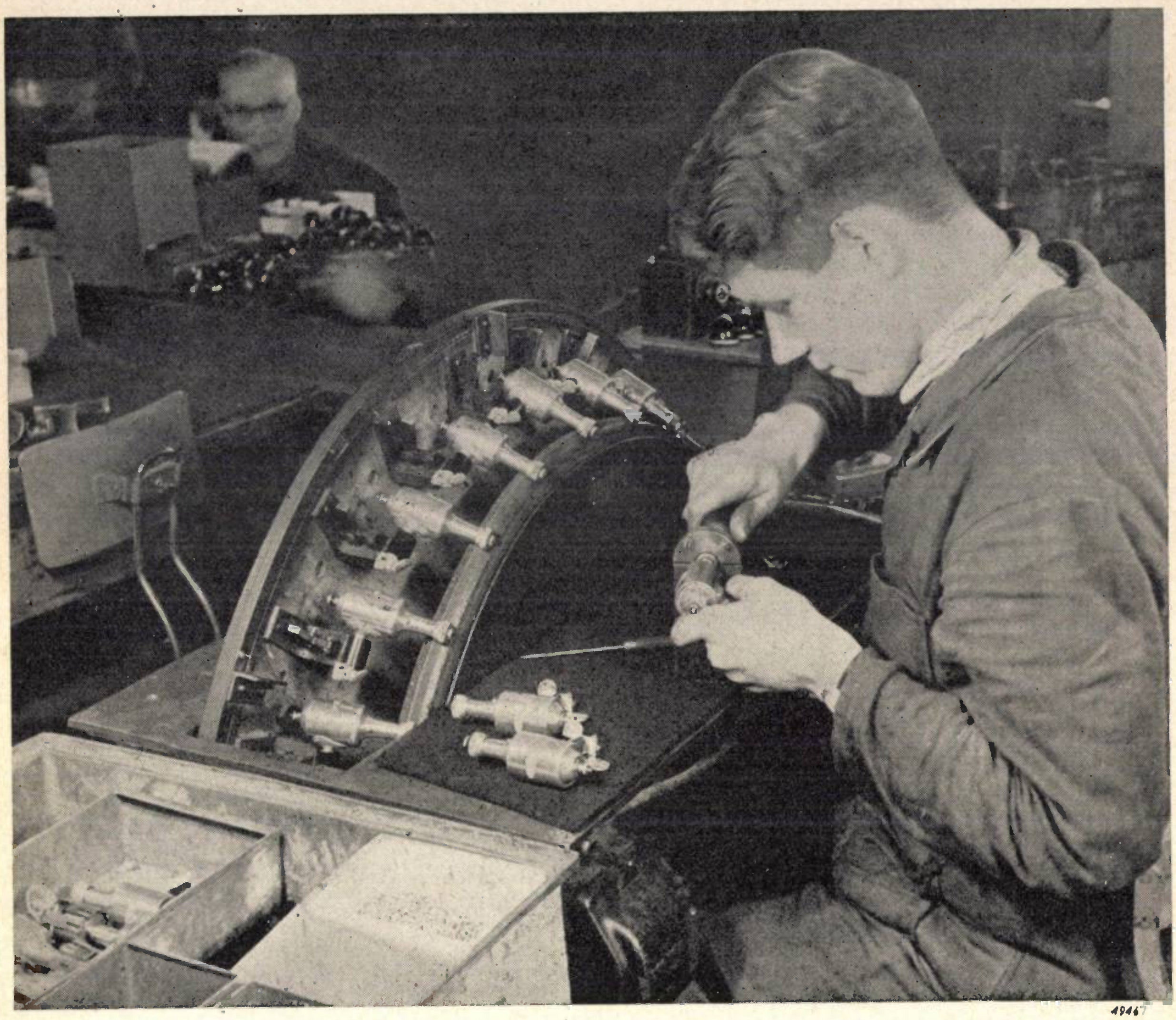
## Practical applications

As regards insulating materials the choice made in practice is such that one does not as a rule use materials in which any deviation from the stoichiometric composition readily occurs. If for some other reason one should, nevertheless, decide to employ such a material one must be careful to take the possibility of such deviations into account. As an example we might mention titanic oxide used as dielectric for condensers, which material has to be sintered by heating to a high temperature. One of the precautions that has to be taken in heating it is that the gas atmosphere and the heat treatment must be such that the final product shows no trace of a reduction to oxides of titanium of a lower valency. The losses in a high-frequency alternating field (dielectrical losses) in particular are very sensitive to the slightest deviation from the stoichiometric composition of $TiO_2$. Quantities of titanium of a lower stage of oxidation which cannot be detected by chemical means may increase the normal value of the dielectrical losses (tg $\delta$ several times $10^{-4}$) by a factor of 10 or 100. Owing to the relatively easy reduction of $TiO_2$ one is somewhat limited also in the use of this material and care must be taken to avoid any contact with organic materials at temperatures higher than a couple of hundred degrees centigrade.

Just as the introduction of potassium atoms (with one electron more than the $K^+$ions) in KCl involves the introduction of conductivity electrons, with an excess of titanium we introduce into the crystal lattice of $TiO_2$ in some way or other weakly bound electrons which cause the conductivity phenomena. One may imagine that part of the $Ti^{4+}$ ions is changed into $Ti^{3+}$ ions by taking up an electron; perhaps, however, this conception is too simple and the state of these bound electrons may be compared to that in coloured KCl.

For the resistance materials mentioned in the beginning of this article use is in fact made of these very properties of some oxides. These materials often contain, in addition to oxides having insulating properties, also a certain percentage of oxides which are easily brought to the conductive state or usually are already in that state. Apart from titanic oxides also iron oxides for instance are used, the conductivity of which is due to the presence of $Fe^{2+}$ions, which again contain one electron more than the $Fe^{3+}$ ions likewise present. We hope to deal with these materials more fully in another article to be published shortly.

# RUNNING IN CYCLE DYNAMOS



The final processes in the manufacture of cycle dynamos are the running in, the adjustment of the driving wheel (reducing the axial play to the minimum) and the mechanical and electrical testing.

A number of dynamos, *e.g.* 24, are run in simultaneously. These are secured to a large disc that can be turned into various positions. They are all driven by a single wheel with a rubber tire of the same diameter as an ordinary bicycle wheel and rotated at a peripheral speed corresponding to the speed of a bicycle of about 15 km per hour. When the dynamo is taken off the disc to be adjusted, another is put in its place, while at the same time the disc is turned to the next position. In this way every dynamo is run in for about 20 minutes, the adjustment taking only about 1 minute.

1947

# INSTALLATIONS FOR IMPROVED BROADCAST RECEPTION

by P. CORNELIUS and J. van SLOOTEN.

621.396.666:621.396.677.1

Broadcast reception is often made unsatisfactory through various causes, the most important of which are: selective fading effect and interference from transmitters on neighbouring wavelengths. Thanks to the fact that fading effect seldom occurs at different places simultaneously, its unpleasant consequences can be successfully counteracted by setting up receivers some distance apart (diversity reception) and connecting to the loudspeaker(s) only that one where the reception happens to be best at the moment. An apparatus has been worked out which brings this about automatically. In practice it appears that three receiving stations about 1 km apart are sufficient. Interferences from other transmitters can be counteracted by applying directional reception with the aid of a frame aerial, preferably in combination with a normal antenna. Thanks to the freedom from disturbances thereby attainable, the bandwidth of the receiving set can readily be increased, thus improving the quality of the sound. Diversity reception and directional reception can easily be combined. The former, however, can only be considered for installations serving a large number of listeners.

## Introduction

For the reception of the transmission from a broadcasting station to be such that the reproduction of the music and the spoken word satisfies reasonable demands, it is necessary that the field of the transmitter has a certain minimum intensity. This minimum field strength is determined theoretically by the noise voltages [1]) present in every receiver and by the atmospheric and local disturbances that almost invariably occur.

Anyone who regularly listens in, or tries to do so, to the programmes of stations a great distance away may find that in practice there are several factors which often make a reasonably satisfactory reception impossible, even though the field strength is greater than the minimum just mentioned. The causes of this are:

1) the so-called selective fading and
2) interferences from other transmitters.

In the following pages it will be explained how the effect from both these causes can be eliminated by employing special methods of reception. Unfortunately these methods are still too complicated and too expensive to be applied on a large scale by individual listeners. On the other hand they are sufficiently simple and reliable to be used wherever one is prepared to go to some expense to get good reception of remote stations, as may be the case in large establishments, such as hospitals for instance, where it is desired to distribute the broadcasting programmes to a large number of listeners.

We will now discuss separately the two abovementioned causes of unsatisfactory reception and indicate how they can be rendered harmless. Although these methods differ for the two causes, it will be seen that there is in principle no objection against their being combined into one system of reception.

## Selective fading and its counteraction

It is a well known phenomenon that after nightfall several distant broadcasting stations can be received which during daytime can hardly be heard at all. This is connected with the fact that the propagation of radio waves, particularly over long distances, depends a great deal upon the state of the ionosphere. The ionosphere is a region at a great altitude in the atmosphere where, owing to the influence of the ultra-violet radiation from the sun, ionisation takes place and one or more electrically conducting layers are formed [2]). When radio waves of a sufficiently large wavelength strike this layer at a favourable angle of incidence they are diffracted back to earth. In the course of a period of 24 hours the condition of the ionosphere shows a periodical change owing to variation in the radiation from the sun. As a consequence in the evening conditions are favourable for bridging long distances, particularly by means of the wavelengths of 200-600 metres.

Besides these diurnal variations (and others of longer periods) there are much more rapid and less regular changes in the state of ionisation. These changes affect the phenomena of interference arising from the fact that the waves from a transmitter may reach a receiving aerial in two ways,

[1]) Regarding noise in receivers see *e.g.* Philips Techn. R. 3, 189, 1938.

[2]) A comprehensive article on the radio investigation of the ionosphere appeared in Philips Techn. R. 8, 111, 1946.

either one way *via* the earth's surface and one *via* the ionosphere, or both *via* the ionosphere. Owing to the irregular changes just mentioned in the ionic density of the ionosphere the maxima and minima of interference are continously changing, and this manifests itself in a receiver as a continuous variation of the field strength. This is what is known as "fading"; the phenomenon occurs on wavelengths of 200-600 metres only at night and during twilight, because it is only then that the ionosphere affects reception.

As long as the paths along which the waves reach the aerial show differences in length only of the order of the carrier wavelength, the fading effect occurs simultaneously for the carrier wave and for the side bands. As a consequence of the automatic volume control usually present in a radio receiver this is little noticed. It becomes a more serious matter when the said differences in the paths assume magnitudes of the order of wavelength corresponding to the audio frequencies of the modulation. Then the fading is highly selective, that is to say it depends to a high degree upon the modulation frequency. Consequently for the components of the frequency spectrum emitted by a broadcasting station the transition at a certain moment may be greatly different. This is apt to lead to serious distortion of the signal, and particularly so when the field strength of the carrier wave is too low compared with that of the side bands, resulting in so-called over-modulation. This is the cause of speech so often becoming quite unintelligible.

This undesirable phenomenon of selective fading can in principle be reduced in two ways. One method consists in artificially amplifying the carrier wave in the receiver, but, apart from the fact that this is technically rather complicated and expensive, it has proved to be inadequate, though undoubtedly the average quality of reception can thereby be appreciably improved.

The second method, which is simpler and also more satisfactory, is based on the fact that the fading varies considerably not only with time but also locally; it is found that receivers set up some distance apart and tuned in to the same transmitting station are seldom troubled with selective fading at the same time. The principle of this method, therefore, lies in a mutual comparison of the signals in two or more such receivers and an automatic selection of the best signal. This is known as diversity reception.

Experiments carried out in Philips Laboratory in 1939 demonstrated, *inter alia*, that by a combination of two receivers about 800 metres apart the reception of a transmitting station greatly subject to selective fading (wavelength 450 m) was so much improved that little trouble remained. When the number of receiving stations was extended to three, at a greater distance apart (about 2 km), selective fading was scarcely noticeable at all. The quality of reception was thereby improved from "very badly disturbed" to "very good".

Although a combination of diversity reception with the aforementioned carrier wave amplification is possible and theoretically may yield still further improvement, in our experiments there was little evidence that such is desirable, a satisfactory improvement being already possible without any such technically objectionable complication.

After these general remarks we will now describe briefly how the method of diversity reception was carried out technically. In each of the three receiving stations (*I*, *II* and *III* in *fig. 1*) there were two receivers, a special broadcasting receiver (*A*) of high quality and yielding a l.f. signal of about 1 volt, and a simple broadcasting receiver (*B*) from which a rectified voltage was drawn to serve as measure for the local signal strength. (Of course both voltages could also be taken from a single receiver specially designed for the purpose.) The two voltages — the l.f. signal and the rectified voltage — were carried through conductors of the local telephone network to a central receiving point,

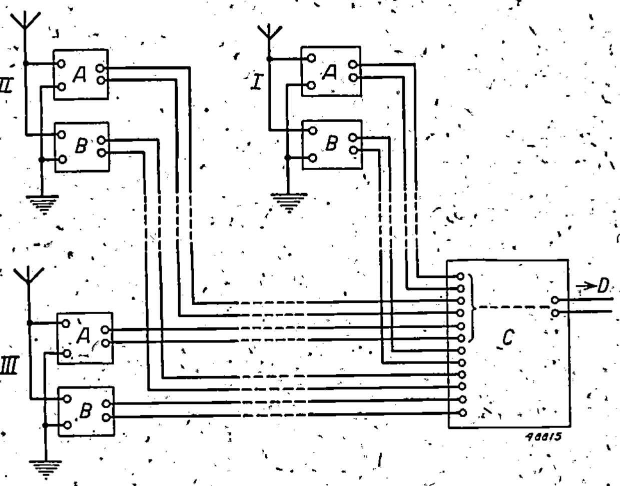


Fig. 1. Block diagram of a receiving system practically free from selective fading. *I*, *II* and *III* are separate receiving stations (at distances of the order of 1 km apart), each with two receiving sets, *A* and *B*, the outputs of which are connected *via* transmission lines with a switching apparatus *C* at a central point. The receivers *A* are high quality sets. At *D* a loudspeaker or l.f. amplifier is connected. Of the three sets *A*, the one where the greatest field strength of the desired transmitting station is received is connected to *D* by means of *C*. This field strength is measured by the simple receivers *B*, which supply the result to *C* in the form of a direct voltage.

where a switching unit ($C$) provided with relays was set up, connecting to the output terminals whichever l.f. signal corresponded with the highest rectified voltage, thus the one coming from the receiving station where the field strength happened to be greatest at the moment. A schematic diagram of this switching unit is given in *fig. 2*.

Briefly this switching system works as follows:

Three triodes (or pentodes), $T_1$, $T_2$ and $T_3$, each contain in the anode circuit the coil of a relay ($Re_1$, $Re_2$, $Re_3$) and an anode resistance ($R_{a1}$, $R_{a2}$, $R_{a3}$); the triodes are fed *via* a common cathode resistance $R_k$ from a voltage source the poles of the valves have a high anode current and one a low anode current. There can only be stability in a condition where one anode current has the full value and the other two are as good as nil[3]). Which valve receives anode current is determined by the voltages which the receivers $B$ (fig. 1) pass to the terminals $B_I$, $B_{II}$, $B_{III}$ and which increase with the local field strength of the transmitting station required. Suppose that at a certain moment the field strength is greatest at station *II*. The voltage on terminals $+B_{II}$ is thus higher than that on $+B_I$ and $+B_{III}$, and *via* the diode $D_2$ it passes almost entirely also across the resistance $R_d$, thereby preventing any current across the diodes $D_1$ and $D_3$ from the lower voltages $B_I$ and $B_{III}$. Therefore $+B_{II}$ has approximately the potential 0 with respect to $-b$; $+B_I$ and $+B_{III}$, however, have negative potentials. As a consequence $T_2$ will carry anode current, while $T_1$ and $T_3$ are dead. Thus $Re_2$ connects the receiver $A$ of post *II* with $D$, while the receivers $A$ of posts *I* and *III* remain out of circuit. If later on post *III* should have the greatest field strength then the anode current passes from $T_2$ to $T_3$ and $A_{III}$ is connected with $D$ by means of $Re_3$.

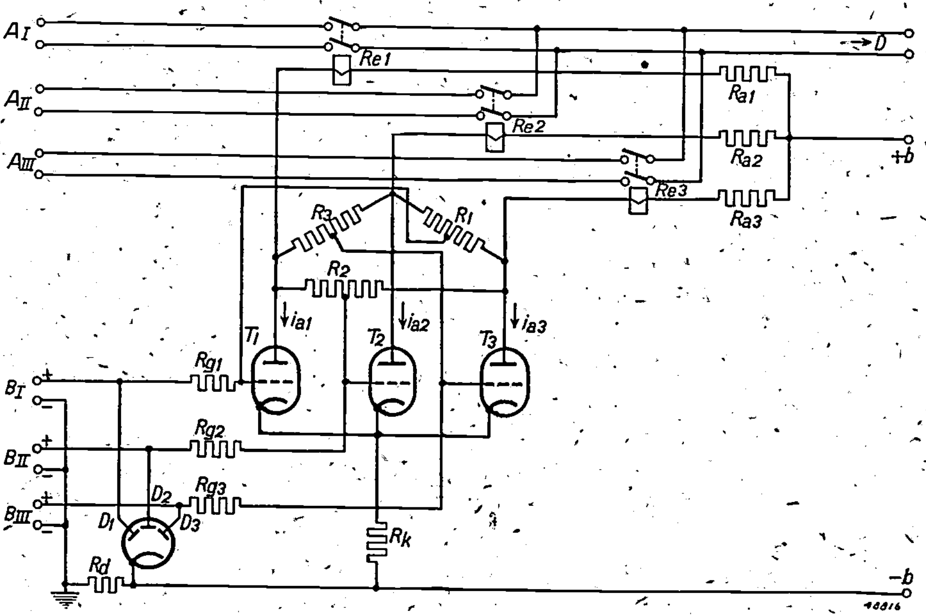


Fig. 2. Diagram of the switching apparatus $C$ of fig. 1. The lines from the receivers $A$ (fig. 1) are connected to the pairs of terminals $A_I$, $A_{II}$, $A_{III}$, one of which can be connected *via* the contacts of relays $Re_1$, $Re_2$ or $Re_3$ to the output terminals $D$. Which pair of terminals is connected to $D$ depends upon the direct voltages on the terminals $B_I$, $B_{II}$, $B_{III}$, which are connected to the receiving sets $B$ (fig. 1). $R_k$ = cathode resistor for automatic negative grid bias of the triodes $T_1$, $T_2$, $T_3$; $R_{g1}$, $R_{g2}$, $R_{g3}$ = current adjusting resistors; $+b$, $-b$ poles of the anode voltage source.

which are $+b$ and $-b$. Between the anodes there are three equal, centre-tapped, delta connected resistances ($R_1$,$R_2$,$R_3$); the grids are connected to the middle of the "opposite sides". Assuming for the moment that the pairs of terminals $B_I$, $B_{II}$, $B_{III}$ are short-circuited, then the grids are at the same time connected to $-b$ *via* the equal resistances $R_{g1}$, $R_{g2}$, $R_{g3}$ and the resistance $R_d$. Considering the symmetry of the circuit, one would expect the three anode currents ($i_{a1}$, $i_{a2}$, $i_{a3}$) to be equal. This situation, however, is unstable, as may readily be understood when supposing, for instance, that $i_{a1}$ exceeds somewhat $i_{a2}$ and $i_{a3}$: in $R_{a1}$ and the coil of $Re_1$ there will then be a greater voltage loss than in $R_{a2}$, etc, the anode of $T_1$ gets a lower potential than the anodes of $T_2$ and $T_3$, the centres of the resistances $R_2$ and $R_3$ (to which the grids of $T_2$ and $T_3$ are connected) likewise drop in potential, so that $i_{a2}$ and $i_{a3}$ will decrease, thus increasing the initial asymmetry, and so on. In a similar manner it may be seen that it is not possible either for a situation to continue to exist where two

It might be feared that the brief interruptions (a fraction of a second) during the relay action and the inevitable small variations in the sound volume would be troublesome, but in our practical experiments it appeared that this was hardly perceptible at all.

[3]) A similar circuit but for only two valves has been given by Eccles and Jordan, Radio Review, 1, 143, 1919; see also Electronics 12, 14, Aug. 1939.

### Counteracting interferences from other transmitters

#### *Nature of the interferences*

Another source of disturbances often making reception very unsatisfactory lies in the transmitting stations working on frequencies in the neighbourhood of the frequency of the station wanted. These interferences can be distinguished as 1) a whistling or, at times, a humming note, 2) the so-called cross-talk, and 3) side-band interference.

Whistling arises when the carrier wave signal from the interfering transmitter is insufficiently suppressed by the tuning circuits of the receiver; the frequency of that note equals the difference in frequencies of the station required and that causing the disturbance. Cross-talk occurs, for instance, when the carrier wave and the modulation of the disturbing station reach the detector in strength; the speech or the music from that station is then heard coming through the programme to which one is tuned in. Finally by side-band interference is understood the interference with a frequency lying between the frequency of the disturbing side-band and the carrier-wave frequency of the desired station; this results in an unpleasant, unintelligible, hissing noise, called "side-band splash" or "monkey-chatter".

Intervals of 9 000 c/s are prescribed for the carrier wave frequencies of broadcasting stations on the medium and long wave ranges. For the sake of sound quality many stations go much higher than half this interval with their modulation frequency (*e.g.* to 10 000 c/s), and consequently at places where their field strength is high they are apt to cause side-band interference in receiving sets tuned in to a station on an adjacent wavelength.

In order to reduce these disturbances one often has recourse to a diminution of the bandwidth passed through, in the high and medium frequency parts of the receiver with the aid of tuning circuits or band-filters, and in the low frequency part with a tone-filter, both of these means being adjustable within certain limits. Obviously this implies that the high notes in the programme from the desired station are suppressed or attenuated.

In good radio sets the sound reproduction is uniform up to frequencies of max. 5 000-6 000 c/s (a compromise between quality of reproduction and the number of broadcasting stations working simultaneously in a certain range). Most listeners, however, adjust by ear their bandwidth and tone controls in such a way that frequencies above say 1500 c/s are already considerably attenuated. The music then sounds muffled and speech is not clear, so that they then try to remedy this by increasing the volume, sometimes to the annoyance of others.

The less the frequency difference between the station desired and the interfering station, the more serious these troubles become.

#### *Receiving antenna with directional effect*

Interferences arising from other stations than the one desired can be reduced much more effectively — and, moreover, without sacrificing quality of sound — by employing directed reception. For this purpose antennae or combinations of antennae are used which have "directional effect", that is to say those which are more sensitive to signals from certain directions than to those from other directions. The antenna system is mounted or adjusted in such a way that its "sensitivity" is great for signals coming from the direction of the station desired and small for those coming from the direction of the interfering station. This method only fails in the rare case where both stations lie in the same direction. Preferably one would like to use an aerial system that receives signals from one direction only, or at least from directions at small angles from each other in the horizontal plane. Such antennae with a "sharp directional effect" are possible, but their dimensions are greater than the wavelength to be received, so that generally they can only be used on the ultra-short wave range. For the ordinary broadcasting wavelength (200-2000 m) the dimensions and cost would be prohibitive, the more so if it should be necessary to turn the aerial system in the directions of the various transmitting stations.

For the broadcasting range in question, therefore, antenna systems for directed reception must be used which are of small dimensions compared with the wavelength. This means that one has to be satisfied with antenna systems having such a directional effect that little or nothing is picked up of transmissions coming from one or two directions (according the system employed), and for that direction (or directions) one chooses that of the interfering station(s). Such systems are: *a*) the frame aerial, and *b*) the frame aerial in combination with an ordinary antenna. We will now consider these two methods of directional reception more closely.

#### *a) The single frame aerial*

The magnetic field strength of an electromagnetic wave (of not too small wavelength) propagated along the earth's surface is directed horizontally and perpendicular to the direction of propagation of the wave. A frame aerial responds only to that component of the magnetic field that is perpendicular to the plane of the frame. This component, therefore, disappears when that plane is perpendicular to the direction of the transmitting station. If the frame is rotated around a vertical axis (other axes are not to be considered at all here), then we get for the amplitude $V_a$ of the voltage excited by the magnetic field the horizontal directional diagram represented in *fig. 3*, which can be expressed by the equation

$$V_a = V_0 \cos a \quad . \; . \; . \; . \; . \; . \; (1)$$

in which $a$ is the angle of the plane of the frame aerial to the direction of the transmitter; if it is turned in that direction ($a = 0$) one gets the maximum amplitude $V_0$.

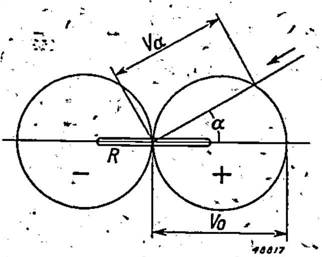


Fig. 3. Horizontal directional diagram of a frame aerial ($R$). $V_a$ denotes the voltage excited in the aerial by a transmitter whose direction is at an angle $a$ with the plane of the frame. This voltage becomes $V_0$ when $a = 0$. The circle marked + is for waves coming from the right, that marked — for those coming from the left.

From fig. 3 it is to be seen that when turning the frame 360° there are two sharp minima in the strength of reception. It is to be supposed, therefore, that by employing a frame aerial one could eliminate entirely the interferences from an outside station, provided the direction of that station makes an angle with the direction of the station desired which does not differ too much from 90°. This, however, is only correct if steps have been taken to attenuate sufficiently the so-called antenna effect of the frame [4]); some of these measures will be mentioned later on.

By antenna effect is understood the phenomenon that unless the said measures are applied the frame aerial may function simultaneously also as an ordinary or capacitive antenna. That is to say, also the electric component of an electromagnetic wave may excite a voltage in the frame regardless of the direction of the frame. In the diagram this finds expression, *i.a.* in the fact that the minima become less deep (thus no longer nil), so that a certain amount of signal of an interfering station always remains. In the articles referred to in footnote [4]) it is indicated how this antenna effect is best avoided.

Here we will briefly mention what is referred to as night effect (so called because it coincides with fading effect, which occurs mostly at night). This manifests itself in the phenomenon that in moments of fading the signals from a transmitter sometimes appear to come from directions deviating more or less from the actual direction of the station. Even if the reception has been cleared of disturbing signals by turning the frame aerial the right way in the absence of fading, those signals may come through again as soon as fading effect occurs. Considering the great irregularity in the occurrence of fading effect there is little sense in trying to correct the position of the frame. Taken on an average, the disturbance-free position, determined while fading is absent, is best.

### *b) Combination of a frame aerial and an ordinary antenna; cardiodal and cycloidal reception*

When the station desired and the interfering station lie in approximately opposite directions with respect to the receiving station, a frame aerial alone will not suffice, but now a combination of a frame aerial and an ordinary antenna can be used, as will be explained in the following.

Let us suppose that a transmitter may again excite a voltage with amplitude $V_0$ in the frame aerial turned in its direction. A circuit connected to an ordinary antenna likewise installed is adjusted in such a way that the transmitter excites a voltage $V_0$ also in that antenna. The two voltages can be added together so as to give an amplitude $2V_0$. Upon the frame then being turned 180° the phase of the voltage is reversed, so that the frame aerial voltage and the antenna voltage neutralise each other and no signals of the tranmitter in question remain. Reception from other directions is still possible; the equation for the combined voltage $V_{\text{res}}$ is:

$$V_{\text{res}} = V_0 (1 + \cos a). \quad . \; . \; . \; . \; (2)$$

This equation is represented by the heart-shaped curve of *fig. 4*, from which the term "cardiodal reception" is derived.

In the foregoing it has been tacitly supposed that the voltage of the frame aerial and that of the ordinary antenna differ 0° or 180° in phase. Actually, however, there is between these two a phase difference of + or —90°. It is therefore necessary to introduce a device that changes the phase of the antenna voltage for instance 90°. This device can be combined with an amplitude control by means of which the antenna voltage output can be adjusted to the right value. An

[4]) M. Ziegler, Philips Techn. R. 2, 216, 1937, and P. Cornelius, Philips Techn. R. 7, 65, 1942.

example of how this can be done will be given farther on.

As may be seen from fig. 4, the directional diagram for the method described has one zero direction. By a slight modification, however, it is possible to get two such directions making any desired angle, simply by making the amplitude of the antenna voltage $V_a$ smaller than that of the maximum voltage $V_0$ in the frame aerial. The directional diagram then obtained is an extended epicycloid, and for that reason this system is referred to as cycloidal reception (the eardiod is a special case of the epicycloid). The directions from which nothing comes through, $\pm\,\alpha_{\min}$, then follow from

$$V_a + V_0 \cos \alpha_{\min} = 0 \quad . \quad . \quad . \quad . \quad (3)$$

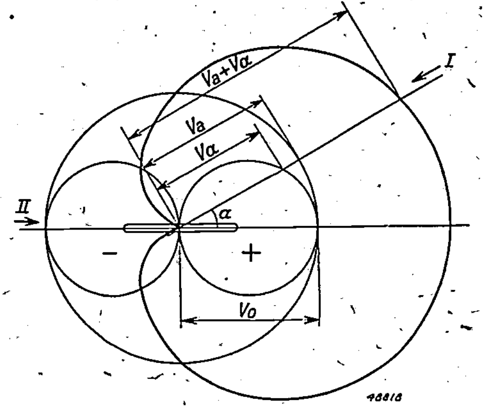


Fig. 4. Directional diagram for a combination of a frame aerial $R$ (circles + and — with diameter $V_0$, *cf.* fig. 3) and an ordinary antenna without directional effect (large circle, radius $V_a$). If $V_a$ is made equal to $V_0$ and steps are taken to ensure that $V_a$ and $V_0$ are in phase or in counter-phase as the case may be, then we get the heart-shaped diagram (cardioidal reception). A signal $V_a + V_\alpha = V_0\,(1 + \cos \alpha)$ is obtained from direction $I$; nothing is received from direction $II$.

By turning the frame so that its plane bisects the angle between the directions of two interfering stations, it is in this way theoretically possible to suppress the interferences from both of them.

In practice it is not quite so simple to suppress two interfering signals. As already stated, one must have a device for deriving from the antenna voltage a voltage shifted 90°. The phase displacement obtained, however, is more or less dependent upon the frequency, and as a rule the two interfering stations will not be working on the same frequency. We cannot here enter upon the complications arising if one nevertheless tries to make the phase displacement 90° for each of the stations, as is necessary for reception free from interference.

Reverting to the case of one interfering station, we may add that according to equation (3) the interfering signals can be reduced to zero also for any arbitrary position $\alpha$ of the frame, by making $V_a/V_0 = -\cos \alpha$. Supposing that this condition is always satisfied, it can be ascertained what value of $\alpha$ yields the strongest signal of the station desired. For that purpose the plane of the frame must be perpendicular to the line bisecting the angle between the directions of the two transmitters, as demonstrated by the following calculation.

In *fig. 5* $\varphi$ denotes the angle between the directions of the desired station ($I$) and the interfering station ($II$). The frame is set in an arbitrary position where its plane forms an angle $\alpha$ with the direction $II$. It is required to ascertain the optimum valve of $\alpha$, that is to say the value which combines complete suppression of the interferences from $II$ with the strongest signal from $I$.

If we use $V_\alpha''$ for the amplitude of the voltage excited by the interfering station $II$ in the frame aerial in the position given, and $V_0''$ for that amplitude when the frame was directed to $II$, then according to equation (1) we get

$$V_\alpha'' = V_0'' \cos \alpha .$$

The ordinary antenna yields a voltage $V_a''$ derived from $II$ and adjustable in size and phase. We assume that $V_a''$ is always adjusted in counter-phase to $V_\alpha''$. The condition necessary for eliminating interferences from $II$ is then

$$V_a'' = V_\alpha'' = V_0'' \cos \alpha \quad . \quad . \quad . \quad (4)$$

Now in order to find the value of $\alpha$ which not only satisfies (4) but also yields the strongest signal from the station $I$ we make use of the fact that in an electromagnetic wave — *i.e.* in a field of irradiation at a sufficiently great distance (*e.g.* more than 10 wavelengths) from the transmitting antenna — there is a fixed relation between the electric and the magnetic field strength [5]). Since

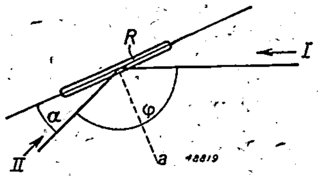


Fig. 5. The directions of the desired transmitter ($I$) and the interfering transmitter ($II$) make an angle $\varphi$; the plane of the frame aerial forms an angle $\alpha$ with the direction $II$. The optimum reception, with a combination of the frame aerial and an ordinary antenna, is obtained when the plane of the frame is perpendicular to the bisectrix of the angle $\varphi$ ($\alpha = 90° - {}^1/_2\,\varphi$).

[5]) In a travelling wave the ratio of the electric to the magnetic field strength is equal to $120\,\pi \approx 377$, if Giorgi's rationalised units are used with absolute volts and amperes, and the said field strengths are therefore expressed in V/m and A/m respectively.

the ordinary antenna responds to the electric component of the wave and the frame aerial to the magnetic component, there is therefore the same relation between the voltages set up by different transmitters in the ordinary antenna and maximally in the frame aerial. By introducing for the voltages derived from the desired station $I$ the symbols $V_0'$, $V_\alpha'$ and $V_a'$, with the same significations as previously given, then we get:

$$\frac{V_a'}{V_0'} = \frac{V_a''}{V_0''} \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

Further, having regard to (1)

$$V_\alpha' = V_0' \cos \{180° - (\alpha + \varphi)\} = -V_0' \cos (\alpha + \varphi).$$

Making use of (5) we may write for (4):

$$V_a' = V_0' \cos \alpha.$$

The resultant voltage of the desired station $I$ is therefore:

$$V_{\text{res}}' = V_\alpha' + V_a' = V_0' \{-\cos(\alpha + \varphi) + \cos \alpha\} =$$
$$= 2\, V_0' \sin {}^1/_2\, \varphi \sin (\alpha + {}^1/_2\, \varphi).$$

This expression reaches its maximum value for $\alpha = 90° - {}^1/_2\, \varphi$, *viz*:

$$V_{\text{res max}}' = 2\, V_0' \sin {}^1/_2 \varphi \quad . \quad . \quad . \quad . \quad (6)$$

The condition $\alpha = 90° - {}^1/_2\, \varphi$ can also be expressed as follows: in order to get the maximum signal strength from the station $I$ without receiving any disturbing signals from station $II$, using cycloidal reception, the frame has to be turned in such a way that it takes a position perpendicular to the line ($a$ in fig. 5) bisecting the angle between the directions of $I$ and $II$.

On the other hand with cardiodal reception (frame turned in the direction of the interfering station $II: \alpha = 0$) the magnitude of the resulting voltage from the desired station $I$ would amount to

$$V_{\text{res card.}}' = V_0' (1 - \cos \varphi) = 2\, V_0' \sin^2 {}^1/_2 \varphi, \quad (7)$$

which is smaller by a factor $\sin {}^1/_2\, \varphi$ compared with the value (6). Therefore the two methods are only equivalent when $\varphi = 180°$.

For the sake of completeness we will now consider the case where exclusively a frame aerial is used. To make the reception free of interferences $\alpha$ must be made equal to 90° (fig. 5), so that the plane of the frame forms an angle $\varphi$ —90° with the direction of the desired station $I$. The latter thus excites in the frame aerial a voltage $V_\alpha'$, given by:

$$V_\alpha' = V_0' \cos (\varphi - 90°) =$$
$$= V_0' \sin \varphi = 2\, V_0' \cos {}^1/_2 \varphi \sin {}^1/_2 \varphi. \quad . \quad (8)$$

In *fig. 6* we have plotted the results of (6), (7) and (8) — omitting the factor $2V_0'$ occurring in all three — as functions of the angle $\varphi$ between the directions of the two stations. As will be seen, the cardioidal reception is better than that with the frame aerial alone when the difference in direction $\varphi$ is more than 90°; if it is smaller the reverse is the case. The cycloidal method, where the frame is perpendicular to the bisectrix of $\varphi$ ($\alpha = 90^0 - {}^1/_2\varphi$), is in every case better than the other two methods. From fig. 6 it is to be concluded that where there is only a small directional difference $\varphi$, there is little advantage in the cycloidal method compared with what can be obtained with a single frame aerial. It is not to be forgotten, however, that with the latter method any small antenna effect remaining will make it impossible to suppress the interferences entirely. Such is indeed possible with the former method, as also with cardiodal reception, which, however, according to fig. 6 gives a considerably smaller signal strength when $\varphi$ is small.

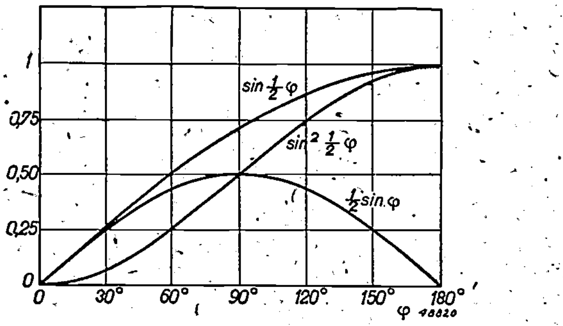


Fig. 6. The resulting signal of the desired station (omitting the factor 2 $V_0'$) is plotted as a function of the angle $\varphi$ between the directions of the desired station and the interfering stations, for three methods of reception. The disturbing signal is in every case entirely suppressed. The curve ${}^1/_2 \sin \varphi$ (= $\sin {}^1/_2\, \varphi \cos {}^1/_2\, \varphi$) applies for a frame aerial alone, the curve $\sin^2 {}^1/_2\, \varphi$ for cardioidal reception, the curve $\sin {}^1/_2\, \varphi$ for cycloidal reception, where, among others, the condition $\alpha = 90° - {}^1/_2\, \varphi$ (*cf.* fig. 5) is satisfied.

## Description of an apparatus for improved broadcasting reception

### *The receiving set and the aerial couplings*

Philips have designed a special receiving set for the methods of reception described (type 4578, *fig. 7*). This set is characterised by a highly uniform amplification in a bandwidth adjustable in four stages and extending to a maximum of 2 × 10 000 c/s (thus much farther than is the case with normal broadcasting-receivers, where it would be of little use owing to the interferences).

There are some variations in design of this model; design 4578/03 is specially constructed for cycloidal and cardiodal reception. *Fig. 8* shows how the first

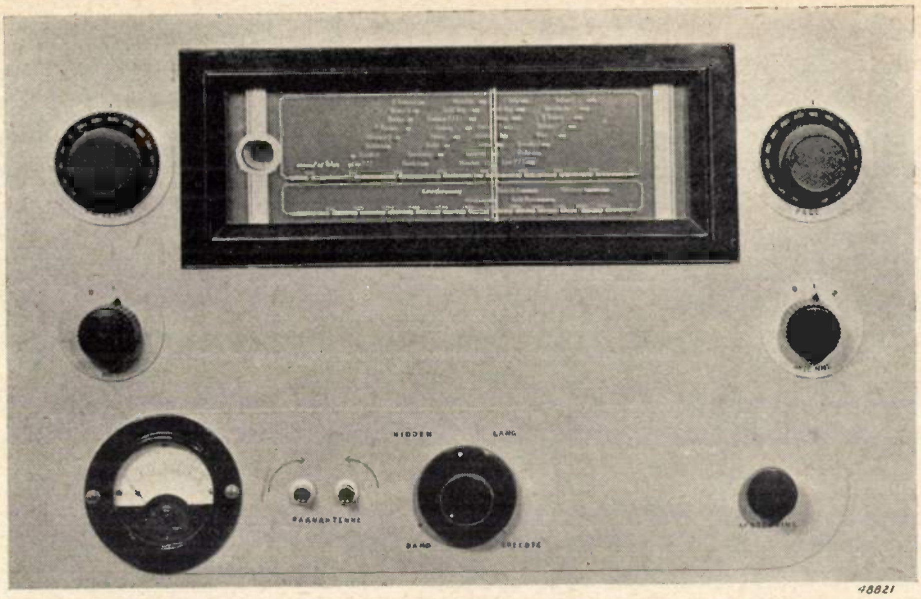


Fig. 7. Receiving set (type 4578/03) for cycloidal and cardioidal reception. On either side of the station dial is a knob with which the signal from the ordinary antenna is given the right amplitude and phase with respect to the signal from the frame aerial. Below these knobs there is on the left the mains switch and on the right a knob which in the position "0" switches off the ordinary aerial and in the positions "1" and "2" switches it on; when changing over from "1" to "2" the aerial coupling coil is commuted, so that the frame aerial need never be turned more than 180°. At the bottom, from left to right: a meter indicating the position of the frame (scale 0-180°); two buttons for the motor turning the frame (two directions of rotation); a double knob, the outer one operating the switch for "medium wave" and "long wave" and the inner one regulating the band-width in four stages (to max. 2 × 10 000 c/s); on the extreme right the tuning knob.

h.f. circuit is connected with the ordinary antenna and with the frame aerial; it is described in further detail in the text underneath the diagram.

*The frame aerial and its connection to the set*

In order to reduce the "antenna effect" of the frame aerial (see note [4]) two measures have been devised: the frame (see *fig. 9*) consists of only one winding, and the primary coil of the transformer $T$ (fig. 8) is earthed in its middle. This transformer is necessary to step up the voltage excited in the frame aerial and thus adapt the self-inductance of the frame (which is relatively very low with a single winding) to the capacity of the normal tuning condenser. The coupling of the coils of this transformer has to be very close. For that reason a ferromagnetic core is used; "Ferroxcube" [6]) is a very suitable material for this.

As to the setting up of the frame, it should preferably be located at some distance from the receiving set, because of the local interferences often transmitted by the mains to which the set is connected [7]). It is desirable, however, that the condenser tuning the frame circuit should be left in the set for the sake of single-knob tuning. This means that a transmission line has to connect the set with the frame aerial.

This transmission line should have an exceptionally low resistance, because owing to the very close transformer coupling this resistance, multiplied by the square of the transformer ratio, appears in the tuning circuit as a loss resistance. In the

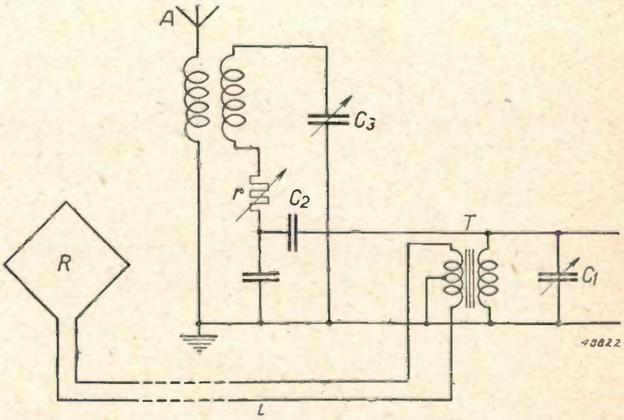


Fig. 8. The frame aerial $R$ is connected *via* a transmission line $L$ of 12 m length to a transformer $T$ and to the variable condenser $C_1$ and forms the first tuned circuit of the receiving set. Across the small coupling condenser $C_2$ the signals from the ordinary antenna $A$ are led in at the same time. These signals can be adjusted in amplitude by means of the resistance $r$ and in phase by means of the condenser $C_3$. The circuit of which $C_3$ forms a part has to be approximately tuned to the wavelength of the interfering station.

[6]) Philips Techn. R. 8, 353, 1946.
[7]) Philips Techn. R. 3, 235, 1938 and 6, 302, 1941.

second place, considering its divided self-inductance and capacity, together with the frame, it should behave approximately like a coil with constant self-inductance and constant self-capacity in the whole of the frequency range to be covered; it forms part of the first h.f. circuit, which, for the sake of single-knob tuning, it is desired to keep tuned together with the other h.f. circuits by means of a multiple variable condenser. This implies that the transformed self-inductance of the frame and of the line — bearing in mind the finite self-inductance and leakage of the transformer — must be equal to the self-inductance in the other h.f. circuits. In the third place the self-inductance of the line should be small compared with that of the aerial, so that the aerial voltage may be conducted to the transformer with the least loss.

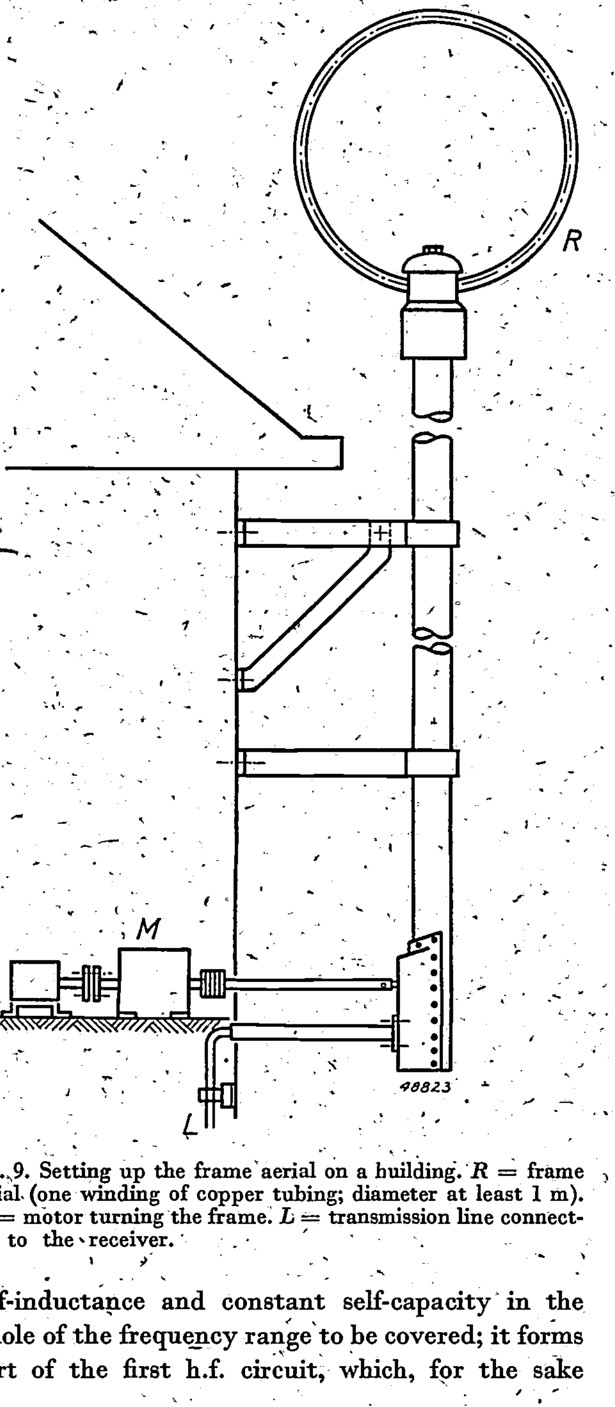


Fig. 9. Setting up the frame aerial on a building. $R$ = frame aerial (one winding of copper tubing; diameter at least 1 m). $M$ = motor turning the frame. $L$ = transmission line connecting to the receiver.

These requirements are partly contradictory, but a satisfactory solution has been found, firstly by making the line of such a length that, while on the one hand the frame aerial is far enough removed from the mains interferences, on the other hand the longest characteristic wave of the line together with the frame aerial is sufficiently smaller than the shortest wave to be received (*i.e.* 200 m); and further by choosing the self-inductance of the line at such a value that it is only a fraction of the self-inductance of the frame aerial. The length chosen on these arguments was 12 meters. The transmission line itself consists of two parallel copper strips 25 mm wide, placed 5 mm apart and mainly separated by air. The losses prove to be small enough; the self-inductance is about one-third of that of the frame aerial. Compared with a concentric line this solution has the advantage that it is symmetrical with respect to the surroundings and consequently does not produce any antenna effect.

### *Diversity reception*

In the beginning of 1940 an installation was set up at Rotterdam on the principles of fig. 1, but no frame was used. As regards reduction of selective fading effect the results were most satisfactory. Owing to the war the development of this method of reception was temporarily suspended. It is to be expected, however, that in the near future more interest will be shown in this method, the more so since now, thanks to the development of directional reception, the main sources of interferences can be successfully counteracted by a combination of the two systems.

# 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.

**1708:** P. C. van der Willigen: Contact arc welding. (The Welding Journal Research Supplement, May 1946).

This paper describes the properties and the advantages of a new type of welding electrode (developed by Philips), which combines the properties of touch welding, self-starting and reignition while also some other advantages are obtained (for full particulars see Philips Technical Review **8**, 161-167, 304-309, 1946).

**1709:** N. G. de Bruyn: On the Zeros of a polynomial and its derivative (Proc. Kon. Ned. Akad. Wet. Amsterdam **49**, 1037-1044 1946).

It is proved that the sum of the absolute values of the imaginary parts of the roots of f′ ($z$) is equal to or less than ($1-1/n$) times the corresponding sum for f ($z$), in which f ($z$) is a real polynomial of degree $n$ in $z$ and f′ ($z$) its derivative. This theorem is extended and a few specializations are considered. It is unknown whether the inequality holds for polynomials with complex coefficients. It does if all the roots are assumed to lie on the imaginary axis.

**1710:** H. B. G. Casimir and D. Polder: Influence of retardation on the London-van der Waals forces (Nature **158**, 787, 1946).

In the course of work on the stability of colloidal solutions, Overbeek arrived at the conclusion that in order to obtain agreement between theory and experiment it is necessary to assume that the London-van der Waals energy decreases more rapidly than $R^{-6}$. He pointed out that the retardation of the electrostatic forced might be responsible for such an effect and that deviations from the $R^{-6}$ law should become effective at a distance comparable to the wavelength $\lambda$ corresponding to the excitation energies of the interacting atoms. Calculations with the aid of quantum electrodynamics prove that this suggestion holds true, the energy being proprtional to $R^{-7}$ rather then to $R^{-6}$ for $R \gg \lambda$. Details of the quantum-mechanical calculation will be published in the Physical Review.

**1711:** P. J. Bouma: Die Grassmannschen Gesetze der Farbmischung (Physica 12, 545-552, 1946) (Grassmann's laws of additive colour mixing).

The author proposes a formulation of Grassmann's laws of additive colour mixing, which is axiomatically as pure as possible and which does not contain more than is necessary for the construction of elementary colorimetry. The correctness of the formulation is proved by working out this construction in rough lines finally the questions of continuity, arising in connexion herewith, are dealt with and are reduced to the existence of thresholds.

---

The last issue of *Philips Research Reports* (Nr 6 of volume 1, December 1946) contains the following papers:

*R27:* N. G. de Bruyn: A note on van der Pol's equation.

*R28:* T. Jurriaanse: The influence of gas density and temperature on the normal cathode fall of a gas discharge in rare gases.

*R29:* J. W. L. Köhler and C. G. Koops: A new alternating current bridge for precision measurements.

*R30:* M. Gevers: The relation between the power factor and the temperature coefficient of the dielectric constant of solid dielectrics, V.

*R31:* K. F. Niessen: A new method for the determination of the electric constants of the earth's surface.

Readers interested in one of the above mentioned articles may apply to the Administration of the Philips Physical Laboratory, Kastanjelaan Eindhoven, where a limited number of copies are available for distribution. For a subscription to Philips Research Reports please write to the publishers of Philips Technical Review.

VOL. 9 No. 3, pp. 65-96 1947

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

# PROBLEMS IN PHOTOGRAPHIC REPRODUCTION, IN PARTICULAR OF SOUND-FILMS

by C. J. DIPPEL and K. J. KEUNING. 778.588.3:771.534.553

The resolving power of a film made by the usual photographic methods is limited by the circle of diffusion formed by the grains of the film. Consequently in order to get sharp pictures on the projection screen the images on the film must be of a certain minimum size. If the film carries a sound track and the speed of the projected film is fixed, then, in spite of a so-called cancellation method being applied when recording and copying, this limited resolving power results in a loss in the amplitude of high frequencies. In order to counteract the circle of diffusion it is desirable that the film should have a high gamma, of say 4 or 5. For picture reproduction, however, Goldberg's rule prescribes a gamma in the neighbourhood of 1 or 2. The compromise that has to be reached when a picture film and a sound track have to be copied on a single film by the usual methods makes its influence felt throughout the whole of the present-day technique of cinematography.

A much simpler and less expensive solution of the problem of copying sound-films is offered by a new method of photographic reproduction that was developed in the Philips laboratories during the war. This method is based on the use of a diazonium compound combined with a mercury salt. The most striking features of this method are the extremely high resolving power (1000 lines per millimetre) and the locally variable gamma (between, *e.g.*, 1 and 8). A more detailed description of this method and its possibilities of application will be given in another article to be published in this journal shortly.

During the war a group of research workers in the Philips Laboratories developed an entirely new method of photographic reproduction. Besides other applications to be referred to later on, some of which are quite new, they had in mind the making of copies of sound-film. In one of the next numbers of this journal we hope to give a concise explanation of this new system of reproduction, its characteristic features and the remarkable possibilities of its application. We deemed it advisable to deal first with the main problems arising in photographic reproduction, in particular of sound-films. In fact in the last decades the photographic methods commonly employed — most of them based on the use of silver halide — have reached such a high state of perfection that one may wonder whether it is not presumptuous and unnecessary to advance a new method. We hope, however, that what follows will make it clear that there are indeed drawbacks attaching to the present methods, drawbacks which have led to this new development. From that it will be seen what place it can take in photographic reproduction.

### The aspects from which a system of reproduction is to be judged

By speaking of photographic reproduction we have already given expression to the fact that it is not the recording but the copying of sound film that forms the subject of this article. Even though both recording and copying are done photographically there are great differences between the two. The exposure time when recording has to be very short in order to get sufficient definition of the variations of picture and sound, whereas copying can be done, in principle, with any length of exposure. For taking the picture (though it may be in colours) the colour sensitivity of the film material has to be about equal to that of the eye to translate the picture in the right shades of grey. For copying the black and white picture the copying material may have any spectral sensitivity provided it is suitable for the colour of the light source employed, which is likewise primarily arbitrary.

Here we have mentioned some properties that are a matter of indifference in reproduction material.

What are, indeed, of importance are mainly the following points.

In the first place there is the quality of reproduction. A good quality picture must be sufficiently sharp and properly graduated in the half-tones. In sound it is desired to retain the right amplitudes of the high frequencies, which in most systems of reproduction are apt to be sacrificed; there should be no distortion, and the pitch of each note should be absolutely constant (no "whining" as heard with gramophone records).

Another condition for the copying of films is that the cost of the copying material and of the processing must be kept low, for in many cases this is done in mass production; as opposed to the taking of the film, where the cost of the film material is as nothing compared with the cost of the acting, scenery, staging, etc. This requirement will be particularly applicable when it comes to using the combined picture and sound film for other purposes outside the cinema.

Further requirements are that the film copy must be durable and, preferably, not easily liable to damage. Safety film is also to be recommended — hence the preference shown in some cases for the 8 and 16 mm sub-standard films, as these are made exclusively of safety material — though it is not regarded as a condition *sinc qua non.*

Finally we have to mention some properties which need not be specifically demanded of a sound track on a film because they are already inherent in a film. These properties are of importance when comparing the sound recording on a film with that on gramophone records. They are: greater length of uninterrupted playing on a single standard reel; ease of transport and storage of the copies (there is little difference in volume of what is recorded, for the same playing time, on gramophone records and on normal film reels); further the automatic synchronisation of pictures and sound when both are side by side on one film. To fully appreciate what this synchronisation means it is necessary to recall that in the beginning of the history of talking films, when the sound was reproduced from gramophone records, it was very difficult while projecting a film to match the sound exactly to the picture on the screen. Modern developments have solved this problem and synchronisation is now done as part of the copying process; having regard to the quality of the sound it has been found impracticable to record sound and picture simultaneously on the film, so that usually the sound is recorded separately, the picture film and sound track then being copied side by side on one film, taking care that the two are properly synchronised.

The aspects of a system of reproduction brought forward here are not all independent of each other. For instance, we have seen that the sensitivity of the copying material is not of primary importance because one is not essentially bound to making short exposures, but considered from the point of view of the cost of the copying process sensitivity is indeed of some account, because the copying process should not take too long. Of still more importance is the relation between the cost factors and the quality of reproduction, as will be seen from what follows.

## The resolving power of the film

The cost of material for a film copy will be all the less according as the size of picture is chosen smaller and the film speed is lower; these are factors allowing of a longer playing time to be compressed into a shorter length of film.

Since, as a rule, it is desired to project a film at the rate of 24 frames per second, the speed of a picture film is fixed by choosing the suitable size of picture (height of frame). In the case of a film carrying only a sound track — as for instance where the Philips-Miller recording system is applied in broadcasting studios [1]) — there is more freedom in the choice of the recording and the projecting speed.

The smaller, however, the frame height and the lower the film speed, the more difficult it becomes to satisfy the requirements for good quality reproduction. This we will explain first in connection with the picture and then in regard to the sound.

The first difficulty encountered in the case of a very small picture is to concentrate on it the high light flux required to produce a sufficient intensity on the screen. Among other requirements is a very high brilliancy of the light source [2]). A second difficulty, however, is one inherent in the film itself, its capability of giving a sufficiently sharp picture on the screen. It is a commonly known fact that it is pointless to go beyond a certain limit in enlarging a photographic picture because if one goes beyond that the enlargement is blurred and no new details are shown. This limit is usually set by the resolving power of the film. To get a very small frame size without sacrificing sharpness of the screen picture the film must possess a very high resolving power.

Since this brings us to a cardinal point in photographic reproduction, it is necessary to give closer

[1]) Philips Techn. Rev. 5, 74, 1940.
[2]) See *e.g.* Philips Techn. Rev. 4, 2, 1939 and 8, 72, 1946.

consideration to what we mean by resolving power. This may be defined in several ways. The most commonly accepted conception is indicated by the number of lines per millimetre that can be printed on the film by copying a test object without their becoming indistinguishable (*fig. 1*). This, however, is rather a subjective measure, depending also upon various objective factors such as the illumination of the test object, width and contrast of the light and the dark parts, etc. This makes it difficult at times to compare the measures given by various investigators for the resolving power of films. The resolving power of standard positive film as used for making film copies is about 55 lines per mm. That of the Kodak Panatomic film for ordinary photography is somewhat higher, *viz.* 60 lines/mm for normal and 70 lines/mm for so-called fine-grain development; still better are the Kodak "High Contrast Positive" films (100 lines/mm) and "Microfile" (160 lines/mm).

Before dealing with the causes and effects of this limited resolving power we will consider how matters stand also in regard to the sound reproduction. We will confine our attention to variable area sound tracks (*fig. 2*), though similar considerations apply also to variable density tracks. It will easily be realised that broadly speaking the position with regard to sound is analogous to that of the picture, since the recording of sound is again a question of sharp reproduction of certain details. The sound track of a frequency of say 9000 c/s recorded at a film speed of 30 cm/sec. contains 30 sines on 1 mm. To record these 30 sines exactly the film would have to have an exceptionnally high resolving power, much higher than the normal values. If the resolving power is inadequate then, as we shall see farther on, the sines are distorted, so that, firstly, a number of harmonics of the original frequency and — where composite sounds are recorded — also sum and difference frequencies arise (non-linear distortion), and, secondly, the amplitude of the original frequencies is reduced.

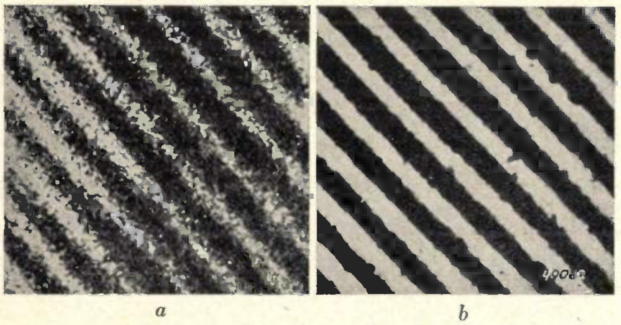


Fig. 1. The limited resolving power of a film. When a grating is pictured on the film the lines flow into each other as soon as the distance between the lines on the film becomes too small. *a*) Grating recorded on a film with low resolving power. *b*) The same grating recorded on a film with a much greater resolving power. (Taken from W. Meidinger. Die theoretischen Grundlagen der photographischen Prozesse, J. Springer, Vienna, 1937.)

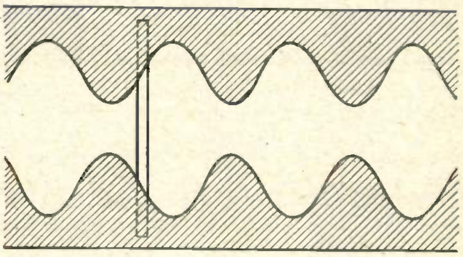


Fig. 2. Variable area sound track recorded on film. This track is played by throwing onto it a narrow beam of light perpendicular to the direction of the film; the light passing through varies with the width of the track and these variations are made audible by means of a photocell and amplifier.

These effects become all the more pronounced according as higher notes are recorded and the speed of the film is reduced. Consequently, for certain admissible values of distortion and of amplitude loss in the high frequencies, the film must have a high resolving power if it is desired to reduce the film speed.

### Cause of limited resolving power

Obviously the limited resolving power of a film is related to the grain of the film material. The granular structure of the emulsion [3]) is in itself so fine that its irregularity would still allow of a grating being resolved of much more than 50 or 70 lines per mm. The disturbing effect lies rather in the diffusion of light (also called the circle of confusion) produced when exposing the film and caused by the scattering of light by the grains in the emulsion. Owing to this diffusion there is light playing in a turbid emulsion layer in places where it ought not to be. As a result "image spread" arises to a greater or lesser degree according to the intensity (and also the wavelength) of the light. With the gelatine emulsion commonly used this phenomenon is, it is true, counteracted to a certain extent by the so-called Ross effect, an "image contraction" due to the slight shrinkage of the gelatine where silver is separated. At a certain density [4]) there is a balance between the two effects.

[3]) The term "emulsion" is not really correct: we have here a colloidal solution, not of liquid globules, but of solid particles, the AgBr grains in the gelatine. It has, however, become so common to speak of emulsion that we will not depart from it here.

[4]) "Density" is defined as $D = \log I_0/I$, where $I$ is the quantity of the incident light $I_0$ that the blackened plate allows to pass through.

For practical purposes, however, this density is much too small (*e.g.* about 0.5), The normally required density is absolutely predominated by image spread, so that the adjacent lines of a too fine grating or, in general, details at too short distances apart are caused to blur.

The magnitude of the image spread can be deduced by experimenting with a circular image of about 0.1 mm in diameter reproduced on a film under different exposures $E$ (intensity × exposure time). Such experiments, first carried out by the astronomer Scheiner, have shown that with not too small and not too large values of $E$ the diameter $d$ of the photographic picture can be represented by

$$d = a + b \log \frac{E}{E_0} \quad \ldots\ldots\ldots \quad (1)$$

where $E_0$ is the exposure under which the image spread is just balanced by the Ross effect. Since for $E = E_0$ the second term on the right disappears, the constant $a$ equals the true image size that would be obtained in the absence of the phenomenon of image spread and contraction. Hence the image spread $s$ may be written as:

$$s = b \log \frac{E}{E_0}, \quad \ldots\ldots\ldots\ldots \quad (2)$$

in which $b$ is an empirically determined constant.

Considering that in recording or in copying a certain density is reckoned with, we express the image spread as a function of the density $D$. The ratio of $D$ to $\log E$ for any photographic negative or positive is indicated by the H & D curve as given in *fig. 3*. This curve is usually characterized by the "gamma", *i.e.* the slope $\gamma$ of the linear portion. If we define $D_0$ as the density corresponding to the exposure $E_0$ (thus where there is no image spread) then:

$$D - D_0 = \gamma (\log E - \log E_0). \quad \ldots \quad (3)$$

From (2) and (3) we derive for the image spread [5]).

$$s = \frac{b}{\gamma} (D - D_0). \quad \ldots\ldots \quad (4)$$

The conclusion of practical importance to be drawn from this is that, other things being equal, a high gamma of the film (of course assuming that $b$ is not appreciably dependent on $\gamma$) is favourable for restricting image spread and, consequently, for increasing the resolving power. This is understandable also without formulae. The circle of light produced by diffusion around an exposed spot is strongest in the middle and gets weaker towards the periphery. If a film with high gamma is used (*e.g.* that of the curve in *fig. 3b*) and the prescribed density $D$ in the exposed spot is aimed at, then in the surrounding circle there will only be a perceptible density where the exposure is not much less than in the spot itself; thus the resulting circle of diffusion is limited to a fraction of the aforementioned circle.

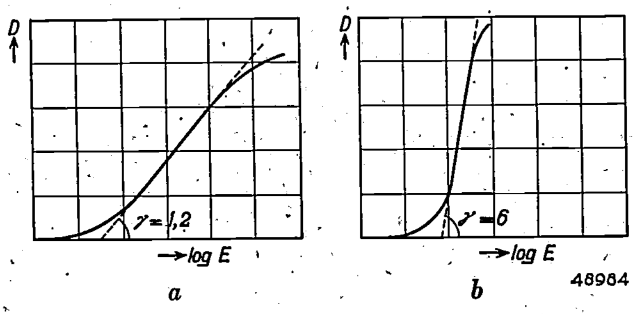


Fig. 3. H & D curve for a photographic negative or positive represented diagrammatically. $D$ = density, $E$ = exposure (= exposure intensity × exposure time), $\gamma$ = slope of the rectilinear portion of the curve. The gamma depends upon the kind of film and the method of developing. *a*) The case of a low gamma; *b*) the case of a high gamma.

In the case of a sound track (fig. 2), where there is essentially only one density value to be reckoned with [6]), it is indeed only right to aim at the highest possible gamma. For the reproduction of pictures, however, this course cannot be followed, and this we will revert to again later on.

## Practical consequences of the limited resolving power

### a) *For pictures*

As we have seen, owing to the limitation of the resolving power the size of the picture in a film for projection cannot be chosen below a certain limit. This limit is easily calculated. Reckoning with the resolving power of 55 lines per mm given above for positive films, then details of about 0.02 mm are still sufficiently separated on the film; smaller details run into each other. If the details do run into each other a little it does not matter so long

[5]) Equation (3) only holds when the density $D_0$ (and $D$) lies on the rectilinear part of the characteristic curve, which it does not as a rule do, because $D_0$ is too small for that. Still one can then write for $s$ the formula (4), where only the constants $b$ and $D_0$ (and in eq. (1) also $a$) are different from those determined experimentally.

[6]) This is not exactly true. The track is recorded by photographing a slit varying in length. Owing to the finite width of this slit one gets on the edge of the track a transitory zone with diminishing density. This implies that in order to avoid distortion it is necessary to apply the Goldberg condition, to which we will refer farther on, as a result of which one would no longer be free in the choice of the gamma. In practice, however, image spread appears to be so much greater that the conclusion still holds that one should aim at a high gamma.

as the cinema patrons do not notice it. Normally the human eye can only discern details and faults in their reproduction when they are viewed under a visual angle of about one minute (at least in the range of brightness with which we are concerned here). This corresponds to 0.1 mm at the distance of clear vision (25 cm), or 6 mm at a distance of 15 metres, which may be taken as the average distance of the patrons from the screen. Therefore the film picture can be projected in a cinema with a linear magnification of 6 : 0.02 = 300 times without the public noticing anything of the consequences of the limited resolving power of the film, except perhaps in the decidedly unfavourable seats close in front of the screen. Reckoning that in a cinema the picture is projected over a width of say 6 metres, the minimum width of the picture on the film will have to be 20 mm. This is in fact approximately the picture width on the 35 mm standard film (the rest is reserved for sound track and perforations).

Besides the 35 mm film, however, the 16 and 8 mm films have become very popular, with corresponding widths and heights of the picture frames as indicated in *fig. 4*. When projecting at the rate of 24 frames per second the speeds of these three sizes of films [7]) are respectively 45.6 cm/sec., 18.3 cm/sec and 9.2 cm/sec.

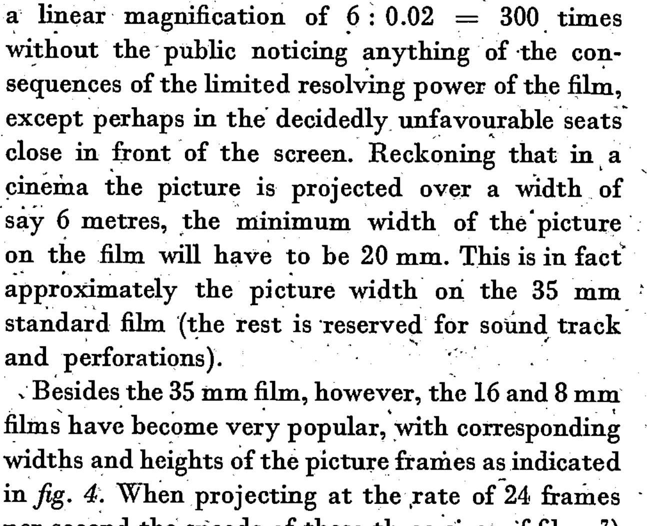


Fig. 4. The three sizes of film commonly used: 35, 16 and 8 mm wide. For each film the width and height of the photographed picture frames are given, the slightly smaller width and height of the film gate in the projector (thus the dimensions of the picture to be projected) and the width of the sound track.

One can see at a glance how much more economical the 16 mm film is than the 35 mm one. For a show of 1 hour 10.5 m² of the 16 mm film is required as against 57.5 m² of the 35 mm film. If, however, the 16 mm film were to be projected under the same conditions as described above then the limit of resolving power of the positive would undoubtedly already be exceeded and the screen picture would not be absolutely sharp. The fact that nevertheless much use is made of the 16 mm film is due partly to the manner of projection, the public seeing the picture on the screen at a smaller angle (less of the undesired properties but also less of the desired features are seen). It is also in part due to the fact that a certain lack of sharpness in moving pictures need not be decidedly troublesome, for one has, so to speak, no time to notice the lack of sharp definition and the movement itself already distracts attention from the observation of details.

With the 8 mm film, however, even these attenuating circumstances are of no help. Although special film material is used ("reversal" film) with exceptionally fine grain, the picture projected from an 8 mm film is never really sharp. It is therefore practically impossible to reduce the already small frame (4.8 mm) to make room also for a sound track on this size of film. Moreover, the low speed necessary for a film with such a small frame height does not allow of proper sound reproduction. This is why 8 mm films never have sound track and their use is limited to the home kino and to amateurs.

b) *Consequences for sound*

Taking the film speed as fixed, then according to the foregoing the limited resolving power sets a limit to the highest frequencies that can be recorded or copied with sufficient amplitude (here we will disregard a similar limitation due to the finite width of the slit in recording and reproduction).

In sound film technique, in order to avoid this restriction as far as possible, the method of cancellation has been developed. The picture spread occurring in the photographic recording of a sound track causes the black-white limit to be shifted perpendicular to the edge of the sound track. The magnitude of this displacement depends upon the curvature of the original edge. The resulting limitation is therefore by no means congruent with that which would occur without image spread. *Fig. 5b* shows that a sine is distorted to a sugarloaf shape: the peaks are rounded off, the valleys become pointed owing to adjacent delineations

[7]) In the case of the 8 mm film one usually projects at the rate of only 16 frames per second, so that the speed becomes 6.1 cm/sec. In this case 16 frames suffice, because the small pictures are generally projected with a relatively low brightness and the eye is then less sensitive to flickering. The lower speed is not attended with the drawback of sound quality because the usual 8 mm film does not carry any sound track; see below.

flowing into each other. The result is a non-linear distortion of the recorded sound, particularly in the high frequencies. However, when a positive copy is made of the distorted negative sound track there is again image spread, this time causing the track edge to be displaced in the direction opposite to that in the original track. By a suitable choice of density and maybe the gammas of the negative and positive films it is possible to cause the two image spreads to compensate each other fairly well; see *fig. 5c*. The non-linear distortion is then practically eliminated, but, as a comparison of figs. 5c and 5a shows, there still remains a perceptible loss of amplitude in the high frequencies. Regardless of the application of cancellation, in order to minimise this loss it is an advantage to reduce image spread by choosing a high gamma for the negative and positive films.

We will devote for a moment special attention to the problems arising in the copying of sound track recorded by the Philips-Miller system, the principle of which is represented in *fig. 6* [8]).

A celluloid film tape is coated first with a layer of gelatine and over that a thin opaque covering layer. This "Philimil" tape is drawn through underneath a wedge-shaped cutter moving up and down in rhythm with the sound vibrations and thus, by removing the covering layer, cutting in the tape a transparent track of varying width. Here, then, the usual photographic method of recording a sound film is replaced by a mechanical method, with the advantages that the film can be played immediately after its recording (done by optical means, the same as ordinary sound films) and in the reproduction of the high frequencies there is no trouble whatever from the feared image spread. Thanks to the ideally sharp delineation of the track, frequencies up to 8000 c/s can easily be recorded and reproduced, even at a film speed of only 32 cm/sec. Thus one gets a very high quality of sound reproduction. This system has, therefore, already found favour in broadcasting studios for recording commentaries and for other special purposes. For its application in cinemas, however, the "Philimil" recording has to be copied. The same applies if a system for sound reproduction in the home is to be based on the Philips-Miller system, in order to get a step ahead of the gramophone record (on which the high frequencies are reproduced only moderately well). For the mass production of copies only a photographic process can be considered: from the "Philimil" tape, on which the sound track is already positive, a negative has to be printed and from this "intermediate negative" any number of positive copies can be made. If the ordinary photographic methods are applied, however, the phenomenon of image spread is again introduced, and it is just this that has been eliminated in the mechanical recording. In the two photographic copying processes one could apply the principle of cancellation, but then the gain in quality is again partly lost.

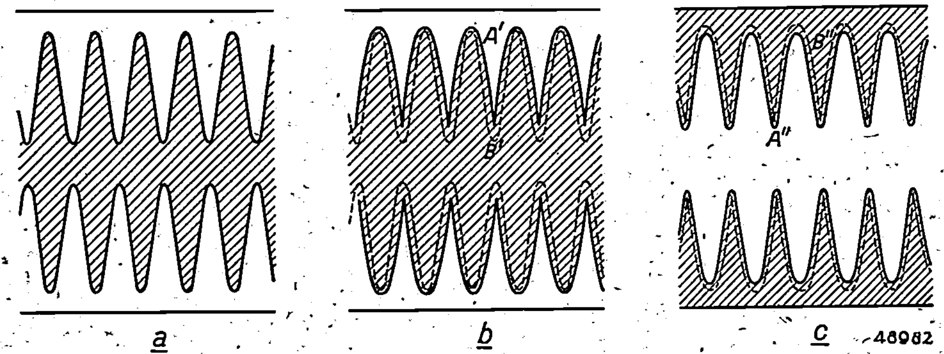


Fig. 5. Cancellation of image spread when recording and copying a sound track, represented diagrammatically.

*a*) Negative track of a certain frequency as it would be recorded if there were no image spread.

*b*) Owing to the image spread the inner parts *B'* of the sines run more or less closer to each other and form sharp peaks, whilst the outer parts *A'* are rounded off. The ideal shape (*a*) is drawn in dotted lines.

*c*) When the distorted track (*b*), indicated here by dotted lines, is copied, image spread again occurs, the peaks *A''* being widened and the rounded parts *B''* narrowed. In this way a practically sinusoidal track can again be produced. It is seen at once, however, that the amplitude of the sine modulation has become smaller than it was in (*a*): the track displacements at *B'* and *B''* are large, those at *A'* and *A''* only very slight.

[8]) For the fundamental principles of this system see Philips Techn. Rev. 1, 107, 1936.

Now in the copying of "Philimil" band a peculiar phenomenon occurs, the so-called lens effect. From *fig. 6* it may be seen that in the recording of the sound track the cutter makes in the transparent gelatine a groove of varying depth and of a triangular cross section. This profile of the gelatine surface acts as a series of lenses which in the copying process break up the homogeneity of the incident light and give rise to alternately very high and very low concentrations of light behind the transparent part of the track. The density differences resulting from this in the negative copy can easily be rendered harmless by seeing to it that in the subsequent copying of the negative also the rather large quantity of light passing through the least dense parts is kept below the threshold of sensitivity of the positive film. The track on the positive will then again be uniformly transparent. The awkward part about this, however, is that in order to get a sufficiently negative density in the parts of the weakest concentration of light one cannot avoid a much more intensive exposure — in practice up to 20 times as much — in the parts with the strongest concentration of light. This results in a marked circle of diffusion and image spread on the track edge where these strong concentrations of light occur.

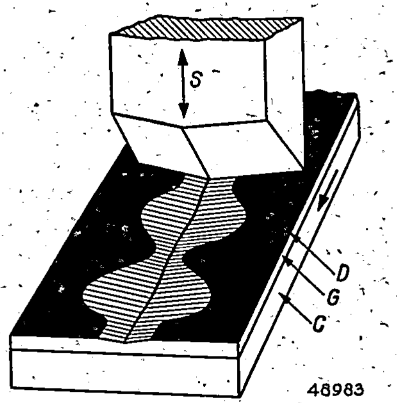


Fig. 6. Principle of the Philips-Miller system of sound recording. The "Philimil" band consists of the celluloid support *C* with a gelatine coating *G* and a very thin opaque covering layer *D*. The cutter *s* vibrating perpendicularly to the band moving in the direction of the arrow cuts out a transparent sound track.

This phenomenon can be counteracted in various ways. For instance, the track could be filled up with a paste or a liquid having the same refractive index as that of the gelatine layer. The easiest way of solving the problem, however, is to choose for the intermediate negative a film with such a high gamma that even with the very intensive light concentrations mentioned above no trouble is experienced from image spread. The desire for a high gamma as already made manifest thus becomes still more imperative. Incidentally, this solution of the problem can only serve if the gamma is not already fixed by the cancellation method first described.

### c) *Combination of picture and sound*

In the case of a sound film sound and picture are placed side by side on the same band. In practice this makes it impossible to counteract image spread by aiming at the highest possible gamma, because the gamma is already determined by the picture reproduction. In order to reproduce in the picture by analogous half-tone contrasts all the visual contrasts of the object photographed it is necessary to comply with Goldberg's condition. This condition requires that the product of the gammas of the positive and the negative must be about 1—1.2 [9]). This leads to the compromise of a picture negative with a gamma of abt. 0.8 and a combined sound-picture positive with a gamma of 1.8—2.5, although it would be much more convenient to print the sound with a gamma of 4—6 and to choose for the picture a gamma much nearer to 1. As a matter of fact the choice of gamma also for the picture alone is a compromise, for a different gamma may be desired for the negative film according to the scene taken, and then one would want to adjust the gamma of the copy from scene to scene. Since the gamma of a certain photographic film can only be influenced to any appreciable extent by the developing times, and this can hardly be varied every time for different parts of the film, one must be satisfied with the choice of one particular mean gamma.

Reviewing the situation as a whole we may say that a high resolving power of the film is desirable both for the picture and for the sound track; that a high gamma is favourable but that in the case of a combination of picture and sound the Goldberg condition makes it necessary to arrive at a compromise in this respect. This compromise makes its influence felt throughout the whole of the present-day technique of cinematography. The new system of reproduction referred to in the beginning of this article offers a better solution to this problem.

## The new reproduction system

This system is based not on an emulsion of AgBr or AgCl in gelatine but on a combination of a light-sensitive diazonium compound with a mercury

[9]) Regarding the Goldberg condition see, for instance, Philips Techn. Rev. 5, 51, 1940. Roughly speaking, this condition expresses nothing more than what is known to any photographer, that a "hard" positive paper must be used for printing from a "soft" negative, and *vice versa*.

salt. The film base is saturated in a solution of this mixture. The system can be applied with different bases. A thin band of cellophane, for instance, is highly suitable for the purpose. After exposure the latent picture obtained is developed by so-called physical development, metallic silver from a solution of a silver salt being deposited on the exposed parts (in contrast to the normal, so-called chemical developing where a silver compound present in the film is reduced to metallic silver in the exposed parts). In this way one obtains a perfectly durable negative image consisting of silver. Compared with the usual photographic negative material, the sensitivity of the new material is low. There can therefore be no question of its competing with silver-bromide emulsion for the taking of films, also because of its sensitivity being restricted to ultra-violet light. For making copies, on the other hand, this mercury diazonium system possesses excellent properties. Since there is here no question of an emulsion but of a homogeneous solution, there are no grains and the film has a resolving power of 1000 lines per mm [10]). The gamma for the dry material is exceptionally high, *viz.* 6—8. A most remarkable feature, however, is that it can easily be adjusted within a wide range already during the exposure, it being possible to reduce it to a very low level.

Thanks to this very high resolving power and the possibility of having a very high gamma, the system is ideal for the reproduction of sound track and also for the photographic multiplication of printed matter and suchlike, where the extremely low cost of the base and of the sensitized material may prove to be of importance. The specific problems arising in the copying of Philips-Miller film are solved by the new system in the simplest possible way. The fact that, if desired, the gamma can be reduced to a low value makes it possible to get good reproductions also of half-tone pictures. Furthermore, the adjustability of the gamma during exposure creates the entirely new possibility of printing picture and sound on the same band with a different gamma and then developing both together.

This brief account of the subject will presumably give rise to more questions than it answers. Our intention here has been only to arouse the reader's interest. Further particulars will be found in one of the next issues of this journal, where we intend to go more closely into the new system of photographic reproduction.

[10]) The question whether the film is capable of resolving still finer gratings could not be determined because the optical system used for the necessarily reduced recordings cannot itself resolve more than 1000 lines per mm.

# ELECTROMAGNETIC CAVITY RESONATORS

by G. de VRIES.

538.565

The forms of oscillation of certain electromagnetic flat cavity resonators are discussed, *i.e.* resonators which may be considered as two-dimensional; namely the forms of oscillation of square plane cavity resonators and the non-rotation-symmetrical forms of oscillation of round plane cavity resonators. Further, the forms of oscillation of three-dimensional cavity resonators are dealt with. The case is then discussed of two coupled cavity resonators, the significanee of which in high-frequency technology is analogous to that of coupled oscillation circuits in the region of lower frequencies, namely to that of a band filter.

As has already been explained several times in this periodical [1])[2])[3]), ordinary oscillation circuits consisting of concentrated self-inductions and capacities are not suitable for the region of very short waves. The quality factor and the resonance resistance of such oscillation circuits are much too small in that region. Recourse is therefore had to the use of other electrical resonators not having these objections. Very important among these resonators are Lecher systems, *i.e.* two parallel conductors close together or concentric, and cavity resonators, *i.e.* empty spaces surrounded by metal walls. Lecher systems have already been discussed in detail in this periodical [2]). Recently a detailed explanation was also given of the properties of a certain group of cavity resonators [3]), namely of the cavity resonators which are in the form of solids of revolution and in which, moreover, the dimension in the direction of the axis of revolution is small compared with the dimensions perpendicular to it (*fig. 1a*). We have called this kind of cavity resonators round flat cavity resonators. We were then concerned exclusively with the rotation-symmetrical oscillations in which the current and the voltage depend only on the distance to the axis of revolution.

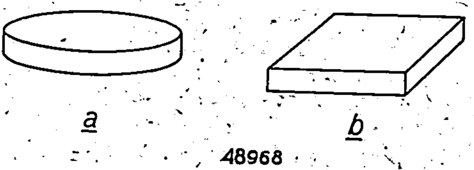


Fig. 1. *a*) A round plane cavity resonator, *b*) a square plane cavity resonator.

In this article we shall go somewhat farther. We shall begin with the consideration of square cavity resonators and especially of such which are not only flat, *i.e.* thin, but also plane, *i.e.* have everywhere the same thickness (fig. 1*b*). The forms of oscillation of square plane cavity resonators are naturally no longer rotation-symmetrical. With our knowledge of square plane cavity resonators we shall then be able to ascertain the characteristics of the non-rotation symmetrical forms of oscillation of the round plane cavity resonators. Moreover, this knowledge will make it possible for us to find the form of oscillation of a cube by an obvious generalization. Thus beginning with a practical two-dimensional case (square plane cavity resonator) we can arrive at conclusions about a three-dimensional case (cube). After a few remarks about the quality factor and the resonance resistance of cavity resonators we shall in conclusion discuss coupled cavity resonators.

## Forms of oscillation of square plane cavity resonators

### *Qualitative considerations*

We shall first mention briefly the result previously obtained (see footnote [3]) for the distribution of current and voltage in the rotation-symmetrical forms of oscillation of a round plane cavity resonator. Such a cavity resonator has a "bottom" and a "cover". We shall consider the voltage between a point of the cover and the point on the bottom lying directly beneath it. The current at two such points is equal in value and opposite in direction. It is thus sufficient to speak about the current, for instance, in the cover, as we shall do here.

The centre of the cover is a point of high voltage [4]); at the outer edge the voltage is zero. The current, on the other hand, is zero at the centre; at all other points it is radially directed and reaches a maximum at the outer edge. The current density is a maximum at a short distance from the edge and then decreases slightly towards the edge, but not very strongly, so that the total current, which is equal to the product of current density and circumference, continues to increase from that point to the edge. This holds for all rotation-symmetrical

[1]) C. G. A. von Lindern and G. de Vries, Resonance circuits for very high frequencies, Philips Techn. Rev. 6, 217, 1941.

[2]) C. G. A. von Lindern and G. de Vries, Lecher Systems, Philips Techn. Rev. 6, 241, 1941.

[3]) C. G. A. von Lindern and G. de Vries, Flat cavity resonators as electrical resonators, Philips Techn. Rev. 8, 149, 1946.

[4]) In the previous article (see footnote [3])) it was mainly a question of flat cavity resonators with a circular hole in the centre of cover and bottom. Here, for the present, we are discussing cavity resonators with no hole.

oscillations which further differ from each other in the number of current maxima and zero points along the radius between the centre of the cavity resonator and its outer edge. In the case of the fundamental oscillation the current increases from the centre towards the edge. In *fig. 2a* several current lines are drawn for that case, whereby the width of the lines has been chosen proportional to the current density. In fig. 2*b* the distribution of the current density is shown for the lowest but one rotation-symmetrical oscillation.

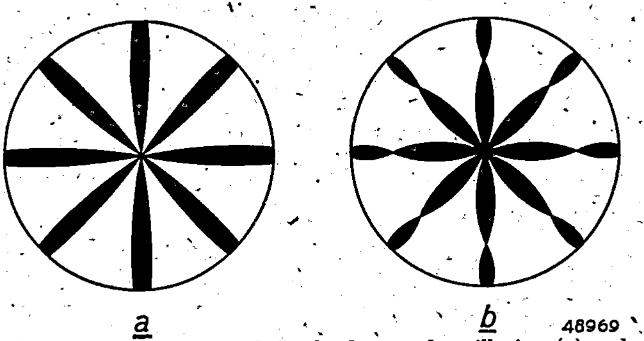


Fig. 2. Representation of the fundamental oscillation (*a*) and of a higher characteristic oscillation (*b*) of a round plane cavity resonator. A few current lines have been drawn whose width is proportional to the current density. Due to the rotational symmetry all the "lines" are congruent.

When we now attempt to sketch a current density distribution for the fundamental oscillation of a square plane cavity resonator it may be assumed that near the centre of the cover the situation will not be very different from that in the case of round plane cavity resonators. The form of the current lines near the edge is determined by the condition that the current lines must be perpendicular to the edge. This condition, whose derivation we shall not give here, enables us to draw the current distribution of a square plane cavity resonator in the neighbourhood of the edge. In this way a picture of the current distribution is obtained which agrees very well with that found by calculation (see *fig. 3*).

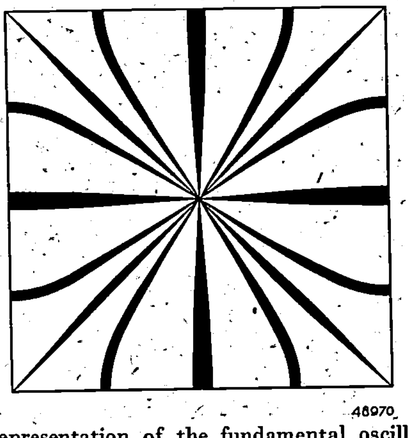


Fig. 3. Representation of the fundamental oscillation of a square plane cavity resonator. Only a few current lines have been drawn whose width is proportional to the current density. The latter differs everywhere from zero except at the centre and at the four corners of the square.

Since over a large part of the square cavity resonator the current distribution does not differ from that of a round cavity resonator, the characteristic frequency will also not deviate too much. In the main the difference between the two kinds of cavity resonator amounts to the fact that in the case of the square plane resonator "corners" have been added. Now close to the edge the voltage is low, so that it is not so much the "capacity" as the "self-induction" of the corners that is decisive for the change in the value of the characteristic frequency. In as far as it is possible to define a conception such as the "average self-induction" of the cavity resonator, such a quantity would become larger by the addition of the corners and the wavelength would thus also become larger. The mathematical theory, which will be discussed in the following, confirms this: while the wavelength of the fundamental frequency of a round cavity resonator is equal to $2{\cdot}61a$ ($a$ is the radius of the round cavity resonator), that of the square one is equal to $2a\sqrt{2} = 2{\cdot}82a$ ($2a$ is the side of the square cavity resonator).

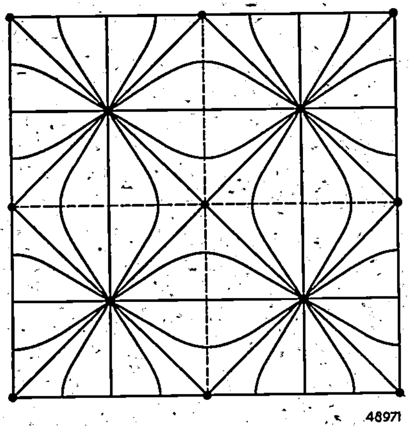


Fig. 4. The occurrence of a higher characteristic oscillation of a square plane cavity resonator. This may be conceived of as the result of placing four smaller square resonators side by side, each of which executes the fundamental oscillation. The points indicate the spots where the current density becomes equal to zero.

Now that we have the figure of the current distribution corresponding to the fundamental frequency of a square plane cavity resonator we can go farther. Let us, for example, place four square plane cavity resonators (side $2a$) side by side in such a way that they form a large square (side $4a \equiv 2b$); see *fig. 4*.

If we then arrange matters in such a way, for example, that the voltage at the centre of the cover of the upper left-hand resonator is positive and that of the upper right-hand resonator negative, the lower right hand positive and the lower left-hand negative, and that these voltages are equal in absolute value, the adjacent currents at a partition are equal and in the same direction. We may therefore omit the partitions without changing the current distribution of the four cavity resonators. From this we conclude that the current distribution in fig. 4 corresponds to a higher form of oscillation of the large square cavity resonator. The longest wavelength for this cavity resonator would be equal to $2b\sqrt{2}$; here, however, we are concerned with a form of oscillation to which the wavelength $2a\sqrt{2} = b\sqrt{2}$ corresponds.

By again placing side by side four square cavity resonators, each of which is in the higher oscillation state just described, we would in a similar manner find the current distribution corresponding to a wavelength four times as small as the wavelength $\lambda_0$ of the fundamental oscillation.

It is clear that in this manner it would be possible to find the current distribution for a whole series of forms of oscillation whose wavelength $\lambda$ would be given by the formula $\lambda = \lambda_0/2^n$ $(n = 0, 1, 2, \ldots)$. It must not, however, be thought that all the forms of oscillation of a square plane cavity resonator would then have been found; from the mathematical treatment, to which we shall now pass on, it will be found that many other forms of oscillation are still possible.

*Mathematical treatment*

The strict theory of electromagnetic cavity resonators is based directly on Maxwell's equations. It is often possible, however, to deduce correct results in a more elementary way, especially for plane cavity resonators.

Thus, for example, in the article already referred to (see footnote [3])) we have developed the mathematical theory of round flat (not only plane!) cavity resonators beginning with the ordinary "cable equations" applying to a Lecher system consisting of two parallel conductors (see *fig. 5a*):

$$\frac{\partial i}{\partial x} = -C\frac{\partial V}{\partial t}, \quad \frac{\partial V}{\partial x} = -L\frac{\partial i}{\partial t}; \quad . \ . \quad (1)$$

in these equations $C$ and $L$ are the capacity and the self-induction, respectively, per unit of length; $V$ is the voltage between a point of the upper conductor and the point of the lower conductor directly below it; $i$ is the current in the upper conductor; that in the lower conductor is then $-i$; finally $x$ is the coordinate in the direction of the conductors and $t$ is the time. By rotating such a Lecher system a round plane cavity resonator (see fig. 5a) is obtained, and by rotating a Lecher system of a general form, such as that of fig. 5b, a round flat cavity resonator is obtained which is no longer plane and which, moreover, has a hole at the centre. It is thus understandable that the rotation-symmetrical modes of oscillation of round flat cavity resonators are also described by equation (1), on the understanding that $C$ and $L$ are, respectively, the capacity and self-induction per ring of 1 cm width, and that in general it will be functions of the radius $r$ which here take over the rôle of the coordinate of length $x$.

In the case of a square plane cavity resonator the method of treatment sketched above could also be used. We shall indicate briefly how this should be done, although it is not of great importance since the rigorous calculation directly from Maxwell's equations is no more laborious. Again we introduce the voltage $V$ between a point on the "cover" and the point on the "bottom" directly beneath it. Further we introduce the current density which at every point, for instance, of the cover of the square resonator has an $x$ and a $y$ component, $i_x$ and $i_y$; we assume at the same time that the coordinate axes are parallel to the sides of the square. The current density components in the bottom will then be $-i_x$ and $-i_y$. Finally we introduce the capacity $C$ and the self-induction $L$ per cm² of base plane. It is then obvious, analogous to equation (1), to write the following equations:

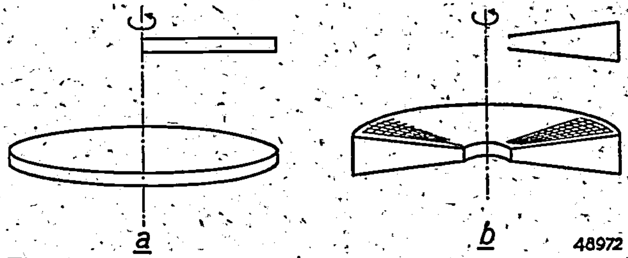


Fig. 5. The formation of a round flat cavity resonator by the rotation of a Lecher system short-circuited at one end: a) plane cavity resonator, b) resonator of a more general form.

$$\left.\begin{aligned} \frac{\partial V}{\partial x} &= -L\frac{\partial i_x}{\partial t}, \\ \frac{\partial V}{\partial y} &= -L\frac{\partial i_y}{\partial t}, \\ \frac{\partial i_x}{\partial x} + \frac{\partial i_y}{\partial y} &= -C\frac{\partial V}{\partial t}. \end{aligned}\right\} \quad . \ . \ . \ . \quad (2)$$

The value to be substituted here for $C$ is obvious: if $h$ is the thickness of the cavity resonator then $C$ is the capacity of a condenser of 1 cm² area and with a plate distance $h$. On the other hand it is difficult to assign to $L$ a similar physical significance.

Further consideration shows, however, that $L$ follows from the relation $LC = 1/c^2$, in which $c$ is the velocity of light.

The conditions at the edge to be satisfied by the solutions of (2) have already been mentioned: at all points on the edge $V$ must equal 0 and the current density must there be perpendicular to the edge. When the side of the square cavity resonator has the length $2a$ and the origin of the $x$, $y$ coordinates is situated at the centre of the square, the second condition means that $i_x = 0$ for $y = +a$ and $i_y = 0$ for $x = \pm a$.

We shall not discuss here all the periodic solutions of (10). As an example we give the following family of solutions:

$$\left.\begin{aligned} V &= V_0 \cos k_1x \cdot \cos k_2y \cdot \cos \omega t, \\ i_x &= -V_0 \frac{k_1}{\omega L} \sin k_1x \cdot \cos k_2y \cdot \sin \omega t, \\ i_y &= -V_0 \frac{k_2}{\omega L} \cos k_1x \cdot \sin k_2y \cdot \sin \omega t, \end{aligned}\right\} \quad (3)$$

with

$$k_1^2 + k_2^2 = \frac{\omega^2}{c^2}; \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

$2\omega$ is here the frequency of the oscillations and $V_0$ an arbitrary constant. In order to satisfy the boundary conditions the following must hold:

$$a\,k_1 = m\,\frac{\pi}{2}, \quad a\,k_2 = n\,\frac{\pi}{2} \quad (m,\ n = 1,\ 3,\ 5\ldots) \quad (5)$$

The mode of oscillation sketched in fig. 3 corresponds to $m = n = 1$. Since the wavelength $\lambda$ is equal to $2\pi c/\omega$, with the help of (4) and (5) it is easy to confirm the statement we made about this mode of oscillation, namely that in this case $\lambda = 2a\sqrt{2} = 2{\cdot}82\ a$.

On the other hand the mode of oscillation represented in fig. 4 is an example of a solution *not* given by a formula of the form (3). This mode of oscillaton corresponds to the following family of solutions:

$$\left.\begin{aligned} V &= V_0 \sin k_1x \cdot \sin k_2y \cdot \cos \omega t, \\ i_x &= -\frac{V_0}{\omega L} \cos k_1x \cdot \sin k_2y \cdot \sin \omega t, \\ i_y &= -\frac{V_0}{\omega L} \sin k_1x \cdot \cos k_2y \cdot \sin \omega t, \end{aligned}\right\} \quad (6)$$

with

$$k_1^2 + k_2^2 = \frac{\omega^2}{c^2}$$

and

$$ak_1 = m\pi, \quad ak_2 = n\pi \quad (m,n = 0,\ 1,\ 2\ldots):$$

For $m = n = 1$ we obtain the mode of oscillation in question. The wavelength is now $\lambda = a\sqrt{2} = 1{\cdot}41\ a$.

## Non-rotation-symmetrical form of oscillation of a round plane cavity resonator

The mode of oscillation of a square cavity resonator represented in fig. 4 could not be derived directly from a rotation-symmetrical form of oscillation of a round cavity resonator. Conversely, however, from this higher mode of oscillation of a square resonator we may now derive a non-rotation-symmetrical form of oscillation of a round plane cavity resonator. It is not difficult to make a sketch on the basis of fig. 4 of the current distribution in a round resonator for the mode of oscillation in question. In the main this current distribution will present the same picture as in the case of the square resonator, but near the edge we must alter the current lines somewhat so that they are perpendicular to it and thus radial in direction. This leads to *fig. 6b.* The application of a more rigorous mathematical theory fully confirms the correctness of this figure.

It is useless to attempt to force the theory for the non-rotation-symmetrical forms of oscillation of round resonators into the mould of the cable equations. The self-induction per cm² was already a rather artificial conception; it would be still more unnatural to speak of the self-induction of a surface element $r \cdot dr \cdot d\varphi$. We shall thus give directly the results produced by the theory based on Maxwell's equations. We introduce the polar coordinates $r$ and $\varphi$ on the plane of the cover of the round cavity resonator. The voltage $V$ is then given by

$$V = V_0\, J_m\,(kr) \cos m\varphi \cdot \cos \omega t, \ (m = 0, 1, 2) \quad (7)$$

where $J_m(x)$ represents the Bessel function of the order $m$ and

$$k = \frac{\omega}{c} = \frac{2\pi}{\lambda};$$

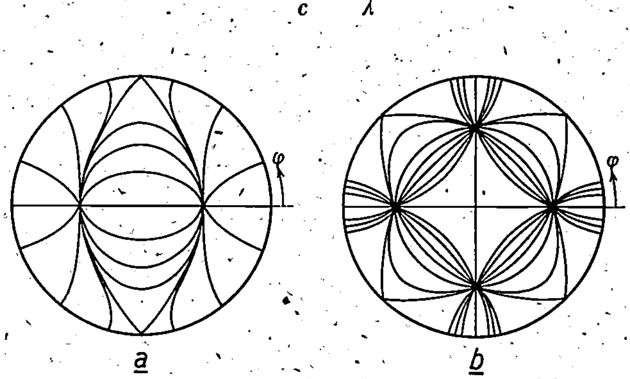


Fig. 6. Diagram of a non-rotation-symmetrical form of oscillation of a round plane cavity resonator. Only a few current lines are drawn here. *a*) The mode of oscillation where the current density near the edge of the resonator is proportional to cos $\varphi$. *b*) The mode of oscillation where the current density near the edge of the resonator is proportional to cos $2\varphi$.

$\lambda$ and $\omega$ are the wavelength and the angular frequency respectively.

At the edge of the cavity resonator the voltage must become equal to zero. When the radius of the resonator is $R$ the following condition must be satisfied:

$$J_m(kR) = 0 \quad (m = 0, 1, 2 \ldots). \quad \ldots \quad (8)$$

Since the Bessel function $J_m(x)$ has an infinite number of roots for every value of $m$, with a given value of $m$ we obtain an infinite number of values of $k$ satisfying equation (8). With a given value of $m$ we may arrange these roots according to increasing size and assign an order number $p$ (1, 2, 3, etc.) to each root. Since $m$ itself may also have an infinite number of values, we thus obtain a doubly infinite series of characteristic oscillations. Values of $kR = 2\pi R/\lambda$ which correspond to several of the lowest characteristic frequencies, *i.e.* to several of the smallest values of $m$ and $p$, are given in *table I*. For $m = 0$ the voltage distribution is rotation-symmetrical, for $m = 1$ the voltage is proportional to cos $\varphi$, for $m = 2$ proportional to cos $2\varphi$, etc.

Table I.

Several values of $2\pi R/\lambda$ for which $J_m(2\pi R/\lambda = 0$; $m$ is the order of the Bessel function $J_m$, $p$ the order of the zero point of $J_m$.

| $m$ \ $p$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 0 | 2.405 | 5.520 | 8.654 | 11.792 |
| 1 | 3.832 | 7.016 | 10.173 | 13.323 |
| 2 | 5.135 | 8.417 | 11.620 | 14.796 |

The formulae for the components $i_r$ and $i_\varphi$ of the current density are as follows:

$$i_r = \frac{V_0}{h} \frac{c}{4\pi} J_m'(kr) \cos m\varphi \cdot \sin \omega t, \quad (9a)$$

$$i_\varphi = \frac{V_0}{h} \frac{c}{4\pi} \frac{J_m(kr)}{kr} m \cdot \sin m\varphi \cdot \sin \omega t; \quad (9b)$$

where $h$ is the thickness of the cavity resonator and $c$ the velocity of light (if $i$ and $V_0$ are expressed in amperes and volts respectively, the lengths in cm, then $1/c = 30$ ohms); the derivative of $J_m$ is $J_m'$.

The current distribution in fig. 6*b* corresponds to $m = 2$ and $p = 1$ (*c.f.* table I). It may therefore be seen that this form of oscillation is not the first following the rotation-symmetrical fundamental oscillation ($m = 0$, $p = 1$). The current distribution corresponding to $m = 1$, $p = 1$ is represented in fig. 6*a*. The forms of oscillation for which $m = 0$ and $p = 1, 2, 3, \ldots$ are all rotation-symmetrical and have already been discussed in detail in a previous article (see footnote [3]).

Because of (8) $i_\varphi = 0$ at the edge of the cavity resonator, *i.e.* the current density is radially directed in agreement with the condition formulated in the treatment of square cavity resonators. From (9a) it then follows that the current density close to the edge is proportional to cos $m\varphi$. If inside the cavity resonator we introduce along its edge a number of loops whose plane contains the axis of revolution of the cavity resonator, currents will be induced in those loops whose intensity is also proportional to cos $m\varphi$. This can easily be demonstrated by connecting small lamps to such loops. For the rotation-symmetrical form of oscillation all the lamps will burn (*fig. 7, left*); for the cos $\varphi$ mode of oscillation two groups will burn: at $\varphi = 0$ and at $\varphi = \pi$ (fig. 7 *middle*); for the cos $2\varphi$ oscillation four groups will burn: at $\varphi = 0$, $\varphi = \pi/2$, $\varphi = \pi$ and $\varphi = 3\pi/2$ (fig. 7, *right*). The form of oscillation prevailing depends upon the frequency applied to the cavity resonator.

With the above the most important facts about the forms of oscillation of square and round flat cavity resonators have been stated. The practical significance of the higher forms of oscillation of cavity resonators in general will be explained presently with reference to an example. First, however, we shall examine the appearance of the forms of oscillation in the case of "three-dimensional" cavity resonators, *i.e.* cavity resonators which can no longer be considered flat.

### Three-dimensional cavity resonators

We shall begin with cavity resonators in the form of a cube. We have seen (see footnote [3]) that the oscillations of a Lecher system of length $a$ and open at one end could be described by the formulae

$$\left.\begin{aligned} V &= V_0 \cos kx \cos \omega t, \\ i &= i_0 \sin kx \sin \omega t, \\ k &= \omega/c, \\ ak &= m\frac{\pi}{2} \; (m = 1, 3, 5, \ldots). \end{aligned}\right\} \quad (10)$$

For a square flat cavity resonator we found above (see equations (3) and (4)) formulae of the form

$$\left.\begin{aligned} V &= V_0 \cos k_1 x \cos k_2 y \cos \omega t, \\ i_x &= i_{0x} \sin k_1 x \cos k_2 y \sin \omega t, \\ i_y &= i_{0y} \cos k_1 x \sin k_2 y \sin \omega t, \\ k_1^2 + k_2^2 &= \upsilon^2/c^2, \\ ak_1 = m\frac{\pi}{2}, \; ak_2 &= n\frac{\pi}{2} \; (m, n = 1, 3, \ldots), \end{aligned}\right\} \quad (11)$$

where $2a$ is the length of the side of the square,

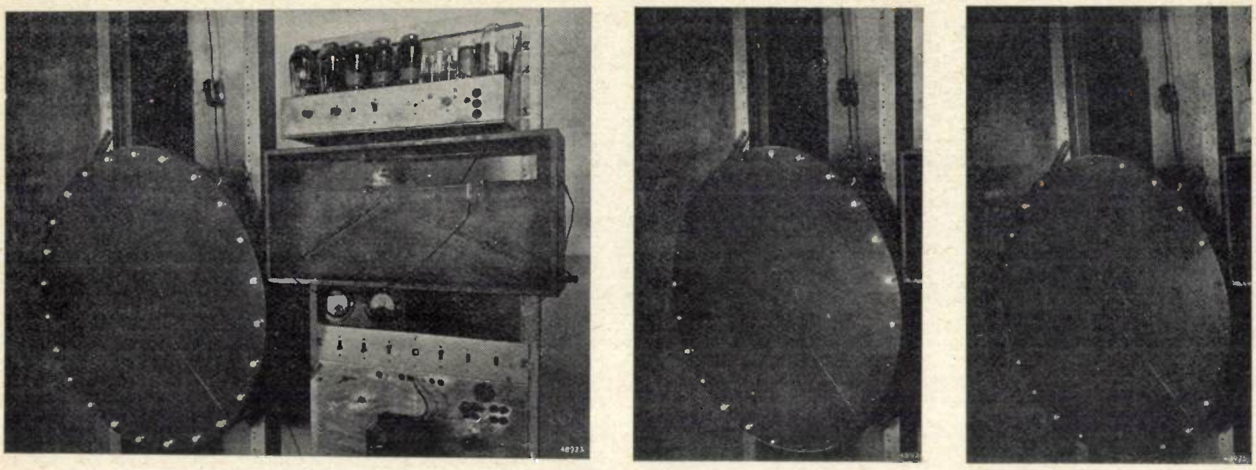


Fig. 7. *Left:* A round plane cavity resonator which executes the fundamental oscillation (*c.f.* fig. 2). This form of oscillation is rotation-symmetrical, as is demonstrated by the fact that the lamps along the edge all burn at the same intensity. These lamps are connected with loops inside the cavity resonator. The cavity resonator is excited with the help of a loop (it projects from the slot in the side wall of the resonator), which is connected with an oscillator (on the right in the figure).

*Centre:* A round plane cavity resonator in a non-rotation-symmetrical state of oscillation in which the current density near the edge is proportional to cos $\varphi$ (*cf.* fig. 6*a*). Only two groups of lamps burn: those at $\varphi = 0$ and those at $\varphi = \pi$. The oscillation takes place in such a way that the loop exciting it is situated at a position of maximum current density on the edge.

*Right:* A round plane cavity resonator in a state of oscillation where the current density near the edge is proportional to cos $2\varphi$ (cf. fig. 6*b*). Four groups of lamps are lighted, those at $\varphi = 0$, $\varphi = \pi/2$, $\varphi = \pi$ and $\varphi = 3\pi/2$.

This was the case for "one-dimensional" and "two-dimensional" systems. It is now reasonable to assume that for a cube, whose sides are also squares, analogous formulae will be applicable. We thus assign to each wall a current density distribution corresponding to equation (11). If the $x$, $y$, $z$ axes are parallel to the edges (length $2a$) of the cube and if the origin of the coordinates is at the centre of the cube, the following formulae should apply (we omit the factor sin $\omega t$): for the walls which are parallel to the $x$, $y$ plane (the $x$, $y$ walls)

$$\left.\begin{aligned}
&i_x = i_1 \sin k_1 x \cos k_2 y,\\
&i_y = i_2 \cos k_1 x \sin k_2 y;\\
&\text{for } y,\, z \text{ walls}\\
&i_y = i_3 \sin k_2 y \cos k_3 z,\\
&i_z = i_4 \cos k_2 y \sin k_3 z;\\
&\text{for } z,\, x \text{ walls}\\
&i_z = i_5 \sin k_3 z \cos k_1 x,\\
&i_x = i_6 \cos k_3 z \sin k_1 x;\\
&\text{and the following equations must}\\
&\text{thereby be satisfied:}\\
&k_1^2 + k_2^2 + k_3^2 = \omega^2/c^2\\
&\text{and}\\
&ak_1 = m\pi/2,\ ak_2 = n\pi/2,\ ak_3 = p\pi/2\\
&\qquad (m, n, p = 1, 3, 5 \ldots)
\end{aligned}\right\} \quad . \; . \; . \quad (12)$$

The current density has the same distribution, except for the sign, in two parallel walls of the cube as in the case of a flat square cavity resonator.

Are the formulae (12) now indeed correct? Before we answer this question we would call attention to the fact that among the equations (12) we have given no formula for the voltage $V$. This was not done without reason. While the current density in the walls of the cavity resonator remains a physical quantity in our somewhat daring generalization from two to three dimensions, it is difficult to picture directly what is to be understood by voltage in the case of a cubic cavity resonator. This brings us to a fundamental question, namely to what extent it is possible to speak of a voltage, *i.e.* of a potential difference, outside the field of electrostatics and direct currents. We must consider that question for a moment.

The voltage $V$, *i.e.* the potential difference between two points (for instance $A$ and $B$), can be defined unambiguously in electrostatics due to the fact that the value of the integral $\int_B^A E_s\, d_s$, in which $E_s$ is the component of the electrical field tangential to the path of integration, is independent of the path of integration. The voltage between $A$ and $B$ is then by definition

$$V_{AB} = \int_A^B E_s \,\mathrm{d}s$$

along any given path. This definition can still be used in the same form outside the field of electrostatics as long as one is only concerned with direct currents. As soon, however, as alternating currents are considered certain new conventions must be adopted in order to be able to continue to use the concept of "voltage".

Let us consider for example the simple case of a conductor in the form represented by *fig. 8*; for the sake of simplicity we shall disregard the resistance of the conductor. When an alternating current flows in the conductor the value of $\int_A^B E_s\, ds$ will certainly be different along the various dotted paths. According to the induction law the integral $\int E_s\, ds$ along a closed path — for instance from $B$ to $A$ along one path and back from $A$ to $B$ along another — is proportional to the derivative with respect to time of the magnetic flux through the area enclosed by this path. And this derivative with respect to time is certainly not equal to zero, since the magnetic field of an alternating current changes with the time. The "voltage" between $A$ and $B$ measured along different paths should then actually be different. Nevertheless, one may usually still continue to speak of voltage, if it is agreed that one means the integral $\int_A^B E_s\, ds$ along the shortest path between $A$ and $B$. This agreement is not merely formal, but usually also practically justified, because it need not by any means be exactly the shortest path. As long as the derivative with respect to time of the magnetic flux enclosed by the shortest path from $A$ to $B$ (dotted line $I$ in fig. 8) and a slightly deviating path (for instance dotted line $II$ or $III$) is small compared with $\int_A^B E_s\, ds$ along the shortest path, the difference between the voltages measured along these two paths will remain relatively small. One will then indeed continue to speak of the voltage. If the wavelength is approximately equal to the distance $AB$ or smaller, the magnetic field strength will in general change very much for distances of this order of magnitude. As a result the difference between the voltage measured along the shortest path from $A$ to $B$ and the voltage measured for instance along the dotted line $II$ may be proportionately large. But in many cases (by no means in all), also in the region of very high frequencies, it is possible to speak of the voltage, in the sense of our convention, when systems are considered whose dimensions are much smaller in one direction than in the other.

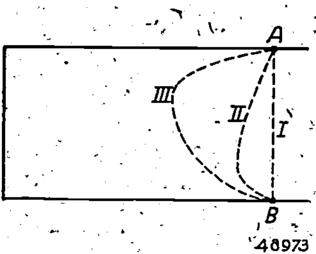


Fig. 8. The voltage between two points $A$ and $B$ of a conductor in which an alternating current flows is not unambiguously defined: the voltage measured along different paths (dotted lines $I$, $II$ and $III$) is different.

Such a case is now met with in the case of flat cavity resonators. Since the thickness of a flat cavity resonator is much smaller than the wavelength which corresponds to not all too high characteristic frequencies, the result is that the value of $\int_A^B E_s\, ds$ along the shortest path between the "bottom" and the "cover" differs relatively little from the value of the integral along a slightly different path. Therefore it was possible to speak of the voltage between a point of the "cover" of a flat cavity resonator and the point in the "bottom" directly under it. Because, however, in the case of a cubic cavity resonator the distance between two such points is not small compared with the other dimensions, it is hardly reasonable to speak of a voltage in that case [5]). This is true, of course, not only for a cubic cavity resonator but also for other "three-dimensional" cavity resonators.

The oscillations must then be described with the help of the electrical and magnetic field strength. The current density in the walls of the cavity resonator remains a reasonable idea: it has the significance of the current density induced by the variable magnetic field. The above formulae (12) actually correspond to certain forms of vibration of the cubic cavity resonator and give us an idea of the distribution of current density in its walls. It is found that there are relations between the constants $i_1$, $i_2$, $i$ . . . $i_6$, which we shall not go into here. In *fig. 9* the distribution of the current density is sketched for two modes of oscillation both of which correspond to $m = n = p = 1$ in equation (12), but to different values of the $i$'s.

For cylindrical cavity resonators with circular cross section formulae are obtained for the current density in the walls which are again a generalization of equation (9) for round flat cavity resonators. Further at the cost of much calculation the current

[5]) For certain forms of oscillation of a cube, namely when the electric field is everywhere perpendicular to a side wall of the cube, the voltage could be defined in a fairly natural manner as the integral $\int E_s\, ds$ along the line of force. A slight deviation from this path of integration would then, however, result in an appreciably different value of the integral, so that this definition would only have a formal significance.

distribution can be studied for spheres, ellipsoids and elliptical cylinders, without enriching the qualitative picture already formed of the possible modes of oscillation of a cube. Rather than enter into these mathematically fairly difficult questions, we shall at the end of this article tell something about a case which in a different respect is more complicated than those considered until now and which does indeed widen our field of vision somewhat.

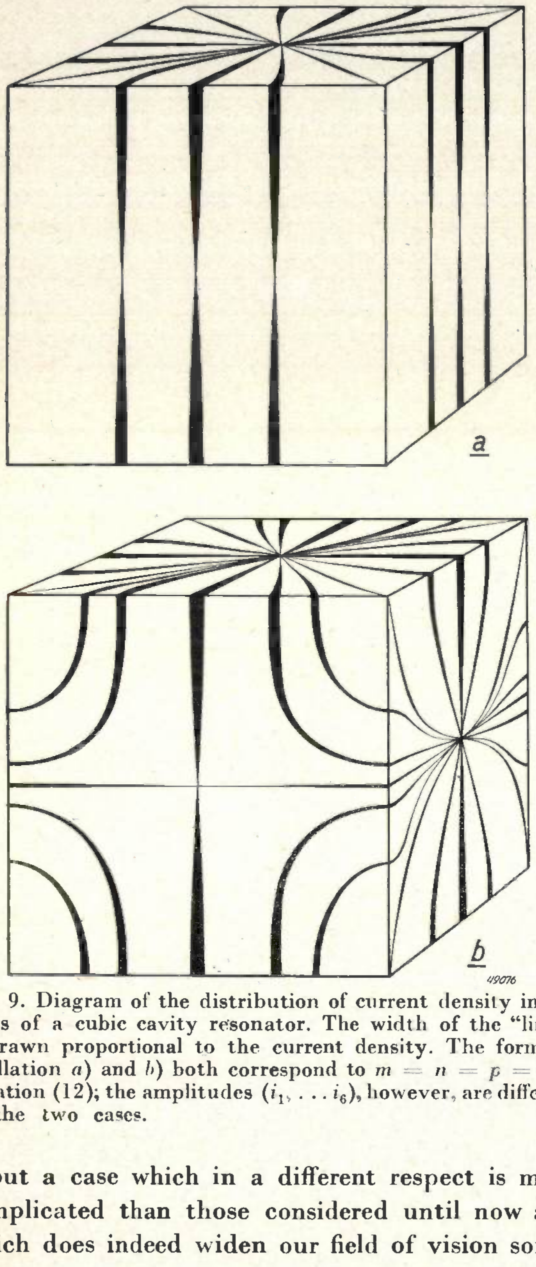


Fig. 9. Diagram of the distribution of current density in the walls of a cubic cavity resonator. The width of the "lines" is drawn proportional to the current density. The forms of oscillation *a*) and *b*) both correspond to $m = n = p = 1$ in equation (12); the amplitudes $(i_1, \dots i_6)$, however, are different in the two cases.

The practical significance of the higher forms of oscillation of the cavity resonators is illustrated by *fig. 10.* In this figure a so-called induction-tube oscillator may be seen, the principle of which has already been discussed in a previous article (see footnote [3])). The cavity resonator here has the function of an oscillator circuit in which the oscillations are excited by the induction tube mentioned. The latter is situated in a hole in the cavity resonator at the position of a voltage maximum. If it is now desired to obtain an oscillator with a high power, two induction tubes could be placed side by side in the hole. Such an arrangement, however, meets with all kinds of objections. Use is therefore made of the fact that with higher forms of oscillation a cavity resonator has more than one voltage maximum. Holes are made at the positions of these maxima and each induction tube is placed in a separate hole.

We shall now discuss the quality factor and the resonance resistance of cavity resonators in general. In practice these two quantities often determine which cavity resonators are to be considered for a given technical purpose. Whether one or another cavity resonator is chosen among those having a suitable quality factor and resonance resistance usually depends chiefly upon the structural requirements which the apparatus involves.

### Quality factor and resonance resistance of cavity resonators

The quality factor of a cavity resonator can be considered as a measure of the sharpness of resonance, *i.e.* of the selectivity of a cavity resonator. The larger the quality factor the greater the selectivity.

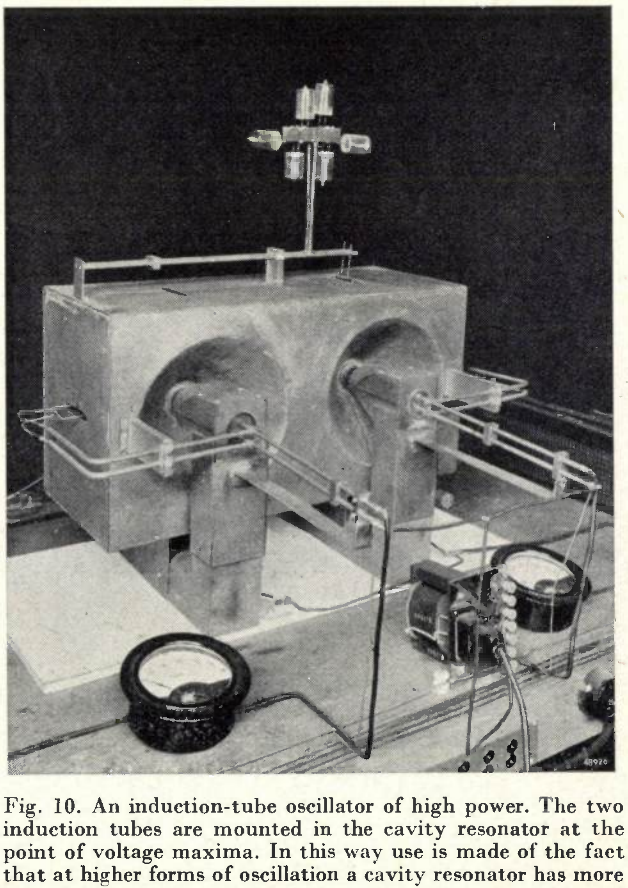


Fig. 10. An induction-tube oscillator of high power. The two induction tubes are mounted in the cavity resonator at the point of voltage maxima. In this way use is made of the fact that at higher forms of oscillation a cavity resonator has more than one voltage maximum.

If only one characteristic oscillation is excited the quality factor $Q$ can be defined by the following relation:

$$Q = 2\pi \frac{U}{W_T}, \quad \ldots \ldots \quad (13)$$

in which $U$ is the average field energy and $W_T$ the heat developed during one period. Definition (13) is very generally valid, not only for ordinary oscillation circuits (*i.e.* with concentrated $L$ and $C$) but also for cavity resonators. If a number of characteristic oscillations are excited simultaneously the quality factor for each characteristic oscillation is also given by (13), provided two characteristic oscillations are never so close to each other that the peaks of the resonance curve coalesce. Equation (13) does not, for example, hold in the case of a cylinder with an inner partition (see fig. 14) to be discussed later.

The significance of definition (13) becomes clearer when it is borne in mind that the average field energy of a free oscillation with frequency $\omega$ decreases owing to losses in the course of the time $t$ according to the formulae

$$U(t) = U\,e^{-\alpha t},$$

where $\alpha$ is the damping constant. If $\alpha$ is small the energy lost during one period is given by

$$W_T = U(0) - U\left(\frac{2\pi}{\omega}\right) = U\left[1 - e^{-2\pi\alpha/\omega}\right] \approx \alpha \frac{2\pi}{\omega} U.$$

By comparison of this result with (13) we find that

$$Q = \frac{\omega}{\alpha}.$$

The quality factor thus actually becomes larger according as the resonance of the cavity resonator is sharper, because the sharpness of resonance of an arbitrarily oscillating system increases when the damping constant decreases.

In order to be able to calculate the quality factor, one must, according to (13), know the average field energy and the energy dissipated during one period.

If $A$ is the magnetic field strength, the field energy is proportional to the integral of $H^2$ over the whole volume of the cavity resonator. The energy dissipated is given by the Joule heat of the currents induced by $H$ in the walls of the cavity resonator (we assume here that the radiation losses may be disregarded). Now the density of the induced current at every point of the wall is proportional to the magnetic field strength at that point and the resistance of the wall is proportional to the depth of penetration $\delta$ due to the skin effect [6]). The energy dissipated is therefore finally proportional to $\delta$ and to the integral of $H^2$ over the whole surface of the cavity resonator. The calculation gives indeed

$$Q = \frac{2}{\delta} \frac{\iiint H^2\,\mathrm{d}\tau}{\iint H^2\,\mathrm{d}\sigma}, \quad \ldots \quad (14)$$

when $\mathrm{d}\tau$ and $\mathrm{d}\sigma$ are volume and surface elements of the cavity resonator respectively. In many cases the order of magnitude of $Q$ is given in sufficient approximation by

$$Q = \frac{2}{\delta} \frac{\text{volume of the cavity resonator}}{\text{surface of the cavity resonator}}. \quad (15)$$

From (15) it follows immediately that for "three-dimensional" cavity resonators $Q$ will in general be larger than for flat cavity resonators. For the latter we have in mind especially the square plane cavity resonators — it may be concluded from (15) that for higher forms of oscillation $Q$ has about the same value as for the fundamental oscillation, if $\lambda$, and consequently $\delta$, which is proportional to $\sqrt{\lambda}$, are kept constant, *i.e.* when cavity resonators of different, suitably chosen dimensions are compared with each other. Such higher forms of oscillation were indeed obtained by placing the cavity resonators side by side and then removing the partitions between them; and for really thin cavity resonators the heat development in the side walls amounts to almost nothing compared with that of bottom and cover.

For the cube $Q$ is equal to $2\,a/3\delta$, where $2a$ is the length of an edge. Since the wavelength $\lambda$ of the fundamental oscillation is equal to $2a\sqrt{2}$ and the depth of penetration for copper is equal to $4 \cdot 10^{-5}$ $\sqrt{\lambda}$ ($\lambda$ in cm), one finally obtains for the quality factor $Q$ of a copper cube oscillating on the longest wavelength $Q = 99000\sqrt{a}$ ($a$ in cm).

We must now consider briefly the resonance resistance of a cavity resonator. We define here the resonance resistance for flat cavity resonators only, because it is then still possible to speak of a voltage $V$, as explained above. The resonance resistance is a quantity $Z$ such that the heat development $W$ in the cavity resonator per unit of time is given by $W = V^2/Z$.

Since the voltage for flat cavity resonators is a function of position, the resonance resistance will

[6]) If $\varrho$ is the specific resistance of the material of the wall of the cavity resonator, the resistance of the layer in which an appreciable current flows is proportional to $\varrho/\delta$. It now follows from the theory of the skin effect that $\delta$ is proportional to $v\varrho$. The resistance of the layer in question is therefore indeed proportional to $\delta$.

in general also depend upon position. Usually by resonance resistance without any indication of position is meant the resonance resistance at the voltage maximum.

We shall now examine how the resonance resistance for higher forms of oscillation behaves compared with that for the fundamental oscillation in the case of square plane cavity resonators, again at constant $\lambda$. When $n$ similar square cavity resonators, each of which executes the fundamental oscillation, are placed side by side, it is clear that $n$ times as much heat is developed as when only one of them is oscillating. If we now remove the intermediate partitions, as described above, and consider the whole as a single cavity resonator in a higher state of oscillation, its resonance resistance is thus $n$ times as small as that of the elementary cavity resonator of which it was built up (we again disregard the heat development in the partitions).

The advantage of the use of the concept "resonance resistance" for cavity resonators is often manifested in the fact that in many arrangements the voltage which will act on a cavity resonator can be calculated as soon as the resonance resistance is known.

We shall not, however, go into that here.

## Coupled cavity resonators

We shall now discuss the lowest characteristic oscillations of a cylindrical cavity resonator divided into two by a transverse partition midway along its length; the partition is provided with a circular hole exactly in the centre.

Such a cavity resonator may be considered in two ways: on the one hand as the limiting case of a cylinder with constriction in the middle (*fig. 11a*), and on the other hand as two separate resonators coupled by a small hole in the partition forming the "cover" of one and the "bottom" of the other (fig. 11*b*). As long as the partition is entirely closed there may be quite independent oscillations, with any arbitrary phase difference, on either side of the partition. When there is a hole that is no longer true, for instead of the characteristic frequency of the fundamental oscillation there are now two characteristic frequencies, at one of which the oscillations to the left and right are in the same phase and at the other in opposite phases.

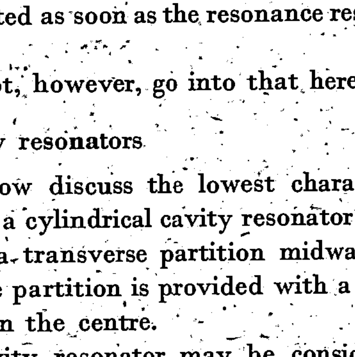


Fig. 11. A cylindrical cavity resonator divided into two by a transverse partition midway along its length. There is a small round hole at the centre of the partition. Such a cavity resonator may be considered: *a*) as a single cylindrical cavity resonator with a constriction at the middle; *b*) as two separate cavity resonators coupled by the hole.

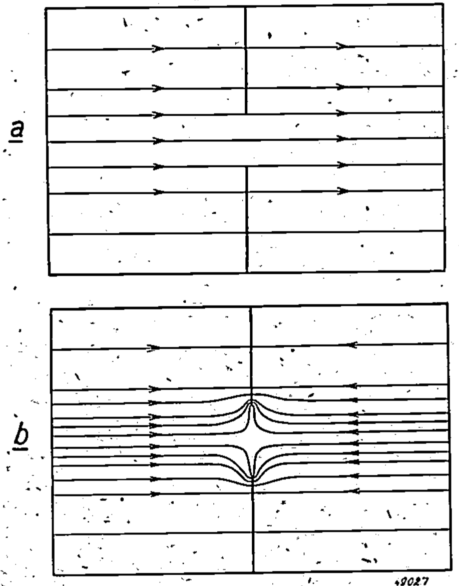


Fig. 12. The course (diagrammatic) of the electric lines of force for the two lowest characteristic freqiencies of the cavity resonator of fig. 11. (The forms of oscillation here having axial symmetry, it is thus sufficient to show a single cross-section.) These characteristic frequencies are both formed from the characteristic (fundamental) oscillation which would be present if there were no hole in the partition.

In the first case, that of the same phase, thus with the electric field similarly directed on the left and right — for the fundamental oscillation the electric field in a cylinder closed at both ends is everywhere parallel to the axis — the situation is always the same with or without partition, with or without a hole in it. Thus when a hole is made in the partition there is no difference in the case of this characteristic oscillation, the characteristic frequency $\omega_0$ of the cavity resonator remaining unchanged and also the course of the lines of force (*fig. 12a*).

It is different in the second case where the oscillations to the left and right of the partition are in opposite phase. Lines of force can indeed end on a conductor, but not somewhere in free space. Thus difficulties are encountered at the hole and we must find out what happens then. Actually the lines of force will curve in the vicinity of the hole, so that they end on the partition or at least near the edge of the hole (fig. 12*b*). The course of the lines of force is thus slightly different from that in the case with no hole in the partition. This is not without consequences for the value of the characteristic frequency. It would take us too far afield to explain how, after having obtained a picture of the variation of the field by means of the above reasoning, by means of a relatively simple calculation a formula can be found giving the frequency correction, at least for a very small hole. The formula in question is as follows:

$$\omega'' = \omega_0 \left\{ 1 + 0.788 \left(\frac{\varrho}{R}\right)^3 \frac{R}{l} \right\}; \quad . \quad . \quad (16)$$

where $\omega''$ is the frequency caused by the presence of the hole, $\omega_0$ the "undisturbed" frequency, $R$ and $2l$ are respectively the radius and the total length of the cylinder and $p$ the radius of the hole.

Thus when the hole is bored in the partition the characteristic frequency $\omega_0$ of the cavity resonator is split into two: $\omega' = \omega_0$, and $\omega$ (given by equation (16)). For a very small hole $\omega'$ and $\omega''$ lie very close together, so that the peaks of the resonance curve corresponding to $\omega'$ and $\omega''$ partially coalesce, exactly as in the case of two coupled circuits with concentrated $L$ and $C$. And just as in the case of two such coupled circuits, the familiar resonance curve with two "shoulders" will occur here as soon as there is any damping in the cavity resonator. In *fig. 13* a photograph is shown of such a resonance curve for a cavity resonator with partition; this

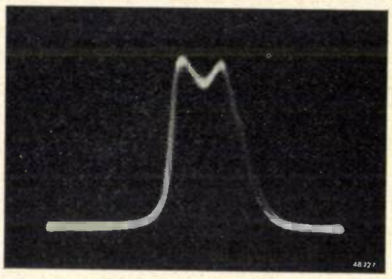


Fig. 13. Resonance curve of coupled cavity resonators recorded with the apparatus of fig. 14. The frequency is here "plotted" in the horizontal direction. The deviation in the vertical direction is a measure of the intensity of the oscillation excited in the cavity resonators.

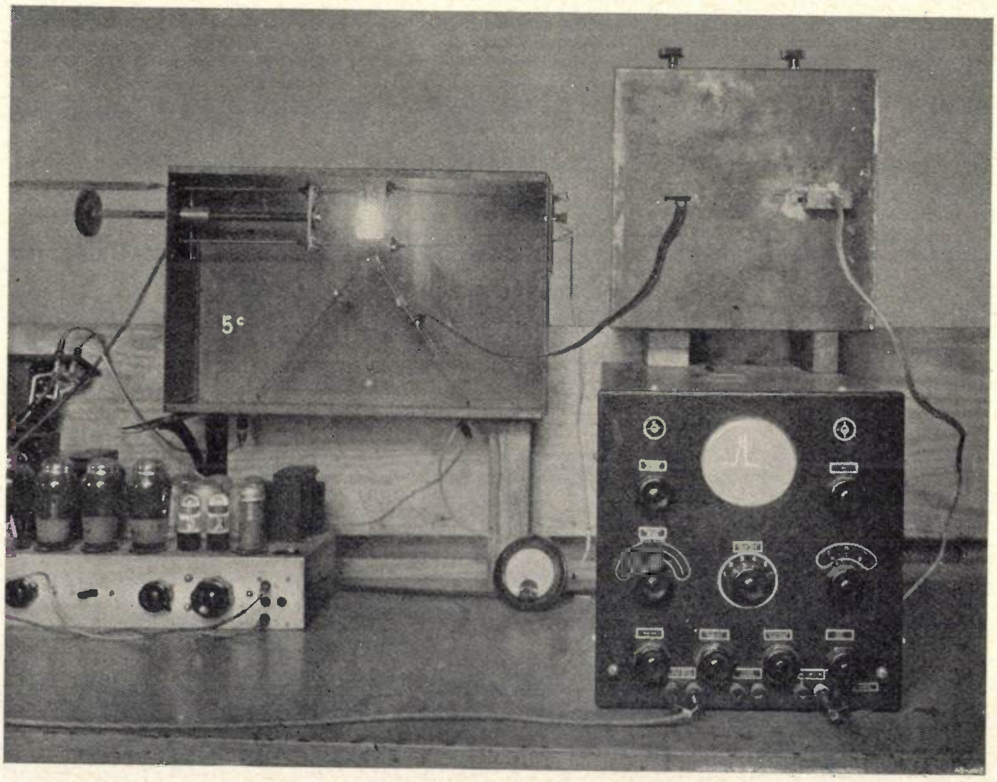


Fig. 14. Apparatus for the recording of the resonance curve of cavity resonators. The resonator visible here on the upper right is divided into two by a partition with a hole in it (the partition is vertical and perpendicular to the plane of the figure), *i.e.* in this case it is actually a question of two coupled cavity resonators. The oscillations are excited with the help of the loop, part of which may be seen protruding from the slot on the left of the cavity resonator, which loop is connected with the oscillator on the left in the figure, whose frequency can be varied. The high-frequency alternating current excited in the loop situated in the other slot is applied to the vertical plates of the oscillograph after rectification. A separate arrangement provides that the horizontal deflection of the electron beam in the oscillograph will be proportional to the variation of frequency.

resonance curve was recorded with the apparatus shown in *fig. 14*. (In the case in question the cavity resonator was square, which of course does not alter the qualitative aspect of the phenomena.) It is thus understandable that the "composite" cavity resonator considered here has an analogous function in high-frequency technology to that of coupled circuits in the region of lower frequencies, in particular to that of a band filter.

Equation (16) was valid for a very small hole ($\varrho \ll R$). What will the frequency correction be when the hole is no longer very small? It is very difficult to answer this question satisfactorily. One thing is certain, however: when the hole is so large that there is no partition left at all, *i.e.* when $\varrho = R$, the two characteristic frequencies $\omega'$ and $\omega''$ must clearly be characteristic frequencies of an undivided cylinder of radius $R$ and length $2l$. The theory of the characteristic frequencies of such a cylindrical cavity resonator shows that the course of the lines of force for the lowest (rotation-symmetrical) characteristic oscillations is as represented in *figs. 15a*, *b* and *c*. We compare these figures with fig. 12*a*, *b* drawn for the case of a small hole. Since the course of the lines of force varies continuously with the size of the hole, fig. 12*a* must correspond to fig. 15*a* and fig. 12*b* to fig. 15*b*. It should therefore be possible to obtain the frequency corresponding to fig. 15*b* from the frequency corresponding to fig. 12*b* when $\varrho$ is increased continuously. The value of the characteristic frequency for the forms of oscillation in fig. 15*b* is familiar from the theory:

$$\omega''_{\varrho=R} = \omega_0 \sqrt{1 + 0.426\frac{R^2}{l^2}} \quad . \quad . \quad . \quad . \quad (17)$$

Of the curve which gives $\omega''$ as a function of $\varrho/R$ we now have a section at the beginning, given by equation (16), and at the end point ($\varrho/R = 1$), given by equation (17) — see *fig. 16*. We shall now guess at the rest by drawing the dotted part of the curve in a reasonable manner. The frequency corresponding to figs. 12*a* and 15*a* is independent of $\varrho$ and $l$, namely in both cases equal to $\omega_0$, *i.e.* it will be represented in fig. 16 by the $\varrho/R$ axis [7]).

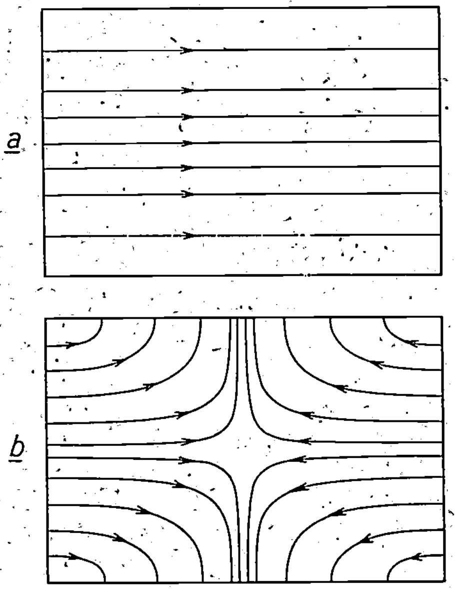


Fig. 15. The (diagrammatic) course of the electric lines of force for the lowest characteristic frequencies of a cylindrical cavity resonator; the forms of oscillation in question are axial symmetrical: *a*) corresponds to the fundamental oscillation; *b*) and *c*) to the two succeeding characteristic oscillations.

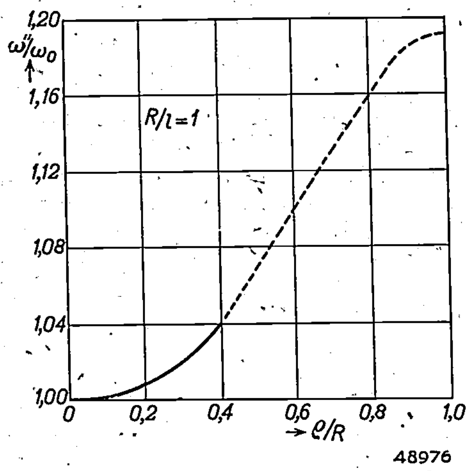


Fig. 16. The lowest but one characteristic frequency $\omega''$ plotted vertically of the cavity resonators coupled by a hole as a function of the ratio $\varrho/R$ (plotted horizontally), of the radius $\varrho$ of the hole and the radius $R$ of the cylinder. The dotted part of the curve is not well known theoretically.

We may not conclude without drawing attention to the analogy which exists between the electromagnetic oscillations in Lecher systems, flat and general cavity resonators on the one hand and on the other the acoustic vibrations of strings, membranes and Helmholtz resonators. This analogy is no identity; there are certain differences between the two groups of phenomena. But the expectation, which is perhaps natural, that there must exist an electromagnetic counterpart of the speaking-tube and the horn is actually confirmed by the electromagnetic "wave guides" and "horn aerials", which we shall not, however, discuss in this article.

[7]) *Cf.* H. A. Bethe, Phys. Rev. **66**, 163, 1944.

# THE PRACTICAL CONSTRUCTION OF VIBRATION-FREE MOUNTING WITH AUXILIARY MASS

by J. A. HARINGX.

621-752

In order to reduce the amplitudes of the forced vibrations of an instrument due to the motion of its surroundings it is usually mounted on a resilient construction which at the same time is provided with a certain damping in connection with the decay of the free vibrations. In a previous article the advantages were shown of introducing the damping between the apparatus and an auxiliary mass attached to it with springs. The features of this system were discussed for the one-dimensional case. In the present article the case of multi-dimensional vibrations is examined (translations parallel to and rotations about various axes). For such a case the mounting should preferably satisfy certain conditions of symmetry. For the resilient support practically only helical steel springs can be considered, use being made not only of their axial rigidity but also of their rigidity with respect to lateral deflection. Further consideration is also given to the practical manner of introducing the damping.

The use of sensitive instruments such as balances, galvanometers and microscopes is often made difficult or even impossible by vibrations of the surroundings. In order to reduce the amplitudes of the forced vibrations of such an instrument due to the vibrations of the surroundings it is placed upon sufficiently weak springs. It is then necessary, however, to introduce a damping to ensure that the decay of the free vibrations of the system caused by slight impulses or initial displacements is rapid enough. In a previous article [1]) we showed the advantages of introducing this damping between the apparatus and an auxiliary mass attached to it with springs. The features of this system were discussed, but only for the case of a one-dimensional movement.

When it comes to putting these ideas into practice we are immediately faced with the problem that the foundations of an apparatus are, in general, apt to execute vibrations in and around various directions. Since in most cases an instrument is sensitive to several of these vibrations, the system of mounting must provide for resiliency in several directions. This cannot be done, however, in an arbitrary manner as regards the choice of the centres of the elastic forces and their directions, for then it is quite likely that once the apparatus begins to vibrate it will show highly complicated and absolutely indeterminable oscillations. It would then be impossible to determine what measures have to be taken to restrict the forced vibration and to stop as quickly as possible the free vibrations of the system after an impulse or an initial displacement. We shall now explain very briefly how this problem has to be dealt with.

A rigid body suspended in space by a set of springs possesses six degrees of freedom, *viz:* translation in three mutually perpendicular directions and rotation about three mutually perpendicular axes. These degrees of freedom are as a rule coupled, that is to say a force acting in one of the directions (for instance the reaction of the springs or the inertia of the mass) results not only in a translation in this direction but also in other translations and rotations. This is the reason why the behaviour of the vibrating system is so complicated. Under certain conditions however the coupling between the different degrees of freedom may disappear, namely when the spring mounting possesses a so-called centre of elasticity with three perpendicular principal main axes of elasticity passing through this centre and coinciding with the principal axes of inertia of the suspended body (the centre of gravity then automatically coincides with the centre of elasticity). A principal axis of elasticity is defined in such a way that a force acting in its direction causes a translation in this direction only, whilst a couple about such an axis likewise causes a rotation only about that axis, so that under the conditions mentioned there are in fact six mutually non-coupled degrees of freedom. Any arbitrary combination of elastic attachments will not as a rule possess a centre of elasticity. Such will be the case, however, if for instance the mounting is symmetrical with respect to two mutually perpendicular planes, like those in *figs. 1a-c,* and if at the same time the elastic elements are of equal rigidity in the directions perpendicular to these planes. For reasons of symmetry the intersecting line of these two planes, *i.e.* the $z$ axis, must be one of the principal axes of elasticity and the two other principal axes must be

[1]) J. A. Haringx, Vibration-free mountings with auxiliary mass, Philips Techn. Rev. 9, 16, 1947.

parallel to the $x$ and $y$ axes. The height of the centre of elasticity on the $z$ axis can be calculated from the rigidity of the springs, an example of which will be given below.

If, now, in our vibration-free mounting the springs are fitted in this two-fold symmetrical manner and care is taken that the principal axes of inertia of the mounting combined with the instrument placed on it coincide with the principal axes of elasticity, *i.e.* the $x$, $y$ and $z$ axes, then the vibration of the system following the six degrees of freedom are approximately[2]) independent of each other and can be treated as six separate one-dimensional problems. This also holds when applying the principle of our vibration-free mounting sketched in *fig. 4* of the previous article[1]) by attaching an auxiliary mass to our apparatus by means of a number of elastic and damping elements. This attachment must also have three principal axes of elasticity coinciding with the same $x$, $y$ and $z$ axes, the principal axes of inertia of the auxiliary mass likewise coinciding with these axes. Moreover, the damping elements must be introduced in such a way that in case of translations parallel to or rotations about the three said axes the damping results in forces along and couples about the same axes.

When these conditions are satisfied — in practice this can be done with sufficient accuracy — we can apply the results of our previous study of the choice of the parameters to each degree of freedom separately and thus endeavour to keep the system free of all disturbing vibrations.

### *The design of the springs*

The mechanical disturbances apt to give trouble in a building may be caused by the regular throbbing of a machine, but they may also be due to all kinds of irregular shocks such as footsteps, the slamming of doors, etc. Consequently any frequencies, even the very lowest, may be represented in the Fourier spectrum of the disturbances. The primary effect of the disturbances is to set in motion the floors, walls and tables in the building, which then execute their natural vibrations. The "vibration systems" of floors, tables, etc. act as a kind of filter, which preferentially transmits the vibrations in the neighbourhood of the resonant frequencies. We shall thus need to analyse the disturbing vibrations more closely with the aid of a vibration pick-up. It has been found that the lowest of the natural frequencies mentioned usually lie between 15 and 20 c/s. If, by means of a vibration-free mounting, the amplitudes of an apparatus are to be reduced, for example, to a few percent of those of the foundation, then under these conditions we must give the mounting the very low resonant frequency of 2 to 4 c/s[3]).

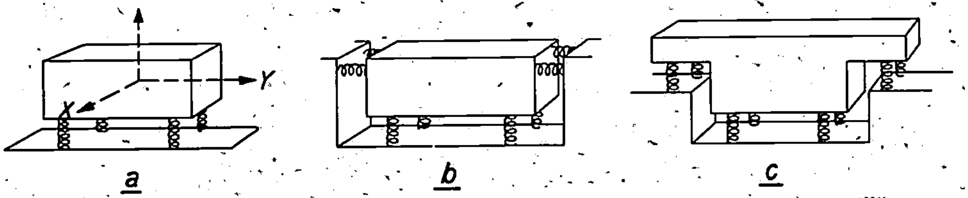


Fig. 1. Three different arrangements for making an apparatus vibration-free where multidimensional disturbing vibrations exist. The mountings are all symmetrical with respect to two perpendicular vertical planes.

Such low resonant frequencies preclude the use of rubber cylinders for the resilient elements of the mounting, because they are too rigid. For the resonant frequency $\omega_0$ of a vibrating system we may write

$$\omega_0 = \sqrt{\frac{g}{f}},$$

where $g$ is the acceleration due to gravity and $f$ the compression of the resilient element caused by the weight of the mass. In order to obtain, for example, a resonant frequency of $\omega_0/2\pi = 3$ c/s, $f$ must therefore be 2·8 cm. If rubber cylinders were to be used as spring elements their length $l$ would have to be at least 10 cm to permit such a large compression. Moreover, to prevent them from buckling, the diameter $D$ would have to be of the same order of magnitude, so that the force necessary for the compression:

$$P = \pi D^2 \cdot E \cdot f/l$$

in the case of the weakest variety of rubber with a

[2]) "Approximately" because the force of gravity exercises a disturbing influence.

[3]) This result is obtained by requiring, in formula (8) of the article referred to in footnote [1]), that for instance $\left|\frac{a}{a_0}\right| = 0{\cdot}04$ and setting $\mu$ equal to 0·5, so that $\omega/\omega_0 = 7$ ($\omega$ = given disturbing frequency, $\omega_0$ = resonant frequency to be chosen).

modulus of elasticity $E = 10$ kg/cm$^2$, would be of the order of 1000 kg. A mounting with four such cylinders should thus have a weight of 4 tons! Obviously such heavy constructions cannot generally be used as a support for a balance or a microscope. A reasonable total weight would be say 50 to 100 kg. Such a weight could be more closely approached if sponge rubber were used, but as this material is not very durable it is not to be recommended for our purpose.

As resilient elements for a vibration-free mounting helical steel springs are much more suitable. These can easily be made so as to take up the desired compression of about 3 cm under a force of about 10 kg. Helical steel springs possess a large number of distinct natural frequencies, with the result that the vibrations having corresponding frequencies are transmitted to the instrument. This fact is in contradiction to the statements previously given and is often considered as a serious objection, but fortunately these natural frequencies usually lie so high that — at least in our case — they do not cause any trouble at all.

As a rule it is only the axial rigidity of such springs that is utilised; take for instance buffers, spring balances, etc. But, in addition to this axial rigidity, springs also possess a lateral rigidity, provided their ends are not both hinged. Therefore, since we have to do with a three-dimensional mounting having six degrees of freedom it is obvious that the construction can be simplified by turning this lateral rigidity to account (see figs. 1*a* and 1*c*).

Let us consider for example the arrangement sketched in *fig. 2*, which may serve as a prototype for many practical applications. A flat plate *p*, on which is placed the balance or other instrument to be rendered free of vibration, is supported by four helical springs. The auxiliary mass, a second plate *P*, is supported by the first plate in a similar manner. In both cases the four springs simultaneously perform the function of a resilient attachment in the vertical direction and in the two horizontal directions.

The question arises as to how the springs must be designed in order to get a certain lateral rigidity in addition to a given axial rigidity. This axial rigidity, *i.e.* the force required per unit of axial compression, can be calculated with the help of the formula

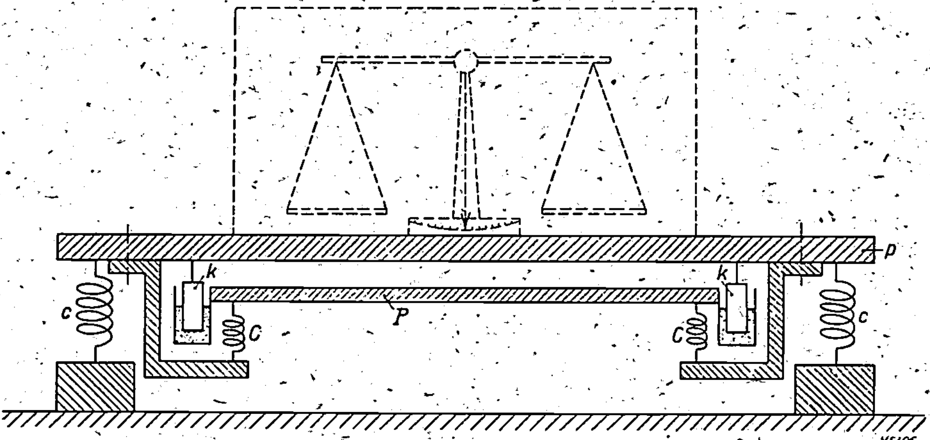


Fig. 2. Model of the practical construction of a vibration-free mounting with auxiliary mass. The plate *p* supported by the helical springs *c* and carrying the instrument represents the main mass, the plate *P* connected below with the helical springs *C* is the auxiliary mass. Between these two plates the damping *k* has been provided.

$$c = \frac{4m}{m+1}\,\frac{EI}{\pi n D^3}.$$

($E$ = modulus of elasticity, $I$ = moment of inertia of the circular cross section of the wire, $n$ = number of turns, $D$ = coil diameter, $m$ = lateral contraction or Poisson's ratio; for steel $m = 10/3$). For the calculation of the lateral rigidity, however, the theory of the lateral deflection of helical springs had first to be worked out in more detail [4]), and at the same time the danger of buckling could also be studied. Here we can give the most important results only very briefly.

A slender helical spring will buckle as soon as the compression load, and thus also the relative

[4]) J. A. Haringx. On the buckling and the lateral rigidity of helical compression springs. Proc. Kon. Ac. v. Wet Amsterdam 45, 533-539 and 650-654, 1942.

compression $\xi$, has reached a certain value. The critical compression $\xi_k$ is found to depend — at least in a first approximation [5]) — only on the ratio of the length $l_0$ of the unloaded spring and its diameter $D$, the so-called ratio of slenderness of the spring. The relation found for the case where both ends of the spring remain parallel after deflection (as in fig. 2) is shown in *fig. 3*. It is seen that when $l_0 < 2{\cdot}6\ D$ the spring can never buckle. This condition must therefore be taken into account when designing the springs.

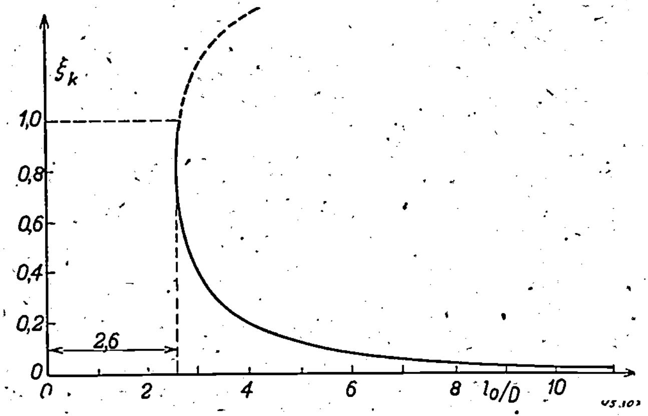


Fig. 3. A loaded helical spring without axial guidance will buckle under a certain force corresponding to a certain (critical) compression $\xi_k$. $\xi_k$ is found to depend only on the ratio of slenderness $l_0/D$ of the spring. From the relation shown here, which holds for a spring with its ends both hinged or remaining parallel, it may be seen that with a ratio of slenderness $l_0/D < 2{\cdot}5$ buckling can never occur. When both ends of the spring are fixed $l_0$ is to be taken as half the length of the spring.

Further, the lateral deflection $y$ of the free end of an axially compressed spring under the simultaneous influence of a lateral force $L$ and a couple $M$ (*fig. 4*) is given by the equations

$$\left.\begin{aligned} L &= c_1 y - c_2 \psi, \\ M &= -c_2 y + c_3 \psi, \end{aligned}\right\} \quad . \; . \; . \; . \; . \quad (1)$$

where $\psi$ is the rotation of the upper end and $c_1$, $c_2$ and $c_3$ are the coefficients of rigidity. These coefficients are related to the axial rigidity $c$ according to the following equations:

$$\frac{c}{c_1} = \frac{m}{2\,(m+1)} \left[1 + \frac{2m+1}{3m}\left(\frac{l}{D}\right)^2\right] \cdot \varepsilon_1, \quad . \; (2a)$$

$$\frac{c_2}{c_1} = \frac{1}{2}\ l \cdot \varepsilon_2, \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; (2b)$$

$$\frac{c_3}{c_1} = \frac{mD^2}{4\,(2m+1)} \left[1 + \frac{4\,(2m+1)}{3m}\left(\frac{l}{D}\right)^2\right] \cdot \varepsilon_3. \quad . \; (2c)$$

Here $l$ is the length of the spring when loaded and $\varepsilon_1$, $\varepsilon_2$, $\varepsilon_3$ are correction factors depending upon the relative axial compression $\xi$ and the ratio of slenderness $l_0/D$. These three factors are represented graphically in *figs. 5a-c*.

By applying the results described to the mounting in *fig. 2* we can first of all calculate the height $h$ of the centre of elasticity above the base plane of the springs.

A horizontal force $L$ through this centre — according to the definitions already given — will cause only a lateral displacement $y$ of the plate; the plate remains horizontal, thus $\psi = 0$. It follows from (1) that in this case the couple $M$ caused by $L$ must be:

$$M = -\frac{c_2}{c_1} \cdot L,$$

while on the other hand this couple is given by

$$M = -(l-h) \cdot L.$$

By eliminating $M/L$ from these two equations and making use of (2b) we obtain:

$$h = l\,(1-{}^1\!/_2\,\varepsilon_2).$$

The centre of gravity of the movable mass of the mounting must therefore lie at this height. Further, under this condition, we have $L = c_1 y$, so that equation (2a) gives us directly the rigidity of the spring $c_1$ required to make our mounting vibration-free as regards transverse vibrations.

It will often be desired to have the mounting behave in the same way in the vertical as in the horizontal direction. When vertical springs alone are used $c_1$ must then be equal to $c$. This establishes a relation between $\xi$ and $l_0/D$ which can be derived

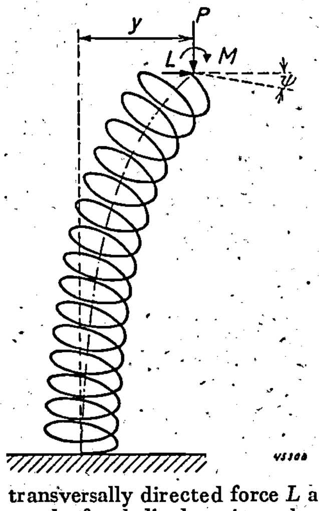


Fig. 4. When a transversally directed force $L$ and a couple $M$ act on the free end of a helical spring whose other end is clamped and which is compressed axially by a weight $P$, a lateral deflection $y$ and a rotation $\psi$ occur at that free end.

[5]) A more precise calculation shows that the pitch and the wire diameter exercise a certain, although slight, influence. This calculation will be published elsewhere in a paper in which other problems discussed in this article are also worked out in more detail.

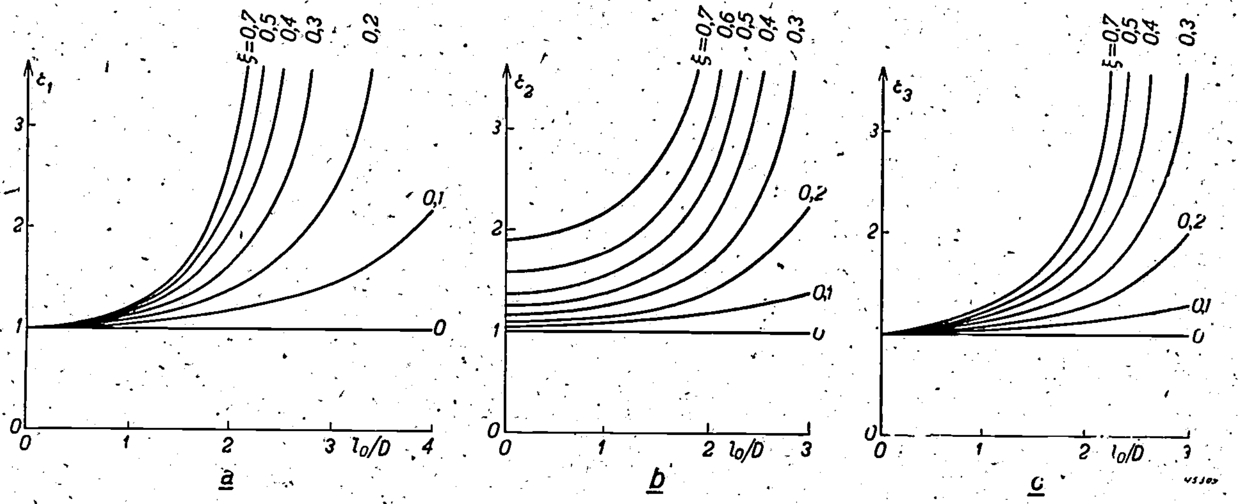


Fig. 5. The correction factors $\varepsilon_1$, $\varepsilon_2$, $\varepsilon_3$ occurring in formulae (2) as functions of the ratio of slenderness $l_0/D$ with the compression $\xi$ as parameter.

from equation (2a) and fig. 5a. This relation is represented graphically in *fig. 6.* With the aid of this graph and the prescribed loaded and characteristic frequencies it is possible to derive all the conditions that have to be satisfied by the parameters of the spring (length, diameter, number of turns, etc.).

Also the rigidity of the mounting with respect to a couple about one of the principal axes of elasticity can be expressed in the quantities $c$, $c_1$, $c_2$, $c_3$; in this case the contributions of the rigidities $c_1$, $c_2$ and $c_3$ can often be disregarded. For counteracting the rotation vibrations connected with this couple the same considerations are valid as were given for translation vibrations, except that the various moments of inertia take the place of the masses. There is nothing new of any importance in this, so that we shall not go deeper into it here.

*The practical application of the damping*

The damping can best be realised by means of four cups $k$ filled with oil of a certain viscosity and affixed to the auxiliary mass (plate $P$ in fig. 2), in which cups four others affixed to the main mass (plate $p$) can move freely in all directions with a clearance of some millimeters.

When it is desired that the damping coefficient $k$ should be the same in the horizontal and in the vertical directions the damping cups must be square and, as calculation has shown, they must be filled to a height $H \approx 1.45\ B$, where $B$ is the side of the basal plane of the inner cup. The damping coefficient in that case is approximately equal to

$$k \approx 18\,\frac{\eta B^4}{d^3}\left(1 + 2.5\,\frac{d}{B}\right),$$

where $d$ is the distance between the walls of the two cups and $\eta$ the viscosity of the liquid.

Eventually the optimum damping can be adjusted experimentally by using oils of a different viscosity, but there is really no very great sensitivity in such an adjustment, as we have seen in the article already referred to [1]). Even if the adjustment is not correct the free vibrations after an impulse die out quickly enough and the forced vibrations of the system due to the motion of its surroundings are scarcely affected at all. It is really fortunate that this adjustment is not critical, for the viscosity of the oil — and thus its damping power — depends closely upon the temperature, so that we can obtain the optimum behaviour only at one temperature.

Instead of the liquid damping that we have so far assumed to be applied, a frictional damping can also serve to dissipate the kinetic energy imparted to the mounting by an impulse; for instance a set of flexible strips can be attached to one of the plates in fig. 2 with a slight pressure

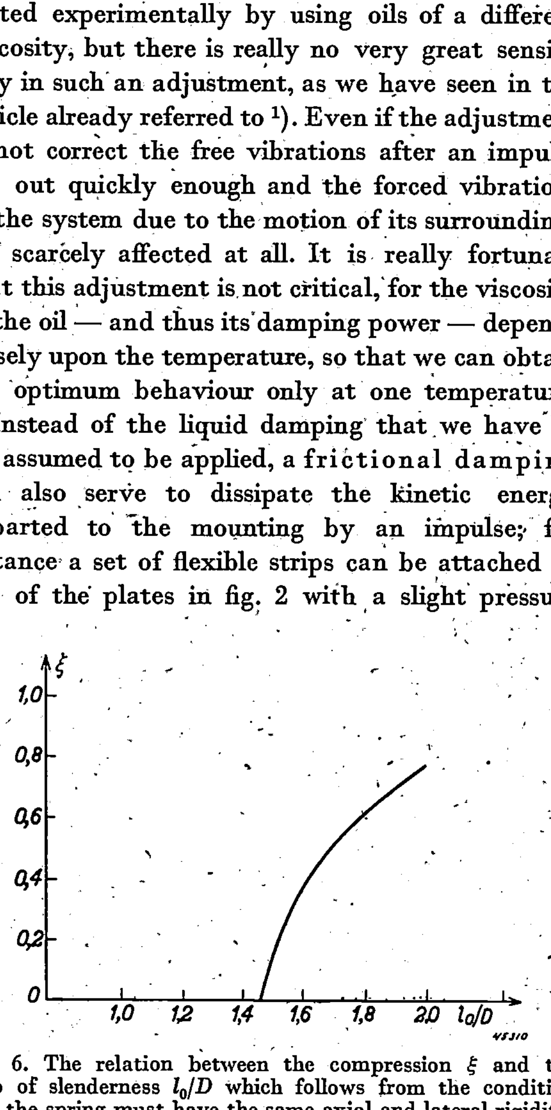


Fig. 6. The relation between the compression $\xi$ and the ratio of slenderness $l_0/D$ which follows from the condition that the spring must have the same axial and lateral rigidity.

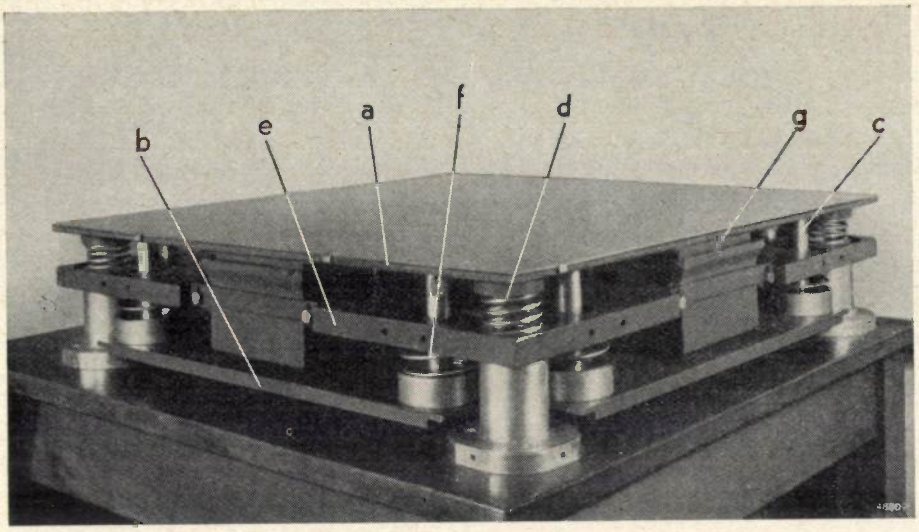


Fig. 7. Practical construction of a vibration-free mounting. The dimensions of this apparatus are 60 × 60 × 15 cm; its weight is 110 kg. The two plates *a* and *b* together form the main mass. They are connected with twelve rods *c*. This main mass rests upon four helical springs *d*. The auxiliary mass *e* is coupled to the main mass with eight springs *f*. The cup *g* is one of the four damping elements.

against the other plate and rubbing over it as soon as a relative displacement takes place. When the amplitudes are sufficiently large it is possible with this very simple construction to obtain a fairly rapid dissipation of the kinetic energy. When, however, the amplitudes are less than a certain small value a more or less unexpected phenomenon occurs; the frictional forces cause one mass to follow the movements of the other. Thus, as soon as the frictional damping has caused the amplitude to drop to this small value it ceases to act and for some time the instrument will continue to vibrate at this amplitude. If, as is usual, the resonant frequency is low, this may not eventually constitute any objection for some instruments, such as dial gauges, microscopes and not too sensitive balances, but for galvanometers and other highly sensitive instruments it will be necessary to use liquid damping.

In *fig.* 7 a vibration-free mounting is shown which is in use in this laboratory and which was designed according to the theory developed here. This apparatus, weighing 110 kg, can be used as a mounting for instruments up to 35 kg. The resonant frequency (with rigidly coupled main and auxiliary masses) lies at about 3 c/sec. This arrangement fully answers expectations both as regards the decay of the free vibrations and as regards the amplitudes of the forced vibrations due to the motion of the surroundings.

# A REMARKABLE PROPERTY OF TECHNICAL SOLID DIELECTRICS

by M. GEVERS and F. K. du PRÉ.

537.226.3:537.226.8

In this article a remarkable property of technical solid dielectrics is discussed. The ratio between the temperature cofficient of the dielectric constant $\varepsilon$ and the quantity $\tan\delta$ ($\delta$ = loss angle) is practically constant for most materials. Expressed in a formula this is $(1/\varepsilon)\,\dfrac{\partial\varepsilon}{\partial T} = A \tan\delta$, where $0{\cdot}04 < A < 0{\cdot}09$ and usually $A \approx 0{\cdot}05$. An explanation of this and other properties is given, viewed from the same aspect. A few exceptions to the above rule are described and explained, while some practical conclusions are drawn.

In electrotechnics solid dielectrics are used chiefly in two ways, namely:

1) for the insulated attachment and leading in of supply leads, and
2) as media in condensers.

For these purposes many different materials are used which may be divided into the following groups according to their structure and chemical composition:

a) natural dielectrics, such as amber, mica, quartz;
b) inorganic materials such as glass, quartz glass, porcelain, steatite (Mg-silicate), cordierite (MgAl silicates), rutile ($TiO_2$) and related products, Mg titanates;
c) organic materials such as condensation products (phenol-formaldehyde resin), polymerization products (polystyrene), rubber and ebonite.

The properties which are required of these dielectrics depend upon the particular use to which they are put. In every case more or less good insulation is required. In case 2) a high value of the dielectric constant $\varepsilon$ is usually also desired.

Another requirement usually made when such materials are used in connection with alternating current technics is that the dielectric should have small dielectric losses. In case 1) the reasonableness of this requirement is obvious, since losses lead to undesired temperature increase and finally to breakdown and destruction of the material. But also in case 2) low dielectric losses are usually desired, namely when the circuit in which the dielectric is used between the electrodes of a condenser has to have a high selectivity. In a previous article [1]), in which dielectric after-effect phenomena were discussed, it was already pointed out that losses have an unfavourable effect on selectivity. This becomes clear when it is kept in mind that the influence of these losses corresponds to the inclusion of a resistance in the $LC$-circuit.

In many cases a second requirement must also be made of an oscillation circuit, namely that the frequency of the circuit shall be more or less independent of the temperature of the surroundings. In this case, therefore, it will be required of the medium that the temperature coefficient of the dielectric constant shall be small.

From the above it is evident that for various technical purposes it is of importance to measure not only the losses (determined by the so-called loss angle $\delta$) but also the temperature coefficient of the dielectric constant (for the sake of brevity called from now on simply "temperature coefficient"). Since in many respects the data from literature were incomplete, special attention has been devoted in the Philips laboratory to the determination of these two properties with the greatest possible accuracy [2]). A remarkable relation between the two quantities has thereby been brought to light, namely that, apart from a few exceptions to be mentioned later, the substances which have a high temperature coefficient have at the same time a large loss angle, while the ratio between the two quantities is practically constant.

This can be understood theoretically when bearing in mind that the losses as well as the temperature coefficient are affected by so-called after-effect phenomena.

In this article some numerical data on the properties in question will be given. Further, beginning with certain conceptions about the cause of the after-effect (namely dipoles which are orientated with a time lag in a field) expressions will be derived for the loss angle and the temperature coefficient, in which the proportionality mentioned appears, while the value of the proportionality factor proves to correspond to the value found experimentally.

## Dielectric losses

Let us suppose that a dielectric is situated in a constant electric field $E$. Under the influence of

[1]) J. L. Snoek and F. K. du Pré, Philips Techn. Rev. 8, 57, 1946.

[2]) *Cf.* M. Gevers, Philips Res. Reports, 1, 197, 279, 361, 447, 1946.

that field a dielectric displacement $D$ occurs where

$$D = \varepsilon E \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

$\varepsilon$ being the dielectric constant.

If the dielectric is not situated in a constant but in a periodically varying electric field $E = E_0 \cos \omega t$, there flows through the condenser an A.C. which is proportional to $\mathrm{d}D/\mathrm{d}t$. When the dielectric is loss-free, the relation between $D$ and $E$ is here given by (1), so that

$$D = \varepsilon E_0 \cos \omega t = D_0 \cos \omega t \quad . \quad . \quad . \quad (2)$$

If, however, the insulator exhibits losses the following holds:

$$D = D_0 \cos (\omega t - \delta), \quad . \quad . \quad . \quad (3)$$

where $\delta$ is called the loss angle. The amplitude $D_0$ will in general differ from that in the ideal case (1), while both $D_0$ and $\delta$ may also still depend upon the frequency.

If a resistance $R_p$ is connected in parallel with a condenser (capacity $C$), a phase difference $\delta$ occurs between current and voltage and

$$\tan \delta = \frac{1}{\omega C R_p}.$$

From this it may be seen that the occurrence of a phase shift is equivalent to the presence of a resistance in the circuit. As a result there is heat development in the dielectric, which is given per unit volume and per second by

$$W = {}^1\!/_2\, E_0 D_0 \omega \sin \delta \approx {}^1\!/_2\, E_0 D_0 \omega \tan \delta, \quad (4)$$

since usually $\delta < 0.01$ and thus the difference between $\delta$, $\tan \delta$ and $\sin \delta$ can be disregarded.

This heat development may sometimes lead to disturbing rises in temperature.

Considering the importance in electrotechnics of having at one's disposal dielectrics which exhibit not only small dielectric losses but also a small temperature coefficient, it is understandable that these quantities have been measured for many substances. The results, as far as the losses are concerned, are usually characterized by giving the value of $\tan \delta$, in which $\delta$ is the phase shift from formula (3), while for the effective dielectric constant the quotient $D_0/E_0 = \varepsilon$ can be chosen [3]).

## Results of the measurements

After these preliminaries we may now formulate the results of the measurements. It was found that many of the technical solid dielectrics investigated possess the following properties:

1) $\tan \delta$ is practically independent of the frequency in a wide region and entirely independent of the amplitude of the measuring field.
2) The dielectric constant $\varepsilon$ a) is practically independent of the frequency, and b) usually has a positive temperature coefficient.
3) Between $\tan \delta$ and the temperature coefficient of the dielectric constant $\varepsilon$ there exists the relation

$$\frac{1}{\varepsilon} \frac{\partial \varepsilon}{\partial T} = A \tan \delta, \quad . \quad . \quad . \quad . \quad . \quad (5)$$

in which $A$ depends only slightly on the nature of the substance and on the frequency and is inversely proportional to the absolute temperature $T$. At room temperature $A \approx 0{\cdot}05$.

In the table below several examples are given. The measurements were taken at room temperature and at a frequency of $1{\cdot}5 \times 10^6$ c/sec. It may be seen that, in spite of the fact that the values of $\tan \delta$ are very varied, the value of $A \times 10^2$ varies only between 4·4 and 9, and usually amounts to about 5.

| Material | $\tan \delta \cdot 10^4$ | $\frac{1}{\varepsilon} \frac{\partial \varepsilon}{\partial T} \cdot 10^6$ | $A \cdot 10^2$ |
|---|---|---|---|
| Quartz glass | 1.0 | 9.0 | 9 |
| Mica | 1.1 | 5.5 | 5.0 |
| Kersima * | 8.1 | 51.0 | 6.3 |
| Boron glass * | 13.1 | 107 | 8.2 |
| Cu-Zn-ferrite | 32 | 141 | 4.4 |
| Mycalex | 38 | 190 | 5.0 |
| G 40 glass * | 40 | 240 | 6.0 |
| Ebonite | 81 | 460 | 5.7 |
| Röntgen glass * | 100 | 580 | 5.8 |
| "Pliofilm" | 300 | 1440 | 4.8 |
| Celluloid | 450 | 2600 | 5.8 |
| "Pertinax" | 550 | 3020 | 5.5 |
| "Novotext" | 850 | 4760 | 5.6 |

*) Products of Philips. "Kersima" is a ceramic product mainly Mg-silicate; G 40 is a NaK-borosilicate glass; Röntgen glass a NaCa-silicate glass.

The remarkable point about the materials having these properties is that they usually possess an irregular internal structure, *i.e.* the substances are in a vitreous state or at least exhibit a strong irregular disturbance of the single-crystal state.

It must be stated here that properties (1) and (2) had already been known for a long time. Already in 1914 Wagner had proposed a theory of the dielectric losses and especially of their variation with the frequency, from which the conclusion could be

[3]) $E = E_0 \cos \omega t$ and the quantity $D$ of formula (3) could also be written as complex quantities. The quotient $D/E = \varepsilon$ formed by analogy with (1) would then also become a complex number, of which $|\varepsilon| = D_0/E_0$ would represent the absolute value. In the notation of $|\varepsilon|$ for the quotient $D_0/E_0$ we have omitted the lines || for the sake of simplification. *Cf.* also M. Gevers, Philips Res. Reports 1, 197, 1946, especially p. 197 p. 198.

drawn that under special circumstances tan $\delta$ might be independent of the frequency in a wide region. Property (3), however, had not yet been observed, and in our opinion it is the most remarkable one. It also has important practical consequences. Search is often made for substances with a low temperature coefficient of the dielectric constant. The property in question here provides the evidence that in general such a search must not be conducted among the substances possessing a high value of the loss angle $\delta$.

It is now our intention to show how not only the first but also the other properties can be made understandable by considerations which are in part very analogous to those of Wagner.

We assume, as is customary, that the dielectric losses in solid insulating substances such as the above are due to one of the following causes:

a) the presence of dipoles, which are only able to follow the alternations of an external field with a certain time lag;

b) the presence of separate semi-conducting regions in the substance, whereby the transport of charge to one side or other of the region exhibits an appreciable retardation;

c) the presence of free ions.

A combination of a), b) and c) can of course also occur. It has been found that each of the causes separately can bring about the occurrence of the properties found experimentally. For the present, however, we shall confine our considerations to the first case.

## Explanation of the results of the measurements

The above-mentioned causes a), b) and c) have this in common, that they lead to after-effect phenomena.

In the article already referred to the influence of after-effect on the phenomena in a dielectric was discussed. It was found that if a constant field $E = E_0$ was suddenly ($t = 0$) applied (*fig. 1*), the dielectric displacement behaves according to the formula:

$$\left.\begin{aligned} D &= E_0\,(\varepsilon_1 + \varepsilon_2 f(t)), \\ f(t) &= 1 - e^{-t/\tau} \text{ for } t > 0,\ f(t) = 0 \text{ for } t < 0. \end{aligned}\right\} \quad (6)$$

The quantity $\varepsilon$ is determined by the polarizability of the atoms; thus all atoms of the substance contribute more or less to $\varepsilon_1$. The quantity $\varepsilon_2$ is due to the mobile dipoles, so that usually $\varepsilon_2 \ll \varepsilon_1$. The quantity $\tau$ is called the relaxation time.

If, instead of a constant field, at the moment $t = 0$ a variable field $E = E_0 \cos \omega t$ is applied the expression for $D$ becomes

$$D = K \cos \omega t + L \sin \omega t + M e^{-t/\tau}, \quad \dots\dots \quad (7a)$$

where

$$K = E_0\,(\varepsilon_1 + \frac{\varepsilon_2}{1 + \tau^2\omega^2}),\ L = E_0\varepsilon_2 \frac{\tau\omega}{1 + \tau^2\omega^2},\ M = \frac{-E_0\varepsilon_2}{1 + \tau^2\omega^2}. \quad (7b)$$

The third term in (7a) dies out quickly and for $t \gg \tau$ it may be ignored, so that (7a) is reduced to

$$D = K \cos \omega t + L \sin \omega t = \sqrt{K^2 + L^2} \cos (\omega t - \text{arc tan}\,\frac{L}{K}), \quad (8)$$

which expression is of the form (3). Here $\varepsilon = \sqrt{K^2 + L^2}/E_0$ is thus the dielectric constant, while tan $\delta = L/K$.

If different relaxation times $\tau_2, \tau_3, \dots$ are present simultaneously, then

$$K = E_0\,(\varepsilon_1 + \sum_m \frac{\varepsilon_m}{1 + \tau_m{}^2\omega^2}) \text{ and } L = E_0 \sum_m \frac{\varepsilon_m\,\tau_m\,\omega}{1 + \tau_m{}^2\omega^2} \quad . \quad (9)$$
$$m = 2, 3, 4, \dots$$

Since the values $\varepsilon_m$ are usually very small compared with $\varepsilon_1$, $L$ is negligible compared with $K$, and in many cases one may confine oneself to the first term in the expression for $K$.

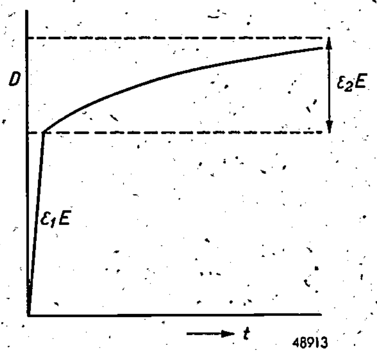


Fig. 1. The variation of the displacement $D$ with the time $t$, when at time $t = 0$ an electric field $E$ is suddenly switched on and causes the displacement $\varepsilon_1 E$. The contribution caused by the adjustment of the dipoles increases gradually from zero to $\varepsilon_2 E$.

For the explanation of the after-effects according to assumption (a) we now make use of considerations which were originally applied by Debije to liquid dielectrics. It is hereby assumed that in an insulating, viscous (thickly flowing) liquid there are a number of molecules, considered to be spherical, which are carriers of a dipole. In the first instance we suppose all molecules to be alike; let the number of them per cm$^3$ equal $n$.

The formula which Debije arrives at and which of course shows great similarity to the previously [1]) derived expressions, is as follows:

$$D = E_0 \left( \varepsilon_1 + \frac{na}{1 + \tau^2\omega^2} \right) \cos \omega t +$$
$$+ E_0 \frac{na \cdot \tau\omega}{1 + \tau^2\omega^2} \sin \omega t, \quad \dots \quad (10)$$

where $\tau$ is the relaxation time and $a$ the maximum possible contribution per particle to the dielectric constant. One thus finds in approximation for the dielectric constant and the phase angle

$$\varepsilon = \varepsilon_1 + \frac{na}{1 + \tau^2\omega^2}, \quad \dots\dots \quad (11a)$$

$$\tan \delta = \frac{na}{\varepsilon_1} \frac{\tau\omega}{1 + \tau^2\omega^2} \quad \dots\dots \quad (11b)$$

As a function of $\omega$ equation (11b) has a maximum at $\omega = 1/\tau$ (*fig. 2*).

The formulae, which were originally derived for a model of a liquid dielectric, are now also applied to the solid substance, although in this case we cannot form such a clear picture of the frictional forces opposing the orientation of the dipoles. It may be assumed that the dipoles present in the solid substance are bound to a number of preferential positions and that as a result, upon the transition from a given orientation to another, a number of "potential peaks" lying between the preferential positions must be exceeded at the expense of a certain energy $q$ ("activation energy").

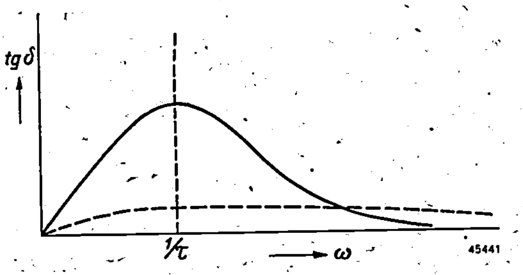


Fig. 2. The continuous line curve represents the variation of the tangent of the loss angle $\delta$ with the angular frequency $\omega$ according to Debije. $\tau$ is the time of adjustment (relaxation time). The dotted line curve relates to the case where different groups of dipoles are present with different values of $\tau$.

It is important for our argument that $\tau$ is very closely dependent on the temperature. From the kinetic theory of heat it follows that

$$\tau = \tau_0 \, e^{q/kT},$$

where $\tau_0$ has the significance of a material constant. Further $q/k$, where $k$ is the Boltzmann constant, is of the order of magnitude 10 000 °K.

Theory as well as experiment show that $\tau_0$ is a maximum of $10^{-13}$ sec but may also be considerably less, for instance $10^{-20}$ sec.

After these preliminaries on the calculations of Debije and the relaxation time $\tau$ — both of which related to insulators with a very regular internal structure (all dipoles alike) — we now return to the technical dielectrics.

The special assumption from which the three above-mentioned properties follow is that in the technical solid dielectrics investigated, which all possess an irregular internal structure, there is a wide scattering in the values of the activation energy $q$. This means, therefore, that not all the dipoles possess the same relaxation time and that the differences which occur are caused by differences in the values of $q$. All values of $q$ may occur in a certain interval.

We are most interested in property (3) and the constant $A$ thereby occurring. Theory based on the special assumption just mentioned gives the following formula:

$$A = \frac{2}{\pi} \frac{\ln \tau_0 \omega}{T} \quad \ldots \ldots \quad (12)$$

We shall here show briefly how the three properties (1), (2) and (3) follow from the fundamental assumption indicated and how in particular the formula for the constant $A$ is derived. To this end we must first discuss the assumption in more detail. Let $n$ be the total number of dipoles present per cm³. We then assume that the part of this number which upon being placed in an electric field shows a value of $q$ lying between $q$ and $q + dq$ can be represented by

$$n \cdot G(q)\, dq \quad (\text{with } \textstyle\int G(q)\, dq = 1),$$

where $G(q)$ in a given region varies only slowly with $q$. We shall be able to formulate this last condition more precisely in the course of our further considerations. The variation of $G(q)$ may, however, differ considerably between one substance and another. Thus per cm³ we encounter in the substance: $n \cdot G(q) dq$ dipoles all possessing the relaxation time $\tau = \tau_0 e^{q/kT}$.

For the sake of simplification we assume that $\tau_0$ and $a$ do not depend upon $q$. It is, however, easy to ascertain that the method of calculation to be followed also remains valid when these quantities do indeed depend upon $q$, but not very closely.

For the total dielectric constant and the phase angle we now find

$$\varepsilon = \varepsilon_1 + na \int_0^\infty \frac{1}{1 + \tau^2 \omega^2} G(q)\, dq \ldots \quad (13a)$$

$$\text{and } \tan \delta = \frac{na}{\varepsilon_1} \int_0^\infty \frac{\tau\omega}{1 + \tau^2 \omega^2} G(q)\, dq\,, \quad \ldots \ldots \quad (13b)$$

where $\tau = \tau_0 e^{q/kT}$, while $\varepsilon_1$ represents the ordinary dielectric constant which the substance would exhibit if the dipoles were unable to move.

*From these expressions, which are nothing but a generalization of formulae (11a) and (11b) of Debije, the three properties in question now follow.*

We shall, however, first reduce these formulae somewhat in order to make it easier to draw conclusions from them. For $\varepsilon$ one may write

$$\varepsilon = \varepsilon_1 + na \int_0^{q_m} G(q)\, dq, \quad \ldots \ldots \quad (14a)$$

where $q_m$ is the value of $q$ for which $\tau\omega = 1$, namely $q_m = -kT \ln \tau_0\omega$, while for tan $\delta$ we may write

$$\tan \delta = \frac{\pi}{2} \frac{na}{\varepsilon_1} G(q_m)\, kT \quad \ldots \ldots \quad (14b)$$

Formulae (14a) and (14b) may be interpreted as follows: Only those dipoles contribute to $\varepsilon$ whose relaxation time is smaller than $1/\omega$ (the others are too "slow"); the losses are caused by a group of dipoles whose relaxation time amounts to about $1/\omega$.

In the derivation of (14a) and (14b) use has been made of the property that $G(q)$ changes "slowly" with $q$. The requirement is that $G$ shall vary little in an interval for $q$ in which $\tau\omega/1 + \tau^2\omega^2$ differs appreciably from zero.

From the expressions (14a) and (14b) for $\varepsilon$ and tan $\delta$ the peculiarities found experimentally now follow, and we shall examine these successively for the properties (1), (2) and (3).

*Property* (1)

In order to understand that tan $\delta$ is practically independent of $\omega$ we note that $q_m \doteq -kT \ln \tau_0\omega$ shows only slight variations with $\omega$. Because, when $\tau_0 = 10^{-14}$ sec, then

$$-\ln \tau_0\omega = 27{\cdot}6 \quad \text{at } \omega = 10^2 \text{ rad/sec,}$$
$$-\ln \tau_0\omega = 11{\cdot}5 \quad \text{at } \omega = 10^9 \text{ rad/sec.}$$

The change in $q_m$ is thus only by a factor 2·5 upon the enormous change in $\omega$ by a factor $10^7$. At smaller values of $\tau_0$ the variation is still smaller. Now because $G$ depends little on $q_m$, in a suitable region tan $\delta$ will depend little on $\omega$. The "suitable region" need thus occupy at the most a factor 2·5 in the values of $q$ in the vicinity of $q_m$.

A dependence of tan $\delta$ on the amplitude $E_0$ does not occur at all.

*Properties* (2a) *and* (2b)

In order to prove the property (2a) it must be recalled that

$$\frac{\partial\varepsilon}{\partial\omega} = \frac{\partial\varepsilon}{\partial q_m} \cdot \frac{\mathrm{d}q_m}{\mathrm{d}\omega} = -\,G(q_m)\,\frac{kT}{\omega} = -\frac{2\varepsilon_1}{\pi\omega}\tan\delta.$$

In the frequency region where tan $\delta$ is practically independent of $\omega$, therefore, as may be found by integration of the above formula, the following is valid:

$$\varepsilon - \varepsilon(\omega_0) = -\frac{2\varepsilon_1}{\pi}\tan \ln \delta\,\frac{\omega}{\omega_0},$$

where $\omega_0$ is the beginning of the frequency region in question. For the sake of example we choose $\omega_0 = 10^2$ rad/sec and $\omega = 10^9$ rad/sec, while for instance we let tan $\delta = 10^{-2}$ at $\omega = \omega_0$. Then $\varepsilon = \varepsilon(\omega_0) - 2\varepsilon_1/\pi \tan\delta \ln 10^7$. Since $\tan\delta = 10^{-2}$, we obtain $\varepsilon = \varepsilon(\omega_0) \cdot (1 - 0{\cdot}1)$. Upon a change in frequency from $10^2$ to $10^9$ rad/sec therefore $\varepsilon$ changes only by 10%.

From (14a) it follows that

$$\frac{\partial\varepsilon}{\partial T} = \frac{\partial\varepsilon_1}{\partial T} + \frac{\partial\varepsilon}{\partial q_m} \cdot \frac{\mathrm{d}q_m}{\mathrm{d}T} = \frac{\partial\varepsilon_1}{\partial T} - na\,G(q_m)\,k \ln \tau_0\omega. \quad (15)$$

In general it is found that $\partial\varepsilon_1/\partial T$, caused by the expansion of the substance upon rise in temperature, is negligible compared with the second term.

At the values of $\tau$ and $\omega$ occurring in practice the second term is always positive. We therefore find that the temperature coefficient is positive, while in a good approximation the following is valid:

$$\frac{\partial\varepsilon}{\partial T} = -na\,G\,(q_m)\,k \ln \tau_0\omega \quad \ldots\ldots \quad (16)$$

*Property* (3)

In order to derive the remarkable property $(1/\varepsilon)\,\partial\varepsilon/\partial T = A \tan\delta$ we note that formula (16), just found by making use of the expression for tan $\delta$, can be written as follows:

$$\frac{\partial\varepsilon}{\partial T} = -\frac{2\,\varepsilon_1 \tan\delta}{\pi T}\ln\,\tau_0\omega,$$

so that

$$\frac{1}{\varepsilon}\frac{\partial\varepsilon}{\partial T} = -\frac{2}{\pi}\frac{\ln\tau_0\omega}{T}\tan\delta = A\tan\delta; \quad \ldots \quad (17)$$

Thus

$$A = -\frac{2}{\pi}\frac{\ln\ \tau_0\omega}{T} \quad \ldots\ldots\ldots \quad (12)$$

The "constant" $A$ thus determined depends only very little on the frequency, because if we choose $T = 300°$ K, at $\tau_0 = 10^{-14}$ sec, we find

$$A = 0.058 \quad \text{at } \omega = 10^2 \text{ rad/sec and}$$
$$A = 0.024 \quad \text{at } \omega = 10^9 \text{ rad/sec,}$$

while at $\tau_0 = 10^{-20}$ sec

$$A = 0.088 \quad \text{at } \omega = 10^2 \text{ rad/sec,}$$
$$A = 0.054 \quad \text{at } \omega = 10^9 \text{ rad/sec.}$$

From this it follows that, although the values of $\tau_0$ for the different substances differ by many times a factor of 10, at room temperature we shall always find only slightly differing values of $A$ for measurements in the customary frequency region. The calculated value of $A$ agrees reasonably well with the measured values which lie in the vicinity of 0·05.

Until now we have specialized on the case (a) where the losses are supposed to be caused by dipoles. We may now give an entirely similar line of reasoning for cases (b) and (c) in which semiconducting regions or free ions are the cause. It would lead us too far to treat this possibility in as great detail as has been done for case (a). The fact that the properties found can be derived not only from cause (a) but also from (b) and (c), however, also implies that for a given substance it is difficult to ascertain whether the losses are caused by (a), by (b) or by (c) or by more than one of these together.

## Exceptions

In the above explanation of property (3) it was assumed that $\partial\varepsilon_1/\partial T$ is small enough to be ignored. In most cases this is true, but there are a few exceptions. Theory shows (see footnote[2]) that for the part $\varepsilon_1$ of the dielectric constant $\varepsilon$ which is independent of the after-effect the following equation holds:

$$\frac{1}{\varepsilon_1}\frac{\partial\varepsilon_1}{\partial T} = -\,\alpha_{\text{lin}}\frac{(\varepsilon_1 - 1)\,(\varepsilon_1 + 2)}{\varepsilon_1}, \quad (18)$$

where $\alpha_{\text{lin}}$ represents the linear coefficient of expansion. From this it follows that $\partial\varepsilon_1/\partial T$ is negative and in absolute value larger according as $\alpha_{\text{lin}}$ and $\varepsilon_1$ are larger.

Now while on the one hand the coefficient of expansion seldom assumes abnormally large values, high values of $\varepsilon_1$ (and thus also of $\varepsilon$) do on the other hand certainly occur. In contrast to most substances having a value of $\varepsilon$ between 1 and 10, for the substance rutile ($TiO_2$), for example, $\varepsilon \approx 100$. For such substances $\mathrm{d}\varepsilon_1/\mathrm{d}T$ can no longer be ignored, as was assumed in the derivation of the formula for $A$ (formula (16)). As a result properties (2b) and (3) do not hold for these substances. Under the influence of the term $\partial\varepsilon_1/\partial T$ they have a lower positive temperature coefficient than would be expected according to formula (5), and in some cases, especially when $\tan\delta$ is small, they even have a negative temperature coefficient.

In cases where a substance is needed with a very small temperature coefficient of the dielectric constant, a choice can thus be made in two ways. If it is desired to use substances with normal values of $\varepsilon$, for which property (3) holds, as was stated at the beginning, the search must be made among the substances possessing a low value of the loss angle $\delta$. The other possibility is that one should use substances with a high value of $\varepsilon$ for which property (3) does not hold. With these substances it is possible, if desired, to obtain negative temperature coefficients.

---

Number 1 of Volume 2 (February, 1947) of *Philips Research Reports* contains the following papers:

*R 32:* B. D. H. Tellegen: Coupled circuits.

*R 33:* W. Elenbaas: On the excitation temperature, the gas temperature, and the electron temperature in the high-pressure mercury discharge.

*R 34:* H. B. G. Casimir: On the theory of eddy currents in ferromagnetic materials.

*R 35:* H. C. Hamaker: Radiation and heat conduction in high-scattering material.

*R 36:* H. A. Klasens: The light emission from fluorescent screens irradiated by X-rays.

Readers interested in any of the above mentioned articles may apply to the Administration of the Philips Physical Laboratory, Kastanjelaan, Eindhoven, where a limited number of copies are available for distribution. For a subscription to Philips Research Reports please write to the publishers of Philips Technical Review.

VOL. 9 No. 4, p. 97-124 1947

# 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

## FUNDAMENTALS FOR THE DEVELOPMENT OF THE PHILIPS AIR ENGINE

by H. de BREY, H. RINIA and F. L. van WEENEN.

621.412

The hot-air engines which were in use in the last century had a low efficiency and a large piston displacement per horse-power. Systematic investigation in the Philips Laboratories has shown that nowadays an air engine with much more favourable characteristics can be constructed. The employment of the heat-resisting steel alloys now available has made it possible to reach a mean effective pressure of 14 atmospheres (maximum pressure 50 atmospheres) and to increase the temperature of the hot chamber to 650° (centigrade). A suitable construction of heater, cooler and regenerator, based upon new insight into the compromise which must be made, in the case of these elements between heat transfer, flow resistance and dead volume, has made it possible to work with speeds of 3000 r.p.m. or more, and at the same time to ensure that the temperature of the work medium in the hot space will be only slightly lower than that of the heater, and in the cold chamber (80°) only slightly higher than that of the cooler. The high pressures and speeds have caused an improvement in the specific power by a factor of more than 100 compared with the old air engines, so that it is now of the same order as in internal combustion engines. The great temperature difference between the hot and cold spaces makes the theoretical efficiency high (like that of the Carnot process). Thanks to a radical improvement of the regenerator, increase in the mechanical efficiency, employment of a suitably constructed preheater and the natural restriction of radiation and other losses due to the small dimensions and the compact construction, the overall efficiency of the Philips air engine is also high; it is now comparable to that of internal combustion engines. Compared with the latter and with other sources of power, however, the air engine possesses a number of fundamental advantages.

### Introduction

In an article in the May, 1946, number of this periodical [1]) the theoretical principles of the air engine were discussed. It was then shown that the theoretical efficiency of an air engine with regenerator is equal to that of a Carnot process taking place between the same temperatures, and thus equal to the maximum possible for a prime mover within those temperature limits. If the absolute temperature of the "hot space" is $T_h$ and that of the "cold space" $T_c$, the theoretical efficiency of the air engine is equal to

$$\eta = \frac{T_h - T_c}{T_h} \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

For $T_h = 920$ °K (about 650 °C) and $T_c = 350$ °K (about 80 °C), for example, $\eta = 62\%$.

This differs very much from the overall efficiencies of the previously constructed air engines, which were usually not over 3%; this value was still given for a 2 h.p. air engine in 1923 in the catalogue of a well-known firm. Moreover, the old models were very large and heavy; 800 kg was given as the weight of the above mentioned engine.

In the article already referred to [2]) it was stated that in the Philips Laboratories during recent years extensive theoretical and experimental investigations of the hot-air process have been carried out, and that it has been determined that this process can be used with a much more favourable result, provided that in the construction of the engines modern knowledge about heat transfer, flow resistance etc. is applied and modern materials are used. In this way it has been found possible to build air engines which, with the same total power

[1]) The volume which takes no part in the expansion and compression and which lowers the specific power (see in I, fig. 5).

[2]) H. Rinia and F. K. du Pré, Air Engines, Philips Techn. Rev. 8, 129-136, May 1946.

and speed, have an efficiency just as high as ordinary internal combustion engines and can develop a still higher power than the latter per litre swept-volume (or per kilogram weight of engine).

We shall now discuss in somewhat more detail several facts which the investigations in question brought to light, as well as several new principles of construction which have been deduced therefrom and which form the fundamentals for the development of the Philips air engine. In the following, when citing the aforementioned article we shall refer to it as "I".

We shall first review in a very general manner the principle of the air engine. A certain quantity of air (or some other working medium, such as hydrogen, helium, argon or the like), which is at a high temperature, is allowed to expand in a cylinder, the hot space, where by the movement of a piston it performs mechanical work. In another cylinder, the cold space, the expanded air is cooled and compressed by the action of a piston to the original volume. It is then heated to the original high temperature. The cycle then begins anew. The compression of the air requires mechanical work; since, however, this compression takes place at low temperature and thus at low pressure, the work required is less than that which the air performs upon expansion at high temperature and thus high pressure. The engine thus produces an excess of work.

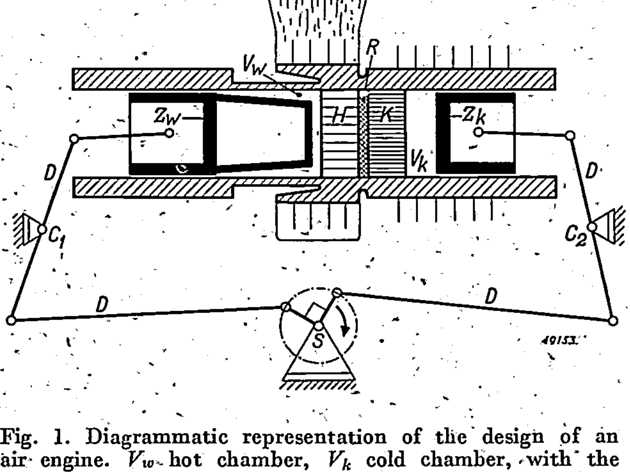


Fig. 1. Diagrammatic representation of the design of an air engine. $V_w$ hot chamber, $V_k$ cold chamber, with the corresponding pistons $Z_w$ and $Z_k$. *B* burner, *H* heater, *R* regenerator, *K* cooler, *S* crankshaft, *D* connecting rods, $C_1$, $C_2$ fixed pivots. The pistons act on the crankshaft in such a way that the volume variations of the hot space are about 90° in phase ahead of those in the cold space.

A possible model of an air engine, very much simplified, is shown in *fig. 1.* It may be seen that the movement of the air to and fro from the hot to the cold space and *vice versa*, is accomplished by two pistons moving approximately sinusoidally, whose motions are about 90° out of phase. Since the four different stages of the process: expansion, cooling, compression, heating, cannot be entirely separated, they partially overlap. In I it was shown that in spite of this the machine functions as an engine, provided the condition is satisfied that the volume changes of the hot space are advanced in phase with respect to those in the cold space. Upon the phase angle depends the power and, to a lesser extent, the efficiency. Both exhibit an only slightly pronounced maximum at a phase difference of about 90° (see I, fig. 5); while if the phase angle is chosen negative, the engine then consumes energy and acts as a refrigerator.

In fig. 1 the so-called regenerator is shown between the parts where the air is heated and where it is cooled. Its function is to store the heat which the air gives off after the expansion when passing to the cold space. When after compression the air is returned to the hot space, it again passes through the regenerator and takes up the heat stored there. As shown in I and as will be seen from what follows, the regenerator fulfils an essential function.

Let us now turn to the investigations made for the improvement of the air engine, which is the subject of this paper. These investigations, as will be clear from the statement at the beginning of this article, had a two-fold object. The first object was to improve the efficiency, the second one to increase the specific power (*i.e.* the power delivered to the main shaft divided by the swept volume). These two objects cannot actually be strictly separated: the measures taken in pursuance of them affect each other in all kinds of ways. In order to give a clear picture it is desirable first to consider the specific power.

## Increase of the specific power

The power of an engine is equal to the product of the engine speed and the work per revolution. The latter, which was investigated in detail in I for the air engine, is given by the area of the $p$-$V$ diagram or indicator diagram, which is represented for a certain case in *fig. 2,* curve *H.* This area is the same as that of the shaded rectangle *abcd,* whose length is equal to the piston displacement $V_a$, while its height is indicated as the mean effective pressure $p_m$.

The specific power, defined as the power divided by the piston displacement $V_a$, is thus equal to the product of the number of revolutions and the mean effective pressure $p_m$. If one desires to increase the specific power, one must try to increase both the factors above mentioned.

### *The mean effective pressure*

In air engines in general a relatively small expansion ratio is used (see I), and the maximum pressure occurring is at most 2 to 2.5 times the minimum. The difference between the two pressures to which the mean effective pressure $p_m$ is roughly proportional, can then only be increased by increasing the whole pressure level in the $p$-$V$ diagram.

All the old air engines worked with a maximum pressure of only a few atmospheres, while the lowest pressure was about 1 atm. (Usually the working medium was brought for a moment into open contact with the external air at the moment of lowest pressure). Therefore, for those engines the mean effective pressure $p_m$ was usually not higher than about 0.6 atm.

The fact that higher pressures were not used was due mainly to the lack of suitable materials for construction. The hot space of an air engine must be permanently at a high temperature. During the time when air engines were in vogue in the last century, only cast iron or bronze was available for the heated parts. The tensile strength of these metals is only small at a high temperature and especially the creep velocity at high temperature and pressure is much too high. It was therefore impossible to raise very high either the temperature (to which we shall return in connection with the efficiency) or the pressure.

It was not until around 1920 that materials became available which had a low creep velocity at high temperatures. In recent years the technology of heat-resisting metals (*i.e.* creep and oxidization-resisting) has made rapid advances. Because of this it was not only possible to get a much better construction of the air engine, but the realization of the gas turbine, the jet engine and similar engines is based upon these advances.

Thanks to the employment of modern materials it has been possible to increase the maximum pressure in the air engines developed by Philips to approximately the same level as in internal combustion engines, namely to about 50 atm. The minimum pressure is then about 22 atm., the expansion ratio 2.3, and the mean effective pressure $p_m$ in that case is about 14 atm. This is more than twice as high as in ordinary internal combustion engines, in which, as a result of the less favourable shape of the indicator diagram, a value of $p_m$ of the order of 6 atm. is obtained. These conditions are diagrammatically rendered in *fig. 2.* In the case of internal combustion engines somewhat higher values of $p_m$ can also be reached, but this requires very complicated auxiliaries such as a compressor with high air yield[3]), possibly combined with an exhaust-gas turbine.

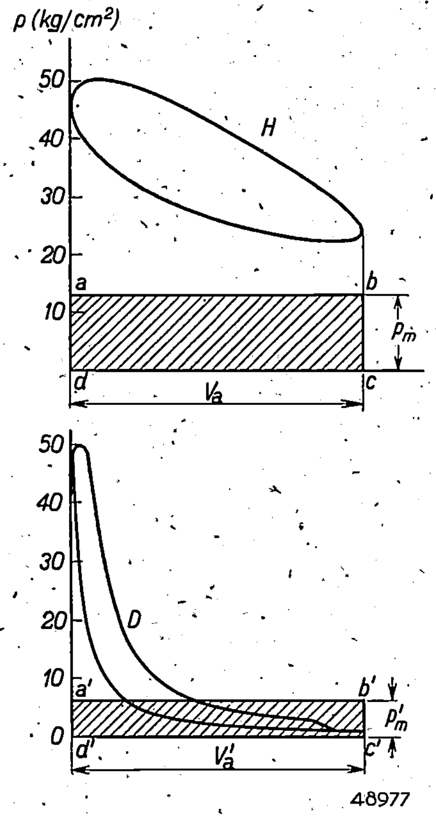


Fig. 2. Comparison of the indicator diagrams of an air engine (*H*) and a two-stroke Diesel engine (*D*). For both the maximum pressure has been chosen equal to 50 atm., while the swept volume $V_a$ is also assumed to be equal in both cases. (Owing to the dead volume of the air engine being much larger than in the Diesel engine, the diagram for the former is actually shifted to the right with respect to the latter.) In the figure the mean effective pressures $p_m$ are also indicated: $p_m$ is the height of the rectangle *abcd* whose base is equal to $V_a$ and whose area corresponds to that of the curve *H*. In the same way $p_m'$ is the height of the rectangle *a'b'c'd'* whose area is equal to that of curve *D*. It is clear that $p_m$ is considerably larger than $p_m'$.

### *Engine speed*

The engine speed is limited mainly by two factors. In the first place the piston speed and thus the inertia forces of the moving parts increase with the engine speed. In the second place, with a higher speed the work medium must flow to and fro between the hot and cold spaces more quickly, thereby not only increasing the losses due to flow resistance, in the heater, the regenerator and the cooler, but making it more difficult to obtain the required heat transfer in these three elements. These latter

[3]) The compressor for an air engine need only be small to reach the abovementioned high minimum pressure of 22 atm. This will be explained in a subsequent article.

questions, which are connected with the efficiency, will be discussed in more detail farther on; here we would only mention that thanks to the new construction of the three elements as developed by Philips; the problems of flow resistance and heat transfer no longer constitute any obstacle in increasing the considerably speed engine. As to the piston speeds, they become smaller the higher the pressure of the air engine is raised; as explained in the foregoing section a given power can then be obtained with a smaller swept volume. As a result it is possible to raise the speed to 3000 r.p.m. or higher, which again leads to a considerable increase in the specific power.

The magnitude of the improvement achieved in this way is made clear from a comparison with the example of the old 2 h.p. air engine already referred to, which had a swept volume of 25 litres. The same power is now attainable with an engine of only about 200 cc constructed for the same, very long life. The swept volume has thus been reduced to 1/125. The weight of the engine has also been very much reduced; in the case of a 2 h.p. engine a reduction by a factor of 50 has been attained [4]).

## Improvement of the efficiency

### *Favourable influence of an increased specific power*

In the foregoing discussion, a point was indicated where the requirements for an increase of the specific power and for the improvement of efficiency conflicted with each other where it was a question of the flow velocity of the working medium. It is remarkable, however, that in many other respects the requirements for the two objects run parallel. For example it is clear that for a given power in the case of a small engine the losses through heat radiation will be smaller than in the case of a larger engine. Moreover, in the case of a small engine these losses can more easily be reduced by heat insulation. In the Philips air engines a still higher efficiency is obtained by a very compact construction, as will be shown from a later description. The conduction losses *via* the walls of the cylinders are of course also smaller in a small engine than in a large one. This is especially true where, as in the case of the large, old engines, in order to obtain sufficient strength, the metal walls of the hot chamber had to be made very thick, a practice no longer necessary with modern steel alloys.

Another point of importance is that the earlier, clumsy air engines had to be equipped with heavy moving parts which caused large friction losses, and the mechanical efficiency, *i.e.* the ratio between the power delivered to the main shaft and the indicated power, was very low. Since the new engines can be built so much smaller, the frictional losses have been appreciably reduced. Thanks to a suitable construction in the Philips air engine they have been kept particularly low. We shall not go into this point here, since the actual construction of these engines will be dealt with in a separate article in this periodical. It can be stated however that for the latest-model of the largest engines, for several hundred h.p., the mechanical efficiency amounts to more than 90%.

### *The temperature of the working medium in the hot and cold spaces*

The losses which we have so far been discussing, all of which formerly constituted an obstacle to the attainment of a high efficiency, are in the main not specific for the air engine. Their decrease is indeed of great importance, but the most important advance in efficiency is made possible by the measures which improved the essential conditions for the hot-air process itself. In the old air engines these conditions were unfavourable in two respects: in the first place the temperature ($T_h$) of the working medium in the hot space was relatively low, in the cold space ($T_c$) relatively high; because of this, according to (1), even the theoretical efficiency was bound to be low. In the second place there was insufficient regeneration, or even no regenerator at all. We shall consider these two points separately.

One of the reasons for the relatively low temperatures formerly employed in the hot space we have already learned: the lack of materials sufficiently resistant to oxidation and at the same time free from creep. Consequently the temperature of the wall of the hot chamber — even at relatively low pressures — was limited. Another reason was the poor heat transfer to the working medium. The transfer of heat to the medium was usually allowed to take place *via* the walls of the hot space themselves. Even at the low speeds of the large old engines this was quite inadequate and as a result the temperature of the air in the hot space was usually much lower than the already relatively low wall temperature. The air temperature was seldom higher than 300 °C.

In order to improve upon this and thus, especially at high speed, to obtain a satisfactory heat transfer, a special heater was introduced, as already stated in I. This heater raises the air at the entrance

[4]) For each type of engine the specific power can be increased and the weight limited if a shorter life (or a lower efficiency) is accepted.

to the hot space to the desired temperature, *i.e.* about 650 °C in the Philips air engine. The construction of the heater is simple in principle: a channel with a large internal wall surface, obtained for example by the introduction of a large number of internal fins of sufficient length. The air is forced to flow through the narrow slits between the fins. The inclusion of such an element in the path of the working medium, however, has its disadvantages: in the first place the flow resistance experienced by the working medium on its path between the cold and the hot space is increased; in the second place the dead volume increases. It is easy to understand that these disadvantages will be all the greater, according as one tries to improve more the heat transfer of the heater. If the channels through which the air flows into the heater are made very long and narrow, it is always possible to raise the temperature of the air at the end to the desired level; at the same time, however, there is the risk that the flow resistance will be so great at the desired high engine-speed that much of the gain in efficiency will be quite lost again through the aerodynamic losses. If one tries on the other hand to make the passages for the air as wide as possible, so that the flow resistance remains sufficiently small, the dead volume again becomes large. A compromise must therefore be sought, which necessitates very careful choice of the dimensions of the heater. One of the most important tasks in the investigation carried out by Philips was to resolve the opposing differences between heat transfer, flow resistance and dead volume, which was not sufficiently understood before, and to determine the optimum conditions. It was possible to profit from the results of the numerous investigations of recent years in the field of heat transfer and theory of flow, supplemented in many respects by the results of our own research work.

Anticipating our discussion of the cooler and the regenerator, it may be stated here that the same problem arises there and a favourable construction has to be sought in a similar manner. The total aerodynamic losses in the engine with the constructions finally obtained, at a speed of 3000 r.p.m., could be limited to only about 10% of the indicated horse-power, while the heat transfer, which at this high speed has to be brought about in less than 1/100 sec., still fully satisfies the demands. Incidentally it may be noted that the decrease in the efficiency attributable to the aerodynamic losses actually amounts to considerably less than 10%, because the flow losses are converted partly into useful heat, namely in the heater and in part of the regenerator.

The construction of the heater is shown for a certain case in *fig. 3,* where the large number of channels formed by the internal fins can be seen.

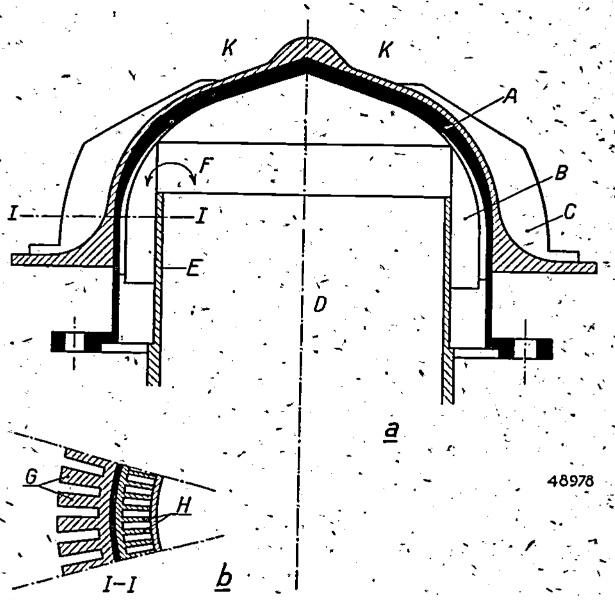


Fig. 3. Cross section of the heater of a Philips air engine; *a*) vertical cross section, *b*) part of horizontal cross section at height *I-I*. *A* bonnet of heat-resisting steel to withstand the pressure of the work medium, *B* fins of aluminium bronze for the heat exchange with the work medium (see also at *H* in the horizontal cross section), *C* fins of aluminium bronze for the heat exchange with the flame of the burner (see *G* in horizontal cross section), *D* hot space, *E* wall of the hot space. The fuel is burnt at *K*. The flame is led through the slits between the fins *G*. These and, as a result of the good heat conduction, also the fins *H*, are thus heated to the required temperature. The heat exchange with the work medium takes place in the slits between the fins *H*. The work medium can enter or leave the hot space only *via* the paths indicated by the arrow *F* (the hot space is closed at the lower side by a piston not shown here).

The outside of the heater is kept at the desired temperature by a burner. In order to ensure the best possible heat exchange also the outside is provided with fins, which are in direct contact with the flame of the burner. The internal and external fins are constructed of a good heat-conducting material, for instance aluminium bronze. This makes the temperature gradient in the fins small. The wall to which the fins are attached must be made of heat- and creep-resisting steel in order to withstand the high pressure. Since, however, this wall is thin, the slight heat conductivity of this steel has little effect.

We may briefly refer to the construction of the cold space. The mechanical requirements made of the walls here can more easily be met because of the lower temperature. As far as the heat transfer is concerned, however, this situation is analogous to that of the hot space. In order to reach and maintain as low a temperature as possible in the cold space $T_c$, the cooler is introduced at the entrance to the cold space. Its construction shows great similarity to that of the heater. Internally it corresponds exactly to the heater. The outside of the cooler is cooled with water or with air.

It should be pointed out that the satisfactory results obtainable with this construction of heater and cooler in practice are in fact due to the presence of the regenerator, which very much lightens the task of the other two elements. Without the regenerator the heater would have to heat the air from, for instance, 100° to 650° C each time, whereas now it has only to be heated from about 500° to 650° C. In the same way, without the regenerator the cooler would each time have to cool the expanded air from about 500° C to for instance 60° C, whereas now it is only necessary to cool from about 120° C to 60° C. This will be clearer from the more detailed discussion of the regenerator, which follows.

*The regenerator*

As was explained in I and again at the beginning of this article, the regenerator has to store the heat liberated as the work medium (air) passes from the hot to the cold space, and it has to give up this heat again when the movement is reversed. The importance of the function of the regenerator for the hot-air process appears most clearly when we calculate the ratio between the amount of heat $q_r$ stored each time in the regenerator and the amount of heat $q$ supplied per cycle by the heater to the work medium, assuming that the regenerator is ideal. With the values of $T_h$, $T_c$ and the expansion ratio which are customary in the Philips air engines, the ratio $q_r/q \approx 3$ or slightly more. *The work medium thus takes up more than three times as much heat in the regenerator as in the heater.* In other words, without a regenerator the heater would have to supply about four times as much heat. The efficiency of the engine could not then be more than 1/4 of what can be attained when an ideal regenerator is present.

For the cycle actually taking place in the engine the calculation of $q_r/q$ is rather complicated. We shall therefore calculate it only for the idealized cycle described in I of two isotherms and two isochores (two verticals of constant volume), for which it is very simple. In that case $q_r = mC_p\,(T_h—T_c)$, in which $m$ is equal to the total mass of the air in gram molecules. Furthermore

$$q = \int_{V_2}^{V_1} p\,\mathrm{d}V = mRT_h \int_{V_2}^{V_1} \frac{\mathrm{d}V}{V} = mRT_h \ln \frac{V_1}{V_2},$$

where $V_1$ is the maximum and $V_2$ the minimum volume of the air. Now for air $C_p$ (per gram molecule) is about $^7/_2R$. If we take $T_h = 920$ °K, $T_c = 350$ °K and $V_1/V_2 = 2$, then

$$\frac{q_r}{q} = \frac{7}{2}\,\frac{T_h - T_c}{T_h}\,\frac{1}{\mathrm{In}\ V_1/V_2} = 3.1.$$

We thus find here a value of the same order of magnitude as that already mentioned, although it is obvious that the manner of calculation followed can only furnish a rough approximation.

From the above it follows that not only is a regenerator absolutely indispensable for good efficiency, but also that it is of the greatest importance that it should be of the best possible construction. If, say, 1% of the heat imparted to the regenerator and subsequently taken back by the work medium is lost, the performance of the heater must be increased by 3% and the efficiency drops accordingly.

It is therefore remarkable that in the earlier constructions of air engines there was often no regenerator at all. The reason for this must be sought in the above-described opposing interests between heat transfer on the one hand and the flow resistance and dead volume on the other. It was in fact quite possible to construct a regenerator with a high efficiency, but in the older constructions this was often accompanied by such a high flow resistance that it was considered preferable to dispense with regeneration.

Let us now study the requirements to be met in order to approximate an ideal regenerator as nearly as possible.

In the first place care must be taken that the heat exchange with the air takes place quickly enough. For this purpose the flow channels in the regenerator must have the correct dimensions, the above-mentioned compromise between heat transfer, flow resistance and dead volume being taken into account. In the case of the Philips air engines with high specific power the choice of this compromise is more critical than it was in the old engines of much larger volume.

In the second place a large heat capacity of the regenerator is required. With an ideal regenerator the process of storing up and giving off heat is "reversible" in every phase. In order to approximate to this it is necessary that, when flowing through the regenerator the air should only differ slightly in temperature from the surroundings. The temperature distribution prevailing in the regenerator must form a certain gradual transition from the heater to the cooler. In order to maintain this temperature distribution and to obtain only slight temperature fluctuations during each revolution of the engine, the heat capacity of the regenerator must be large compared with the heat capacity of the air which flows periodically back and forth through the regenerator.

Finally the heat conductivity of the regenerator in the direction of flow must be slight, otherwise a continual flow of heat occurs from the heater to the cooler, which means a loss.

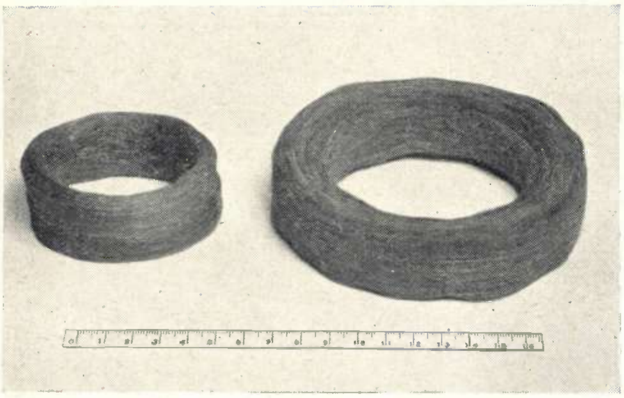


Fig. 4. Two regenerators for engines of $3^1/_2$ and 10 h.p. respectively. In the Philips air engines the air flows from the hot to the cold space through an annular shaped channel in which the regenerator is placed. (The scale is in centimeters.)

It has been found possible to construct a regenerator which comes very close to the ideal. In *fig. 4* a photograph is reproduced of two regenerators used for models of the Philips air engine of different power. These regenerators consist of a porous coil of thin metal wire. As was already stated in I, regenerator efficiencies of 95% and more are attained. It is remarkable that such a coil of metal wire is able to raise the temperature of the quantity of air flowing through it from about 100° C to about 600° C within 1/100 sec, or the reverse, and that temperature gradients of several hundred degrees per centimeter in the direction of flow can exist in it without appreciable loss of heat.

### *The preheater*

In conclusion we will discuss a loss in efficiency which, like the losses considered at the beginning, is not specific for the air engine but which played a very large part in the case of the old engines, *viz.* the loss of heat in the exhaust gases from the burner used for heating, a loss encountered with every burner. In the case of the Philips air engine the heater has a temperature of more than 650° C. The flame and thus also the products of combustion, the exhaust gases, necessarily have a still higher temperature. If the exhaust gases are allowed to escape unused a considerable amount of heat is lost. This loss can be very much restricted by the use of a heat exchanger in which the exhaust gases give off their heat in counterflow to the air entering for the combustion.

Such a preheater has also been worked out for the Philips air engine. For example a certain model of it has the form of a pleated collar of heat-resisting sheet material, the exhaust gases being passed along one side of the pleats while combustion air flows along the other side. By this means a very intense heat exchange is obtained and the exhaust losses can be restricted to about 30% of the original value. *Fig. 5* shows a preheater for gaseous fuel constructed in this way and adapted especially to the design of the air engine.

### Conclusion

Summarizing the results of the development sketched here, we may say that the air engine in its new form as designed in the Philips Laboratories possesses many favourable properties which until now had been considered only possible in internal

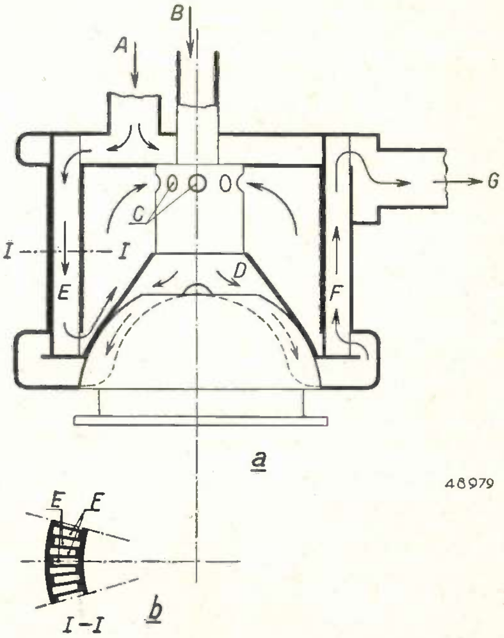


Fig. 5. Cross section through the preheater of a Philips air engine; *a*) vertical cross section, *b*) part of horizontal cross section at height *I-I*. The air for combustion enters at *A*. It follows the path of the arrows, thereby passing through the channels *E* and entering the space *D via* the openings *C*. The fuel gas — which enters through *B* — is mixed with the air in *D*, where the combustion takes place. The heater (fig. 3) along which the flame is led is indicated by dotted lines. The products of combustion are passed off *via* the channels *F* and the opening *G*. In the annular-shaped space, of which *E* and *F* represent a vertical cross section, is a pleated collar (*cf.* also the horizontal cross section). In this lie the channels *E* and *F* alternately side by side; this ensures a very good heat exchange between the air entering and the combustion products escaping.

combustion engines. It has proved possible to build compact, light, high-speed air engines of high power with an efficiency comparable to that of internal combustion engines. At the same time, however, all the fundamental advantages of the air engine over the internal combustion engine or other prime movers have been retained in the new form. We may mention here the possibility of using all kinds of fuel, low wear (since the surfaces of the cylinders do not come into contact with any corrosive gases), low consumption of lubricating oil, little noise (since there are no valves and no periodic explosions) and uniform torque thanks to the favourable shape of the indicator diagram (fig. 2). Other important advantages can only be discussed after a more detailed treatment of the design and characteristics of the new air engines. In our opinion, however, it is clear enough from the above that we are at the beginning of a very promising development.

# THE IGNITION MECHANISM OF RELAY TUBES WITH DIELECTRIC IGNITER

by N. WARMOLTZ.

621.3.032.43:621.314.653

Following a short account of the most usual methods of producing an arc discharge with a mercury cathode, closer consideration is given to the capacitive method of ignition. A positive voltage of several kilovolts is applied to a conductor separated from the mercury cathode by a thin insulating wall. The action is explained on broad lines by assuming a field emission from the convex surface of the mercury against the wall of the igniter, along which the electrons move towards an auxiliary anode, this possibility being accompanied by secondary emission. The field strength at the mercury surface as deduced from measurements, however, is much too low to be able to cause field emission, so that this would seem to disprove the conception just given. A solution to the problem lies in the theory of Tonks, according to which at a certain field strength (actually reached here) the mercury surface becomes unstable and small irregularities are drawn out to sharp points, where the field strength is indeed sufficient to produce field emission. This drawing out of the mercury surface to points requires time, and this explains the measured time lag. The size of the "humps" (which must be assumed to be initially present on the mercury surface) correlating the measured time lag with the "drawing out" time calculated by Tonks agrees with the calculated thermal irregularity of a mercury surface. In conclusion several applications are discussed of relay tubes with dielectric ignition.

## Survey of the methods of ignition of a mercury arc

Discharge tubes with a pool of mercury as cathode have been used for many years, for instance in the form of mercury-cathode rectifiers. Here the arc discharge emanates from a small part of the mercury surface, the so-called cathode spot. This intensely luminous spot, in which the current density is very high, moves around over the mercury in no defined course. Generally, special means have to be employed to bring it about. Formerly, and not infrequently still now, this ignition was initiated by interrupting the contact between the mercury cathode and a starting electrode forming part of an electric circuit.

In the course of time various constructions have been invented, of which we shall mention only a few here, namely those based upon:

*a*) the tipping of the whole discharge vessel, thereby breaking a mercury bridge between the starting electrode and the mercury cathode;

*b*) the raising of the starting electrode out of the pool of mercury (by means of an external magnetic coil);

*c*) the flowing back of mercury previously raised some way or other until it had made contact with the starting electrode.

Since a cathode spot can only exist as long as the current is above a certain limit (of the order of 5 amp.), mercury cathode rectifiers are usually provided with one or more auxiliary anodes whose current maintains the cathode spot even when the current of the main anode(s) falls below the critical limit. This makes it unnecessary, in the latter event, for the ignition mechanism to come into action again. A permanent auxiliary discharge, however, has its disadvantages, such as:

1) the increased risk of arc-back (passage of current in the wrong direction) owing to the continual presence of ionization, and

2) the not inconsiderable energy dissipation in the auxiliary discharge, at least in rectifiers for low power.

In the last decades great interest has arisen in discharge tubes in which the passage of current can be made to begin at precisely determined moments (relay tubes). Such tubes are practically inertia-free switches. In principle the above mentioned rectifiers with mercury cathode can be used if they are provided with control grids, but then, in addition to the objection just mentioned in connection with the permanent auxiliary discharge, there arises a third drawback, in that the ions from that discharge create the possibility of discharges (in the right direction) at moments when they are undesired.

For these reasons it was desirable not to have a permanent cathode spot but to seek means of causing the spot to arise only when it is needed. This we call controlled ignition. It can be used, for instance, for starting a relay tube at each cycle of the A.C. voltage used. In principle control grids thus become superfluous, because their function of opening the passage for current at certain moments is performed by the ignition arrangement.

### Controlled ignition

The above-mentioned methods of ignition are unsuitable for controlled ignition because of the mechanical inertia and the waves present on the surface of the mercury. However, two other methods have been developed which are free of these drawbacks. In the older of these an external ignition band is applied, which method was made publicly known in 1901 by Cooper Hewitt.[1] A modified version of this dielectric ignition will be discussed in this article. The other method, which we shall not discuss, was discovered in 1933 by Slepian and Ludwig[2]; it makes use of a semiconductor partly immersed in the cathode mercury.

*Fig. 1a* is a diagram showing the principle of a valve with the original ignition band. A conducting band is laid round the glass envelope of the discharge valve at the level of the mercury meniscus. When applying to this band a sufficiently high voltage impulse, positive with respect to the mercury cathode, a spark occurs near the place where the mercury touches the glass wall, thereby forming a cathode spot, after which the discharge is taken over by the anode. For smooth ignition, especially where the anode voltage is low, it is advisable to use an auxiliary anode connected to a point of high voltage, for instance the ignition band itself.

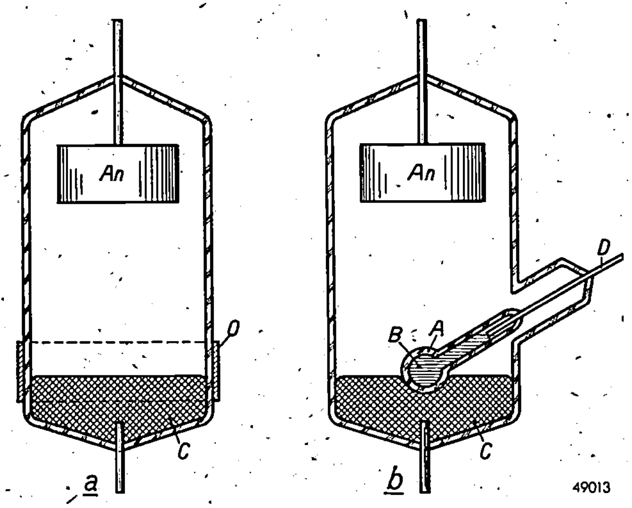


Fig. 1. *a*) The ignition band (*O*) of Cooper Hewitt is a metal band laid around the outside of the glass discharge vessel at the height of the meniscus of the mercury cathode *C*. A voltage impulse of about 10 kV applied to *O* and positive with respect to *C* causes a cathode spot on the mercury near the glass wall, after which a discharge to the anode *An* develops which further maintains the discharge.

*b*) Sketch of the internal dielectric igniter. *A* insulating wall, *B* conductor, *C* mercury cathode, *D* voltage lead to the starter.

Later on, several investigators recommended the form of an internal dielectric igniter, or starter, as illustrated in fig. 1*b*. An insulating tube, for instance of quartz, filled with a conductor is partially immersed in the cathode mercury. A positive voltage impulse of sufficient strength applied to the internal conductor causes the formation of a cathode spot on the outside of the starter. Here, too, an auxiliary anode may be desirable; the non-insulated part of the supply lead (fig. 1*b*) may serve as such.

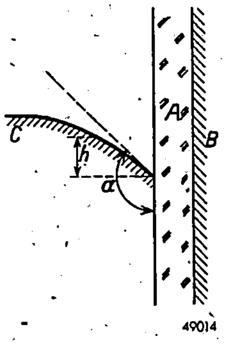


Fig. 2. Shape of the mercury meniscus at the igniter. The letters *A*, *B* and *C* have the same meaning as in fig. 1*b*. *α* is the so-called boundary angle, which varies for different substances and for the combination mercury-quartz amounts to about 135°. The position of a point on the convex mercury surface is determined by the height *h* above the boundary line between the mercury and the wall of the igniter.

The spark from which the cathode spot originates is formed in the gap between the convex mercury meniscus and the insulating wall of the starter. *Fig. 2* is an enlarged diagram of this gap. We shall revert presently to the phenomena taking place in this gap.

### Sketch of the process of ignition

In broad lines one might picture the ignition process as follows: the positive voltage on the conductor in the starter sets up in the gap mentioned above an electric field which, when it is strong enough, causes a field emission (also called cold emission) of electrons from the mercury. This current of electrons flows to the wall of the starter, which is at a positive potential, and is conducted along it (probably due to the occurrence of secondary emission) to higher parts of the wall at a higher positive potential until it reaches the auxiliary anode. Along their path the electrons ionize mercury atoms, thus converting the electronic discharge into an arc discharge. The formation of an arc from a glow discharge[3] must be considered out of the question; the pressure prevailing under these conditions (saturated mercury vapour at about room temperature) is too low, the distance between the electrodes is too small and the voltages (less than 10 kV) are too low for the occurrence of a glow discharge.

[1] U.S. patent 682 691.

[2] Trans. A.I.E.E. 52, 693, 1933.

[3] See for example M. J. Druyvesteyn and J. G. Mulder, Philips Techn. Rev. 2, 122, 1937.

This picture of the mechanism may be true in the main if the two following questions can be answered in the affirmative:

1) In an arrangement like the one used but in a vacuum (in order to avoid the complications of the gas discharge), can a field emission cause a current to flow to an auxiliary anode?
2) In the customary capacitive starters, is field emission in mercury possible with the field strengths occurring there?

We shall now describe some experiments [4]) from which it is to be concluded that the answer to the first question is, indeed, in the affirmative. As to the second question, the answer to that seems at first sight to be in the negative, but given a certain mechanism preceding the field emission this difficulty is solved and at the same time the ignition lag, which will be discussed below, is explained.

## Verification of the assumed process of ignition

### *Electron emission in a vacuum*

*Fig. 3* is a sketch of the valve with which we demonstrated the occurrence of field emission.

Fig. 3. Experiment for the demonstration of field emission in vacuum. The invaginated tube *A* of a glass envelope is covered at the end with a layer of graphite, *G* making contact with a helical spring. The valve is exhausted through the pump tube *P*. A cable *B* of stranded copper wire is inserted in the inner tube *A* and connected *via* a galvanometer with the anode *An* and *via* a resistance *R* of 100 megohms with the positive terminal of a source of D.C. voltage. The negative terminal is connected to *C*. At a voltage of several kilovolts a current is measured across *An* which continues to flow when the conductor *B* is taken out of the tube *A*. This current, however, begins to flow only when *B* is in *A*.

The invaginated part *A* of the glass envelope corresponds functionally to the starter wall *A* in fig. 1*b*. The internal conductor consists of a cable of stranded copper wire inserted loosely in the tube *A*. Since in this experiment we wish to avoid the presence of gas, there must be no mercury in the tube. The cathode therefore consists of a layer of graphite deposited on the inner wall and brought into contact with a helical spring. An anode is connected to the internal conductor. The tube is evacuated to a pressure of $10^{-5}$ mm Hg or less. When a D.C. voltage of the order of some thousands of volts is applied between the graphite and the internal conductor (the latter positive), a current can be detected across the anode. This current continues when the conductor is taken out of the inner tube *A*; without the conductor having been placed in the inner tube, however, there can be no current.

This is explained in the following way: when the conductor is inserted in the inner tube and comes to lie immediately opposite the rough graphite surface, then at that surface there is a field strength high enough to cause field emission, the electrons emitted move towards the vacuum side of the glass inner tube, and would give this a negative charge and thus screen off the field if there were no secondary emission to keep the wall surface at a high positive potential. This process may be repeated one or more times until finally the secondary electrons reach the anode.

The fact that the current still continues to flow after the conductor is removed supports the assumption of secondary emission. How, otherwise, could the graphite continue to emit electrons if the wall *A* (fig. 3) were not kept at a high potential. This is only possible where there is secondary emission. But the internal conductor is needed to start the discharge, because otherwise the cathode cannot be given the field strength necessary for the field emission.

Let us now turn to the second question: whether with the usual starters the field strengths are sufficient to cause field emission.

### *The field strength at the cathode*

The experiment just described demonstrates in principle the possibility of the occurrence of field emission, perhaps followed by secondary emission from the starter wall. There are, however, considerable differences between the situation existing in this experiment and that of a starter immersed in mercury. As to the field strength at the cathode surface in the latter case, the experiment of fig. 3 cannot teach us much. In the first place, in the situation given in fig. 3 there was no gap such as occurs where there is a mercury meniscus (fig. 2),

[4]) See for a more detailed description of these experiments (and others in the same field) N. Warmoltz, On the mechanism of dielectric igniter and of the semi-conductive igniter in mercury-vapour rectifiers, thesis, Delft 1946.

and in the second place the graphite layer was rough, whereas the mercury, at least apparently, is smooth. It is known that higher field strengths occur locally on a rough surface than on a smooth one.

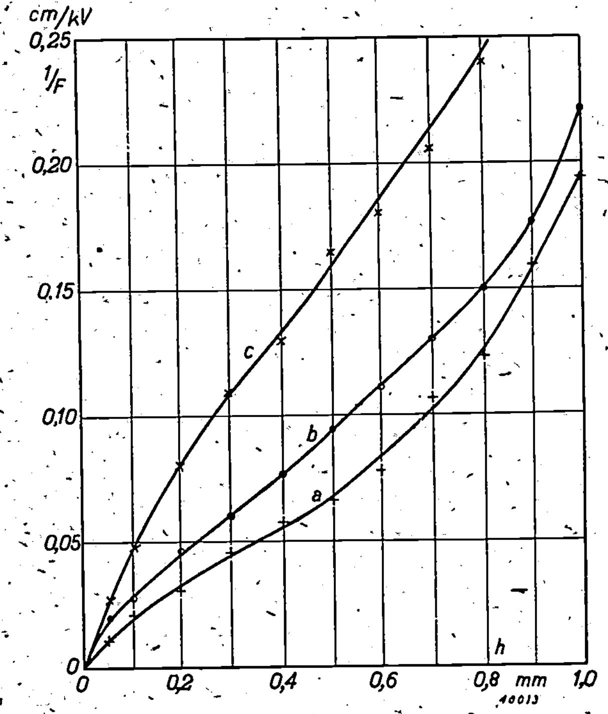


Fig. 4. Field strength $F$ at the mercury surface at a voltage of 1 kV on the igniter. $1/F$ ($F$ in kV/cm) is plotted as a function of the height $h$ above the boundary between the mercury and the starter wall. Curves $a$, $b$ and $c$ are for walls with $\varepsilon = 4.4$ respectively 0.35, 1.0 and 3.0 mm thick. They are derived from measurements described in the article cited in footnote [5]).

The field strength on the supposedly smooth mercury surface can be deduced from the distribution of potential in the space between the meniscus and the starter. Measurements carried out on models in an electrolytic tank have already been described in this journal [5]). From these measurements the field strength $F$ at the mercury surface has been derived as a function of the height $h$ above the boundary line between the mercury and the starter wall, and for a given potential difference (1 kV) between the mercury and the interior of the starter. The result depends also upon the thickness and the dielectric constant $\varepsilon$ of the starter wall. Curves for quartz ($\varepsilon = 4.4$) in thicknesses of 0.35 mm, 1.0 mm and 3.0 mm are given in *fig. 4*, where $1/F$ (for easier extrapolation to small values of $h$) is plotted as a function of $h$.

The next step is to find where the spark occurs. This place (fixed by the height $h$ above the boundary line) is determined directly with the aid of a magnifying glass. We used an ignition band as shown in fig. 1*a*. This band consisted of a layer of silver so thin as to be transparent, so that the height at which the spark occurred could easily be seen. With a given starter this height was found to depend upon the voltage $V$ applied to the starter. *Fig. 5* shows this relation between $h$ and $V$ for a given case (wall of 0.35 mm quartz). The higher the voltage $V$, the deeper the spark occurs in the gap.

The field strength on the mercury at the place where the spark occurs is found from a combination of figs. 4 and 5. From fig. 5 we find for a wall of 0.35 mm quartz and for values of $V$ from 3.00 to 6.55 kV, the values of $h$ given in *table I*, to which correspond, according to fig. 4, the values of the field strength $F$ given in the last line of the table.

Can field emission be expected with these field strengths? Haefer has determined the field strength necessary for field emission in the case of tungsten [6]). His measurements are in agreement with the theoretical deductions. He found, for example, that with a field strength of $3 \times 10^4$ kV/cm there is a current density of 45 amp/cm$^2$, but that with a field strength three times as small it is only $3 \times 10^{-16}$ amp/cm$^2$. The same applies to mercury, where the work function (4.4 V) is practically the same as for

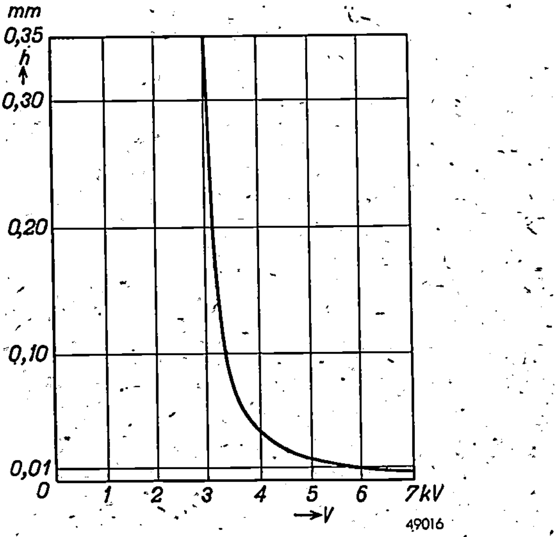


Fig. 5. Measured height $h$ at which sparking takes place on the mercury cathode (in the case of a given igniter) as a function of the voltage $V$ applied between the igniter (+) and the mercury cathode (—).

Table I. Height $h$ at which the spark occurs and field strength $F$ on the mercury for several values of the voltage $V$ on a starter with a wall of 0.35 mm quartz.

| | | | | | |
|---|---|---|---|---|---|
| $V$ (in kV) = | 3.00 | 3.15 | 3.55 | 4.50 | 6.55 |
| $h$ (in mm) = | 0.35 | 0.21 | 0.065 | 0.025 | 0.008 |
| $F$ (in kV/cm) = | 50 | 94.5 | 255 | 750 | 2500 |

[5]) N. Warmoltz, Potential distribution at the igniter of a relay valve with mercury-cathode. Philips Techn. Rev. **8**, 346, 1946.

[6]) Z. Phys. **116**, 604, 1940.

tungsten. The field strength of $10^2$—$10^3$ kV/cm cannot, therefore, cause any appreciable field emission. We shall presently see how this difficulty can be satisfactorily solved.

### The time lag of the ignition

If the ignition does indeed take place according to the process sketched above, one would not, *a priori*, expect any larger time lag (the time elapsing between the discharge and the moment when the ignition voltage is applied) than the time taken by the current of electrons to pass over into an arc (*i.e.* the time for building up the gas discharge). Sparking measurements taken by various investigators in an evacuated tube have shown that this building-up time is very short, of the order of $10^{-7}$ sec. Nevertheless, oscillograms of the voltage on the starter show that much longer time lags occur, as has been demonstrated with the apparatus of *fig. 6*; tbis is explained further in the text underneath the diagram.

As a check, the time lag of a spark was first measured between two tungsten spheres in a vacuum. With the spheres 0.25 mm apart sparking occurred at 15 kV. The time lag amounted to $2 \times 10^{-7}$ sec, in agreement with the results of the investigators just mentioned.

In *figs. 7a-d* some oscillograms are reproduced which were obtained from measurements on the discharge tube sketched in fig. 6; the time scale is fixed by fig. 7*e*, the oscillogram of an A.C. voltage with a frequency of $10^6$ c/sec. As may be seen, the oscillograms of the voltage on the discharge tube are roughly trapezoidal; in a very short time (here about $5 \times 10^{-8}$ sec.) — depending mainly upon the capacity of the starter and the resistance of the circuit — the voltage on the starter rises to a value where it remains constant for a certain time, which time we shall consider as the real time lag *T*, after which it falls in $16^{-6}$—$10^{-7}$ sec to the arc voltage in mercury vapour at low pressure. The higher the voltage *V* on the starter, the shorter the time *T*. In *fig. 8* *T* is plotted as a function of *V* for four different thicknesses of the starter wall.

The oscillograms show an entirely different picture when the starter is wetted by the mercury, *i.e.* when the mercury meniscus at the starter is

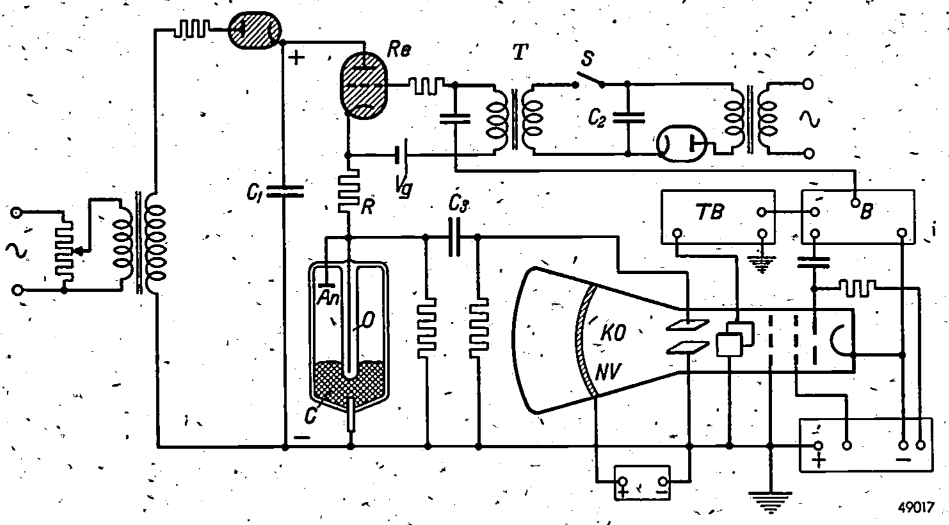


Fig. 6. Circuit for measuring the time lag. The capacitor $C_1$, charged by a rectifier to a variable voltage, can be discharged across the relay valve *Re*, the resistance *R* and the valve under investigation with the anode *An*, igniter *O* and mercury cathode *C*. The discharge is brought about by depressing the key *S*, whereupon the capacitor $C_2$ is discharged through the primary coil of the transformer *T*. The voltage impulse thereby excited in the secondary coil causes the relay valve *Re* to ignite, this valve having been previously blocked by the negative grid bias $V_g$. The variation of the voltage at the starter is observed on the cathode-ray oscillograph *KO*, which is connected with *An via* the capacitive coupling $C_3$. Since the light spot only describes the oscillogram once, in order to obtain a sufficiently bright image a cathode-ray tube with post-acceleration (*NV*) is employed [7]). In a manner not indicated here, the voltage impulse of *T* at the same time sets in action the generator *B* supplying a voltage of a rectangular form. This voltage releases the electron beam and at the same time causes a calibrated time base (*TB*) to function.

[7]) A cathode-ray tube with post-acceleration is described in Philips Techn. Rev. 5, 257, 1940.

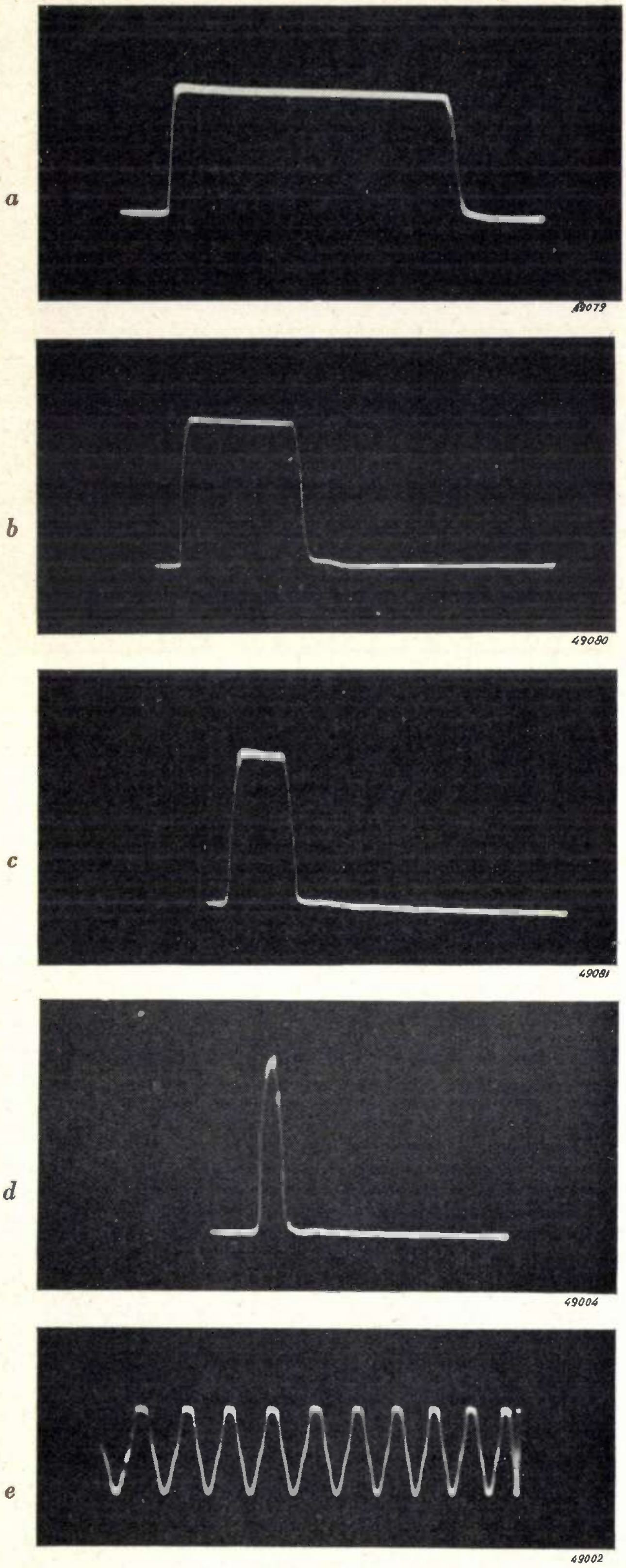


Fig. 7. *a-d*) Oscillograms of the voltage between the starter and a mercury cathode with convex meniscus. The voltage $V$ applied to the starter amounted in ($a$) to 5.5 kV, in ($b$) to 6 kV, in ($c$) to 8 kV and in ($d$) to 10 kV. The length of the horizontal part is the real time lag ($T$). The time scale is given by the oscillogram ($e$) of an A.C. voltage with the frequency $10^6$ c/sec. The scale of the ordinates in the figures $a$ to $d$ is not exactly the same. The oscillograms, like those of fig. 9, were recorded with a cathode-ray tube working with an acceleration voltage of 10 000 V.

not convex but concave, as may happen when there are certain contaminations on the starter wall. Even so, the starter is found to function. *Fig. 9a* shows that even at the low voltage of 1.2 kV there

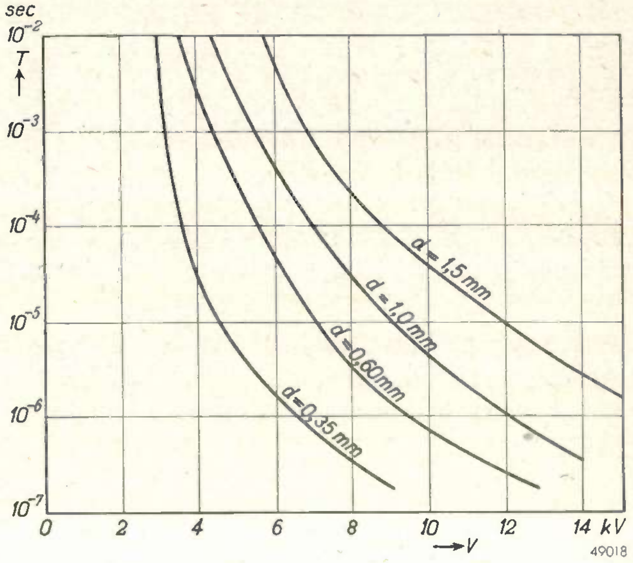


Fig. 8. The oscillographically determined real time lag $T$ as a function of the voltage $V$ applied to the starter; $d$ is the thickness of the wall of the starter.

is no real ignition lag (it is to be noted that the time scale here is different from that in fig. 8; it is given by fig. 9*b*, the oscillogram of an A.C. voltage with the frequency $10^7$ c/sec.)

### The microscopic deformation of the mercury surface

We have interrupted the verification of the ignition mechanism described, with the statement that, assuming the surface of the mercury to be smooth, the field strengths found are much too small for any appreciable field emission of electrons. The ignition lag just discussed furnishes the key to the solution of this difficulty. This time lag may be interpreted as the time required for the force exerted by the electric field to cause the formation of the sharp points on the originally macroscopically smooth mercury surface, at which points the field strengths required for field emission are actually reached.

Tonks has calculated the minimum field strength and the time taken to draw out small irregularities on the mercury surface into points [8]); such irregularities are always present owing to the Brownian motion (thermal irregularities). For this minimum field strength Tonks found a value of 53 kV/cm. With higher field strengths the mercury surface becomes unstable, since upon a slight disturbance

[8]) Phys. Rev. 48, 562, 1935.

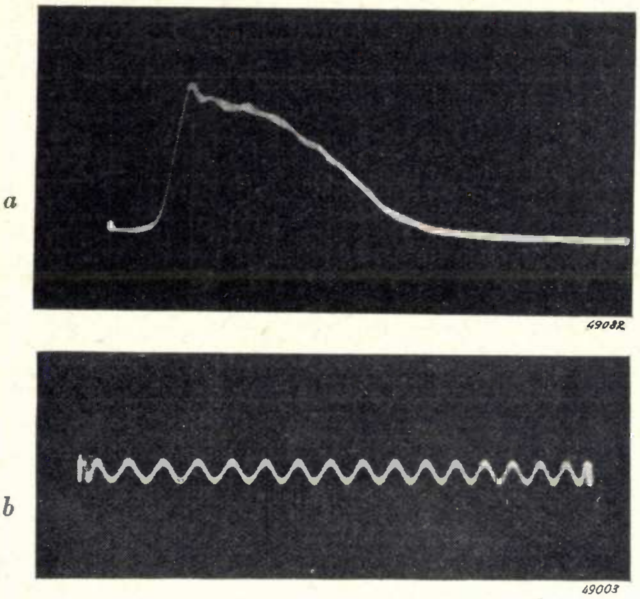


Fig. 9. *a*) Oscillogram of the voltage between starter and mercury cathode with a concave meniscus. There is no time lag here, although the voltage applied is very low, *viz.* 1.2 kV. (N.B. The time base is different from that in fig. 8 *a-d*!)

*b*) This oscillogram of an A.C. voltage with a frequency of $10^7$ c/sec gives the time scale for oscillogram (*a*).

of the equilibrium the drawing force increases more rapidly than the opposing force (capillary force and force of gravity). The time needed depends upon the field strength and the magnitude of the original irregularities. Thus, for example, with a field strength of 2 MV/cm a "hump" $10^{-7}$ cm in height and with a radius of $10^{-3}$ cm would be drawn out to a point in $2.3 \times 10^{-6}$ sec.

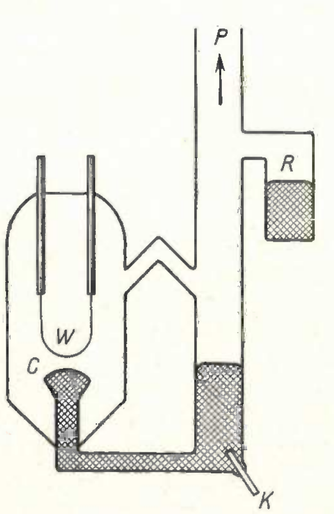


Fig. 10. Experimental tube for measuring the ignition time lag under simplified conditions. The mercury level at *C* is so adjusted (roughly by the addition of mercury from the reservoir *R*, more finely by slightly tipping the tube) as to lie at a distance of 0.25 to 0.125 mm from the tungsten strip *W*. The strip *W* is electrically heated to incandescence before each measurement in order to remove any mercury. *K* current lead, *P* pump tube.

We have investigated these phenomena with a simplified arrangement in which the field strength is directly known and the sparking is brought about in a simpler way due to the absence of the dielectric of a starter. The cathode spot was formed by applying a high positive voltage to an electrode placed at a short distance above the mercury.

In practice this method of ignition is unsuitable, since, in order to obtain the necessary field strength, either very high voltages or a very short distance must be chosen. The latter is quite out of the question owing to the changes in the level of the mercury caused by evaporation or condensation and also the occurrence of waves on the mercury surface caused by mechanical vibrations.

*Fig. 10* is a diagram of the experimental tube used. A fine strip of tungsten is mounted at a short distance (0.25—0.125 mm) above the mercury surface; the whole set-up has to be free of vibrations [9]) to avoid any variation in this distance due to ripples on the surface of the mercury.

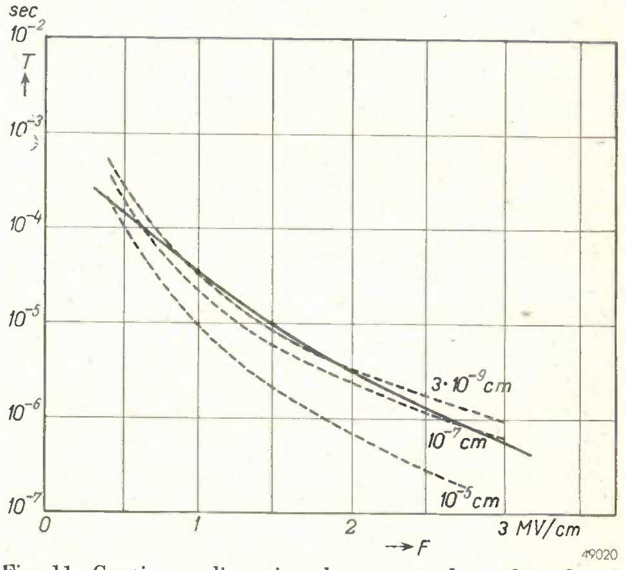


Fig. 11. Continuous line: time lag measured on the tube of fig. 10 as a function of the field strength *F* at the (still smooth) mercury surface.
Dotted lines: times calculated by Tonks for mercury humps to be drawn out into points as a function of *F*, with the height of the humps as parameter. These dotted lines correspond to the continuous line when the height of the humps is about $10^{-8}$ cm.

The time lags measured according to fig. 7 are represented in *fig. 11* by the continuous line as a function of the field strength *F* at the (still smooth) surface of mercury. The dotted lines in the same graph represent the times taken, according to Tonks, for the mercury "humps" to be drawn out to points; the parameter is the height of the humps.

[9]) A very suitable arrangement is for instance that described by J. A. Haringx in Philips Techn. Rev. 9, 16, and 85, 1947 (Nos. 1 and 3).

If we interpret the measured time lags entirely as "drawing-out" times, our measurements agree with Tonks' results when we set the original height of the humps at $10^{-8}$ cm, the height which Gans also arrived at in his calculation of the irregularity of a mercury surface caused by Brownian motion [10] [11].

This gives a satisfactory explanation of the time lags observed with the tube of fig. 10. In the case of tubes with dielectric ignition, with which we are really concerned, the insulating igniter wall does indeed form a complication, but it is not to be expected that this wall will essentially affect the mechanism on the mercury surface, as will be evident from the following.

Measurements have been carried out with a number of starters with wall thickness $d$ varying from 0.35 to 3.0 mm to determine the field strength $F$ at which time lags $T$ of $5 \times 10^{-3}$, $10^{-3}$, $10^{-4}$, $10^{-5}$ and $10^{-6}$ sec. occur. As *fig. 12* shows, the thickness of the wall has very little effect.

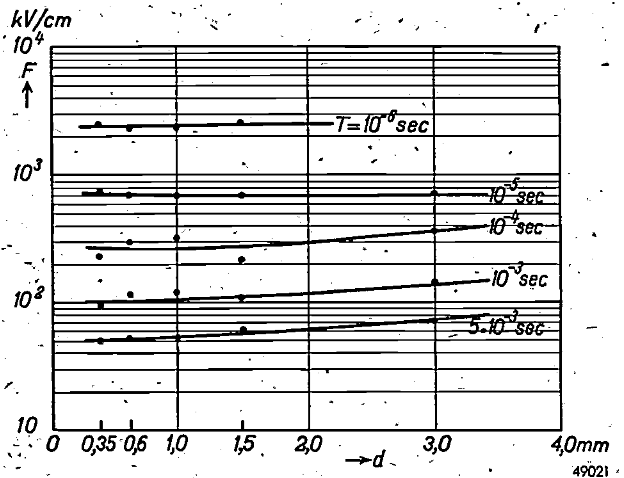


Fig. 12. The field strength $F$ necessary to obtain a given time lag $T$ is found to depend only little on the thickness of wall $d$ of the dielectric igniter.

If for each time lag the average ($F_1$) of the only slightly varying field strength for the five wall thicknesses is taken and compared with the field strengths $F_2$ required for the same time lags $T$ in the tube of fig. 10, the two series of results (see *table II*) are found to be of the same order of magnitude.

From this it may be concluded that the field emission on a mercury surface near the wall of a dielectric igniter occurs in the same manner as in the case of an anode electrode placed above the mercury.

Table II. Field strength $F_1$ in the case of a valve with dielectric igniter compared with the field strength $F_2$ in the valve of fig. 10, both so chosen that there is the same time lag.

| $T$ (sec) = | $5{\cdot}10^{-3}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | $10^{-6}$ |
|---|---|---|---|---|---|
| $F_1$ (kV/cm) = | 58 | 117 | 286 | 725 | 2462 |
| $F_2$ (kV/cm) = | 60 | 125 | 625 | 1500 | 2700 |

We must now return for a moment to the measurements of fig. 9*a*, which referred to a concave mercury meniscus. The fact that no time lag was found there is by no means contradictory to what has been stated above. Since the boundary angle $\alpha$ (fig. 2) is acute, it is reasonable to assume that the edge of the mercury is already so sharp as to bring about a field strength sufficiently high for field emission straightaway. Tonks' mechanism of drawing out to points and the accompanying time lag do not then enter the picture.

Summarizing, the course of the process of ignition sketched at the beginning may be defined as follows.

The positive voltage on the starter causes an electric field which, when the meniscus is convex, is not strong enough to bring about field emission on the mercury but strong enough to make that surface unstable. Irregularities due to thermal causes are drawn out to sharp points, for which a certain time is needed (the time lag). At these points the field strength attains values which lead to electron emission. The electron current flows along the starter wall, probably assisted by secondary emission, to an anode. The current concentrated in the mercury points causes Joule heating, as a result of which some mercury vapour is formed which promotes the transition of the discharge into an arc.

## Application of relay tubes with dielectric igniter

In the beginning of this article the rectifier was mentioned as an object in which controlled ignition offers certain advantages over other methods. There are, however, many other possibilities of application of relay tubes as switches, for (periodic or non-periodic) current impulses of great intensity and short duration. Such current impulses are required in various fields of technology, for instance for stroboscopic illumination, for spot and seam welding, for transmitters of radar installations, etc. A mercury cathode can safely emit currents of thousands of amperes. In many applications the current impulses are of such short duration compared with the intervals between the impulses that the average value of the current over a length of time

[10] Ann. Phys. (Leipzig) 74, 231, 1924.
[11] A more detailed description of these phenomena will appear in Philips Research Reports.

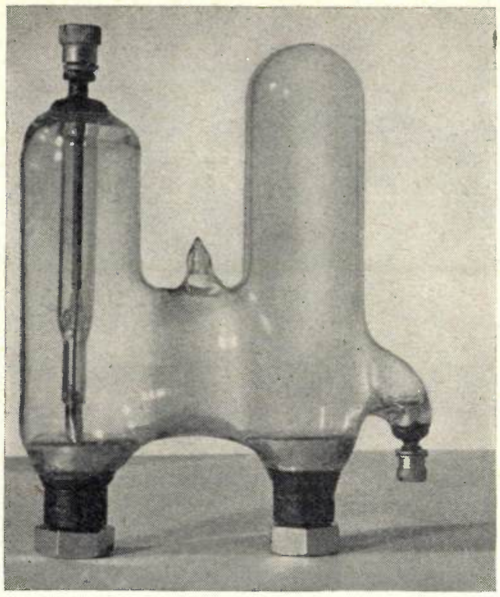


Fig. 13. Relay valve with mercury cathode (left) and dielectric igniter. The anode and the auxiliary anode (right) also consist of mercury in order to avoid contamination of the igniter. The valve, which is about 18 cm in height, can transmit current impulses of 2000 amp. In view of the permissible temperature, the average current is limited to about 1.5 amp. with natural cooling of the valve, and to about 3.5 amp. when a fan is used.

is many times smaller than the impulse current. Since it is the average current, together with the arc voltage, that determines the heat development in the tube, its dimensions can be kept relatively small.

An example of a relay valve for a high peak current but for a much lower average current is shown in *fig. 13.* For this valve, which serves, *inter alia*, as switch in a stroboscope [12]), the peak current is 2000 amp. and the average current only a few amperes. The peculiar shape of the tube is due to the fact that not only the cathode but also the anode and the auxiliary anode consists of mercury. In principle the two latter electrodes could be made of metal or carbon, but then in course of time the surface of the starter would inevitably become contaminated with a film of sputtered anode material, which may lead to irregular ignition. It was therefore made a condition that the inner surface of the tube should consist only of mercury, glass and quartz, so as to ensure that the starter is kept free from contamination.

[12]) Described in Philips Techn. Rev. 8, 25, 1946.

# DETERMINING THE LIGHT DISTRIBUTION AND LUMINOUS FLUX OF PROJECTORS

by J. BERGMANS and H. A. E. KEITZ.

535.245.24:628.932

In every illuminating engineering laboratory instruments are used for registering the distribution of light from light sources emitting rays in all directions. The light source is suspended in a certain fixed position, the luminous intensity (candle power) is measured in several meridian planes, a polar graph of the light distribution is plotted and from that one calculates the total flux (usually with the aid of a Rousseau diagram). In the case of projectors this method is not directly applicable. If it is a small projector emitting a very narrow beam, the light source can be set up in a fixed position and the light distribution determined by means of a number of isolux curves on a projection screen perpendicular to the optical axis of the beam, the flux then being found by integration. In all other cases the photometer has to be set up in a fixed position and the light beam moved in such a way that each part of the beam is successively thrown on the photometer; for this purpose two different types of rotating apparatus are used. From the results of the measurements taken an iso-candle diagram can be obtained. Before the luminous flux can be determined that diagram has to be reproduced on a flat plane, employing the sinusoidal method or Lambert's azimuthal reproduction, both of which are area-proportional. It has advantages when the measurements are so carried out that the successive directions of measurement lie on a conical surface around the optical axis of the light beam. To make this possible Philips Laboratory for Illuminating Engineering have constructed and are using a special rotating apparatus, which is described in this article.

## Introduction

To judge the properties of a light source correctly one should consider not only the total quantity of light emitted but also the manner in which the light is distributed in various directions.

The lamp — the light source proper — is usually mounted in a fitting. In this article we shall use the term light source both for the lamp and for the combination of lamp with its fitting.

Generally one has to do with light sources which emit their light in all directions, or at least within a very wide solid angle. For a long time already appliances have been known and methods have been worked out for determining the light distribution and luminous flux from these light sources.

In this article, however, we are concerned mainly with light sources from which the light flux is emitted in a fairly narrow beam. These are called "projectors". To give an example of such a projector we would recall the fitting with a water-cooled mercury lamp for airfield illumination which has previously been described [1]). That projector produces a flat and fan-shaped beam: vertically the angle of divergence is 2 times approximately 1/2° and horizontally 2 times approx. 40°. Projectors are also widely used for illuminating (floodlighting) façades and suchlike.

We shall explain in this article how in the case of projectors the methods mentioned above need some modification, and we shall describe methods by which it is possible in these cases to register the light distribution and determine the total flux.

## Determination of the light distribution and total flux of light sources with multi-lateral radiation

We shall begin by recalling briefly how the light distribution and total flux are usually determined with light sources that radiate in all directions.

The most obvious way is to suspend the light source in a fixed position, preferably in the position in which it is normally used, and then to measure successively the light distribution in a number of vertical planes passing through the light source. In analogy with the usual geographical terminology, these planes are called meridian planes. With the aid of a mirror movable in a meridian plane, or if necessary by means of more than one mirror, the light from the fixed light source is thrown onto a photometer.

Since the introduction of the photocell this is also done by mounting this instrument on a long rotatable arm and turning it in the meridian planes around the light source.

The lux values found in a certain meridian plane can then be plotted in a polar diagram. The light source is taken as lying in the centre and the length of the radius vector indicates what intensity of *I* has been measured in a certain direction. *Fig. 1* gives a specimen of such a diagram.

[1]) Th. J. J. A. Manders, Philips Techn. R. 6, 33, 1941.

When the light source is rotation-symmetrical every meridian plane will produce the same polar diagram.

The total flux of a light source is the flux (*i.e.* the quantity of light passing through per second) passing through a spherical surface enveloping the light source. The radius $r$ of that sphere must be chosen in such a dimension that from any point of that spherical surface the light source can be regarded as being situated in the centre of the sphere. If $d\Phi$ is the flux within an infinitesimally small solid angle $d\omega$ bounding a minute area $r^2 d\omega$ of the sphere, then $d\Phi = I d\omega$.

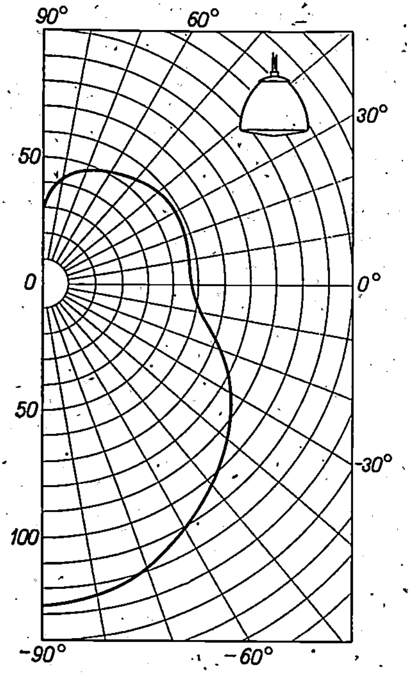


Fig. 1. A polar diagram of the light distribution of a light source in a meridian planc. The radius vector in a certain direction is proportional to the luminous intensity measured in that direction, expressed in candles. The figures given apply for a light source with a flux of 1000 lumens. In agreement with the annotation used in the text, we have taken a horizontal line as the zero direction and set out from that angles up to 90 degrees below and about the zero line. In illuminating engineering one mostly uses a vertical zero line and sets out the angles downward from the top from 0 to 180 degrees. In the top corner is an illustration of the fitting to which this diagram relates.

In the foregoing we have introduced the term meridian planes. We shall now continue to follow the analogy with geographical terms by calling the angular coordinate in a plane perpendicular to the axis, corresponding to the geographical longitude, the longitude $l$ and the coordinate perpendicular thereto, which corresponds to the geographical latitude, the latitude $b$.

The total flux of a light source is the integral of $I d\omega$, for which we thus find:

$$\Phi = \int_{-\pi/2}^{\pi/2} \cos b \, db \int_{0}^{2\pi} I \, dl. \quad (1)$$

In the case of a rotation-symmetrical light source $I$ is independent of $l$, so that (1) becomes:

$$\Phi = 2\pi \int_{-\pi/2}^{\pi/2} I \cos b \, db. \quad (2)$$

This integral is usually calculated by a graphical method, that of the Rousseau diagram. In *fig. 2* it is shown how this diagram is plotted from a polar diagram. A line $pq$ is drawn parallel to the vertical axis of the polar diagram and to the right of it a curve is drawn with $r \sin b$ and $I$ as coordinates. The area contained between this curve, the perpendicular just mentioned and the two drawn horizontal lines $pp_1$, and $qq_1$, is proportional to the luminous flux $\Phi$. This area can be determined with a planimeter.

When the beam of light is not rotation-symmetrical the Rousseau diagram cannot be employed in this form, because each meridian plane then produces a different polar curve. If it is, nevertheless, desired to use this diagram for determining the total luminous flux then one must have available candle power measurements found in conical surfaces described about a certain axis selected for the measurement. The average candle power $I_b$ for such a conical surface has then to be determined and substituted for $I$ in formula (2). We shall revert to this later.

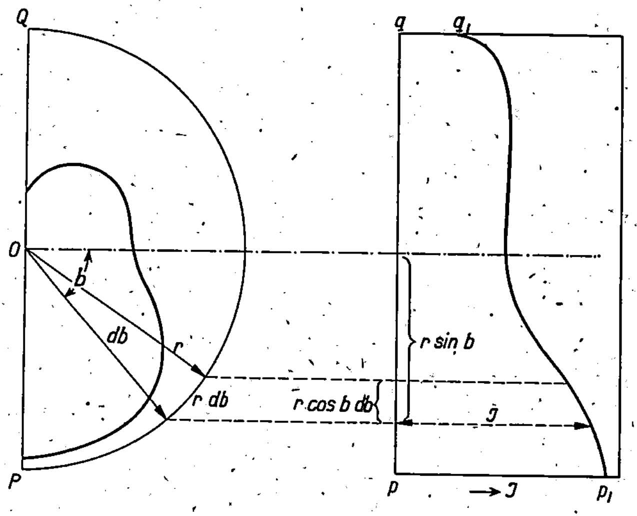


Fig. 2. The calculation of the total luminous flux of a rotation-symmetrical light source with the aid of the Rousseau diagram from the polar graph of the candle power distribution measured in a meridian plane. Here $r$ is the radius of the sphere imagined around the light source, $b$ the latitude in the system of coordinates and $I$ the measured candle power. The area of the figure $pp_1q_1q$ is proportional to the total luminous flux $\Phi$.

Particularly with non-rotation-symmetrical light sources another method is often applied. One then uses an iso-candle diagram. To produce such a diagram one again imagines around the light source a sphere with such a radius that the light source can be regarded as a light-emitting point from any point on the surface of the sphere. One imagines that the candle power of the light source is determined from every point of the sphere and that an iso-candle curve connects the points on the spherical surface from which the same luminous intensity is found. For practical purposes a number of meridian planes can again be measured up with a photometer.

By integration it is possible to derive the total luminous flux from such a diagram by multiplying the area of the strip between two iso-candle curves by the corresponding lux and adding up the result for all strips. But when the planimeter is used to measure the area of a strip one is faced with the difficulty that these iso-candle curves lie on a sphere. It is therefore necessary first to make from that diagram a projection on a flat plane, and this projection has to be "area-proportional". We just mention it here because we shall be reverting to it later on in this article.

We now put the question whether the methods described can be applied to projectors.

### Measuring with a stationary projector and a movable photometer

Is it possible to register the light distribution of a projector by fixing this light source and measuring several meridian planes successively with a photometer? Theoretically this is undoubtedly possible. And a polar diagram could then be drawn for each of these planes. But since as a rule in one or more directions the beam is very narrow, in practice it is difficult to apply this method. Think for a moment of the example previously given of a projector where the angle of divergence in the vertical direction is twice approx. 1/2°. But if in one or more meridian planes no reliable polar curve can be drawn then no Rousseau diagram can be employed either, and neither can an iso-candle diagram be constructed. Consequently the methods outlined above are not generally applicable for projectors.

Since projectors emit their luminous flux only in a narrow beam, another measuring method has had to be thought of, where the light source is also set up in a fixed position. By this method a plane is placed perpendicular to the optical axis of the projector (this plane is usually called the projection screen — see *fig. 3*) and the distribution of illumination in this plane is measured and plotted. As the photocell with which this measuring can be done gives the illumination as a direct result, it is logical in this case to record the light distribution from the projector by means of a number of isolux curves (lines of equal illumination) on the projection screen.

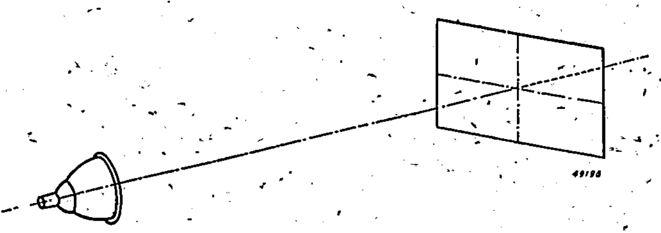


Fig. 3. A projector with a "projection screen" set up perpendicular to the optical axis.

The illumination being equivalent to the luminous flux per unit area, it is easy to determine the flux emitted by the light source. All that is needed is to measure with a planimeter the area between every two isolux lines and multiply by the average illumination of that area.

The method of measuring with a projection screen is applied in practice. It is employed, for instance, with bicycle and motorcar lamps. Sufficiently accurate results can be obtained when the screen is placed at a distance of 6 to 8 metres from the light source.

This method, however, can only be used for small light sources which, moreover, give a beam with a small angle of divergence. The reason is that the luminous body is not a point but has finite dimensions. What is the effect of this upon the beam? A projector consists of a system of reflecting or refracting surfaces, in many cases a rotational paraboloid, as indicated in *fig. 4*. The light source, in this case spherical, is set up in such a way that its centre coincides with the focus of the paraboloid. The part of the light coming from the sphere which strikes the reflector $R$ at the point $P_1$ is reflected in the form of a conical beam. None of the light from this beam strikes the axis within a certain distance $a$ from the reflector. For a point $P_2$ on the circumference of the reflector the distance $a$ reaches a maximum value $g$, the so-called photometrical boundary distance. At points on the axis lying on the projector side of the beam cross-over point the light does not come from the whole reflector but only from its more centrally located parts. When the light distribution of a projector is to be determined the projection screen has to be set up in such a position that the candle power measured there is independent of the distance from the light source. We shall not go farther into the question what the minimum distance between light source and projection screen has to be to achieve

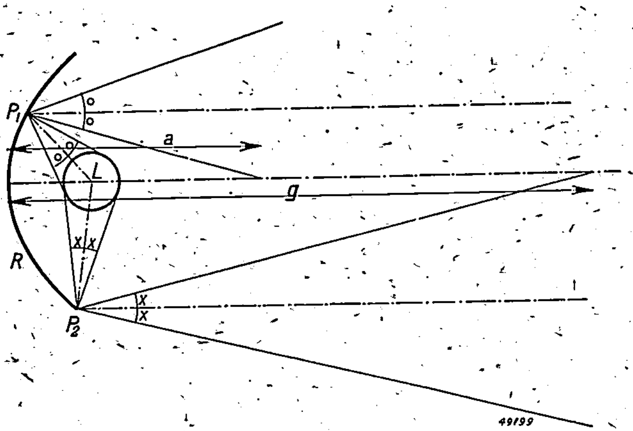


Fig. 4. A spherical light source $L$ with its centre in the focus of the reflecting rotational paraboloid $R$. None of the light in the beam reflected at $P_1$ will reach the axis within a distance $a$ from the reflector. For a point $P_2$ on the edge of the reflector $a$ reaches its maximum value, $g$, called the photometrical boundary distance.

this, but suffice it to say that it depends upon the dimensions of the luminous body and the diameter and focal distance of the reflector, and that in any case it must be located beyond the photometrical boundary distance $g$.

In the case of projectors of a large size (*e.g.* searchlights) the distance of the photometrical boundary distance $g$ to the projector is of the order of several hundreds of metres. It is obvious that even for a beam with a relatively small angle of divergence the measuring method described cannot be applied for such large projectors. The projection screen would be far too large. And this applies, *a fortiori*, when we have to do with a projector having a wide divergence in a certain direction (*e.g.* horizontally).

## Measuring with a stationary photometer and a moving projector

In cases where the method with a projection screen cannot be employed another method is followed. The photometer is set up in a fixed position and the projector is turned about a horizontal and a vertical axis, in such a way that every part of the beam is cast in turn upon the photometer. Here again one must take the photometrical boundary distance into account. Consequently in many cases the light distribution from searchlights, such as are used for instance on airfields, cannot be determined in the laboratory on account of the great distance required for measuring it. This must then be carried out in the open field.

From the readings of the photometer an iso-candle diagram can then be constructed as explained in the beginning of this article. These iso-candle curves lie on a spherical surface. As long as the beam divergence does not exceed 5 degrees in all directions around the axis one may safely regard these curves as being lines in a flat plane. By integration one can then determine the luminous flux from that diagram in the manner described above.

If, however, the beam has a wider divergence the results obtained in this way would be too inaccurate. In such a case, before using the planimeter, one has to make a projection of the iso-candle diagram on a flat plane by one of the methods to be discussed below.

In literature there are two types of rotating apparatus described. It may be so constructed that the horizontal axis of rotation is a shaft contained in bearings in the fixed frame. The apparatus then turns first around this fixed horizontal axis and then also around an axis starting in the vertical position and following the rotation about the horizontal axis (*fig. 5*). It is also possible, however, to have the vertical shaft in fixed bearings and the horizontal axis following the movement around the vertical axis (*fig. 6*); this is the second type of apparatus. Both types of apparatus are used in laboratories for illuminating-engineering. There is no international agreement as to which type is to be taken as standard. This is to be regretted, because when an iso-candle diagram is given in literature one is at a loss to know with what type of apparatus the light distribution was measured, unless it is

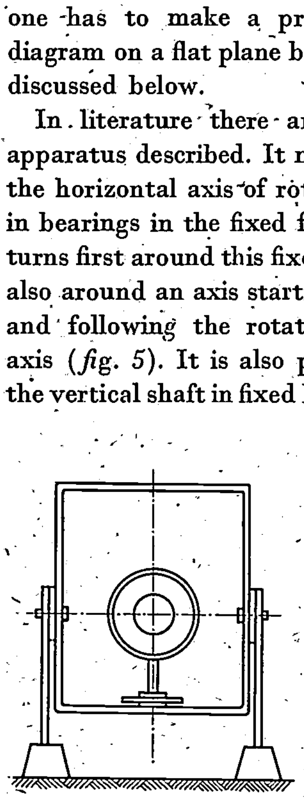


Fig. 5

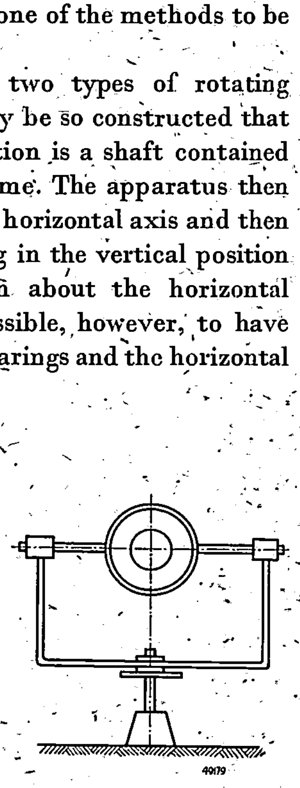


Fig. 6

Fig. 5. A measuring apparatus where the projector turns about a fixed horizontal axis and also about an axis which in the position of rest is vertical and upon rotation follows the movement of the horizontal axis.

Fig. 6. A measuring apparatus where the light source turns about a fixed vertical axis and also about a horizontal axis following the movement of the vertical axis.

specifically quoted, and as a consequence in some cases, as will appear farther on, the results given may be misconstrued.

To ascertain what difference it makes in the shape of an iso-candle diagram whether an apparatus is used of one type or of the other, we must bear in mind that what we want to construct is a diagram showing the candle power of a projector emitted at a certain angle to its axis. In the construction of the diagram we therefore imagine the projector as fixed and plot the angle deviations at which the measurements have been taken. When the actual measurements are being taken, however, it is in fact the photometer that is fixed and the projector is rotated. When, therefore, the projector is turned "to the right upward" the direction of measurement with respect to the axis of the beam has a deviation "to the left downward". And then we must first consider how the diagram just referred to can be constructed in a flat plane.

## Projecting the spherical figures onto a flat plane

The manner in which an iso-candle diagram as determined on a spherical surface can best be projected onto a flat plane is not a new problem in illuminating engineering. It is in essence the same problem with which any geographer is faced when producing a map. Euler has already proved that it is not possible to project spherical figures onto a flat plane so as to give a true reproduction both of the shape and of the area of the figures. There are, however, several methods in use by which a projection is obtained that gives either one or the other feature of the figure in its true dimensions.

### *The sinusoidal projection*

A familiar method in geography is the so-called sinusoidal projection. When we take a point on a sphere (with radius $r$) and indicate that with the coordinates longitude $l$ and latitude $b$, then by this method that point is determined on the flat plane by the flat coordinates:

$$x = rl \cos b, \quad y = rb.$$

This sinusoidal projection is widely used in photometry because it has the advantage of being area-proportional.

We shall use this sinusoidal projection method to compare the iso-candle diagrams obtained from observations and with the apparatus illustrated in figs. 5 and 6.

With the apparatus of fig. 5 we have to do with a system of spherical coordinates, where we may regard the angle of rotation around the horizontal axis as longitude and the rotation around the other axis as a movement in latitude. The poles of this system lie in the horizon in a direction perpendicular to the optical axis of the projector and the equator passes through the zenith. As the apparatus turns about the vertical axis the measuring direction describes meridians, and as it turns about the horizontal axis it describes circles of latitude. In *fig. 7* we see how these meridians and circles of latitude are represented in a sinusoidal projection.

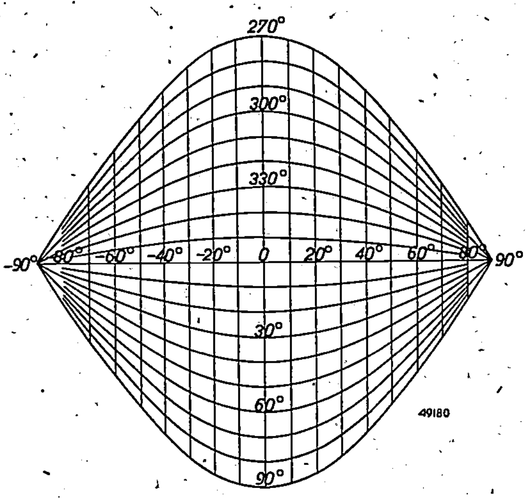


Fig. 7. The sinusoidal projection of the grid belonging to the measuring apparatus of fig. 5.

With the apparatus of fig. 6 we have to regard the rotation around the fixed vertical axis as a movement in longitude and the angle of rotation around the other axis as latitude. The poles coincide with the zenith and the nadir, and the equator coincides with the horizon. As the apparatus turns around the horizontal axis the measuring direction describes meridians and as it turns around the vertical axis

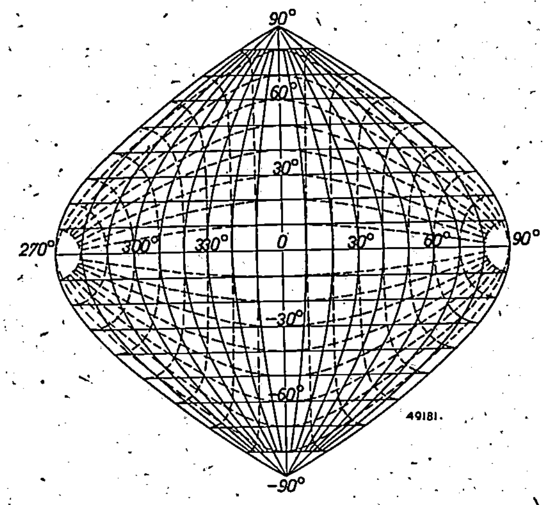


Fig. 8. For comparison some meridians and small circles of fig. 7 are shown as dotted lines on the sinusoidal projection of the grid belonging to the measuring apparatus of fig. 6.

it describes circles of latitude. In *fig. 8* we have the grids corresponding to the movements in this system projected sinusoidally, whilst for comparison some of the meridians and circles of latitude from fig. 7 are interposed in dotted lines.

When the angle of divergence of the beam is small, it will not make much difference as regards the calculated luminous flux whether or not we know which rotating apparatus has been employed in determining the light distribution. But when we have to do with a divergence angle of say 20 degrees it would be an intolerable mistake to interpret measurements taken with an apparatus of fig. 5 as having been taken with such an apparatus as illustrated in fig. 6.

To give an idea of what this mistake would mean we have reproduced in *fig. 9* on a magnified scale the middle part of the diagram of fig. 8. It is seen that the bearing, for instance, indicated with the apparatus of fig. 6 as lying 30 degrees north and 30 degrees west (point $A$ in fig. 9) makes an angle of approximately 5 degrees with the position obtained from the same bearings indicated with the apparatus of fig. 5 (point $B$ in fig. 9). Obviously, then, when measuring beams with a wide divergence it is essential always to state with what type of apparatus the measurements have been taken.

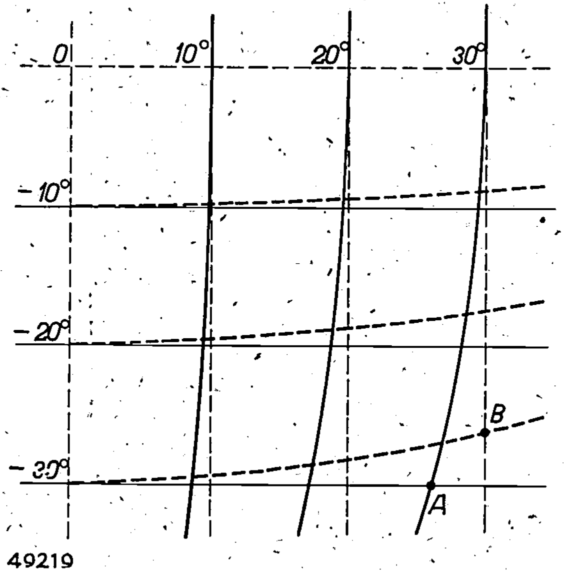


Fig. 9. The middle part of the diagram illustrated in fig. 8 on a magnified scale. The bearing indicated by 30° north and 30° west (point $A$) by means of the apparatus according to fig. 6 makes an angle of about 5 degrees with the same bearings obtained with the apparatus of fig. 5 (point $B$).

*Fig. 10* is a sinusoidal projection of an iso-candle diagram of a light source emitting its light in a rather narrow beam. As already indicated, the total luminous flux from this light source can be calculated by measuring with a planimeter the area between two iso-candle curves at a time and multiplying it with the estimated average candle power of that area.

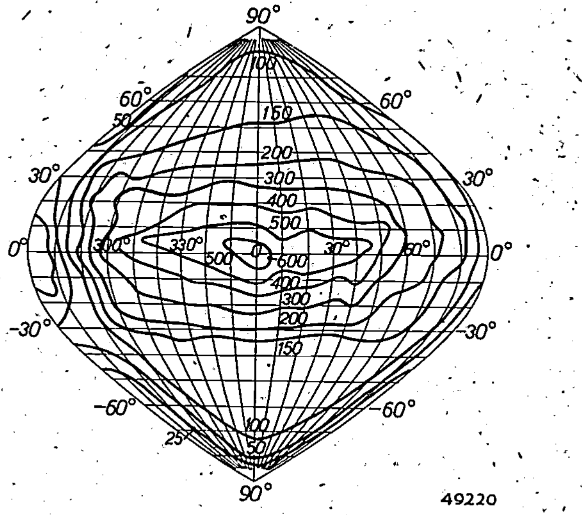


Fig. 10. The iso-candle diagram of a "Cornalux" lamp in sinusoidal projection. The luminous intensity is given in candles, whilst the total luminous flux of this light source amounts to 1320 lumens. From the diagram it may be read that this lamp, which is used for ceiling lighting, emits its light in a rather narrow beam.

*The azimuthal projection*

If it is desired to represent in a flat plane a figure lying on a sphere one does not necessarily have to use the method of sinusoidal projection. Another method, which is not only employed in geographical work but can also be applied in photometry, is the azimuthal projection according to Lambert. This, too, is area-proportional.

With this method of projection the area of a hemisphere is projected with a circle. The meridians

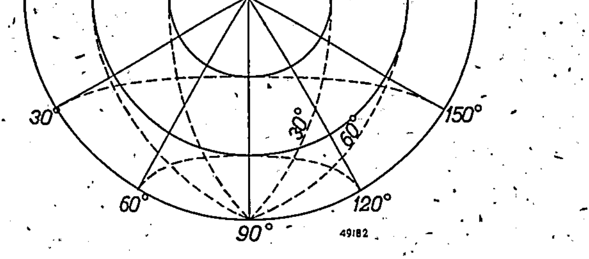


Fig. 11. The azimuthal projection of a spherical grid according to Lambert. The full lines represent some meridians and circles of latitude of a system of coordinates whose polar axis is perpendicular to the drawing plane. The dotted lines are some meridians and circles of latitude of the system of coordinates corresponding to the rotating apparatus of fig. 6.

are drawn as one sees them when looking at a sphere in the direction of the polar axis, thus as diameters of the edge circle or the equator. The lines of latitude are represented by circles concentrical with the edge circle.

In discussing the systems of spherical coordinates as applying to the apparatus of fig. 5 the polar axis is a horizontal line perpendicular to the optical axis of the projector and in the case of fig. 6 a vertical line likewise perpendicular to that optical axis.

There is, however, a third system of coordinates that can be used, where the polar axis is coincident with the optical axis of the projector in its zero position. We shall now first consider this system of coordinates. It is illustrated in *fig. 11*. There we see some meridians drawn (diameters of the outer circle) and some circles of latitude. When we represent in this system of spherical coordinates the longitude as $l'$ and the latitude as $b'$, and the radius of the spherc as $r$, we find for the coordinates of the plane figure:

$$x = 2\,r \sin\,(45°—1/2b')\cos l'$$

and

$$y = 2\,r \sin\,(45°—1/2b')\sin l'.$$

One can draw in the same figure also the spherical grids appertaining to the two other systems of coördinates mentioned above. In fig. 11 we have indicated in dotted lines some meridians and circles of latitude of the spherical system belonging to the apparatus of fig. 6. The poles of this system, as explained before, coincide with the zenith and the nadir. When we represent the longitude and latitude of this system as $l$ and $b$, as before, then for the coordinates of the plane figure we find:

$$x = r\sqrt{2}\,\frac{\sin l \cos b}{\sqrt{1 + \cos l \cos b}}$$

$$\text{and } y = r\sqrt{2}\,\frac{\sin b}{\sqrt{1 + \cos l \cos b}}.$$

In the Lambert projection we may now — just as by the sinusoidal method — indicate the measured candle power in the directions of the points of intersection of the meridians and circles of latitude every 5 or 10 degrees and from that derive the iso-candle curves by interpolation. To arrive at the luminous flux the area of each segment between two iso-candle curves is again measured with the planimeter and the result multiplied by the corresponding average candle power. For this we need a new interpolation. The result, therefore, is not always sufficiently accurate, especially where the light distribution is most irregular.

In *fig. 12* we see the light distribution of the Philips fitting ZE 30 [2]) as used for road-lighting represented in *a*) a sinusoidal projection and *b*) an azimuthal projection.

A comparison of these two shows that the latter has some advantages. In the Lambert projection there is little distortion of the figures; a small circle on the sphere is projected as a circle round about the centre and as an ellipse close to the edge, with the axes in a ratio of 2 : 1. This is due to the fact that the scale on which one degree in a radial

[2]) See Philips Techn. R. 5, 231, 1940.

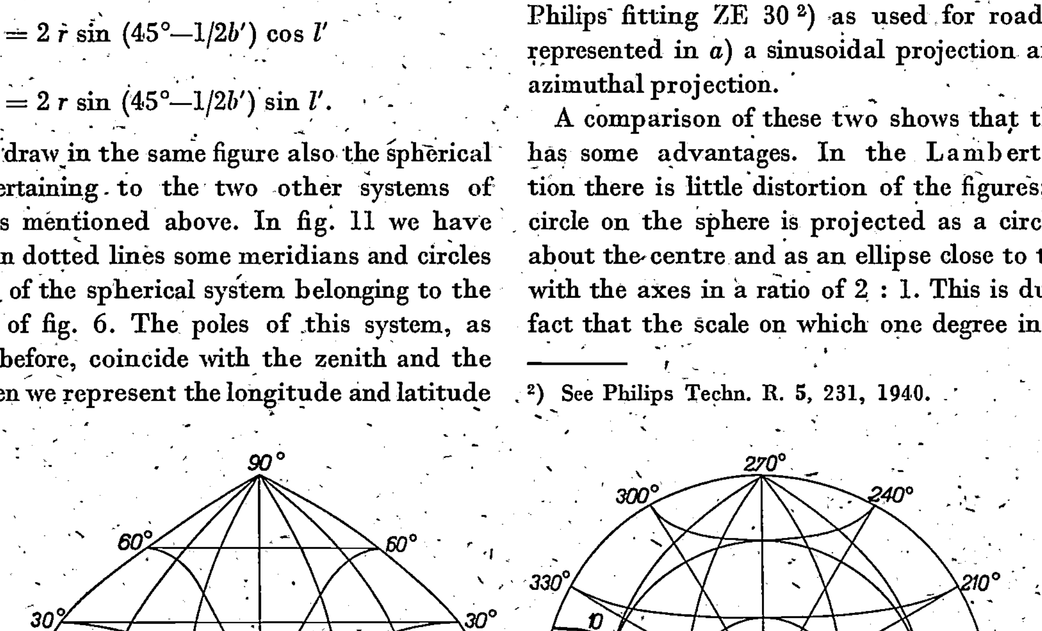


Fig. 12. The iso-candle diagram of a light source for road lighting (fitting Philips ZE 30) in *a*) sinusoidal projection and *b*) azimuthal projection according to Lambert. The other hemisphere is symmetrical with the one represented here. The luminous intensity is given in candles for a light source with total luminous flux of 1000 lumens.

direction has to be plotted differs but very little for the various parts of the diagram. Within the circle of a radius of 30 degrees about the axis (thus between $b' = 60°$ and $b' = 90°$) this scale is even practically constant. It is likewise an important advantage that this greatly facilitates interpolation in the diagram. It is true that in the Lambert projection the meridians and circles of latitude of the network belonging to the apparatus of fig. 6 do not as a rule intersect exactly perpendicular to each other, but the deviations from 90° are not so great as is the case with the sinusoidal projection. Furthermore, with the latter projection one cannot see whether a beam is rotation-symmetrical, as is in fact possible with the Lambert projection.

Upon studying the shape of projection of the network in fig. 11 we notice that all bearings lying on one conical surface, having as its axis the optical axis of the projector, are represented in the diagrams as points on one circle (concentrical with the edge circle). This leads us to put the question whether it would not be desirable to arrange the measuring in such a way that the successive bearings lie on a conical surface around the optical axis. This requires a rotating apparatus making it possible to obtain these results quickly and concisely. Such an apparatus we shall now discuss.

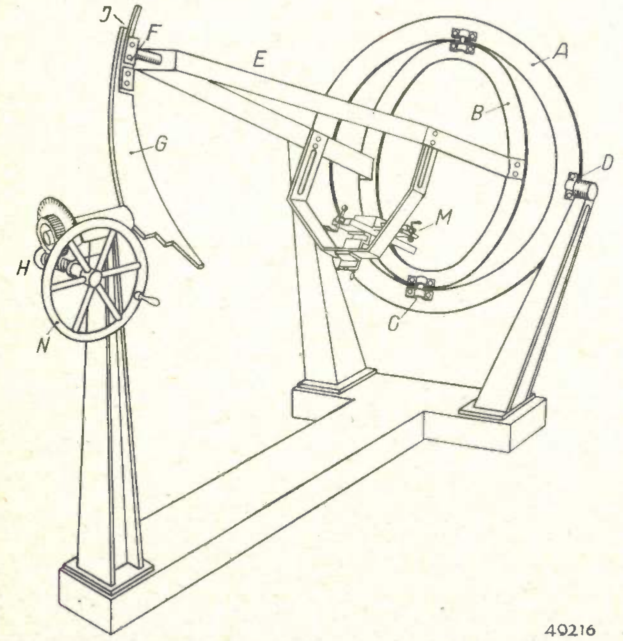


Fig. 13. Diagrammatic representation of the rotating apparatus described for projectors. The projector is suspended on the principle of a cardan joint by fixing it inside the inner one of two rings $A$ and $B$ rotatable about the gudgeons $C$ and $D$. Attached to the inner ring is a fork $E$. By means of a sliding joint $F$ this fork can be moved and fixed on a graduated segment $G$ that can be turned about its other end by means of a worm gear $H$. The sliding joint and the fork are connected by a ball fitting exactly in a cylindrical hole in the sliding joint. By means of a *yernier* $J$ adjustments can be made accurately to within one-eighth degree. Further the fork $E$ is fitted with a bracket to which an adjusting device $M$ is attached, consisting of turning-lathe supports and making it possible to adjust the projector accurately.

### The Philips rotating apparatus for measuring projectors

In Philips laboratory for illuminating engineering an apparatus has been constructed which makes it possible to carry out measurements in a simple way in conical planes around the axis. This will be described with the aid of the illustration in *fig. 13*; two photographs of the apparatus are reproduced in *figs. 14* and *15*.

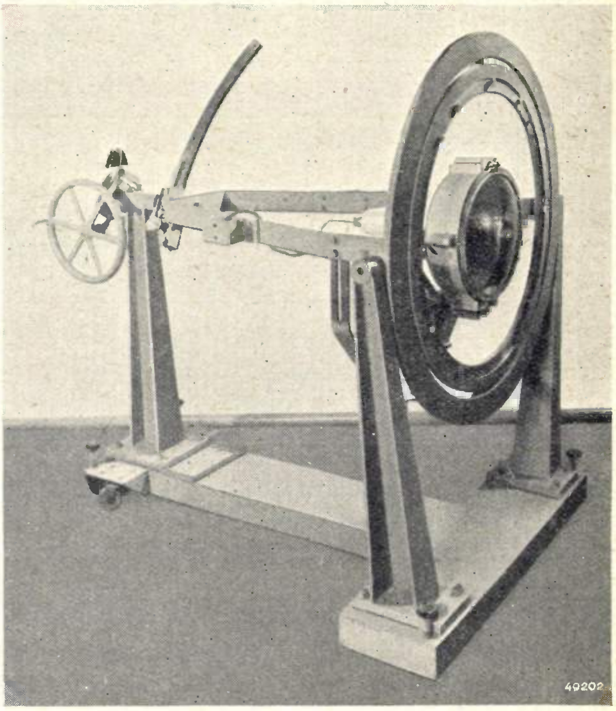


Fig. 14. The Philips measuring apparatus, with a "Philiflood" floodlight mounted in it.

The projector is suspended by an arrangement on the principles of a cardan joint, by fixing it inside the inner one of two rings rotatable about two pairs of gudgeons. Attached to the inner ring is a fork $E$. By means of a sliding joint $F$ this fork can be moved and fixed on a strong graduated arc made rotatable at the other end by means of a worm gear. The sliding joint $F$ is connected with the fork by a ball fitting exactly into a cylindrical hole in the sliding joint. A vernier allows of adjustments accurate to one-eighth degree. Further, on the fork $E$ is a bracket to which an adjusting device is attached consisting of turning-lathe supports and making it possible to adjust the position of the projector accurately.

By adjusting the sliding joint $F$ to a certain number of degrees on the graduated arc and turn-

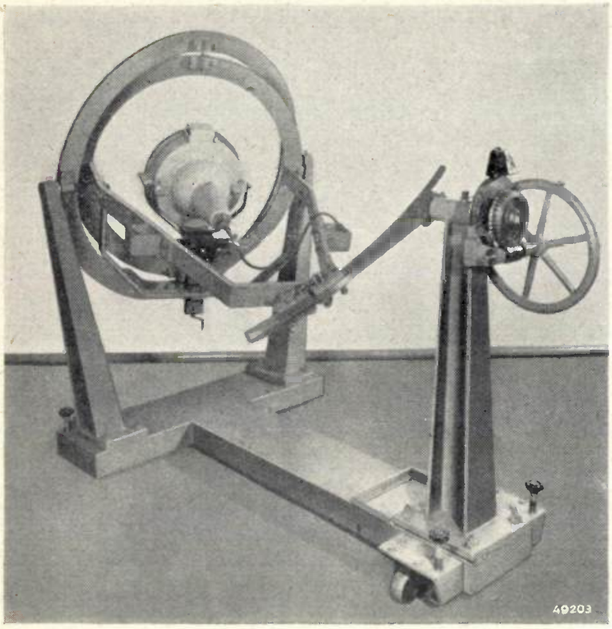


Fig. 15. The rotating apparatus when registering the light distribution of the floodlight in a different position.

ing this arc by means of a handwheel and the worm gear, a large number of measurements can be taken which all give the candle power values in directions making the same angle with the axis of the light source.

This rotating apparatus offers advantage first and foremost when applying the azimuthal method of projection. We have already remarked that in this form of projection all bearings lying on the same conical plane are represented in the diagram as points on a circle. Further, all bearings measured while the segment $G$ is in the same position will be represented in this diagram (fig. 11) by points lying on the same radius from the origin ($O$).

In the second place it is advisable to use this instrument when the luminous flux of projectors is to be determined with the aid of the Rousseau diagram. The arithmetical mean then has to be determined of all measurements relating to the same angle of deviation from the axis. In this way it is possible to find accurately the average candle power from the projector in a given conical plane. This is then plotted in the Rousseau diagram in the manner explained in the beginning of this article.

The rotating apparatus described above therefore makes it possible, with the help of the Rousseau diagram, to determine the luminous flux of a projector more quickly and more accurately than would be possible with one of the apparatuses of fig. 5 and fig. 6 when employing the sinusoidal or the azimuthal method of projection.

# 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.

**1712:** E. J. W. Verwey (+ J. E. Asscher): Lattice structure of the free surface of alkali-halide crystals (Receuil Trav. Chim. Pays-Bas **65**, 521-528, 1946).

Calculations have been made of the lattice distortion at the free crystal surface of the cube face of alkali halide crystals on the basis of Born's lattice theory, with the result that the displacement of the positive surface ions in a direction perpendicular to the surface differs considerably from that of the negative. A kind of ionic double layer is formed with the negative ion pointing away from the crystal. The corresponding dipole is more or less compensated by the dipole moment induced in the negative ions. This distortion lowers the surface energy as is illustrated by the calculation for the case of NaBr.

**1713:** J. Th. G. Overbeek and P. W. O. Wijga: On electro-osmosis and streaming potentials in diaphragms (Rec. Trav. Chim. Pays-Bas **65**, 556-563, 1946).

It is stated that the streaming potential E/P and the volume of liquid transported by electro-osmosis v/i are equal, independent of the structure of the diaphragm and independent of its surface conductance. The value of these ratios which, in tbe absence of surface conductance, equal $\varepsilon\zeta/4\pi\eta\lambda$ ($\varepsilon$ = dielectric constant, $\eta$ = viscosity, $\lambda$ = electric conductivity) diminishes when surface conductance is present. This decrease can be accurately taken into account when the diaphragm consists of a single capillary of constant diameter. In the case of real diaphragms, consisting of a network of capillaries of different shape and dimensions the correction factor for surface conductance cannot be computed! It has been the common practice to estimate the correction factor from the ratio of the observed to the calculated electric resistance of the diaphragm. This procedure is shown to be erroneous, leading to values of the ö-potential that are essentially too low.

**1714:** J. M. Stevels: The physical properties of glass in relation to its structure (J. Soc. Glass Technology **30**, 31-53, 1946).

A fairly extensive account is given of the knowledge we have at present of the structure of glass in general. On this basis some physical properties of glass *viz.* the density and the electric conductivity, are discussed in detail. The density of „normal" glasses can be calculated satisfactorily by means of a formula containing only two constants. One of these constants gives valuable information about the structure of the glass, especially about the way in which the „excess" of oxygen is taken up by the network. The electric conductivity-temperature relationship is briefiy discussed.

**1715:** J. A. Haringx: The Notch-Impact Test according to Schnadt („De Ingenieur" **58**, Mk 15-17, 1946).

A new type of notch-impact bar has been introduced by Schnadt. With this bar the bending pressure does not act upon the material itself but on a hardened steel pin. Since the centre line of this pin functions more or less as the axis of rotation and the material itself is thus entirely subjected to a tensile load, this is really a sort of impact-tensile test. The greatest advantage lies in the fact that every test bar breaks, even if no other notch is made in it. This makes it possible, with the same cross section of fracture ($3\times10$ mm$^2$), to make several notches with greatly different radius of curvature. Consequently a series of widely divergent stress conditions can be created.

**1716*:** H. A. Klasens: The light output of zinc sulphide on irradiation with alpha rays (Trans. Faraday Soc. **42**, 666-668, 1946).

A critical review of the literature dealing with the light output of ZnS excited with alpha rays shows that, in most measurements values are found of 10—15% and of the same order of magnitude as those found with cathode rays, in accordance with the theories of Thomson, Bethe and others.

The author refutes the much higher value (80%) given by Riehl without giving accurate data about his measure monte.

**1717:** J. M. Stevels: The physical properties of glasses.

III. The density of borate glasses

IV. The density of phosphate and germanate glasses.

(J. Soc. Glass Techn. **30**, 173-197, 1946).

The theory given is an earlier communication *), on the density of silicate glasses (as a fuction of their composition) is extended to borate, borosilicate and boro-aluminate glasses. It is shown that in these three cases there is an "accumulation region" (where the O ions form tetrahedral instead of triangular configurations round the $B^{3+}$ ions) and a „destruction region" (where the $O^{--}$ ions are taken up by the breaking of linkages between the network formers). For the boro-aluminates a „crystalline region" (where at least two phases are formed, of which at least one is crystalline) occurs as well. Methods are given for calculating the density. The values so derived are in satisfactory agreement with those obtained by experiment. The theory given is found to hold also for phsophate and germanate glasses.

*) J. M. Stevels, Rec. trav. chim. Pays Bas **60**, 85, 1941; **62**, 19, 1942.

Number 2 of Volume 2 (April, 1947) of *Philips Research Reports* contains the following papers:

*R 37:* J. L. Meyering and M. J. Druyvesteyn: Hardening of metals by internal oxidation; part I.

*R 38:* H. C. Hamaker: Radiation and heat conduction in light-scattering material, II General equations including heat conduction.

*R 39:* H. C. Hamaker: Radiation and heat conduction in light-scattering material, III Application of the theory.

*R 40:* A. van Weel: An improved method of coupling valves at ultrashort waves.

*R 41:* F. L. H. M. Stumpers: Interference problems in frequency modulation.

Readers interested in any of the above mentioned articles may 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 please write to the publishers of Philips Technical Review.

VOL. 9. No. 5, pp. 125-160 1947

# 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 CONSTRUCTION OF THE PHILIPS AIR ENGINE

by F. L. van WEENEN.

621.412

A number of possible designs are discussed by which the hot-air process can be realised. The fundamentally simplest design of a single-cycle engine has two loaded pistons, while the hot and the cold spaces can be either both in one cylinder or in two separate cylinders (V-construction); the merits of these two types are compared. The old air engines had in general only one loaded piston and a transfer piston. The principle of that design is explained and two possible models are discussed; in one the cold space is divided between two cylinders, while in the other there is only one cylinder. It is on the latter principle that the Philips single-cycle air engine has been developed, with all the previously discussed improvements incorporated in it. A description is given of a number of structural details of this engine, which in efficiency and other properties is so much better than the old air engines as to be quite comparable with modern internal combustion engines, while in some respects it possesses a number of fundamental advantages. For higher powers the Philips multi-cylinder air engine has been developed, in which a number of hot-air cycles take place simultaneously. By a suitable combination of the systems necessary for the cycles, only one piston per system is needed (thus no separate transfer piston). This not only simplifies construction but also has the advantage that the direction of rotation can easily be reversed while the engine is running. Moreover, the pistons are double-acting, so that the driving mechanism is very lightly loaded. As a practical example a four-cylinder engine with a wobble-plate mechanism is illustrated, which is found to be very suitable for not too high powers (at most 20—30 h.p.) For higher ratings other types have been developed with a crankshaft mechanism. The mechanical solutions found for these different engines will be discussed in subsequent articles.

The theoretical principles of the air engine were discussed in the May 1946 number of this Review [1]). In a second article [2]), which we shall refer to as "II", further communications have recently been made about the knowledge gained from research carried out by Philips on the hot-air process. From this new insight into the process several new principles of construction have been derived which have made it possible to develop the Philips air engine with all its specific properties. We shall now describe various models of the engine in more detail. First we will deal with single-cycle engines, which may be compared with a single-cylinder internal combustion engine. In order to develop higher powers and to obtain balanced constructions, several single-cycle systems acting on a common shaft can be combined to form a multiple engine, as is the case with multi-cylinder internal combustion engines. (For certain reasons the term "multi-cylinder engine" will not yet be used for the multiple air engine.) In the development of the multiple air engine it has been possible to bring about a radical simplification and improvement allowing of very elegant designs.

### Single-cycle engines

#### *The principle*

*Fig. 1* is the single diagram that was also used in the previous articles to illustrate the practical realization of the hot-air process. The "hot space" $V_w$ and the "cold space" $V_k$ may be seen in open connection with each other *via* heater, regenerator and cooler. In the text underneath the diagram the four stages of the process are described: expansion of the working medium, cooling, compression and heating. The working medium is usually air, but it

[1]) H. Rinia and F. K. du Pré, Air Engines, Philips Techn. Rev. 8, 129-136, 1946.
[2]) H. de Brey, H. Rinia and F. L. van Weenen, Fundamentals for the development of the Philips Air Engine, Philips Techn. Rev. 9, 97-109, 1947.

may also be hydrogen, helium, argon or some other gas. For a proper understanding of what follows it should be recalled that with sinusoidally moving pistons the four stages are not strictly separated but partially overlap, but that the machine continues to function as an engine with hardly any change in power and efficiency, provided the motions of the two pistons differ in phase by about 90°, in such a way that the volume variations in the hot space are in advance of those in the cold space. Furthermore, it must be pointed out that thanks to the open connections the momentary pressure is in principle the same everywhere in the engine: as the working medium is transferred from the cold to the hot space and *vice versa* so the pressure in the engine rises or falls as a whole.

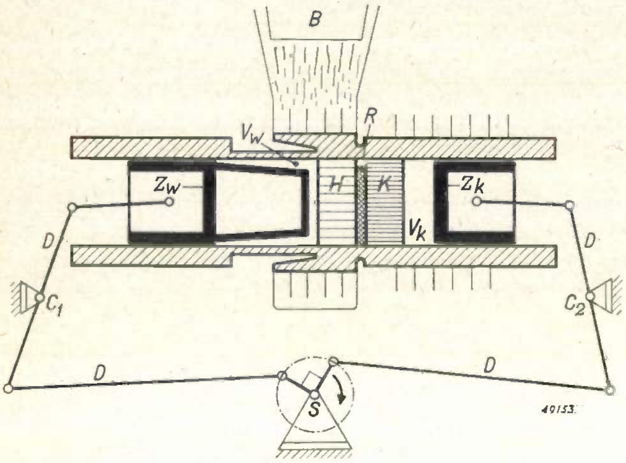


Fig. 1. Diagram of an air engine. The hot working medium (air for instance) in the hot space $V_w$ expands, thereby performing work on the hot piston $Z_w$. It is then transferred, *via* the regenerator $R$ where its heat is stored, to the cold space $V_k$ and there compressed by the cold piston $Z_k$ to its original volume. This requires work, but owing to the lower temperature and thus lower pressure the work thus expended is less than what is gained in the expansion, where the air is at a high temperature and thus under a high pressure. The air is then returned to the hot space *via* the regenerator, where it again takes up the heat stored there, and the cycle begins anew. The cooler $K$ and the heater $H$, the latter heated by the burner $B$, ensure that the desired difference in temperature between the hot and cold spaces is maintained. The driving mechanism (connecting rods $D$, fixed pivots $C_1$, $C_2$, crankshaft $S$) is so constructed that the volumes $V_w$ and $V_k$ vary with about 90° phase difference.

*Single-cycle model*

It is possible to build an air engine exactly according to the scheme of fig. 1. It is even tempting to try it in this way, because theoretically, as far as the process in the cylinder itself is concerned, this is the most favourable form that can be imagined. In II it has been explained that in the construction of the heater, regenerator and cooler a compromise has to be made between the heat transmission on the one hand and the flow resistance and dead volume on the other, and that the right choice of this compromis3 is essential for high efficiency. In fig. 1 the situation is very favourable for the air flow: a short, straight connection with a large diameter between $V_w$ and $V_k$; it is also possible to limit the dead volume to a minimum. For the purpose of our research, which in the beginning was directed mainly upon the points mentioned, this actually led us to build an engine to the design of fig. 1. A photograph of this experimental model is given in *fig. 2*.

The drawbacks depriving this design of any great practical importance are of a mechanical nature. The transmission is rather complicated. The housing of the driving mechanism, in a crankcase, as is customary in internal combustion engines for many obvious reasons (*i.a.* for good lubrication), involves elaborate and expensive constructions. Furthermore, the construction with the two pistons moving along the same line but with a certain phase difference cannot be balanced in a simple way — an objection which arises in the case of every normal single-cylinder engine and which weighs all the more in the designing of an engine with a high speed and or a high power. Finally, for the mechanical efficiency it is a great disadvantage that the construction outlined here works

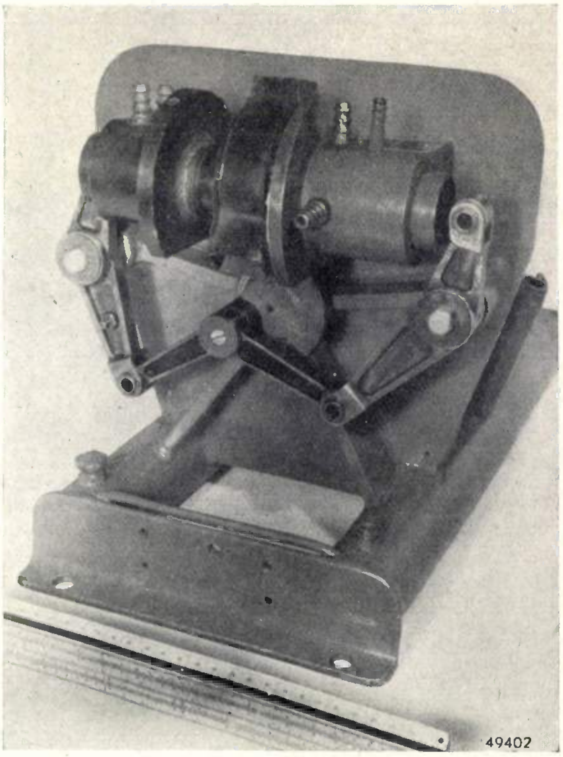


Fig. 2. Experimental model of an air engine constructed according to the principles of fig. 1. The position of the parts is the same as in fig. 1.

with two loaded pistons: there are different pressures on either side of each piston. In order to minimise leakage along the pistons, they have to fit tightly and this results in frictional losses. Owing to these losses it was hardly possible to apply the construction according to fig. 1 for the old air engines with their already very low efficiency. In nearly all the old models an entirely different design was used, with only one loaded piston; we shall return to this later. For our test model the situation was somewhat more favourable: various new ideas had already been incorporated in it, some of which have been discussed in II, and its efficiency was thereby so much improved that the engine did indeed work. But it offered little perspective, partly because of the objections already mentioned, but especially because of the two loaded pistons. The object of developing higher powers with a reasonably high efficiency and high specific power made it desirable to increase the pressure in the engine; this, however, involves still higher requirements as to the close fitting of the two pistons and thus still higher frictional losses result.

*The V-construction*

A number of the mechanical drawbacks mentioned in the foregoing are avoided by modifying the design of fig. 1 to that of *fig. 3*. The two halves of the cylinder in fig. 1 containing the hot and cold spaces have been turned into separate cylinders with their axes placed at an angle of 90° to each other. In this "V-construction" the driving mechanism is much simpler, a closed crankcase is possible and it is found that the engine can easily be almost completely balanced. If there is only one piston, by its motion to and fro transmitting to the driving mechanism an alternating force of acceleration parallel to the cylinder axis, that force can be wholly compensated with the aid of a counterweight rotating around the crank shaft [3]). This, however, introduces a new force, of equal magnitude and alternating at the same frequency, perpendicular to the original force; the centrifugal force of a rotating counterweight can be resolved into two perpendicular alternating forces with a mutual phase difference of 90°. This "compensation", therefore, really brings us little further. If, however, there are two identical pistons moving in a direction perpendicular to each other and with a phase difference of 90° — which is the case in our V-construction — then conversely the two alternating forces of inertia of the two pistons can be compounded to a single rotating force of constant magnitude. This resulting force on the crankshaft can therefore be wholly compensated by a counterweight acting in the opposite direction.

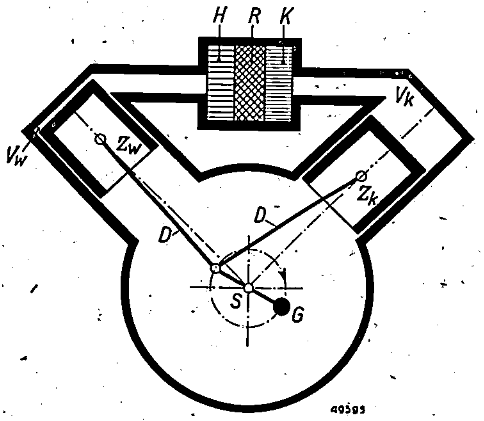


Fig. 3. Diagram of an air engine according to the so-called V-construction. The letters have the same meaning as in fig. 1. G = counterweight. Advantages: simpler drive, closed crankcase (possibly with elevated pressure), balanced inertia forces of the pistons.

The disadvantage of the two loaded pistons as mentioned in connection with fig. 1 also exists in the case of the V-construction, but here it does not count so much because in the case of an elevated pressure, where the frictional losses occur most, the crankcase can be put under pressure, thereby considerably reducing those losses. This method will be discussed in detail below in connection with another construction. Suffice it to say here that thanks to this and the other improvements described the V-construction is very suitable for high-speed engines of low power (for instance less than 1 h.p.) where the efficiency is not decisive.

*Constructions with transfer piston*

We have already mentioned that the old air engines nearly always had only one loaded piston. This was made possible by the use of a transfer piston: a body situated between the hot and the cold spaces and occupying (at least) as much volume as the maximum of the hot or cold space. By moving this body to and fro periodically the air is made to flow back and forth between the hot and cold spaces. The transfer piston is not loaded, because, as explained above, the pressure in the cold space is in principle the same at every

[3]) Strictly speaking this holds only in the case of a driving mechanism with an infinitely long connecting rod. With a rod of finite length the piston does not move purely sinusoidally, its motion containing higher harmonics which, because of their higher frequency, can never be balanced by a weight rotating at the fundamental frequency.

moment as that in the hot space [4]). The cold space is shut off from the outside air by an ordinary piston. This single piston now serves both for the gain of work upon expansion and for the supply of work upon compression.

Since there is no difference in pressure on either side of the transfer piston it is not necessary for that piston to be made to fit tightly in the cylinder, so that there need be no great frictional losses there. It is even possible — although, according to our present insight, not at all advisable — to leave channels open along or through the transfer piston for the free passage of thc air in passing to and fro between the cold and hot spaces; we shall presently discuss a design used about 80 years ago when such was generally applied. But also in the designs where the hot and cold spaces have an "external" connection, the sealing of the transfer piston still only needs to satisfy moderate requirements, so that again there need not be any high frictional losses. This advantage is all the greater when the transfer piston drive is of such a construction as to avoid any appreciable lateral forces acting upon the transfer piston.

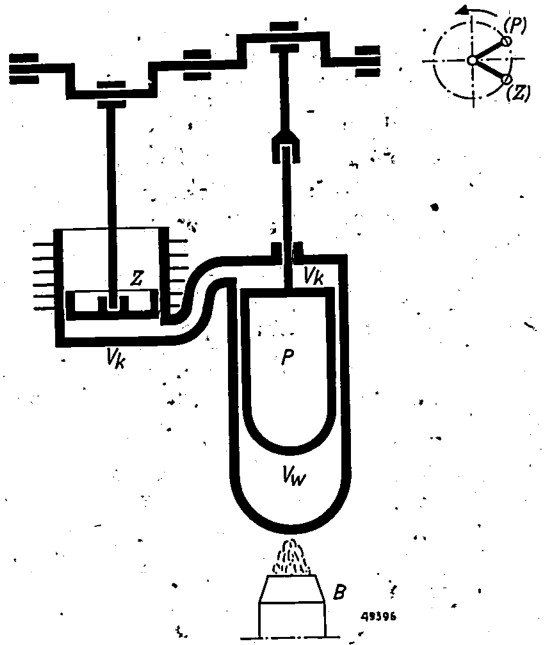


Fig. 4. Diagram of a classical air engine with only one loaded piston ($Z$) and a transfer piston ($P$). The work medium flows along the transfer piston to and fro between the hot space $V_w$ and the cold space $V_k$. The source of heat $B$ corresponds to the burner in fig. 1.

The principle of an air engine with transfer piston can be practically realized in various ways. An arrangement which was formerly commonly employed is represented diagrammatically in *fig. 4*. It was employed by Stirling as early as 1816. But the same type is still seen in the small air engines built at the present day and sometimes used as prime mover for stirring machines in chemical laboratories, as toys, etc.

In article II it has already been made quite clear why these old engines had such unsatisfactory properties as to cause them to lose the contest with internal combustion engines. We shall mention here only a few of the defects dealt with in that article. The heater and the cooler were very primitive (*cf.* fig. 4). Heating was done by means of a fire placed under the hot space, the walls of which were unable to withstand a very high temperature. Cooling took place by heat exchange with a cooling-water jacket *via* the wall of the cold space. In many of the old constructions there was no regenerator; thus, for instance, in the design of fig. 4, the work medium has to flow past the transfer piston to get from the hot space to the cold space and *vice versa*. Only the cylinder wall opposite the transfer piston plays, very ineffectively, the rôle of regenerator: the air coming from the cold space gets back at least part of the heat it gave off to the wall half a cycle earlier when coming from the hot space.

It would have been possible to apply the new ideas resulting from our investigations to the design of fig. 4. This was not done because of a fundamental drawback pertaining to that design: the cold space is divided between two cylinders. Since the piston and the transfer piston can certainly not move with a phase difference of 180° — the volume variations in the hot and cold spaces would then differ 180° in phase [5]) and the engine would not function — the volumes of the two parts of the cold space are never both zero at the same time. The smallest value of the cold volume thus being larger than zero, a high specific power can never be reached.

A better arrangement of an engine with transfer piston is shown in *fig. 5*. Here the cold space is not divided between two cylinders and by a suitable choice of the lengths of the stroke and the mutual phase difference between working piston and transfer

[4]) Actually there is a small difference in pressure due to the flow losses. There is also a slight difference in the areas upon which the forces act on either side of the transfer piston: at the side where the connecting rod is affixed to the transfer piston the rod cross-section has to be subtracted from the area. However, the remaining force in the connecting rod is very small. There is, therefore, practically no loss (nor gain) of work accompanying the motion of the transfer piston.

[5]) Except at 180° the phase difference between the volume variations of the hot and cold spaces is always larger than the phase difference between the working piston and transfer piston movements. At the angle chosen in fig. 4 between the cranks the latter phase difference is 60°, the former 120°.

piston both the hot and the cold volumes can be made equal to zero once per revolution [6]).

The conclusion to be drawn from the above is that the design given in fig. 5 must be considered to offer the best possibilities. In the Philips laboratories most of the experiments were in fact carried out on this type of engine, and these have led to the building of a model incorporating all the fundamental improvements described in II and several other incidental improvements. It is this new engine that will now be described.

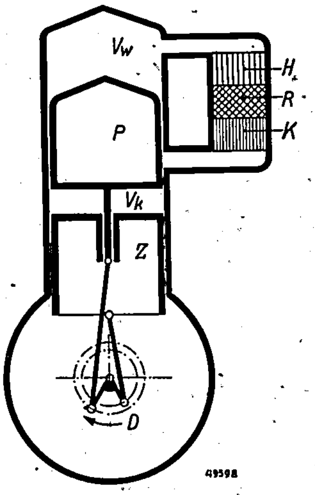


Fig. 5. Diagram of an air engine with transfer piston where only one cylinder is used. For the meaning of the letters see figs. 4 and 1.

### *The single-cycle Philips air engine*

We will first give a brief summary of the new principles of construction discussed in II which have been applied in this engine: transition to high pressures and to a high temperature of the hot space, made possible by the employment of heat and creep-resisting steel alloys; effective construction of the heater, cooler and regenerator, so that a sufficiently rapid heat transfer is obtained without involving too great dead volume and flow resistance (at the high speeds of the Philips engines); reduction of chimney losses by the introduction of an efficient heater; general limitation of heat losses by small compact design; general reduction of frictional losses by small compact build.

The most important details of the engine are to be seen in *fig. 6*. The heater, regenerator and cooler, instead of being placed in a separate communication channel between the hot and cold spaces, as may seem the obvious way and as is indicated in fig. 5, are given an annular shape and placed around the cylinder. This produces a very compact arrangement, while at the same time, with the absence of inlet channels, unnecessary flow resistance and dead volume are avoided. The heater, which is enclosed on the inside by a thin-walled cylinder of heat-resisting metal, opens directly into the hot space. The heater is heated externally by means of a burner. The cold space is in communication with the cooler *via* a ring of ports in its wall. For further details of the construction of the heater, cooler and regenerator see II.

The lower part of the transfer piston is provided with a bearing surface while the upper part in the hot space consists of a thin-walled cap leaving some clearance in the cylinder (see fig. 6). This cap, which is made of heat-resisting material with low heat conductivity (nickel-chromium steel), fulfils an important function. This thin, poorly conductive wall of the cap ensures that only very little heat flows through it from the hot top end to the cold bottom end. Furthermore, there is only a minimum transfer of heat through the thin layer of gas in the annular space between the cap and the

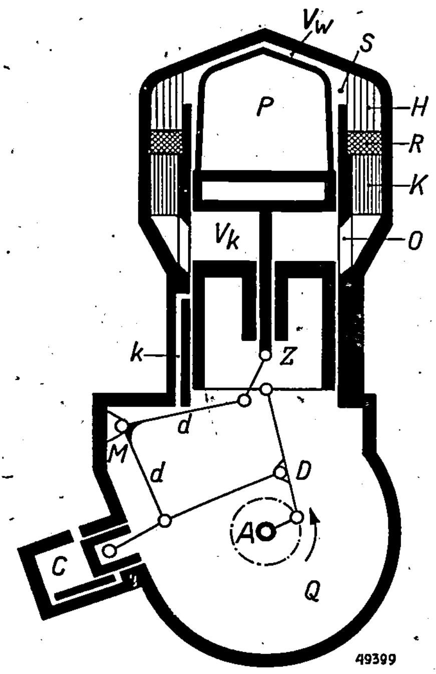


Fig. 6. Cross section of the single-cycle Philips air engine. The hot and cold spaces, respectively $V_w$ and $V_k$, are in one cylinder and separated by transfer piston $P$ with insulating cap. The heater $H$, regenerator $R$ and cooler $K$ are annular in shape and built around the cylinder. $S$ outlet of heater into $V_w$, $O$ ports between cooler and $V_k$. The connecting rod $D$ of the piston $Z$ and the driving mechanism $d$ of the transfer piston are housed in a closed crankcase $Q$, in which air is maintained at the minimum pressure of the cycle in the cylinder. The drive $d$ with fixed pivot $M$ is such that only extremely small lateral forces act on the transfer piston. Air which has leaked into the crankcase is returned to the cylinder *via* the channel $k$. $C$ pressure pump; $A$ engine shaft.

[6]) A design with work piston and transfer piston in one cylinder is fairly old; it was employed around 1860 by Lehmann.

cylinder wall. Thus any loss of heat that might in principle be due to the transfer piston is limited. Still more important, however, is the fact that thanks to this insulating action of the cap the bearing surface of the transfer piston is kept cold. This avoids all the difficulties of tight-sealing, wear, etc. which are encountered with hot sliding surfaces.

It is to be pointed out that it is thanks to the nature of the hot-air process that the cap construction described gives no difficulties. In internal combustion engines the material of such a cap would be too quickly destroyed by the very high temperatures of combustion. Especially in the case of petrol engines the hot cap would furthermore lead to premature combustion of the mixture of fuel and air. Moreover, the clearance between the cap and the cylinder wall would soon become clogged up by the deposition of carbon or other products of combustion, thus increasing friction and causing damage to the running surface of the piston.

The engine is provided with a closed crankcase. The air in the crankcase is kept at the desired pressure by means of a small pump (*C* in fig. 6) which pumps air from the outside into the crankcase and is driven by the engine itself. The pressure in the crankcase is kept equal to the minimum pressure occurring in the cycle of the engine. If, for instance, the engine is to work with a top pressure of 20 atm. and a minimum pressure of 8 atm. the crankcase is brought to 8 atm. This has important advantages. The greatest differential pressure at the piston is thereby considerably restricted (in the example mentioned to 12 atm; while with a crankcase at atmospheric pressure it would be 19 atm.), leakage along the piston is reduced, piston sealing is simplified and there is less friction. With this construction the air leaking along the piston at maximum pressure is returned from the crankcase to the working space through a small channel (*k* in fig. 6) momentarily making an open communication between the work space and crankcase when the piston is in its lowest position (state of approximately minimum pressure). In the case of the normal construction with a crankcase at atmospheric pressure the air to be returned would first have to be compressed to the minimum pressure (here 8 atm.), but with this design the return of air requires much less work. Any leakage from the crankcase to the outside is now only possible along the engine shaft (*A* in fig. 6), and this is counteracted by applying the usual kind of shaft packing.

We have just stated that the pressure in the crankcase is made equal to the minimum pressure in the cycle. Actually it is just the reverse: through the open connection at the lowest position of the piston the minimum pressure of the cycle is automatically made equal to the crankcase pressure. Therefore, by adjusting the crankcase pressure with the help of the pump the pressure and thus the power of the engine can be regulated in a very simple way.

When the engine has not been used for a long time the pressure in the crankcase may have dropped to atmospheric pressure due to the possible slight leakage to the outside. Upon being started up again, therefore, the engine will at first develop only a low power, but this rises in a very short space of time as the pump quickly raises the pressure in the crankcase to the desired level.

Finally the elevated pressure in the crankcase offers another great advantage with respect to the mechanical efficiency of the engine. Owing to the fact that the alternating difference in pressure on either side of the piston is lowered by a constant amount, the forces of the piston rod that have to be transmitted to the crankshaft are reduced accordingly. There is thus a lighter load on the bearings

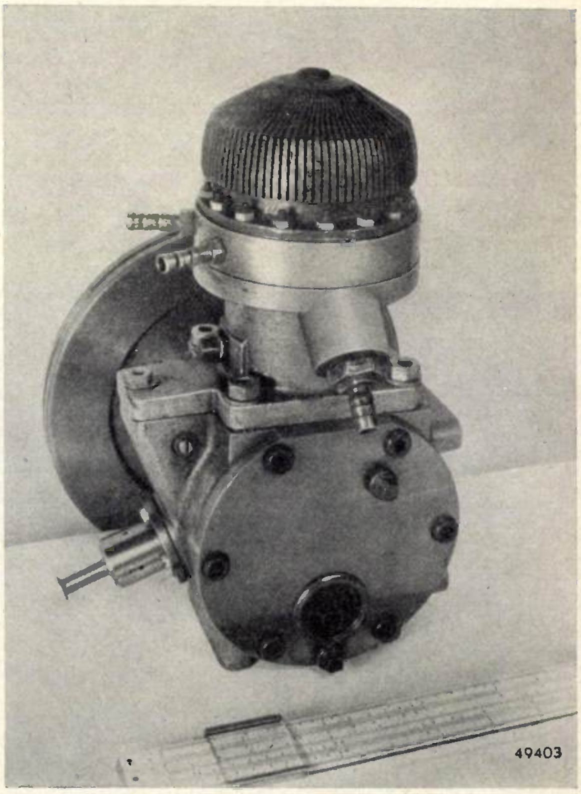


Fig. 7. Photograph of an experimental model of the single-cycle Philips air engine built according to the diagram of fig. 6. At the top is the heat exchanger of the burner, built around the heater (*cf.* article II). At the bottom on the left is the pump with screw for regulating the pressure. Behind the engine is the flywheel.

and the frictional losses are therefore less. A similar effect in a still more pronounced form will be found in the multiple engine.

A photograph of an engine of the type described is reproduced in *fig. 7*. With this model a large number of test measurements have been taken which will be dealt with in another article. We would only state here that the efficiency of this type of engine is now quite satisfactory, many times better than that of previous air engines. This type is suitable for a rating of several h.p. and not too high speeds (*e.g.* up to 2000 r.p.m.). The limited speed is due to the fact that, just as in the case of the primitive model in fig. 2 and in all kinds of single-cylinder engines, the engine cannot be balanced in a simple manner. The need for a proper balancing and higher powers brings us to the multiple engines.

## Multiple engines

Although the results obtained with the single-cycle engines were already very satisfactory, further development was greatly stimulated when it was found that in multiple or multi-cylinder engines the construction could be greatly simplified.

We must pause a moment to consider the term "multi-cylinder engine". In the foregoing we have seen that some types of single-cycle engines have two cylinders (figs. 3 and 4), while others have only one cylinder (figs. 1, 5, 6). All these engines we could call single-cycle engines, because they possess only one "system" consisting of cold space, cooler, regenerator, heater, hot space. In the case of the multiple Philips air engines about to be described, however, the number of cylinders is in every case equal to the number of systems, so that for instance the term "four-cylinder engine" can no longer cause any misunderstanding.

The simplification in construction mentioned above lies in the manner in which several systems are combined to form a multiple engine. In *fig. 8* the principle is illustrated for the case of a four-cylinder engine. In each of the four cylinders there is a hot space (above) and a cold space (below). Contrary to what might be expected, however, the hot and cold spaces in each cylinder do not form a coordinated system, for here the hot space of one cylinder is connected — *via* a heater, regenerator and cooler — with the cold space of the next cylinder[7]). Such an arrangement could be compared, for instance, to that of fig. 3, for in both cases it is immediately to be seen that upon expansion of the working medium the two pistons between which the system is situated will deliver more work to the engine shaft than what they have to perform for compression of the working medium within the system, provided the volume variations of the hot space (in the first cylinder) are advanced in phase with respect to the variations in the cold space (in the second cylinder) according to the conditions already given. It is easy to see that this is ensured when the piston of the second cylinder is in advance of that of the first — a condition that can be satisfied in a very natural way, since in a multi-cylinder engine the series of pistons will never be made to move in phase.

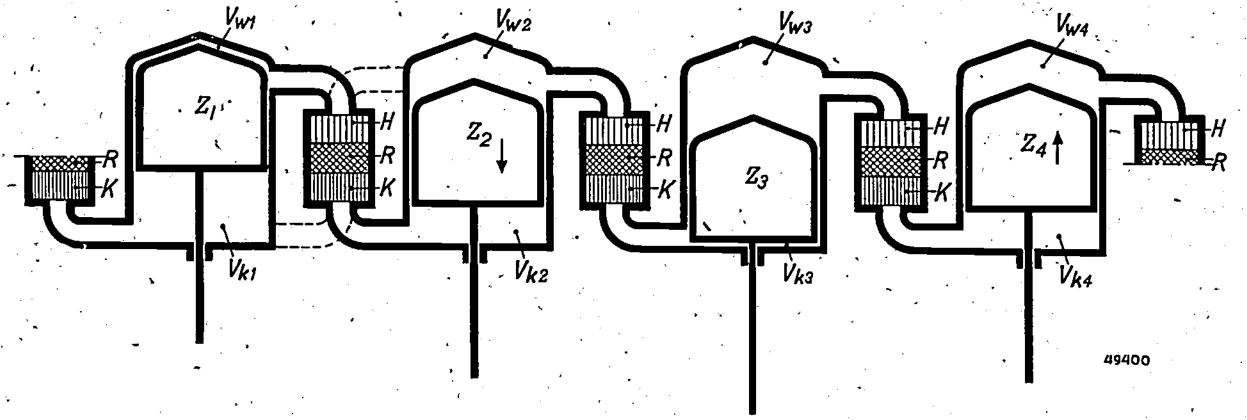


Fig. 8. Principle of the Philips multi-cylinder air engine. In each of the four cylinders is a hot space at the top, $V_{w1}$ to $V_{w4}$, and a cold space at the bottom $V_{k1}$ to $V_{k4}$. The hot space of each cylinder is connected *via* a heater *H*, regenerator *R* and cooler *K*, with the cold space of the next cylinder. The pistons $Z_1$-$Z_4$ of the successive cylinders move with a suitably chosen phase difference (in the case shown with four cylinders this is 90°).

[7]) This principle has already been described by H. Rinia in his article: New possibilities for the air engine, Proc. Kon. Ned. Akad. Wet., Amsterdam, 49, 150-155, Febr. 1946.

What has been stated here for one pair of pistons and the system between them also holds for the rest. If, therefore, each of the four pistons in fig. 8 is made to move in advance of the preceding one, each of the four systems works as an air engine. *With this simple combination of the systems the transfer pistons with their driving mechanism are dispensed with.* Needless to say, this means a very important mechanical simplification and allows of a considerable increase in mechanical efficiency.

As to the phase differences of the pistons, it is obvious that with four pistons there should be a phase shift of 90° in their motions. The volume variations of corresponding hot and cold spaces then likewise differ 90° in phase. Combinations can also be made with more than four systems, of course with different phase differences. Within certain limits this has little effect upon the efficiency of the engine, since the curve representing the efficiency of the hot-air process as a function of the phase difference between the hot and cold spaces is fairly flat in the neighbourhood of the maximum.

With the method of communication between the hot and cold spaces as shown in fig. 8 each piston must be in advance of the preceding one. This order of piston movements determines the direction of rotation of the engine. If the connections between the cylinders are interchanged, each hot space being connected with the cold space of the preceding cylinder (in fig. 8 this has been indicated diagrammatically for one cylinder by dotted lines), each system again works as an air engine, provided the order of the piston motions is reversed. This means that the engine then runs in the opposite direction. *This provides a very simple method of reversing the direction of rotation of the engine while it is running.* The reversal of the connections between the cylinders is brought about with the aid of a slide for each cylinder, which can be fitted on the cold side of the engine (see below).

It is also of importance that in a multi-cylinder engine each piston is double-acting. To understand this properly let us first consider fig. 6. There, as in all the engines so far dealt with, the piston is single-acting, because it transmits energy from the work medium to the crankshaft only during the down stroke, while on the up stroke work is performed on the working medium. The first-mentioned work is larger than the latter, the difference being the mechanical energy gain. But the frictional losses bear a definite relation to the total energy conversion in the driving mechanism, no matter whether the various energy contributions count in a positive or a negative sense for the consumer. By way of contrast now let us turn to fig. 8: there each piston transmits energy from the working medium to the engine shaft both on its down stroke and on its up stroke, for each piston is situated between two systems which, for convenience, we shall call the left-hand and the right-hand systems. The upward stroke of the piston coincides for a large part with the expansion of the left-hand system and with the compression of the right-hand system; conversely the down stroke is for the greater part simultaneous with the expansion of the right-hand and the compression of the left hand system. In both cases, therefore, the compression in one system is brought about for a large part directly by the expanding medium in the neighbouring system, only *via* the body of the piston; the driving mechanism takes no part in this, only transmitting the excess of work furnished by each expansion. The energy conversion in the driving mechanism

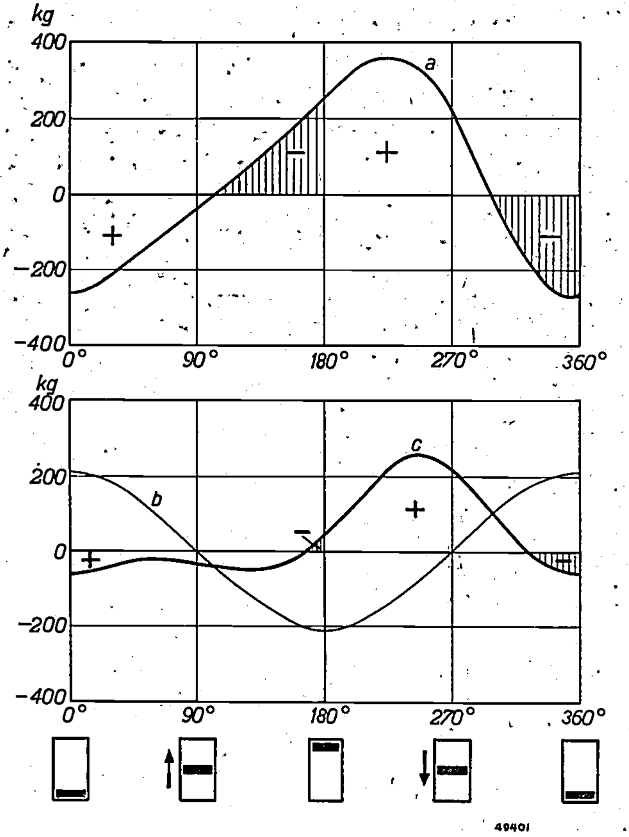


Fig. 9. Diagram for the forces in the connecting rods for a given 20 h.p. four-cylinder air engine with a speed of 3000 r.p.m. at full load. The force (in kg) of the connecting rod of a cylinder is plotted as a function of the crankshaft angle. The left-hand half (0-180°) corresponds to the up stroke, the right-hand half to the down stroke. Forces directed downward are reckoned to be positive; + and — indicate that during the intervals in question the piston delivers positive or negative work to the crankshaft. Curve $a$ = gas force, $b$ = inertia force, $c$ = resulting force in the connecting rod. The curve $c$ shows that in this four-cylinder engine each piston is practically purely double-acting.

is therefore no greater than what corresponds to the power of the engine, and consequently the frictional losses are also limited.

This rough picture needs correcting on two points. In the first place there is the fact that for each piston in fig. 8 the expansion of the left-hand and the compression of the right-hand system coincide only "for the greater part" with the up stroke (and *vice versa*). Secondly, in the energy conversion in the driving mechanism we have still to take into account a "wattless component", *viz.* the inertia forces of the moving masses of piston, connecting rod, etc. In regard to these forces we will consider the diagram in *fig. 9* for the forces in the connecting rods. Curve *a* gives the variation of the "gas force", *i.e.* the force exerted by the working medium (of the two adjacent systems together) on each connecting rod, as a function of the crankshaft angle, for a given four-cylinder engine. A force directed downwards is here plotted positive; the abscissa 0° corresponds to the lowest position of the piston; the left-hand half of the diagram thus corresponds to the up stroke and a positive gas force here performs negative work on the engine shaft, whilst a negative force yields positive work. In the right-hand half this is reversed. It is in this sense that the plus and minus signs in the diagram, referring to the work performed on the shaft, are to be understood. It may be seen that the gas forces do indeed require work of the connecting rod during two intervals of time per revolution. It is found, however, that it so happens that the greater part of this work can be supplied by the inertia forces; see curve *b* and the resulting rod force curve *c* in fig. 9. Of course no energy is thereby gained, because the inertia forces consume a corresponding amount of energy again in the rest of the period. But, as curve *c* shows, the piston has now practically become purely double-acting, so that our original conclusion remains valid.

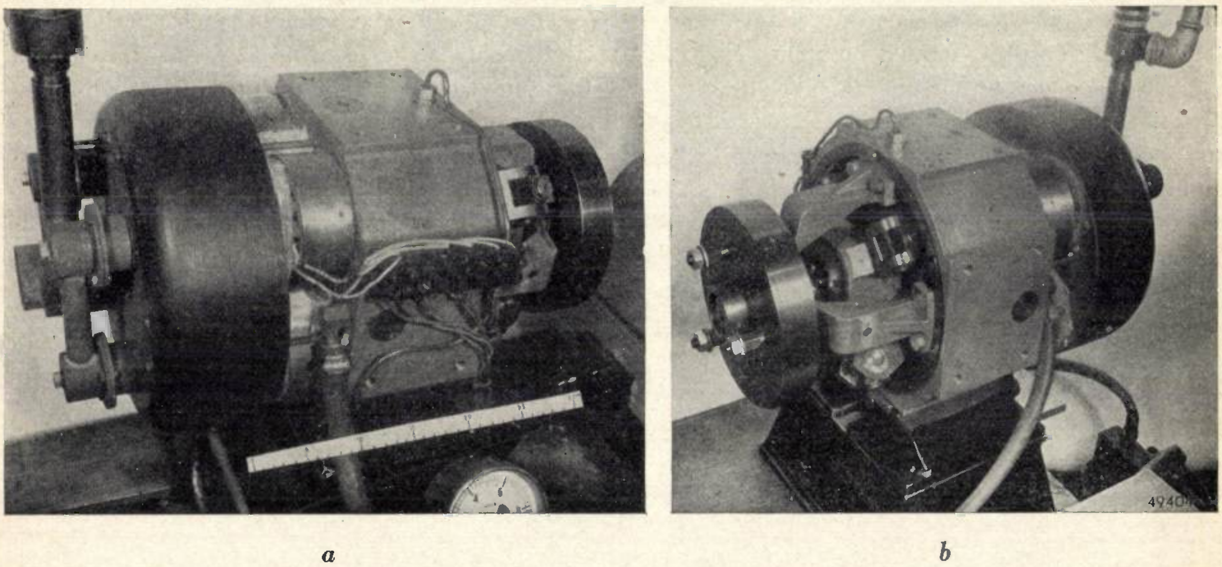


Fig. 10. Photographs of a practical model of the Philips multi-cylinder air engine: four-cylinder engine with parallel cylinders placed in a square with a wobble-plate mechanism for the transmission of the motion of the four pistons to the engine shaft. The engine can deliver 15 h.p. at 3000 r.p.m. The model shown is run with gas as fuel.

*a*) In this photograph the hot side of the engine is on the left and the cold side on the right. The pipe for the fuel supply is on the extreme left. The cap on the left contains the heater. Part of the jacket of three cylinders is visible. On the extreme right are the wobble-plate mechanism and the flywheel. With this mechanism the engine shaft is parallel to the axes of the cylinders. The driving mechanism is in a closed crankcase, the cover of which has been removed here for clearness. The wires connected to the terminal board in the middle are connections to a number of thermo-elements introduced at various points in the engine for taking test measurements. (The scale is drawn in centimetres.)

*b*) The same engine seen from the other side. It gives a better view of the wobble-plate mechanism between the flywheel and the engine block. Note the compact construction obtained with this mechanism.

Owing to the relatively small forces in the connecting rods the multi-cylinder air engine compares favourably with an internal combustion engine of the same power per cylinder, the same swept volume and the same speed. In fig. 9, which relates to an air engine of 5 h.p. per cylinder and a speed of 3000 r.p.m., the peak value of the resulting force in the connecting rod is about 250 kg. In a comparable internal combustion engine, on the other hand, there is a maximum force in the connecting rod of about 1200 kg. These smaller connecting rod forces in the air engine mean less load on the bearings, a factor of great importance for the construction.

As to the practical construction of the multi-cylinder engine, it has been possible to apply practically unaltered many of the structural elements described in connection with the single-cylinder engine, such as the annular-shaped heater, regenerator and cooler around each cylinder, the insulation caps on tbe pistons serving to keep the running surfaces cool, etc. A new problem, however, was how to arrange the position of the various cylinders and the transmission of all the piston forces to one shaft. In this article we can only briefly touch upon these points.

The cylinders will preferably be arranged so that all the cold spaces lie on one side of the engine, with all the driving mechanism, etc. on this cold side. Such an arrangement has already been assumed in the diagram of fig. 8.

In that diagram the four cylinders are in a row. This arrangement has the objection that a long connecting channel is necessary between the last and the first cylinder, which involves losses. In order to avoid this the four cylinders can be placed in star formation, as is customary in aircraft engines (the cold side is then in the centre of the star), or in two V's one behind the other, as is done in some automobile engines. Another possibility is to place the four cylinders parallel to each other and bundling there as it were) in a square. Whereas with the first two arrangements a crankshaft construction can be used for converting the linear motion of the pistons into the rotating motion of the engine shaft, in the last case mentioned a wobble-plate mechanism is indicated. The practical model of the Philips multi-cylinder air engine illustrated in *fig. 10* is constructed in the last manner, but this design is considered only suitable for air engines of some 20—30 h.p. For higher powers other driving mechanisms have been worked out, based on the abovementioned arrangement of the cylinders in two V's one behind the other.

It would lead us beyond the scope of this present article to go deeper into the mainly mechanical details of these different designs and the reasons for their choice. These will be left to be dealt with in further articles in this periodical.

# VOLTAGE IMPULSES IN RECTIFIERS

by Tj. DOUMA.

621.314.65.015.33

The valves in a rectifier may be regarded as switches opening or closing electric circuits at certain moments. This is accompanied, as a rule, by transient phenomena, which manifest themselves, *inter alia,* in voltage impulses on self-inductances occurring in the circuits; when valves with a variable ignition moment are employed these impulses or peaks are exceptionally high in cases where the rectified voltage has been stepped down more or less by delaying the moment of ignition. Owing to the presence of parasitic capacities, oscillating circuits are formed and, moreover, the voltage peaks are concentrated on the outermost windings of the transformer and any other coils. In rectifiers for high tensions, such as for high power transmitters, this is apt to endanger the insulation. Various methods are indicated for avoiding this danger, such as the introduction of anode choke coils and of condensers with damping resistances.

## Introduction

When a sudden change takes placc in an electrical system carrying currents — for instance through the short-circuiting or opening of a branch of the network — in course of time a stationary condition of the currents and voltages will set in which in general differs from the condition that prevailed immediately before the change took place. The difference existing between this ultimate condition and the actual condition as from the moment of switching is called the transient phenomenon. Owing to the dissipation of energy it gradually dies out and thus forms a bridge between the old state and the new state.

Such transient phenomena are brought about not only by switching operations, for in rectifiers, for example, they continually occur during normal working. This will be readily understood when it is considered that a relay valve bears a great resemblance to a switch: so long as the anode is negative with respect to the cathode the relay does not allow any current to pass (or at least only a negligible current), so that it corresponds to an open switch; when the relay allows a current to pass through (in the right direction) this is accompanied by a voltage drop which, though greater than that with a normal switch, is generally small compared with the working voltage, so that in this case the relay acts as a closed switch. While a rectifier is working its relays "open" and "shut" at certain moments, each time giving rise to a transient phenomenon. Not only do these transient phenomena partly govern the moments when the relays become dead (thereby influencing the general picture of the working of the rectifier), but at those moments in certain cases voltage impulses arise with a steep front and large amplitude, which impulses may seriously endanger the insulation of the transformer or choke coils forming part of the rectifier. Such phenomena occur especially when the moment (recurring for each cycle of the mains frequency) at which the relays begin to pass through current —, the so-called ignition moment — has been artificially delayed with respect to the earliest possible ignition moment in order to reduce the rectifier voltage. This may be done by using relay valves, *e.g.* rectifying valves, with gas filling and control grids, to which latter an alternating voltage is applied which in phase is behind the anode alternating voltage, or else by employing mercury cathode valves with, *e.g.*, a capacitive starter, to which ignition impulses are applied at the desired moments [1]).

In rectifiers one has to do not only with the regularly recurring transient phenomena just mentioned, resulting from the periodical opening and shutting of the valves, but also with the transient phenomena caused by the "ordinary" switching operations. These latter phenomena will be left out of consideration here.

The voltage impulses referred to above are excited in the self-inductance of the rectifier. In the rectifiers with which we are mainly concerned here — those for supplying the anode voltage of transmitters — there is generally a self-inductance in two places, *viz.* in the choke coil — which serves to reduce the ripple of the rectified current — and the leakage inductance of the transformer coils, *i.e.* the inductance corresponding to the magnetic lines of force in the transformer that are connected only with a primary or a secondary coil (or a part of these coils).

[1]) A rectifier provided with relay valves having a filament cathode and a control grid was described in Philips Techn. Review 1, 161-165, 1936. An article on the action of relay valves with a mercury cathode and a capacitive igniter has been published in Philips Techn. Rev. 9, 105-113, 1947.

However, also the capacities of the windings are of great importance, in two respects: 1) they govern, at least at the moment that the voltage impulse arises, the distribution of the impulse between the windings connected in series; 2) in combination with the above-mentioned inductances they form oscillating circuits which are activated by the transient phenomenon.

These voltage impulses will now be examined and means will be indicated for counteracting them successfully. We will start by considering a simple rectifying circuit working on only one phase of the alternating voltage, after which we will proceed to discuss the more common circuits with several phases, each case being treated first without consideration of the influence of the capacities and then by taking this influence into account.

### Rectifier working on one phase of the alternating voltage

In the simple circuit of *fig. 1* the leakage inductance is represented by the separate inductance $L_a$ in series with the secondary circuit, which also comprises the secondary transformer coil in which at no load the alternating voltage $v = V_{\max} \sin \omega t$ is induced, the valve *Re* (with gas filling and a control grid, thus a relay valve), a choke coil with the self-inductance $L_0$, a resistance $r$ and the parallel circuiting of the load resistance $R$ and the condenser $C_0$. The object of this condenser is to reduce the voltage ripple on the terminals of $R$; so as not to make the matter unnecessarily complicated we will assume that this ripple is negligible.

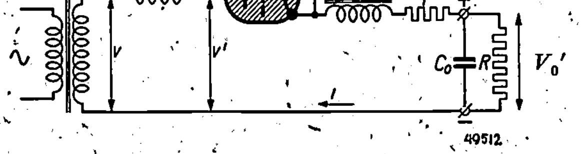


Fig. 1. Rectifying circuit fed by a monophase alternating voltage. Given a sufficiently large capacitor $C_0$, the ripple of the voltage $V_0'$ on the load resistance $R$ may be ignored. $Re$ = relay valve, the ignition moment of which is regulated by the phase of the grid alternating voltage $v_g$; $L_0$ = self-inductance of the choke coil; $L_a$ = leakage inductance of the transformer; $v = V_{\max} \sin \omega t$ is the secondary voltage under no load; $v'$ = terminal voltage; $r$ = resistance of the choke coil and the transformer.

A pulsating current $i$ will flow through the circuit. The current impulse starts at a moment $t_0$ which can be fixed between certain limits, for instance by regulating the phase of the grid voltage $v_g$; this impulse ends at a moment $t_d$, which will be further defined presently. At these two moments the current curve makes a bend and consequently the voltage at the self-inductances $L_a$ and $L_0$ will undergo a sudden change of the order of $L_a \mathrm{d}i/\mathrm{d}t$ and $L_0 \mathrm{d}i/\mathrm{d}t$ respectivcly, where $\mathrm{d}i/\mathrm{d}t$ applies for the moments $t = t_0$ and $t = t_d$. In order to calculate these voltage jumps we have to consider the differential equation of the circuit:

$$L_a \frac{\mathrm{d}i}{\mathrm{d}t} + v_{\mathrm{arc}} + L_0 \frac{\mathrm{d}i}{\mathrm{d}t} + ri + V_0' = v = V_{\max} \sin \omega t.$$

where $V_0'$ is the condenser voltage assumed to be constant and $v_{\mathrm{arc}}$ the momentary value of the arc voltage of the relay valve. By a good approximation this may be taken to be independent of the current, so that it may be regarded as a rectified voltage. If we combine this and the condenser voltage to form together one direct voltage $V_0 = v_{\mathrm{arc}} + V_0'$ and add the two self-inductances to make one self-inductance $L = L_a + L_0$, and further by substituting $\tau$ for $\omega t$, then the equation may be written as:

$$\omega L \frac{\mathrm{d}i}{\mathrm{d}\tau} + ri = V_{\max} \sin \tau - V_0. \quad . \quad . \quad (1)$$

At the moment $t_0$ at which the current begins to flow — owing to the grid voltage of the relay valve then exceeding the critical value — we have the "ignition angle" $\omega t_0 = \tau_0$. The terminal voltage $v'$ of the secondary transformer coil has at that moment the value

$$(v')_{t_0} = V_{\max} \sin \tau_0 - (L_a \frac{\mathrm{d}i}{\mathrm{d}t})_{t_0} =$$

$$= V_{\max} \sin \tau_0 - \frac{L_a}{L} (\omega L \frac{\mathrm{d}i}{\mathrm{d}\tau})_{\tau_0}.$$

At the moment of ignition there is thus a jump in the terminal voltage of the order of

$$(\Delta v)_0 = \frac{L_a}{L} (\omega L \frac{\mathrm{d}i}{\mathrm{d}\tau})_{\tau_0}.$$

Substituting the value of $\omega L \; \mathrm{d}i/\mathrm{d}\tau$ from (1) and bearing in mind that at the moment $t = t_0$ the current $i = 0$, then we find:

$$(\Delta v)_0 = \frac{L_a}{L} (\sin \tau_0 - \frac{V_0}{V_{\max}}) \cdot V_{\max} =$$

$$= \frac{L_a}{L_0 + L_a} \cdot (\sin \tau_0 - q) \cdot V_{\max}, \quad . \quad . \quad (2)$$

The amplitude of a relatively small ripple voltage is about $p\%$ or less ($p \ll 100$) of the direct voltage when $C_0$ and $R$ satisfy the equation

$$\omega C_0 R \geqq 100 \, \pi/p.$$

In the case of the rectifiers with $m$ phases to be dealt with later this condition is:

$$\omega C_0 R \geqq 100 \, \pi/mp.$$

where $q$ represents the voltage ratio $V_0/V_{\max}$. The jump is clearly seen in *fig. 2* giving the oscillogram of the secondary transformer voltage in a still further simplified circuit.

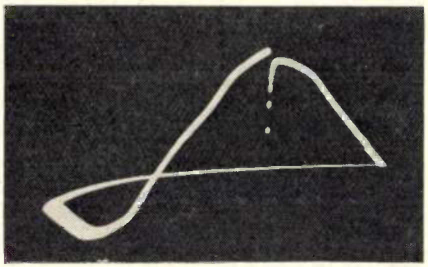


Fig. 2. Oscillogram of the secondary terminal voltage $v'$ in a rectifier on the principle of fig. 1 with $L_0 = 0$, $C_0 = 0$, ignition angle $\tau_0 =$ appr. 90°. Upon ignition a voltage impulse arises in the leakage inductance $L_a$.

Another bend occurs in the current curve when $i$ is zero, thus when the valve is extinguished ($t = t_d$, to which corresponds $\omega t_d = \tau_d =$ the "extinction angle"). We then find for the jump in the terminal voltage:

$$(\Delta v)_d = \frac{L_a}{L_0 + L_a} \cdot (\sin \tau_d - q) \cdot V_{\max}. \quad . \quad (3)$$

The fact that nothing is seen of this jump in fig. 2 is due to the fact that this oscillogram was recorded under circumstances where $q = 0$, $\tau_d \approx 180°$, so that $(\Delta v)_d$ was very small.

Similar jumps occur for the self-inductance $L_0$ and these can be expressed by substituting $L_0$ for $L_a$ in the numerator of equations (2) and (3).

As to the size of these jumps, from equations (2) and (3) one sees in the first place that the two self-inductances $L_a$ and $L_0$ function as voltage dividers and further that for a given peak value $V_{\max}$ of the alternating voltage the jumps are the greatest when $q$ is as small as possible, thus when the condenser voltage is zero (which is the case with short-circuiting, $R = 0$). Moreover, there is a big jump upon ignition when this takes place at the moment at which the alternating voltage reaches its peak value ($\sin \tau_0 = 1$).

To investigate this more accurately we have to consider more closely the quantities $q$ and $\tau_d$, occurring in (2) and (3), which in general are functions of $\tau_0$, $R$, $r$, $L$ and $\omega$. Here the resistance $r$, which in practical cases consists only of the resistance of the transformer, of the choke coil and of the conductors, is usually of little significance. Ignoring this resistance, it appears that $\tau_d$ and $q$ may be written as the functions of two variable quantities $\tau_0$ and $\varphi$, the latter being defined by $\operatorname{tg} \varphi = \omega L/R$. Briefly this calculation is arrived at in the following way:

For $r = 0$ equation (1) becomes

$$\frac{di}{d\tau} = \frac{V_{\max}}{\omega L} (\sin \tau - q), \quad . \quad . \quad . \quad . \quad (4)$$

to which correspond the limiting conditions

$$i = 0 \text{ for } \tau = \tau_o \quad . \quad . \quad . \quad . \quad . \quad (5)$$

and

$$i = 0 \text{ for } \tau = \tau_d \quad . \quad . \quad . \quad . \quad . \quad (6)$$

Equation (4) is easily integrated and the integrating constant can be determined with the aid of (5). With the aid of (6) we then find a relation between $\tau_0$, $\tau_d$ and $q$. A second equation of these quantities follows from the condition that per cycle the charging of condenser $C_0$ equals its discharge. These charges amount to

$$\int_{t_o}^{t_d} i \; dt \text{ and respectively, } \frac{V_0'}{R} \; T,$$

where $T$ is the duration of a cycle. Thus

$$\int_{\tau_o}^{\tau_d} i \; d\tau = 2\pi \frac{V_0'}{R} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

By substituting in this equation the expression found from (4) and (5) for $i$, then we get a second relation between $\tau_0$, $\tau_d$ and $q$, from which, in combination with that already found, $\tau_d$ and $q$ can be calculated separately as functions of $\tau_0$; $\varphi$ is then the parameter.

Proceeding in the manner outlined above we find from (4) and (5):

$$i = \frac{V_{\max}}{\omega L} \left\{ \cos \tau_o - \cos \tau - q (\tau - \tau_o) \right\} . \quad . \quad . \quad (8)$$

and from (6) and (8) it follows that:

$$0 = \cos \tau_o - \cos \tau_d - q (\tau_d - \tau_o) . \quad . \quad . \quad . \quad . \quad (9)$$

By now introducing two auxiliary angles $x$ and $y$:

$$x = {}^1/_2 (\tau_d - \tau_o), \quad y = {}^1/_2 (\tau_d + \tau_o), \quad . \quad . \quad . \quad (10)$$

we may write (9) in the following form:

$$q = \frac{\sin x}{x} \sin y. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

If we ignore the arc voltage with respect to the direct voltage $V_0'$, that is to say if we ignore the difference between $V_0'$ and $V_0$, then by substituting (8) in (7) and using this in combination with (10) we arrive at

$$\operatorname{tg} y = \frac{\operatorname{cotg} \varphi}{\pi} \cdot x \cdot (x \operatorname{cotg} x - 1). \quad . \quad . \quad . \quad . \quad (12)$$

For various values of the parameter $\varphi$, (12) yields corresponding values for $x$ and $y$, thus of $\tau_0$ and $\tau_d$. Finally from (11) we get the corresponding values of $q$.

The result, $\tau_d = f(\tau_0)$ and $q = f(\tau_0)$, with $\varphi$ as parameter, is represented in *figs. 3* and *4* respectively. From $q$ follows immediately the direct voltage: $V_0' = qV_{\max}—v_{\text{arc}}$. Since the arc voltage is generally small compared with the rectified voltage, the curves of fig. 4 may be taken as the so-called regulating characteristics, representing the trend of the rectified voltage as a function of the ignition angle.

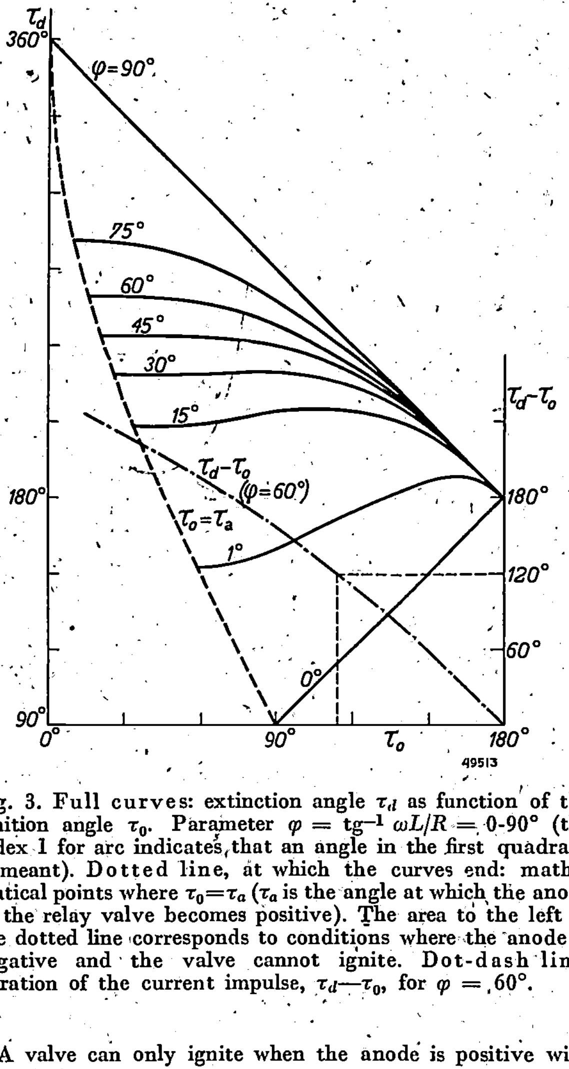


Fig. 3. **Full curves**: extinction angle $\tau_d$ as function of the ignition angle $\tau_0$. Parameter $\varphi = \text{tg}^{-1}\,\omega L/R = 0$-90° (the index 1 for arc indicates that an angle in the first quadrant is meant). **Dotted line**, at which the curves end: mathematical points where $\tau_0 = \tau_a$ ($\tau_a$ is the angle at which the anode of the relay valve becomes positive). The area to the left of the dotted line corresponds to conditions where the anode is negative and the valve cannot ignite. **Dot-dash line**: duration of the current impulse, $\tau_d—\tau_0$, for $\varphi = 60°$.

A valve can only ignite when the anode is positive with respect to the cathode. If the valve is a relay valve with a control grid then a second condition has to be satisfied for the ignition, *viz.* the grid voltage must lie above a certain critical value, which generally is a function of the momentary value of the anode voltage and which is only of importance when the anode voltage is positive. Now the anode voltage is only positive at the moment $t_a$ at which the transformer voltage exceeds the direct voltage at the condenser, thus at an angle of $\omega t_a = \tau_a$ in the first quadrant, for which applies: $V_{\max} \sin \tau_a = V_0'$, or $\sin \tau_a = V_0'/V_{\max}$, which is approximiately equal to $q$ when $v_{\text{arc}} \ll V_0'$. The valve cannot be ignited earlier than what corresponds to $\tau_a$. This explains why the curves in figs. 3 and 4 end on the dotted limit curves on the left-hand side, these limit curves being defined by $\tau_0 = \tau_a$; in fig. 4 this curve is part of a sine line.

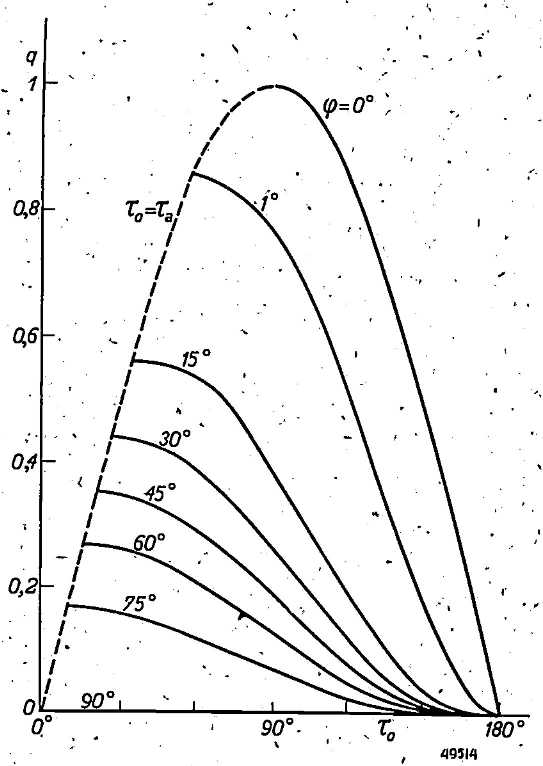


Fig. 4. The relation $q$ (approx. equal to the direct voltage $V_0'$ divided by the top value $V_{\max}$ of the alternating voltage) as a function of $\tau_0$, with $\varphi$ as parameter. The dotted line again corresponds to the earliest possible ignition ($\tau_0 = \tau_a$).

The factors ($\sin \tau_0—q$) and ($\sin \tau_d—q$) occurring in (2) and (3) respectively can now be calculated from figs. 3 and 4 as functions of $\tau_0$, with $\varphi$ as parameter. In this way we find the curves given in *figs. 5* and *6*. The size of the voltage impulses

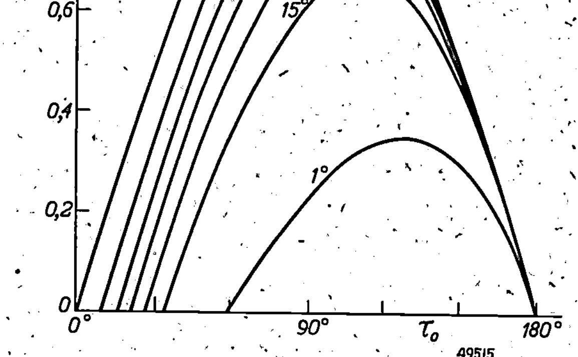


Fig. 5. The factor ($\sin \tau_0—q$) by which the voltage impulses upon **ignition** are proportional, as function of the ignition angle $\tau_0$, with $\varphi$ as parameter. Large values occur at $\varphi \approx 90°$ (strongly inductive circuit) and $\tau_0 \approx 90°$ (ignition at the top of the alternating voltage).

upon ignition and extinction is found by multiplying the values taken from *figs. 5* and *6* respectively by a factor of $V_{\max}L_a/(L_0 + L_a)$ or $V_{\max}L_0/(L_0 + L_a)$, according to whether we are concerned with the transformer coil or the choke coil.

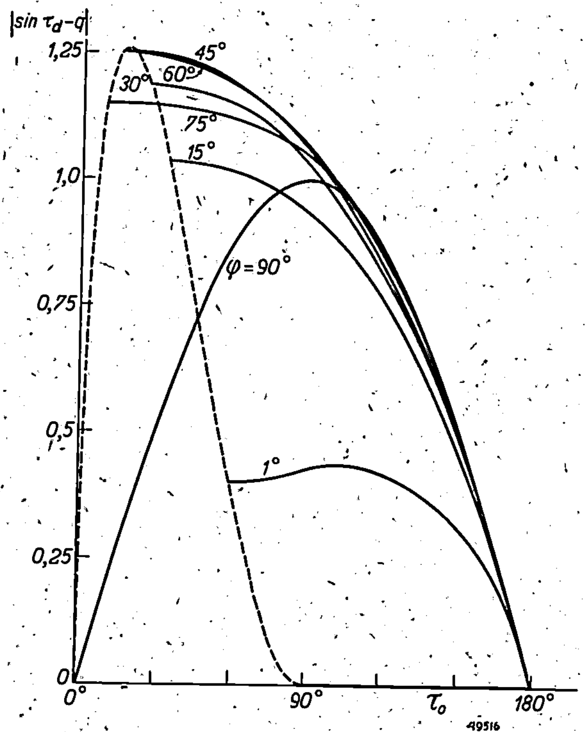


Fig. 6. The factor $|\sin \tau_d—q|$ by which the voltage impulses upon extinction are proportional, as function of the ignition angle $\tau_0$, with $\varphi$ as parameter. The curves end to the left on the dotted limit line, for which the equation (13) applies; this has a maximum of 1.26 at $\tau_0 = 18°$. The extinction impulses are large with many combinations of $\varphi$ and $\tau_0$ where $\varphi \approx$ 15°-90°, $\tau_0 \approx$ 15°-120°.

From fig. 5 it may be seen that the ignition impulses are large when in a strongly inductive circuit ($\varphi \approx$ 45°-90°) the ignition takes place roughly at the peak of the alternating voltage ($\tau_0 \approx 90°$). The factor $\sin(\tau_0—q)$ is maximum 1, which value occurs for $\varphi = 90°$, $\tau_0 = 90°$.

In fig. 6 the curves end at the left on the dotted limit curve, the equation for which reads[2]):

$$\cos \tau_d — \tau_d \cdot \sin \tau_a = \cos \tau_a — \tau_a \cdot \sin \tau_a. \quad (13)$$

The function given by (13) shows a maximum, amounting to 1.26, at $\tau_0 = 18°$, $q = \sin 18° = 0.31$, to which correspond $\tau_d = 252°$, $\varphi = 53°$. The voltage jumps occurring upon extinction of the valve are therefore greatest for this combination. With many combinations within wide ranges of $\varphi$ and $\tau_0$, however, — $\varphi \approx$ 15°-90°; $\tau_0 \approx$ 15°-120° — they are not much less than this maximum.

Finally *fig. 7* shows the trend of the quantities $v$, $v'$, $i$ and $\omega L\ \mathrm{d}i/\mathrm{d}\tau$ as functions of $\tau$ for the case where $\varphi = 30°$, $\tau_0 = 60°$. Here the two voltage jumps are clearly seen.

## Rectifier working on an *m*-phase system of alternating voltages

The system just described (fig. 1) can hardly be considered for a power supply to transmitters, because, among other reasons, in view of the strong pulsation of the rectified current the self-inductance $L_0$ and the capacity $C_0$ must have high values to reduce the ripple voltage at the output terminals

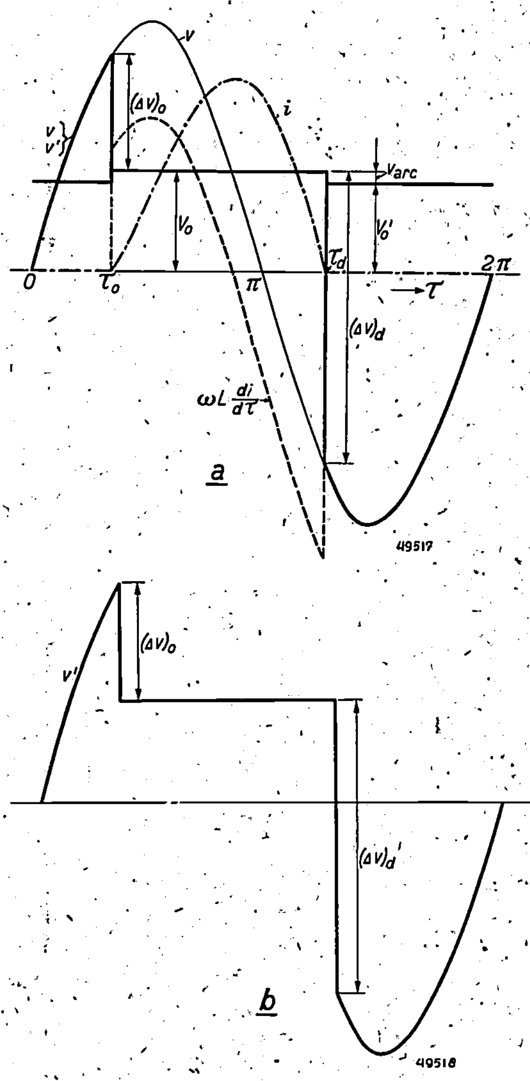


Fig. 7. Voltages and currents as functions of $\tau$ in a rectifier according to fig. 1, with $\varphi = 30°$, $\tau_0 = 60°$, to which correspond, according to fig. 3, $\tau_d = 229°$ and, according to fig. 4, $q = 0.39$.

*a*) Light sine line: feeding alternating voltage $v$; dot-dash line: current $i$; dotted curve: voltage at the self-inductance, $\omega L\,\mathrm{d}i/\mathrm{d}\tau$. It is here assumed that $L_0 = 0$, thus $L = L_a$. The terminal voltage $v'$ (heavy line) is then, during the passage of current ($\tau_0 < \tau < \tau_d$), equal to $v—\omega L \mathrm{d}i/\mathrm{d}\tau = V_0 = V_0' + v_{\mathrm{arc}}$; before and after the passage of current ($\tau < \tau_0$ respectively $\tau > \tau_d$) $v' = v$. At $\tau = \tau_0$ an ignition impulse $(\Delta v)_0$ arises in the terminal voltage, and at $\tau = \tau_d$ an extinction impulse $(\Delta v)_d$.

*b*) Here the trend of $v'$ is shown separately.

[2]) This follows from the relation (9) between $\tau_0$, $\tau_d$ and $q$ as applied to the limit case of the earliest possible ignition: $\tau_0 = \tau_a = \sin^{-1} q$, where the index for arc indicates an angle in the first quadrant.

below the admissible maximum. In such cases one always chooses a rectifying circuit fed from a symmetrical multi-phase system. The number of phases $m$ is usually 2, 3, 4 or 6. For $m = 3$ a circuit is given in *fig. 8*. The grid alternating voltages must likewise form a system of $m$-phases in the same order of succession as the system of anode voltages.

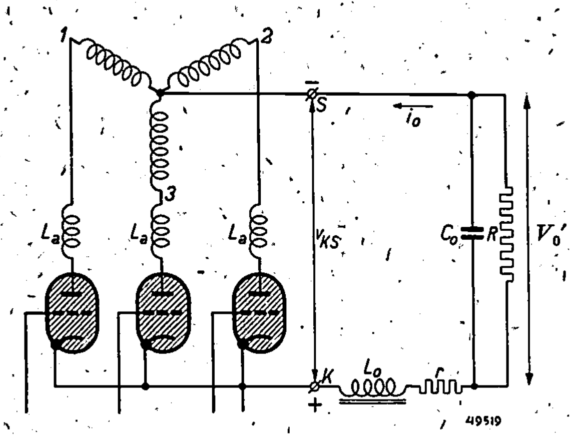


Fig. 8. Circuiting of a rectifier fed from a three-phase system of alternating voltages. $L_a$ = leakage inductance per branch.

*Simplified case: no leakage and only resistance in the external circuit*

We will begin the study of this circuit by considering a very simplified case. For the present we will assume that in the three-phase circuit of fig. 8 the self-inductances $L_a$ and $L_0$ and the capacity $C_0$ are zero, so that there is only the resistance $(R + r)$ in the external circuit. The transformer voltages are assumed to be sinusoidal, also when the system is loaded; the arc voltage is taken to be negligible.

We will start with a "late" ignition moment, *i.e.* with the ignition angle $\tau_0$ but little less than 180°, and then see what happens when the ignition is gradually advanced.

In *fig. 9* the light sine lines indicate the voltage variation of the anodes with respect to the star point $S$ (the negative pole). The heavy curves represent the trend of the voltage of the cathodes $K$ (the positive pole) with respect to $S$, in particular in the case ($a$) for $\tau_0 = 135°$. Considering that this voltage is proportional to the current flowing through the resistances $R$ and $r$, these curves likewise give a graphical representation of that current, which, as is seen, consists of three impulses per cycle. Through each of the valves a current flows consisting of only one such impulse per cycle.

If, now, we cause the ignition gradually to take place earlier then for $\tau_0 = 60°$ we reach the situation sketched in fig. 9*b*, where the dead interval between the impulses is reduced to zero, so that each of the impulses covers exactly 1/3 of the cycle. Upon $\tau_0$ being further reduced (fig. 9*c*) the duration of the impulse remains 1/3 of a cycle, this being explained in the following way. After the ignition of a valve, for instance that in the branch *1*, the voltage $v_1$ of the respective transformer coil is greater than the voltage $v_3$ supplied by the coil preceding *1* in phase (see fig. 9*c*, *e.g.* at $\tau'$). Since the higher voltage $v_1$ prevails also between the direct current terminals $S$ and $K$, the anode voltage of the valve in branch *3*, $v_3$—$v_1$, is negative. This valve must therefore have become dead as soon as the valve in branch *1* was ignited. In each valve of the rectifier the live period is thus exactly 1/3 cycle (in general $1/m$). The extinction angle $\tau_d$, which up to $\tau_0 = 60°$ was constant (180°), now becomes equal to $\tau_0 + 120°$.

This continues to apply until at $\tau_0 = 30°$ we get the situation outlined in fig. 9*d*. At $\tau < 30°$ (*e.g.* $\tau''$ in fig. 9*d*) the anode voltage of the valve in branch *1*, $v_1$—$v_3$, is still negative. Fig. 9*d* therefore represents the case where the ignition is advanced as far as possible.

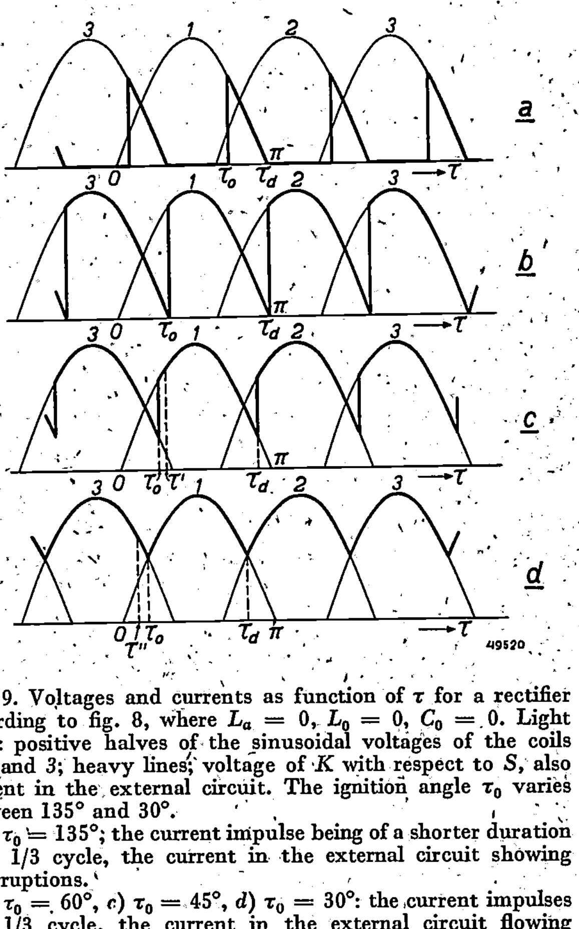


Fig. 9. Voltages and currents as function of $\tau$ for a rectifier according to fig. 8, where $L_a = 0$, $L_0 = 0$, $C_0 = 0$. Light lines: positive halves of the sinusoidal voltages of the coils *1*, *2* and *3*; heavy lines: voltage of $K$ with respect to $S$, also current in the external circuit. The ignition angle $\tau_0$ varies between 135° and 30°.

*a*) $\tau_0 = 135°$; the current impulse being of a shorter duration than 1/3 cycle, the current in the external circuit showing interruptions.

*b*) $\tau_0 = 60°$, *c*) $\tau_0 = 45°$, *d*) $\tau_0 = 30°$: the current impulses last 1/3 cycle, the current in the external circuit flowing without interruption. The case (*b*) is the border line between interrupted and continuous current. At (*d*) the ignition is as early as possible and the direct voltage obtained is as large as possible.

The area enclosed between the heavy curves in figs. 9*a*-*d* and a part of the $\tau$ axis of the length of one cycle is the greatest possible in fig. 9*d*. Since this area is a measure for the average voltage supplied by the rectifier, this voltage is highest with the earliest possible ignition.

One can therefore distinguish two working conditions of the rectifier according to the shape of the current in the external circuit:

1) that with interrupted current: duration of impulse $\tau_d - \tau_0 < 2\pi/m$; ignition "late" ($\tau_0 > \pi - 2\pi/m$);

2) that with continuous current: duration of impulse $\tau_d - \tau_0 = 2\pi/m$: ignition "early" ($\tau_0 < \pi - 2\pi/m$).

In this simple case where there is no self-inductance there are no voltage jumps at the transformer.

*The case with leakance and with self-inductance and inverse voltage in the external circuit*

We will now return to the circuit of fig. 8 more closely approaching the reality, where the leakage inductances $L_a$ and the equalising self-inductance $L_0$ and capacity $C_0$ can no longer be assumed to be zero. We will take $C_0$ of such a value that the voltage at the terminals of $C_0$ and $R$ may be regarded as a ripple-free direct voltage $V_0'$. We can again make the above-mentioned distinction between working conditions with interrupted current and with continuous current, although it is not so easy to indicate as in the case of fig. 9 at what value of $\tau_0$ the transition takes place between the two conditions. It will be evident, however, that with interrupted current the $m$ branches act entirely independently of each other, so that what has been deduced above for fig. 1 applies equally to each of them. Consequently ignition peaks again arise according to equation (2) and extinction peaks according to equation (3), such once per cycle at each self-inductance $L_a$, but $m$ times per cycle at the self-inductance $L_0$ (with a factor $L_0$ instead of $L_a$ in the numerator). The extinction angle is now no longer equal to $\pi$, as in the simple case of fig. 9*a*, but a function of $\tau_0$ and $\varphi$ as represented in fig. 3. Instead of $\tau_d$ one may plot the impulse duration $\tau_d - \tau_0$ as a function of $\tau_0$; in fig. 3 this has been done by way of example for one value of $\varphi$ (= 60°). When for a given value of $\varphi$, $\tau_0$ is reduced from $\pi$ onwards then the impulse duration $\tau_d - \tau_0$, increasing from 0, reaches at a certain point the value $2\pi/m$; here the successive valve currents join up to each other. The value of $\tau_0$ at which this happens is a function of the number of phases $m$ and of $\varphi$. From fig. 3 it is to be seen that for $m = 3$ and $\varphi = 60°$ the limit lies at $\tau_0 = 114°$.

Upon $\tau_0$ being still further reduced the impulse duration $\tau_d - \tau_0$ does not remain constant at $2\pi/m$ as in the leakance case of fig. 9 but may increase slightly. That this is possible is due to the leakage inductances $L_a$. Let us consider any one of the $m$ branches, which we will call *1*, at the moment that it takes over from the preceding branch (this then being the $m$th). Whereas in the case of $L_a = 0$ the current suddenly passes over from branch $m$ to branch *1* (*fig. 10a*), with infinite values of $L_a$ this commutation is gradual (fig. 10*b*): upon the ignition of the relay valve *1* the current $i_1$ in branch *1* begins to expand and the current $i_m$ in branch $m$ decreases. The commutation is ended when the diminishing current (in this case $i_m$) has reached zero.

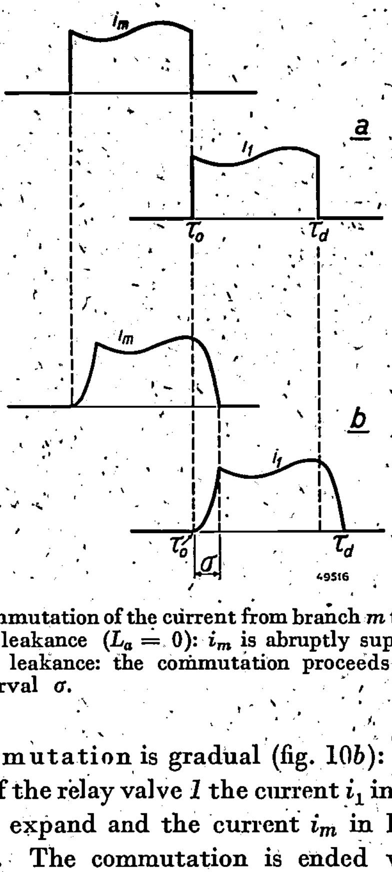


Fig. 10. Commutation of the current from branch *m* to branch *1*. *a*) without leakance ($L_a = 0$): $i_m$ is abruptly superseded by $i_1$; *b*) with leakance: the commutation proceeds gradually, in an interval $\sigma$.

To calculate the voltage jumps occurring we have to analyse the situation existing during the commutation. In order to avoid unnecessary complications we will again introduce an approximation which is usually justified in practice: we assume that the current $i$ in the external circuit is constant, that is to say free of ripple, and thus equal to the mean value $I_0$ of the rectified current.

During the commutation, which, expressed in angles, occupies an interval $\sigma$, the equation applying for the circuit formed by the transformer coil *1*, the valve *1*, the valve $m$ and the coil $m$ is

$$V_{\max}\sin\tau - \omega L_a\frac{\mathrm{d}i_1}{\mathrm{d}\tau} + \omega L_a\frac{\mathrm{d}i_m}{\mathrm{d}\tau} - $$

$$- V_{\max}\sin\left(\tau + \frac{2\pi}{m}\right) = 0, \quad (14)$$

$$(\tau_o < \tau < \tau_o + \sigma).$$

Since, further, $i_m = i - i_1$ and $\mathrm{d}i/\mathrm{d}\tau$ according to the approximation just introduced may be taken as equal to zero, after a small reduction it follows from (14) that:

$$\omega L_a\frac{\mathrm{d}i_1}{\mathrm{d}\tau} = - V_{\max}\sin\frac{\pi}{m}\cos\left(\tau + \frac{\pi}{m}\right). \quad (15)$$

During the commutation interval referred to the terminal voltage of coil *1* has a value of $V_{\max}\sin\tau - \omega L_a \mathrm{d}i_1/\mathrm{d}\tau$. Immediately before this interval there was no load on coil *1* and thus it had a terminal voltage of $V_{\max}\sin\tau$. Immediately after the interval it is loaded with the full current $i = I_0$, which, however, owing to its being constant does not give rise to any voltage loss in $L_a$; thus the terminal voltage of coil *1* is then also given by $V_{\max}\sin\tau$. The magnitude of the jumps at the beginning ($\tau_0$) and at the end ($\tau_0 + \sigma$) of the commutation from branch $m$ to branch *1* amounts according to (15) to:

$$\left.\begin{aligned}(\Delta v_1)_{\tau_o} &= V_{\max}\sin\frac{\pi}{m}\cos\left(\tau_o + \frac{\pi}{m}\right),\\ (\Delta v_1)_{\tau_o+\sigma} &= -V_{\max}\sin\frac{\pi}{m}\cos\left(\tau_o + \frac{\pi}{m} + \sigma\right).\end{aligned}\right\} \quad (16)$$

We notice that the ignition impulses are only dependent upon $\tau_0$ and not $L_a$; they are zero for $\tau_0 = \pi/2 - \pi/m$, *i.e.* with the smallest possible ignition angle. At the end of the commutation the impulses are indeed dependent upon $L_a$, since $\sigma$ increases with $L_a$; under normal conditions, however, $\sigma$ is a small angle (of the order of a few degrees), so that these impulses, except for their polarity, are approximately equal to those at the beginning of the commutation.

Equally large jumps but of opposite sign are found in the terminal voltage of the same coil *1* for the commutation from this branch to branch *2*. The terminal voltages of coils of the other $m$—1 branches show the same picture but displaced in phase a whole number of times $2\pi/m$.

We will now consider the voltage $v_{KS}$ on the external circuit, thus between the points $K$ and $S$ (fig. 8). This voltage has the mean value $(R+r)I_0$ but possesses also a large ripple component that is taken up by $L_0$. One might suppose that from the assumption of a ripple-free rectified current $i$ in the external circuit it would follow that $L_0\mathrm{d}i/\mathrm{d}t = 0$, but generally this is not so, because $i$ (a very special case excepted) can only be constant when $L_0$ is infinitely large. $L_0\mathrm{d}i/\mathrm{d}t$ therefore does differ from zero.

Immediately before the commutation from branch $m$ to branch *1* the voltage between $K$ and $S$ is

$$v_{KS} = v_m \quad \text{(17a)}$$

and after it:

$$v_{KS} = v_1 \quad \text{(17b)}$$

During the commutation $v_{KS} = v_1 - \omega L_a \mathrm{d}i_1/\mathrm{d}\tau = v_1 - {}^1\!/_2\,\omega L_a\,\mathrm{d}(i_1 - i_m)/\mathrm{d}\tau$, because $i_m = i - i_1$. Substituting here for $\mathrm{d}(i_1 - i_m)/\mathrm{d}\tau$ the value obtained from (14), we get

$$v_{KS} = {}^1\!/_2\,(v_1 + v_m) \quad \text{(17c)}$$

thus exactly the average of the voltages of the two branches concerned in the commutation.

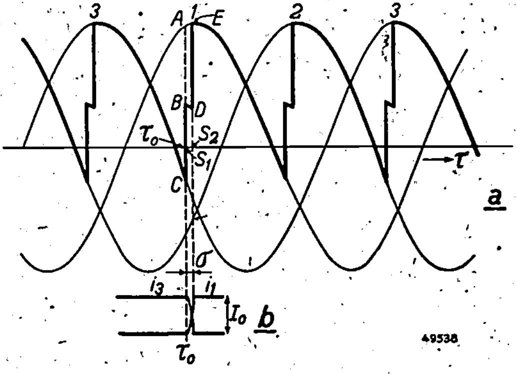


Fig. 11. *a*) voltages, *b*) currents, represented as functions of $\tau$, for a rectifier according to fig. 8, with $\tau_0 = 75°$. Heavy lines in *a*): voltage of $K$ with respect to $S$. At the beginning of the commutation from branch *3* to branch *1* the following voltage impulses arise: $AB$ in the terminal voltage $v_1'$, $CB$ in the terminal voltage $v_3'$ and also in the voltage $v_{KS}$. At the end of the commutation $v_1'$ and $v_{KS}$ make a surge $DE$ and $v_3'$ makes a surge $D$ in the curve *3*.

This is all made clear by *fig. 11*, which shows the trend of voltage and of current as functions of $\tau$ for the case where $m = 3$, $\tau_0 = 75°$. The heavy line in fig. 11*a* represents the voltage $v_{KS}$ according to the equations (17a, c and b). Fig. 11*b* illustrates the current commutation from branch *3* to branch *1*. (The shape of the currents is found by integrating eq. (15) and making use of the boundary conditions $i_1 = 0$ for $\tau = \tau_0$; $i_1 = I_0$ for $\tau = \tau_0 + \sigma$.) At the moment that the relay valve *1* is ignited ($\tau = \tau_0$) the terminal voltage of coil *1* drops by an amount $AB$ from $S_1A$ to $S_1B$ and the voltage $v_{KS}$ rises by an equally large amount $BC$ from $S_1C$ to $S_1B$. At the end of the commutation from branch *3* to branch *1* ($\tau = \tau_0 + \sigma$) both the voltages mentioned rise from $S_2D$ to $S_2E$.

The alternating components of the pulsating voltage $v_{KS}$ are entirely absorbed by $L_0$ (if we again ignore $r$). Consequently the voltage impulses found at $L_0$ are just as large as those at $L_a$ (though sometimes of a different sign), contrary to the case with interrupted current, where the simultaneous voltage jumps at $L_a$ and $L_0$ are in the ratio of $L_a : L_0$. This is due to the fact that at the moments of the jumps with interrupted current we have the same value of $di/dt$ at $L_a$ as at $L_0$, whereas with non-interrupted current these values differ.

**Influence of parasitic capacities**

In the foregoing we have shown that in the normal working of rectifiers voltage impulses occur at the self-inductances. The results found have been recapitulated in *table I*. It is to be noted that not only have some expressly mentioned approximations been introduced but that we have also tacitly passed over the various details. For instance we have not considered conditions where more than two valves are working simultaneously. Impedances on the primary side of the transformer have been assumed to be zero; nothing has been said about the manner in which the primary coils are circuited (in delta, in star, etc). More complicated connections, such as those of Grätz, which are often applied in practice (see for instance the first of the articles cited in footnote [1]), have not been dealt with, because a closer investigation shows that they may be replaced by equivalent circuits of the type in fig. 8. There is no need to go into all these finesses here.

*Table I.* Summary of the magnitude of the voltage impulses on $L_a$ and $L_0$ in the absence of parasitic capacities. No attention has been paid to the polarity because this is of no consequence for our considerations.

| Working condition | | Ignition | Extinction |
|---|---|---|---|
| Interrupted current | In $L_a$: | $\frac{L_a}{L_0 + L_a}$ (sin $\tau_o - q$) $V_{\max}$ | $\frac{L_a}{L_0 + L_a}$ (sin $\tau_d - q$) $V_{\max}$ |
| | In $L_0$: | $\frac{L_0}{L_0 + L_a}$ (sin $\tau_o - q$) $V_{\max}$ | $\frac{L_0}{L_0 + L_a}$ (sin $\tau_d - q$) $V_{\max}$ |
| Continuous current | In $L_a$: | $\sin\frac{\pi}{m}\cos\left(\tau_o + \frac{\pi}{m}\right) V_{\max}$ | $\sin\frac{\pi}{m}\cos\left(\tau_o + \frac{\pi}{m} + \sigma\right) V_{\max}$ |
| | | and equally large impulses a time $\frac{2\pi}{\omega m}$ later | |
| | In $L_0$: | The same as in $L_a$ | |

A consideration of the formulae given in table I will show that none of the voltage impulses can be much greater than $V_{\max}$ and that in most cases they may even be much smaller. Notwithstanding these impulses, the potential difference between the ends of a secondary transformer coil is never greater than $V_{\max}$. Nevertheless, disruptions have occurred in transformers whose insulation should be taken to be absolutely proof against a top voltage $V_{\max}$. How are these to be explained? The answer is to be sought in a factor not yet taken into account in our considerations, namely the presence of parasitic capacities.

These capacities affect the situation in various respects, although generally speaking they are so small that they need not be considered in the equation (1) for the main current $i$. In the first place they form oscillating circuits with the self-inductances ($L_a$ or $L_0$) to which they are connected in parallel. These circuits are excited by the voltage impulse and oscillate in their own frequency, which is usually much higher than the mains frequency. Consequently it is no longer correct to say, as in the preceding paragraph, that the voltage across a transformer coil cannot be greater than $V_{\max}$. In a fraction of a cycle of the mains frequency the voltage on the oscillating circuit is opposite in polarity and combines with the momentary value of the transformer voltage. *Fig. 13* gives the oscillogram of the circuit of *fig. 12*, where the oscillations are clearly seen. Since the voltage peak is rapidly repeated a number of times this involves the risk of a step-like disruption.

A second effect of the presence of the parasitic capacities is their influence on the distribution of the voltage impulses, in particular the distribution among the individual windings of the secondary transformer coils or of the choke coil. The fact that a voltage impulse may be distributed very unevenly among the windings was known already in the times when it first became the practice to employ overhead cables for high tensions and it

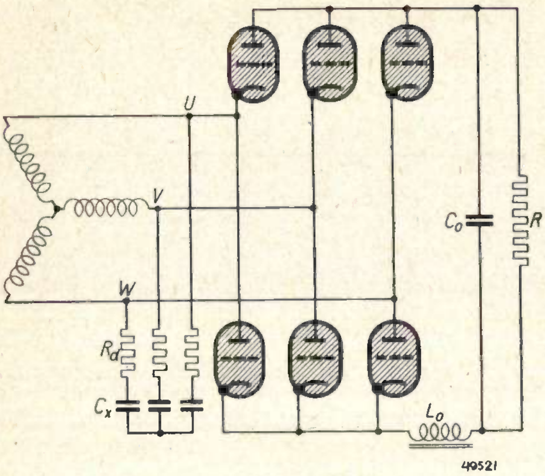


Fig. 12. Circuit according to Grätz, the voltage on which between the points *U* and *V* is represented in the oscillograms of figs. 13 and 16. The functions of the capacitors $C_x$ and the damping resistance $R_d$ will be discussed farther on.

was found that when struck by lightning the insulation of the outermost windings of the transformers connected to these overhead cables most frequently showed the worst damage [3]).

To get an idea of this effect it does not suffice simply to regard the parasitic capacities as capacities connected in parallel to the coils. Rather we have to consider that adjacent windings have a capacity both mutually and in respect to earth. Thus we arrive at a replacement system as given in *fig. 14*, where these capacities are indicated by *K* and *C* respectively. The result is that a sudden change in voltage at the end *a* of the coil does not by any means at first distribute itself evenly over the whole coil, but that the windings at the end *a* receive a disproportionately large part of the voltage impulse, so that there the insulation is endangered. The voltage distribution is all the more irregular according as the voltage at the point *a* changes more quickly.

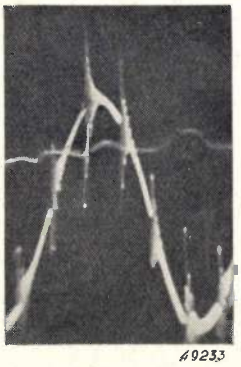


Fig. 13. Photo of an oscillogram of the voltage between *U* and *V* in a rectifier according to fig. 12, for $C_x = 0$ and $R_d = 0$. The voltage surges occurring when commuting excite oscillating circuits formed by leakage inductance and parasitic capacity.

That there is no need to worry about the polarity of the voltage impulse (compared with the sign of $V_{\max} \sin \tau$ at the moment at which the impulse occurs) will be evident, in view of the fact that a few windings between which there is normally only a small fraction of the voltage $V_{\max} \sin \tau$ are loaded with the full voltage impulse.

It is not only this "internal" distribution of the impulse that is affected by the capacities, but also the fractions which the choke coil or the transformer each take for their account. In the non-capacity case, with interrupted current, these fractions were $L_a/(L_0 + L_a)$ and $L_0/(L_0 + L_a)$. Without going into it more closely here it will be evident that particularly at the first moment the capacities may entirely change these ratios.

## Precautions against the danger of voltage impulses

We will now discuss some measures that should be taken to minimise the danger of voltage impulses.

A most obvious means is the extra insulating of the outermost windings of the secondary transformer coils, known from olden times, but this is not always practicable. The aim will therefore be to keep the impulses themselves as small as possible. It will then be necessary to distinguish between the condition existing while the transmitter is being started up, when the direct voltage is gradually raised from zero, and the normal working condition when the direct voltage is practically the maximum that can be supplied by the rectifier. As the direct voltage is increased, the ignition angle $\tau_0$ changes, passing through a series of values from about $\pi$ to a certain minimum, thereby passing the value where the condition of interrupted current changes into that of a continuous current. Let us first give attention to the normal working condition where the current is not interrupted. At the beginning and end of each commutation impulses occur of an amplitude which according to eq. (16)

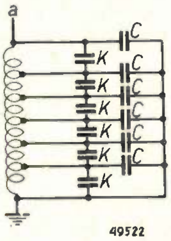


Fig. 14. Replacement diagram of a coil with parasitic capacitors *K* between adjacent windings and parasitic capacitor *C* of the windings with respect to earth.

[3]) See for instance K. W. Wagner, Das Eindringen einer elektromagnetischen Welle in eine Spule mit Windungskapazität, E.u.M., 33, 89-92 and 105-108, 1915.

is proportional to cos $(\tau_0 + \pi/m)$ and cos $\{\tau_0 + (\pi/m) + \sigma\}$ respectively.

As already remarked, the impulses occurring at the beginning of the commutation become zero when $\tau_0 = \pi/2 - \pi/m$, thus when the ignition takes place as early as possible, *i.e.* when the direct voltage is as high as possible. It is therefore advisable to design the rectifier in such a way that this highest direct voltage is at the same time the voltage required. The voltage impulses at the end of the commutation are only small when the ignition takes place as early as possible, being proportional to cos $(\pi/2 + \sigma)$; $\sigma$ is usually an angle of only a few degrees.

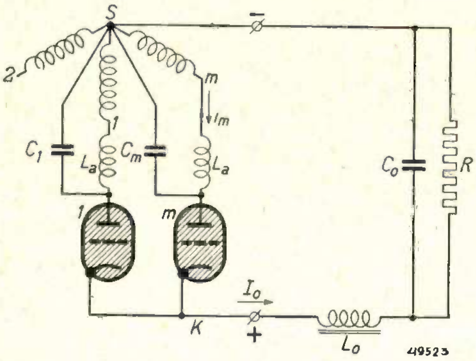


Fig. 15. Two in phase successive branches ($m$ and $1$) of a rectifier with $m$ phases provided with capacitors $C_1 \dots C_m$. These capacitors reduce the ignition peaks.

Circumstances may arise, however, which make it necessary to use temporarily a lower working voltage, in which case adjustments have to be made to less favourable values of $\tau_0$. Furthermore, these unfavourable $\tau_0$ values occur every time the voltage is stepped up. Precautions will therefore have to be taken to guard against the accompanying voltage impulses.

The means available, apart from the extra insulation already mentioned, for rendering these voltage impulses harmless may be divided into three groups:

a) condensers parallel to the transformer,
b) chokes in series with the valves,
c) a combination of a) and b),
which should be supplemented with
d) damping resistances.

### a) *Condensers parallel to the transformer*

An effective means of attenuating the commutation impulses is to introduce condensers ($C_1 \dots C_m$ in *fig. 15*) between the star point and the anodes of the valves. Just before the valve *1* is ignited the condenser $C_1$ has a higher voltage than the condenser $C_m$ in the preceding branch. Upon the valve *1* being ignited, a sudden discharge takes place from $C_1$, mainly because of the circulation of a current through the circuit formed by $C_1$, the valves *1* and $m$, and $C_m$. This circuit has practically no self-inductance nor resistance; the current mentioned — which in the valve $m$ is opposed to the current $i_m = I_0$ still flowing there — will therefore rise very rapidly and almost immediately reach the value $I_0$. The current in valve $m$ will then have become zero; it cannot reverse its direction and the valve $m$ is therefore extinguished; the current $i_m$ in the transformer then flows to $C_m$. As a result the voltage on this condenser $C_m$ soon rises high enough for the valve $m$ to be ignited again, after having been dead for only a small fraction of a cycle. During this second current impulse, which gradually drops to zero, the valves $m$ and *1* work simultaneously. Given a sufficiently large

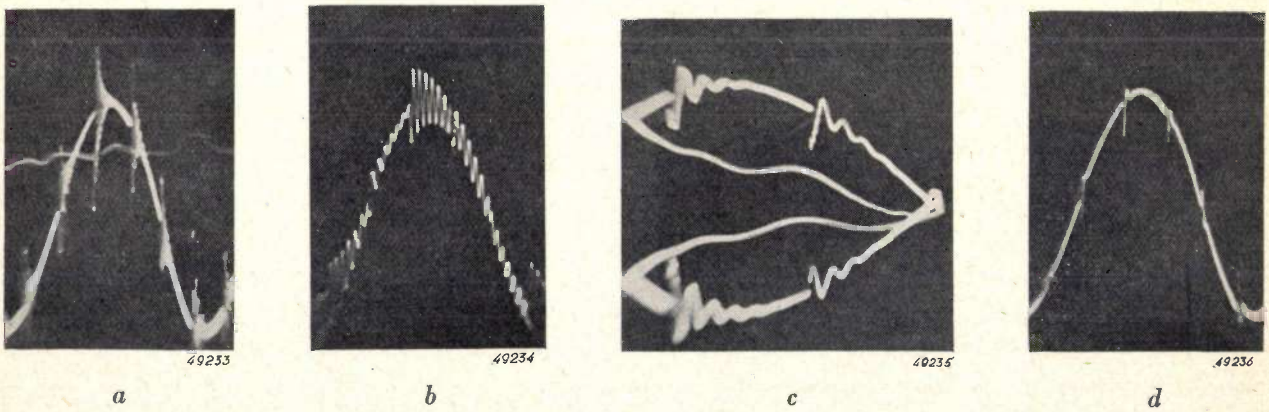


Fig. 16. Oscillograms of the voltage between $U$ and $V$ in the circuit of fig. 12, for various values of $C_x$ and $R_d$.
*a*) $C_x = 0$. $R_d = 0$: high voltage peaks with a steep front, followed by a damped oscillation.
*b*) $C_x = 7000$ pF, $R_d = 0$: lower, less steep peaks, followed by a damped oscillation.
*c*) $C_x = 7000$ pF, $R_d = 3000$ ohm: front steeper than for (*b*); but damping much greater. (N.B. Time scale differs from that of the other oscillograms, so that the negative half wave is pictured underneath the positive half).
*d*) $C_x = 7000$ pF, $R_d = 10\,000$ ohms: steep front; damping still greater than for (*c*).

capacity, $C_x$ of the condensers $C_1 \dots C_m$, the voltage surges will be smaller than the normal commutation impulses which would arise at $L_a$ if these condensers were not present. This may be seen from the oscillograms of *figs. 16a* and *b* taken from the circuit of fig. 12 for different values of $C_x$: in (b), where $C_x = 7000$ pF, the largest amplitude of the superposed oscillation is less than that in (*a*), where $C_x = 0$. Of course there is a limit to the value chosen for $C_x$, because account has to be taken of the cost of these condensers and the load they constitute for the transformer.

The oscillations seen in fig. 16*b* occur in the circuit formed by $L_a$ and $C_x$. A rough calculation shows that the voltage amplitudes of these oscillations are approximately given by $(1/R) \cdot \sqrt{L_a/C_x} \cdot V_0$. Allowing for a maximum value equal to $kV_0$, then it follows that

$$C_x \geqq L_a/k^2R^2,$$

in which $k$ may amount, for instance, to 0.2.

The upper limit for $C_x$ follows from the consideration that for reasons of economy it is advisable to keep the component with the mains frequency of the currents flowing through $C_1 \dots C_m$ small in amplitude compared with $I_0$, say at a maximum of $hI_0$. Hence it follows:

$$C_x \leqq \frac{h}{\omega R},$$

with $h$, for instance, $= 0.1$.

b) *Choke coils in series with the valves*

Another method of safeguarding the transformer consists in connecting up a choke coil (self-inductance $L_x$) between the ends of each of the secondary transformer coils and the anode of the corresponding valve. In the absence of parasitic and the other extra capacities mentioned under (*a*), then only the fraction $L_a/(L_x + L_a)$ of the voltage impulse excited in the anode circuit will come to lie on the transformer. Actually, however, the parallel capacities that are always present will govern the distribution of the voltage surge. This is all the more favourable, because the parasitic capacities of the choke coil can easily be kept in relation to those of the transformer coil, for instance by making the choke coil without a core and with a greater length than its diameter. By this means it is also easy to make it capable of withstanding the voltage impulses, of which it now has to bear the lion's share.

c) *Combination of parallel condensers and series choke coils*

It is an obvious solution to combine methods (a) and (b). Increasing the parasitic transformer capacity by connecting a capacity $C_x$ in parallel enhances the action of the choke coils, which consequently can be made smaller and cheaper.

d) *The application of damping resistances*

As seen from figs. 16*a* and *b*, owing to the introduction of the capacities $C_x$ the parasitic oscillating circuits die out more gradually. Since in general these oscillations are undesirable, an endeavour will be made to provide for ample attenuation. For the circuit of which $L_a$ forms a part this can be done by connecting the resistance either in series with $L_a$ or in series with $C_x$, or parallel to $C_x$. The first would involve larger losses, since the main current would flow through the resistance. The second is less satisfactory because a resistance in series with $C_x$ would prevent this capacitor from supplying the sudden current impulse required for the effect aimed at. This can be seen in figs. 16*c* and *d*, compared with fig. 16*b*: owing to the introduction of the resistance the damping is increased, but there are again stcep fronts, which are absent in fig. 16*b*.

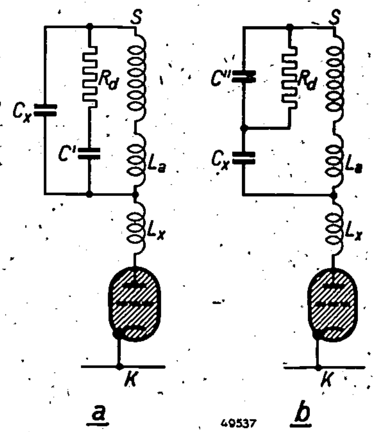


Fig. 17. Two methods of providing the oscillating circuit originally consisting of $L_a$ and $C_x$ with a damping resistance $R_d$ without involving any large loss of energy. In *a*) in series with $R_d$ is a capacitor $C'$ the reactance of which at the frequency $f_0$ of the circuit is of the order of $R_d$, but at the mains frequency $f$ large compared with $R_d$. In *b*) $C''$ is so chosen that at the frequency $f_0$ ample damping is provided by $R_d$ and at the frequency $f$ a voltage distributor is formed by $C_x$ and $C''$ which reduces the voltage with this frequency at $R_d$.

The connecting of a resistance in parallel with $C_x$ has the drawback that again large losses would occur in the damping resistance, owing to the high tension with the mains frequency present at $C_x$. This can be avoided by applying the methods of *figs. 17a* or *b* as explained in the text below those diagrams.

The oscillating circuit formed by the choke coil $L_x$ with its parasitic capacity can be damped by connecting a resistance to it in parallel. In the case of the filter choke coil it will be preferable to use a damping resistance in series with a capacitor.

The measures discussed here have been successfully applied in several transmitters, with the result that former troubles from breakdowns owing to disruption of insulations no longer occur.

# PHOTOGRAPHING COOLING CURVES OF HARDENING OILS BY MEANS OF A CATHODE-RAY OSCILLOGRAPH [1])

621.785.65:
621.317.35

Steel is usually hardened by first heating it to about 800 °C and then rapidly cooling it by immersion in a liquid (quenching). In this cooling process three successive stages can be distinguished:

1) slow cooling, owing to the metal being more or less insulated by a vapour layer around it;
2) rapid cooling, the vapour having disappeared and the heat being bound by the boiling of the liquid immediately around the metal;
3) slow cooling, after the temperature has dropped to below the boiling point.

For the success of the hardening process it is of great importance that the first two phases should be of short duration to prevent conversion of the austenite. Furthermore, the boiling point of the hardening liquid should be about 400 °C.

It appears that these two requirements can be met by using colza oil, but this, like all vegetable as well as animal oils, has the disadvantage that it changes in composition under the action of oxygen, becoming heavy and sticky and losing much of its originally good cooling property. For this reason mineral oils are usually preferred, although the hardening process is less satisfactory than with colza oil, their greater chemical durability being the decisive factor. Tests have shown that colza oil absorbs more than six times as much oxygen in a given time than does ordinary mineral oil.

The question then arose whether a mixture could not be compounded with a mineral oil as the main constituent and having cooling properties closely approximating those of colza oil but without its drawbacks.

First of all a method of measuring has to be thought out for studying the course of the cooling process. It was found that the method described below proved to be quite satisfactory.

Use was made of a solid silver ball 20 mm in diameter, inside which is a thermo-couple to which the heat of the silver is well conducted. The ball is heated to about 800 °C, the temperature being measured by means of a millivolt meter connected to the thermo-couple. The ball is then immersed in the oil to be tested for its cooling properties. In principle it would then only be necessary to record the reading of the millivolt meter at short intervals and to plot them, converted into temperature equivalents, as a function of time. However, as will be seen presently, the process takes place rather too quickly to allow of this being done with the necessary degree of accuracy; certain phases sometimes last no more than 2 or 3 seconds. This difficulty has been solved by visually recording the electromotive force of the thermo-couple on the screen of a cathode-ray oscillograph and photographing the picture.

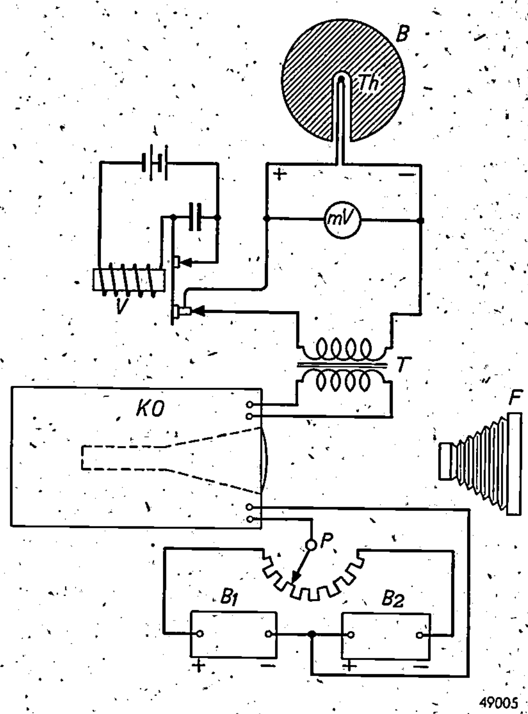


Fig. 1. $B$ = silver ball which, after heating to about 800 °C, is immersed in the hardening medium to be tested. $Th$ = thermo-couple, whose e.m.f. is read from a millivoltmeter and converted into an alternating voltage by means of a vibrator $V$ and a transformer $T$. This voltage is conducted to the amplifier for the vertical deflection of the cathode-ray oscillograph $KO$. The voltage for the time base is supplied by means of a potentiometer $P$, whose arm makes one stroke in the time required (*e.g.* 25 secs) and which is connected to the two batteries $B_1$ and $B_2$. $F$ = camera for photographing the oscillograms.

Since the thermo-electric potential is too low to give a direct indication on the oscillograph by direct connection to the plates, it has to be amplified. So as to be able to use the normal amplifier of the oscillograph for amplifying this direct voltage, the latter is converted into an alternating voltage with the aid of a vibrator and a transformer, as illustrated in *fig. 1*; the transformer supplies the voltage conducted to the amplifier of the oscillograph and

[1]) The data contained in this article have been taken from an article by B. Levy of Aktiebolaget Wahlén and Block, Stockholm, published in "Electronic Measuring" (1, No. 4, 1946), a Philips periodical dealing with the applications of electronic-measuring technique. The investigations referred to were carried out in cooperation with Svenska Aktiebolaget Philips, Stockholm.

causing the vertical deflection. To get reliable results the vibrator has to be of a robust construction and attention has to be paid to the contacts. Tungsten contacts proved to be unsatisfactory. An alloy of 93% gold and 7% platimun [2]) was found to be highly suitable. Furthermore, the circuit feeding the vibrator has to be well shielded to avoid induction of disturbing voltages in the measuring circuit.

As to the horizontal deflection, the time base of the usual oscillographs cannot be used for this because the phenomenon to be recorded may last 30 to 60 seconds. For this reason the method shown in fig. 1 was employed, where the arm of a potentiometer is rotated by a gramophone motor. Between the motor and the potentiometer spindle is a variable gearing and also an electromagnetic coupling, which, *via* a relay, starts up the potentiometer at the same moment that the silver ball falls into the cooling medium.

Upon the vibrator being started a vertical line appears on the screen of the oscillograph on the extreme left when the potentiometer is in its initial position. The rotation of the potentiometer arm causes this line to change into a more or less sinusoidal line, which gradually diminishes in amplitude as the silver ball cools down. The peaks thus form two symmetrical cooling curves. The picture can easily be photographed, as may be seen from *fig. 2*.

Fig. 2*a* shows the cooling curve of colza oil. The maximum vertical deflection corresponds to 800 °C and the horizontal movement to 23 seconds. One can clearly distinguish here the three phases mentioned in the beginning; the first and second phases, before the temperature drops to about 400 °C, each lasting from 2 to 3 seconds, which is satisfactory for the hardening of steel. In the case of an ordinary mineral oil, however, the first phase takes 12-13 seconds (fig. 2*b*), which is too long for proper hardening.

The recording of a large number of such curves as these has resulted in hardening oils being compounded which have better cooling curves than have been known hitherto. Fig. 2*c* shows the curve for such an oil.

Finally, attention is drawn to another factor of great importance in hardening processes. In this technique it is generally known that hardening oil should be free of water. The trend of the cooling curves clearly shows the effect of traces of water in the oil. Compare the curve of a water-free mineral oil (fig. 2*b*) with that of the same oil but containing 0.2% water (fig. 2*d*); in the latter case the cooling in the second phase is much more intensive, and with such intense cooling cracks in the workpiece will almost inevitably result.

Thus the application of the cathode-ray oscillograph with simple accessories has not only resulted in a better product — here a hardening oil — but it has also made it possible to carry out a simple and sensitive quality test.

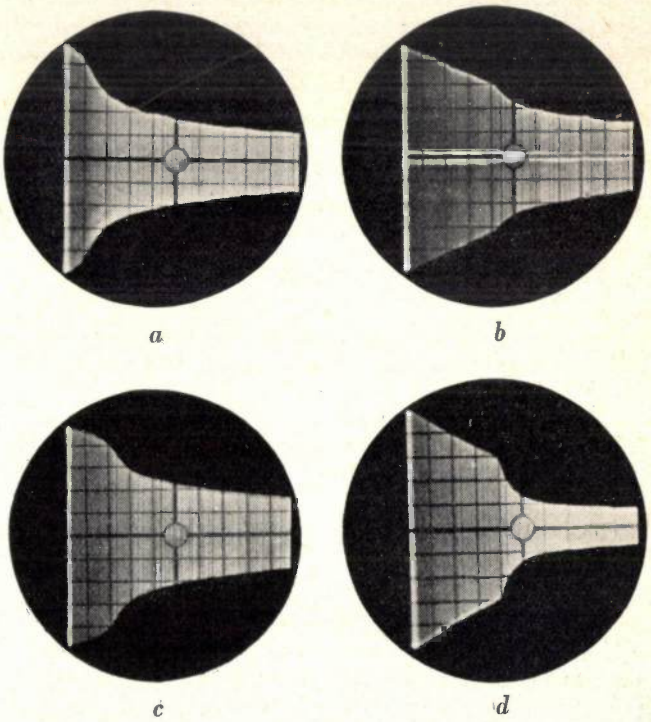


Fig. 2. Photographs of cooling curves obtained with different oils. The max. vertical deflection corresponds to 800 °C and the horizontal to 23 secs.

*a*) Colza oil. The first and second phases (respectively slow and rapid cooling) take together about 5 secs, which is favourable for hardening steel. Owing to their lack of durability, however, vegetable oils are not suitable in practice.

*b*) A mineral oil. The first phase takes too long, about 13 secs, for steel hardening.

*c*) The investigations resulted in the compounding of a mixture, with a mineral oil as the main constituent, which combines a satisfactory cooling curve with great durability.

*d*) Traces of water in the oil affect the cooling curve: the same oil which in the dry state gave the cooling curve (*b*), when 0.2% water is added gives the curve (*d*); in the latter case the cooling in the second phase is so rapid as to cause cracks in the steel workpiece.

[2]) Benedicks and Härdén, Z. techn. Phys. 13, 71, 111 166, (particularly 113), 1932.

## IN MEMORIAM Dr. P. J. BOUMA.

We should like to pay tribute to the memory of Dr. Pieter Johannes Bouma, born in Amsterdam on 14th April 1908, who died at Eindhoven on 19th January 1947. Bouma was connected with the Philips Laboratory in Eindhoven as an assistant from 1928 until 1930 and as a research physicist from 1933 until his death. From 1930 until 1933 he studied at Utrecht, where in November 1933 under Ornstein he received his degree on his thesis: "Beitrag zur Dynamik der flüssigen Kristalle". Bouma specialized on the theory of illumination and colorimetry, in which fields various publications from his pen have appeared in Physica and in the Proceedings of the "Kon. Ned. Akad. van Wetenschappen". He also wrote many articles for the Philips Technical Review. Furthermore he rendered valuable services to the Netherlands Foundation for Illumination Technology (Ned. Stichting v. Verlichtingskunde).

Since 1940 Bouma had been suffering from a disorder of the central nervous system, which, though it fortunately did not affect his acute intellect, made speaking and writing more and more difficult for him. Nevertheless, between September 1944 and April 1945 he was able to complete the manuscript of his book "Colour Stimuli and Colour Sensations", as well as three other publications.

During 1946 his strength failed him, but he still had the satisfaction of seeing his book in print (it was published in November 1946, and a review of it is given on page 159 of this number).

The following article is a practically unaltered version of a lecture given by Bouma at Utrecht in 1940 for a Vacational Course on "Road and Street Lighting".

THE EDITORS.

# PERCEPTION ON THE ROAD WHEN VISIBILITY IS LOW

by P. J. BOUMA †. 628.971.6:612.843.6

This article deals with the factors affecting perception, when visibility is low, especially as applied to road lighting. The author discusses in turn: contrast sensitivity, distinguishability of the object, speed of perception, glare, uniformity of illumination, the Purkinje effect, and the influence of colours upon perception.

The purpose of road lighting is to make it possible to see objects on the highways. Therefore in planning their illumination we have to take into account the properties of the eye. As will be shown later, the performance of the eye depends very closely upon the level of brightness: generally speaking, the higher this level the better we can see. Therefore it would be desirable to employ very high intensities of illumination on the highways, but this is limited by considerations of economy and in general we are not able to go beyond a few score lux. Compared with the luminous intensities of several times ten thousand lux that we often have during the daytime, this is extremely low. The fact that we can distinguish anything at all is due to the wonderful adaptability of the eye (expansion of the pupil, adaption of the retina); anyhow our vision is still much poorer than it should be with high luminous intensities. Hence it follows that whatever light is available must be used very economically. In other words:

1) We must try to get as much energy as possible (usually electrical energy) for our money's worth.
2) We must get as many lumens as possible per kilowatt input, *i.e.* we must use lamps with a high efficiency.
3) The lumens obtained must be projected on the highway with little loss and in the best possible manner; this means that the most suitable fittings must be used, taking into account the properties of the road surface.
4) In the designing of the lighting system allowance must always be made for the properties of the eye, giving consideration to the two following questions:
   a) under what conditions does the eye function best, and
   b) in how far can these conditions be realized?

# IN MEMORIAM Dr. P. J. BOUMA.

We should like to pay tribute to the memory of Dr. Pieter Johannes Bouma, born in Amsterdam on 14th April 1908, who died at Eindhoven on 19th January 1947. Bouma was connected with the Philips Laboratory in Eindhoven as an assistant from 1928 until 1930 and as a research physicist from 1933 until his death. From 1930 until 1933 he studied at Utrecht, where in November 1933 under Ornstein he received his degree on his thesis: "Beitrag zur Dynamik der flüssigen Kristalle". Bouma specialized on the theory of illumination and colorimetry, in which fields various publications from his pen have appeared in Physica and in the Proceedings of the "Kon. Ned. Akad. van Wetenschappen". He also wrote many articles for the Philips Technical Review. Furthermore he rendered valuable services to the Netherlands Foundation for Illumination Technology (Ned. Stichting v. Verlichtingskunde).

Since 1940 Bouma had been suffering from a disorder of the central nervous system, which, though it fortunately did not affect his acute intellect, made speaking and writing more and more difficult for him. Nevertheless, between September 1944 and April 1945 he was able to complete the manuscript of his book "Colour Stimuli and Colour Sensations", as well as three other publications.

During 1946 his strength failed him, but he still had the satisfaction of seeing his book in print (it was published in November 1946, and a review of it is given on page 159 of this number).

The following article is a practically unaltered version of a lecture given by Bouma at Utrecht in 1940 for a Vacational Course on "Road and Street Lighting".

THE EDITORS.

# PERCEPTION ON THE ROAD WHEN VISIBILITY IS LOW

by P. J. BOUMA †.

628.971.6:612.843.6

This article deals with the factors affecting perception, when visibility is low, especially as applied to road lighting. The author discusses in turn: contrast sensitivity, distinguishability of the object, speed of perception, glare, uniformity of illumination, the Purkinje effect, and the influence of colours upon perception.

The purpose of road lighting is to make it possible to see objects on the highways. Therefore in planning their illumination we have to take into account the properties of the eye. As will be shown later, the performance of the eye depends very closely upon the level of brightness: generally speaking, the higher this level the better we can see. Therefore it would be desirable to employ very high intensities of illumination on the highways, but this is limited by considerations of economy and in general we are not able to go beyond a few score lux. Compared with the luminous intensities of several times ten thousand lux that we often have during the daytime, this is extremely low. The fact that we can distinguish anything at all is due to the wonderful adaptability of the eye (expansion of the pupil, adaption of the retina); anyhow our vision is still much poorer than it should be with high luminous intensities. Hence it follows that whatever light is available must be used very economically. In other words:

1) We must try to get as much energy as possible (usually electrical energy) for our money's worth.
2) We must get as many lumens as possible per kilowatt input, *i.e.* we must use lamps with a high efficiency.
3) The lumens obtained must be projected on the highway with little loss and in the best possible manner; this means that the most suitable fittings must be used, taking into account the properties of the road surface.
4) In the designing of the lighting system allowance must always be made for the properties of the eye, giving consideration to the two following questions:
   a) under what conditions does the eye function best, and
   b) in how far can these conditions be realized?

The latter problem is often difficult to solve because there are a number of factors over which the illuminating engineer has no control (reflection of the road surface and the objects upon it, surroundings of the road, size of the sources of light, etc.)

We will confine our considerations to point 4) and in particular to what is to be found in literature on point (4a).

For the study of the performance of the eye there are widely divergent methods of approach. This is best illustrated by mentioning two extremes. The one consists in observing the traffic along the lighted roads and from the accident statistics attempting to determine what system of illumination offers the best possibilities of vision. The other extreme consists of a theoretical and laboratory analysis of the process of vision into all its components, then subjecting each component to an experimental investigation and building up from the components a theory of perception on the road; this can be done even without taking a look at a lighted highway. Since both of the methods mentioned have great obvious disadvantages, we shall here attempt to find a golden mean. In our opinion the best compromise is to investigate the components of vision in the laboratory according to the second method and to test the results by repeated practical measurements on the road.

It is in this way that we shall now try to survey a number of important problems connected with vision at the low levels of brightness occurring in road lighting.

## Perception

An object on the road is noticed because there is a contrast, *i.e.* because the brightness of an object differs from that of its background, usually the road surface.

Objects can therefore most quickly be perceived when care is taken to provide strong contrasts everywhere, for instance by trying to make all objects appear very dark against a bright background. Since, however, there are a number of factors over which we have no control, this is not always possible. From time to time there will inevitably be such small contrasts as to be hardly perceptible, if at all.

In order to minimise the number of cases where through this cause an object is not noticed at all or else too late, we shall therefore have to provide for conditions where the smallest possible contrast is still just perceptible. By a contrast still just perceptible we mean the ratio $\Delta B : B$, where $B$ is the brightness and $\Delta B$ the smallest perceptible difference in brightness. $B : \Delta B$ is called the contrast sensitivity, and it must therefore be made as large as possible.

Upon what condition does contrast sensitivity depend?

In the first place upon the level of brightness. In *fig. 1* the continuous curve gives the relation between contrast sensitivity and brightness [1] [1]). In the problems of road lighting we are mainly concerned with the interval [2]) between 0.1 and 1 candle/m². In this interval the quotient $B : \Delta B$ is found to increase rapidly with the brightness; it varies from 18 to 33, which means that contrasts of 3-5.5% lie on the limit of perceptibility. This continuous curve of fig. 1, however, was measured under the most favourable laboratory conditions, with large light spots on a completely uniform background, with an unlimited time of observation and with the possibility of calm and full concentration, with no disturbing effect from glare.

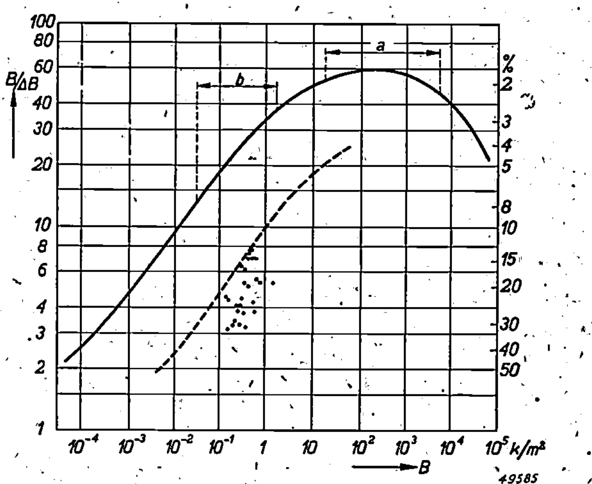


Fig. 1. Dependence of the contrast sensitivity $B/\Delta B$ on the level of brightness $B$. In road lighting (brightness range $b$) this is less than in daylight (range $a$).
Continuous-line curve: laboratory experiment under most favourable conditions.
Dotted-line curve: outdoor measurement taken as dusk was falling.
Scattered dots: measurements taken with good road-lighting installations.

Under practical conditions the contrast sensitivity is therefore much smaller. The dotted line of fig. 1 represents the results obtained under conditions more closely approaching reality: just at dusk, with a uniformly clouded sky, on a road with no artificial illumination the contrast sensitivity was measured [2] for small objects as observed by a

[1]) The numbers in square brackets refer to the publications listed at the end of this article.

[2]) 0.1 and 1 candle/m² correspond to the brightness of a dry, light-coloured concrete road surface (reflective power 30%) when it is illuminated with respectively 0.1 and 1 lux (level of illumination with full moon = 0.2 lux).

pedestrian at a distance of 150 metres; nature provided automatically for the transition through the different levels of brightness. Under these conditions the contrast sensitivity is lower by a factor of more than three.

The dots shown in fig. 1 represent the results of measurements taken under conditions still more closely approaching the reality. Each dot gives the contrast sensitivity under a given lighting installation for the above-mentioned small objects on the brightest part of the road surface. All these measurements refer to "good" lighting installations (with sodium and mercury lamps); they were taken when the road surface was dry [2]. Compared with the dotted line all these contrast-sensitivity points are still lower, as is understandable considering that the non-uniformity of the distribution of brightness and a certain amount of glare play a part as disturbing factors (see below). The fact that some of these points lie almost on the curve and others far below it shows that these disturbing factors are of a greater effect with one installation than with another. Given an installation that is good in every respect, the contrast sensitivity will not lie more than a factor 1.5 below the dotted line, so that at an average level of brightness it will amount to at least 4.5 (*i.e.* contrasts of about 22% will still be visible), while at higher levels of brightness (such as occur with light-coloured concrete roads) it may rise to 6 (17% perceptible contrast), and under favourable conditions even higher.

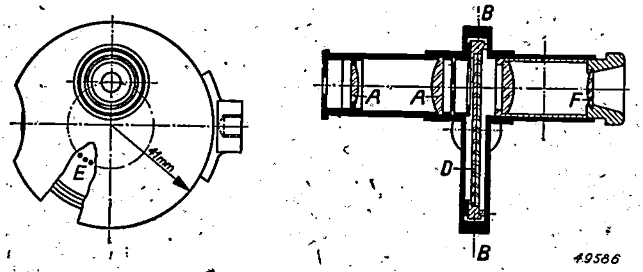


Fig. 2. Visibility meter for road lighting according to Holst and Bouma.

The above measurements with small objects were taken with the help of a so-called visibility meter [3], a diagram of which is given in *fig. 2.* A lens system $AA$ focuses the road surface in the plane $BB$. In this plane there is also a rotating glass plate $D$, upon which a number of dots $E$ of varying density have been produced photographically. By placing the eye in front of the ocular $F$ the road surface and dots are seen focused simultaneously. Upon rotating the disc $D$ it is possible to pick out from the different dots passing the field of vision the one that is still just visible, in that way determining the contrast sensitivity. The instrument is so constructed that with the meter in front of the eye the road is observed under practically the same conditions as with the naked eye (enlargement 1 : 1, large field of vision, including any glare from light, small objects, contrast dark on light background, etc.).

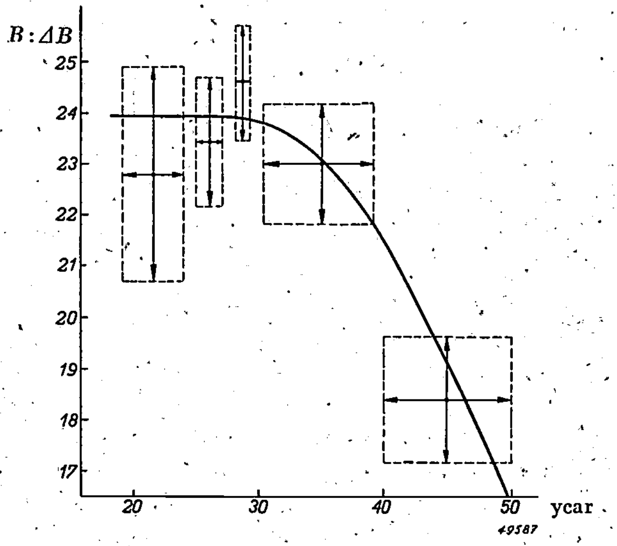


Fig. 3. Average contrast sensitivity as a function of age, measured at a level of brightness of 0.3 candle/m$^2$, with no source of glare.

Finally it must be noted that the contrast sensitivity is also dependent on age, as is illustrated by the curve in *fig. 3,* taken from an investigation with 100 observers under laboratory conditions [4]. The horizontal arrows give the averaged age groups and the vertical arrows the individual deviations. This dependence on age is even more pronounced in the curve of *fig. 4,* representing the contrast

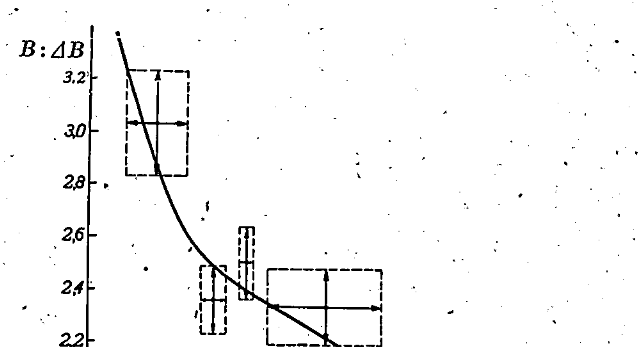


Fig. 4. Average contrast sensitivity as a function of age, measured at a level of brightness of 0.3 candle/m$^2$ in the presence of a source of glare. The glare causes a much greater decrease in sensitivity to contrast in the older age groups.

sensitivity in the presence of a strong source of glare.

So far we have been speaking only of "just perceptible" contrasts. In practice, however, it is also of great importance that a certain contrast, given by two different brightnesses, lying far above this threshold should be made as striking as possible. Thus the objects are more readily noticed and driving on the road becomes much less of a strain. This can be achieved by eliminating all disturbing influences as far as possible, while also the choice of the kind of light plays a part (sce the end of this article).

### Recognition

Once an object has been noticed we also want to recognize it as quickly as possible, because our reaction (braking, swerving, sounding the horn) very often depends upon the nature of the object. This recognition is closely related to the visual acuity, *i.e.* the ability of the eye to determine the shape of the object by the look of small details (it is usually only a question of the outline of the object). As a measure of the visual acuity $G$ we may take the inverse ratio of the minimum value of the angle of vision within which a detail must be seen to be observed, or a quantity proportional thereto. This value $G$ is limited on the one hand by the structure of the retina of the eye, thus by the size of the separate light-sensitive elements, and on the other hand by the focusing errors of the eye (spherical and chromatic aberration, diffraction, etc.). It is remarkable how all of these effects cause errors of the same order of magnitude: in this respect the eye makes a wonderful compromise.

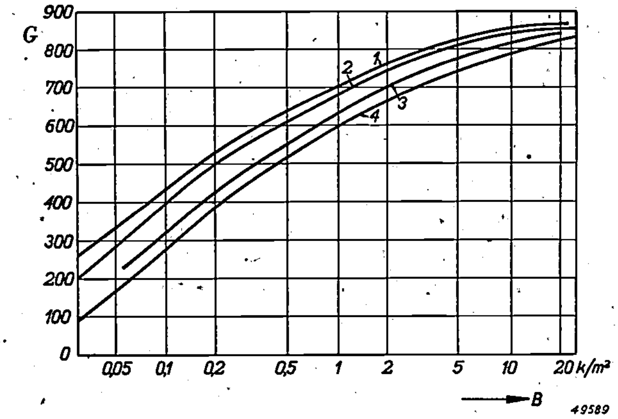


Fig. 5. The dependence of visual acuity $G$ on the brightness $B$ for different kinds of light, *viz* *1* for mercury light, *2* for sodium light, *3* for neon light, *4* for incandescent lamp light. The value of the visual acuity $G$ is expressed here as the distance in cm at which circles and squares 1 cm in diameter can just be distinguished. The value $G = 690$ thus corresponds to the capability of just seeing a detail within an angle of vision of about 1 minute.

The effects mentioned have, for example, the result that when looking at a figure consisting of parallel alternating black and white stripes, the parts of the retina upon which the image of the black stripes is formed also receive some light, so that the contrast may be only 70% instead of 100%.

When the level of brightness is lowered then $G$ is reduced [5]. *Fig. 5* illustrates this decrease for different kinds of light.

The cause of the decrease lies partly in the expansion of the pupil; with levels of brightness such as occur in road lighting the pupil diameter is about $2^1/_2$ times as large as in daylight and the effect of the focusing errors is then greater.

At very low levels of brightness visual acuity is affected also by the transition from cone vision to rod vision (see below).

Under laboratory conditions and at high levels of brightness it is just possible to observe details within an angle of vision of about 1 minute. With road lighting, even under favourable conditions, it will not be possible to reach much lower values than 3 minutes. (This is the angle within which an object of 13 cm is seen at a distance of 150 metres.)

### Quick recognition

It is particularly of importance for fast traffic that objects on the road can be recognized quickly. It is in fact a question of the time elapsing between the moment when the object comes into the field of vision, under sufficiently favourable conditions for observation, and the moment of reaction to what is seen (braking, swerving, etc.). This period of time may be divided into four parts:

a) The time taken for the image to be built up on the retina. This depends (just as in the case of a photographic plate) very closely on the intensity of the light. We can obtain some insight into this by measuring the time $\tau$ during which an object must have been visible in order to recognize it again; $\tau$ is thus analogous to the exposure time for a photographic plate. The value $1/\tau$ is often called the velocity of observation. *Fig. 6* gives the relation between $1/\tau$ and the brightness $B$ : curve *1* for sodium and mercury light and curve *2* for incandescent electric light. It is seen that at high levels of brightness we have extremely small values of $\tau$, while at the level of brightness on the road (0.1-1 candle /m²) the time $\tau$ usually lies between 1/4 and 1/20 sec, in the case of objects with little contrast, or where there is any intense glare considerably longer times may occur.

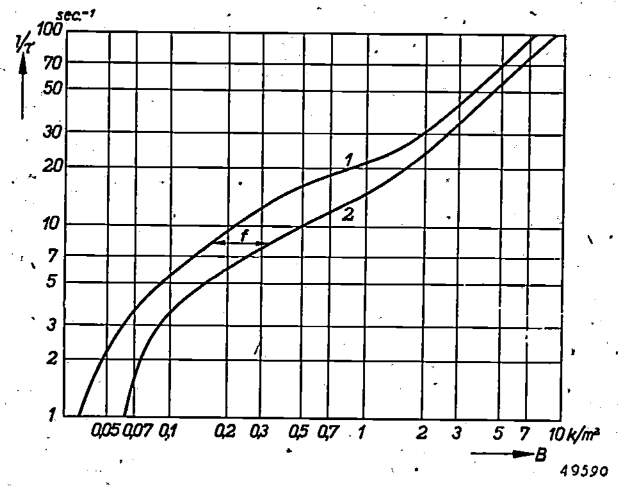


Fig. 6. The dependence of the velocity of observation $1/\tau$ on the brightness in the case of stationary objects:
*1* for sodium and mercury light,
*2* for incandescent lamp light.
For the same velocity of observation the brightness in case *2* must be a factor $f \approx 2$ times as large as in case *1*.

Fig. 6 holds for stationary objects [6], while *fig. 7* gives analogous curves for moving objects, those which flash past a slit (Weigel's observation [7]); since the velocity of motion was increased proportionally with $1/\tau$ we cannot expect to find such low values of $\tau$ here. In fig. 7 curve *1* represents the result with sodium light, curve *2* with incandescent electric lamps, curve *3* with mercury light. The fact that here the order differs from that in fig. 6 is not surprising, because quite different factors begin to play a part (after-image, extinction of the retinal image, etc.)

b) The time it takes to become conscious of the image and to recognize the object. This depends upon the clearness and the "striking" power of the image and especially whether it is anything common or uncommon.

As to the order of magnitude of these periods of time, measurements of the velocity of reading teach us something. *Fig. 8* shows the number of letters of an easily legible text and good print that can be read per second without any particular strain [6] (curve *1* under sodium light and curve *2* under incandescent electric light). At very high levels of brightness a rate of 30 letters per second can be reached. What is actually measured is the variation of the sum of the times (a) and (b) according to the brightness. At very high levels of brightness the time (b) plays the main role, while at low levels (road lighting) both components are of importance.

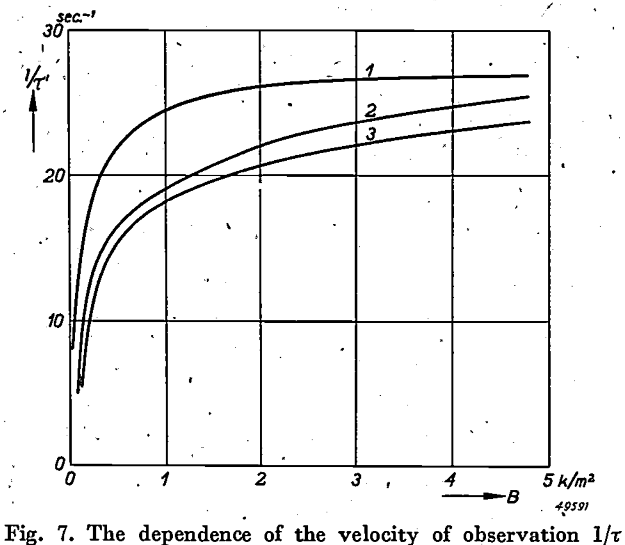


Fig. 7. The dependence of the velocity of observation $1/\tau$ on the brightness in the case of moving objects (according to Weigel).
*1* for sodium light,
*2* for incandescent lamp light,
*3* for mercury light.

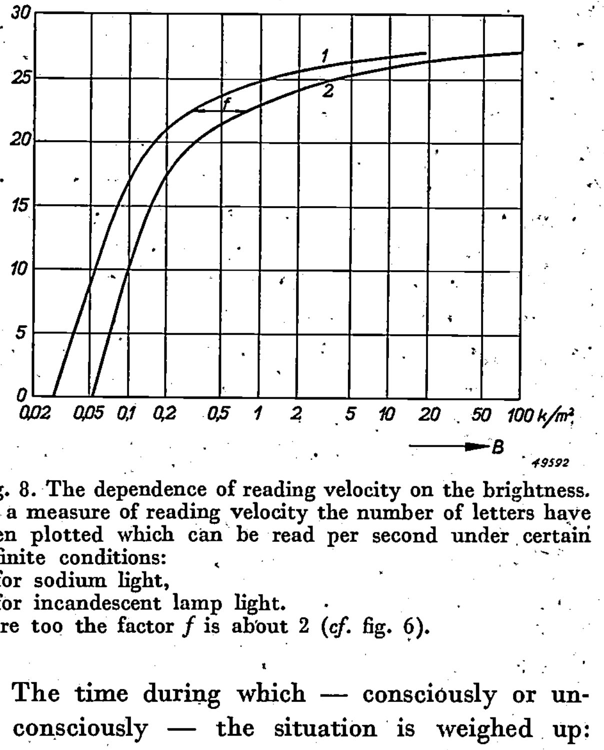


Fig. 8. The dependence of reading velocity on the brightness. As a measure of reading velocity the number of letters have been plotted which can be read per second under certain definite conditions:
*1* for sodium light,
*2* for incandescent lamp light.
Here too the factor $f$ is about 2 (*cf.* fig. 6).

c) The time during which — consciously or unconsciously — the situation is weighed up: what must I do?

d) The time taken to set the reaction in motion. These last two periods of time have nothing to do with the eye and are influenced, among other things, by psychological factors (shock, concentration, distraction, difficulty of the problem, fatigue, etc).

## Disturbing influences: glare

In general the disturbing effect of a given source of glare is all the greater according as the general level of brightness is lower and the visual task to be performed is more difficult. From this it follows that at the low brightnesses of road lighting and especially on fast traffic highways we must try to limit this disturbing influence to a minimum. Glare may be manifested in several ways:

a) *Glare depreciates the performance of the eye.*

The following are some examples of this form of disturbance (while the source of glare is present):

Example 1: At a distance of 2 metres to the side of an object 30 metres away there is a source of light radiating 810 candles in the direction of the observer and thus casting 0.9 lux on the eye. According to accepted standards such a source of light is considered "just troublesomely glaring" [8]. Because of this glare, with a normal brightness of the road surface the contrast sensitivity is found to have fallen by a factor 2.3. The seriousness of this is evident when it is considered that we should reach the same depreciation if the level of illumination were decreased by a factor of not less than 15! Moreover, it is assumed that the eye is kept fixed on the object and does not chance to look for a moment directly into the source of light. The reduction in visual acuity due to this degree of glare is much less, *viz.* about 15%. This is a general phenomenon: contrast sensitivity decreases much more due to glare than does visual acuity. Glare does not so much obscure the shape of the object but rather often causes it to fade entirely into the background.

Example 2: In order to give an idea of the extent to which glare may still occur with a modern road-lighting installation with well shielded light sources, we assume that the horizontally directed eye is struck by the light radiated at an angle of 70° to the vertical by a source of light (sodium lamp in an enamelled reflector) with a light intensity of 850 candles in that direction. The light falling on the eye in this case reduces the contrast sensitivity by about 10%, equivalent to the effect that would be obtained by reducing the level of brightness by a factor $1^1/_2$. This gives the degree of glare considered permissible with well-planned installations if the illumination is not to be too uneconomical. With many installations, however, the degree of glare encountered is many times greater.

Table I.

| Smallest perceptible contrast | | Ratio |
|---|---|---|
| With oncoming traffic | Without oncoming traffic | |
| $29^{50}$% | $16^{50}$% | 1.8 : 1 |
| 44% | $12^{50}$% | 3.5 : 1 |
| 25% | $14^{50}$% | 1.7 : 1 |
| $28^{50}$% | 10% | 2.8 : 1 |

Example 3: *Table I* [2] gives an idea of the disturbing influence of oncoming traffic — even with dimmed headlights. For four different installations the smallest perceptible contrast was measured (with the above-described visibility meter) both with and without oncoming traffic. It appears that the glare from oncoming traffic causes the contrast sensitivity for small objects to fall on an average by a factor 2.

b) *After-effect, after-images*

After the source of glare has disappeared from the field of vision, its disturbing influence may still continue for some time (successive glare). We shall return to this later.

c) *Hindrance*

A third manifestation of the effect of glare is of a more psychological nature: the general feeling of hindrance, annoyance, and distraction of attention. These factors are difficult to measure and express in figures. We shall not, therefore, go more deeply into them here.

**Disturbing influences: non-uniformity**

The non-uniformity of the distribution of brightness on the road surface may also have a very disturbing effect that manifests itself in a depreciation of the performance of the eye [2]. This is illustrated by the figures in *table II.* For a given installation (sodium lighting, asphalt road surface) the smallest perceptible contrast was measured with the visibility meter at different states of wetness of the road surface: $c_1$ on the brightest spot of the road and $c_2$ on the darkest spot. It is found that with increasing wetness of the road surface the dark spots become darker and darker and the light spots brighter and brighter but also smaller. It is therefore understandable that $c_2$ increases, but from the table we see that $c_1$ also becomes larger with increasing wetness. It is thus clear that the unfavourable effect of the increasing non-uniformity of the distribution of brightness exceeds the favourable effect of the increase in the brightness itself:

Table II.

| Condition of the road surface | Smallest perceptible contrast | |
|---|---|---|
| | $c_1$ | $c_2$ |
| Dry | 17 | 31 |
| Slightly damp | 24 | 43 |
| Damp | 28 | 48 |
| Wet | 36 | — |

We may say that on an average with an installation with shielded sources of light and with a dry road surface the disturbing effect of the glare of the fixed light sources on the contrast sensitivity is the same as that due to a lowering of the level of brightness by a factor $1^1/_2$, while the disturbing effect of non-uniformity is equivalent to a lowering of the level of brightness by a factor 2. This agrees with fig. 1, where the measured points lie, on an average, a factor 3 to the right of the dotted-line curve.

Finally, non-uniformity may also cause various disturbances of a more psychological nature, which we shall not go into, here. As examples we may mention certain light spots giving mistaken impressions of objects, whether or not we can see the road surface as a whole, distraction. fatigue, etc.

### The part played by each component

In the foregoing the process of vision at low level of brightness has been resolved into a number of components. The question arises as to which components are of the greatest importance for visibility on the road. In general we may say that all the factors dealt with play an important part.

There is a lack of unanimity in literature as to whether the visual acuity and the closely related velocity of observation are also important factors. This depends very much on the speed at which one is driving on the road. As the speed increases it is necessary to be able to recognize objects at greater distances and thus within smaller angles of vision, so that the visual acuity begins to play a more and more important part. Since at high speeds delays in reaction of a fraction of a second may have fatal results, the speed of observation is also of much greater importance. The task of vision becomes more and more difficult as speed is increased, so that especially on fast-traffic highways every endeavour must be made to minimize the various disturbing elements.

It should also be noted that in the measurements with the visibility meter described above a number of the factors studied (contrast sensitivity, visual acuity, glare) are expressed in their correct relations.

### Purkinje effect

In several respects it is of great significance for vision on the road that the eye possesses two different kinds of light-sensitive elements: the rods and the cones. These elements are distinguished not only by their shape, to which they owe their names, but also according to the following aspects:

a) By their different location on the retina. In the centre of the retina within a circle of some tenths of a millimeter there are only cones, while closer to the edges of the retina there are only rods, with a gradual transition between the two. The part of the retina containing cones is the part with which we see most sharply. It is for that reason that we stare directly at objects we wish to see sharply.

b) By their capacity of distinguishing colours. The cones can observe colour differences, the rods cannot.

c) By the difference in the range of brightness within which they function best. At high levels of brightness we see almost exclusively with the cones and at very low levels of brightness only with the rods. Between these extremes there is a transitory region where both elements make important contributions to vision. The following may serve as a general guide:
Level of brightness in daylight: exclusively cone vision.
Level of brightness artificial illumination indoors: usually exclusively cone vision.
Level of brightness road lighting; mixed vision, mainly with cones.
Level of brightness with full moon: mixed vision, mainly with rods.
Level of brightness with stars and no moon: exclusively rod vision.

d) By the difference in their eye-sensitivity curve. In *fig. 9* curve *A* is the eye-sensitivity curve for cone vision, *B* that for rod vision. The curve for mixed vision would lie intermediate. The cones are apparently most sensitive to yellowish green, the rods to bluish green. This difference in estimation of the brightness of colours has an important consequence. When a number of colours are given, all of which make an impression of equal brightness, and the amount of light

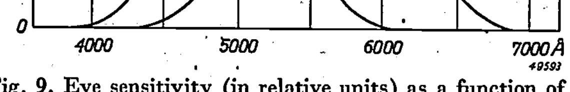


Fig. 9. Eye sensitivity (in relative units) as a function of the wavelength (in Å) for cone vision (*A*) and for rod vision (*B*).

radiated is then reduced by the same large factor for all colours, they are found no longer to give the impression of being "equally bright". The changed estimation of brightness for the different colours results in the bluish colours appearing to be brighter and the reddish colour less bright than the others (Purkinje effect).

If we were to reduce the brightness to such a low level as to get purely rod vision then the reddish kinds of light would ultimately become less efficient. In road-lighting practice, however, we work in the upper part of the transitory region; when a reddish light source is employed (for instance sodium light), it will therefore be "inferior in brightness" only for the darkest parts of the field of vision and not for the brighter parts. The result is that the dark objects appear extra black against the much lighter road surface (greater contrast).

Finally it may be noted that the distinction between rods and cones is also important in the study of the phenomena of glare, since we usually try to see as much as possible with the cones, while usually the glaring light mainly affects the rods.

## Influence of colour on vision

In the following we will briefly review the influence of the colour of light on the various factors of vision, confining our considerations to the three most important kinds of artificial illumination: incandescent lamp light, mercury light and sodium light.

a) Influence on contrast sensitivity. Little difference between the three kinds of light.

b) The "striking" power of greater contrasts. In this respect sodium light offers important advantages (see above).

c) Visual acuity. In this respect mercury and sodium light are superior to incandescent lamp light; see fig. 5 (*1* mercury, *2* sodium, *4* incandescent lamps).

d) Velocity of observation of stationary objects. Here, too, mercury and sodium light are superior to incandescent lamp light; see fig. 6 (*1* mercury and sodium, *2* incandescent lamps).

e) Velocity of observation of moving objects. In this respect sodium light is superior and mercury light inferior to incandescent lamp light; see fig. 7 (*1* sodium, *2* incandescent lamps, *3* mercury).

It must be noted that on the road one is concerned more with case (*d*) than with case (*e*), since the movements on the road are for the most part towards or away from us; transversally there is little displacement of objects in our field of vision.

f) Glare; influence on visual acuity and contrast sensitivity. There is little difference here between the three kinds of light.

g) Glare; after-effect. *Fig. 10* gives measurements of the time of recovery of the visual acuity, *i.e.* the time elapsing after the disappearance of the source of glare until the visual acuity has again attained a certain value. This recovery time has been plotted in the vertical direction for the case where the glare was due to incandescent lamp light, and in the horizontal direction for an experimental case of glare due to sodium light under otherwise the same conditions. Each point recorded thus gives the recovery times $t_1$ and $t_2$ in analogous experiments for "white glare" and for "sodium glare". If there were no difference between the two kinds of light all the points would lie on the dotted line ($t_1 = t_2$). From the diagram it may be seen that taken on an average the recovery times $t_1$ for white light are a factor 1.8-2.0 longer than the corresponding times $t_2$ for sodium light [9].

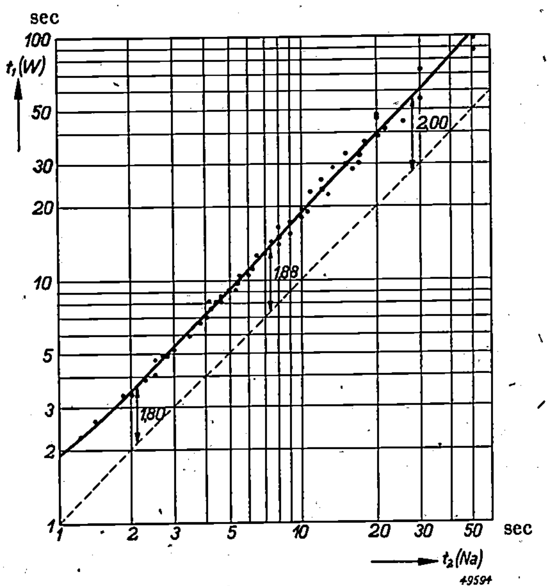


Fig. 10. The recovery time of visual acuity after glare is dependent on the colour of the light. With glare from white light ($t_1$) it is about twice as long as with glare from sodium light ($t_2$).

h) Glare; psychological effects. In general yellow light is found to be less disturbing than white light, though there are individual differences.

i) Experiments with the visibility meter.

*Table III* gives the average results obtained by measurements with the visibility meter with a large number of installations [2]. $c_1$ again indicates the just perceptible contrast (in %) for the brightest part of the road surface, $c_2$ for the darkest part.

It also gives the wattage used per kilometre and the lumens radiated by the lamps per metre of road length: the latter quantity was practically the same for the different kinds of light.

Table III.

| Kind of light | State of road surface | $c_1$ (%) | $c_2$ (%) | $\frac{\text{kW}}{\text{km}}$ | lm/m |
|---|---|---|---|---|---|
| Sodium | dry | 17.2 | 31.2 | 3.3 | 200 |
| Mercury | dry | 26.0 | 35.4 | 6.4 | 220 |
| Blended light | dry | 21.4 | 35.3 | 10.8 | 230 |
| Sodium | damp | 24.1 | 44.3 | 3.4 | 210 |

It may be seen from the table that the results for the sodium installations exceed those for the mercury and blended light installations notwithstanding the much smaller power consumed. It also appears again that both $c_1$ and $c_2$ increase as the road surface becomes wet and thus the contrast sensitivity decreases both for the light and for the dark parts of the road surface.

Table III does not give any figures for incandescent lamp installations. Data have in fact been collected also for these installations — the results were by far less satisfactory than those given in table III — but a true comparison was impossible because most incandescent lamp installations already differ considerably from the sodium and mercury installations also in other respects.

## BIBLIOGRAPHY

[1] P. J. Bouma, Properties of the eye in connection with their significance for road lighting, Philips Techn. Rev. 1, 102-106, 1936.

[2] P. J. Bouma, Measurements carried out on road-lighting installations, Philips Techn. Rev. 4, 304-313, 1939.

[3] G. Holst and P. J. Bouma, Ein neues Meßgerät zur Beurteilung der Güte einer Straßenbeleuchtung, Physica 3, 1159-1163, 1936.
G. Holst and P. J. Bouma, How to judge the quality of road lighting. Philips Techn. Rev. 1, 353-356, 1936.

[4] P. J. Bouma, The problem of glare in road lighting, Philips Techn: Rev. 1, 225-229, 1936.
P. J. Bouma, Aanvullende mededeelingen over eenige eigenschappen van het oog, De Ingenieur 55, G. 10-12, 1940.

[5] P. J. Bouma, Gezichtsscherptemetingen bij diverse lichtsoorten, De Ingenieur, 49, A243-246, 1934.

[6] P. J. Bouma, Gezichtsscherpte en waarnemingssnelheid bij wit licht en natriumlicht, De Ingenieur 49, A31-34, 1934.

[7] R. G. Weigel, Untcrsuchungen über die Sehfähigkeit im Natrium- und Quecksilberlicht insbesondere bei der Straßenbeleuchtung, Das Licht 5, 211-216, 1935.

[8] P. J. Bouma, Verblinding, Natuur en Mensch 56, 217-221, 1936.
P. J. Bouma, Verblinding, Polytechn. Weekbl. 29, 625-629, 1935.

[9] P. J. Bouma, Contrastrijkheid bij natriumlicht, kwiklicht en wit licht, De Ingenieur 49, A290-294, 1934.

# BOOK REVIEW

**„Kleuren en Kleurindrukken"**[1], by P. J. Bouma, 320 pages, 113 illustrations and 15 tables, Philips Technical Library, Meulenhoff & Co. N.V., Amsterdam, 1946.

In this comprehensively written book (the reader is presumed to have a secondary school knowledge of mathematics and physics plus a certain amount of zeal!) the problem of colorimetry is approached from the angle of experimental physics and the theory of illumination. Other aspects of the many-sided field of the theory of colours which are connected with physiology and psychology are discussed only in so far as they are necessary for a proper understanding of the subject; also these discussions are kept on an experimental basis as far as possible. The aesthetic aspect has not been touched upon at all, neither have the physics and chemistry of dyes and pigments been dealt with.

Of the fourteen chapters, each divided into about ten sections, the first six are devoted to the theory of colour proper: the part played by the eye, the concept of brightness, the colour triangle and colour space. In contrast to the usual procedure and in accordance with Schrödinger, the author has restored colour space to a place of honour as basis for the theory of colour. The laws of Grassmann for the additive mixing of colour are dealt with at length in connection with colour space. Then follows an explanation of the international (C.I.E.) system of coordinates *XYZ* illustrated by a number of examples of calculation[2].

The other chapters are devoted to special subjects. Chapter VII treats of several special cases such as the colours of black-body radiators, the boundary colours occurring upon refraction by a prism, the ideal remission curves, the so-called optimum colours (emission colours of maximum brightness) and the characteristic or most "colourful" colours (full colours).

Chapters VIII and IX give a survey of objective and subjective methods of colorimetry and of the instruments thereby employed.

In Chapter X the deviations from the normal sense of colour are described, dealing with the laws which these deviations obey, as well as their inheritability.

In Chapter XI the foremost personalities in the history of the science of colour are reviewed, with a short description of their contributions to this science: Newton, Goethe, Young, Grassmann, Maxwell, Helmholtz, A. König, Hering, Guild, Wright, Schrödinger, Ostwald.

In Chapters XII and XIII a number of questions are discussed which do not belong directly to the field of colorimetry, namely those cases where one is rather concerned with the concept of "colour sensation". Among these questions is the perception of colour differences and the study of the character of colour sensation, in particular the phenomenon of simultaneous contrast and the chromatic adaptation of the eye to coloured surroundings or to an illumination with more or less coloured light.

Finally Chapter XIV deals with several important fields of application, namely the theory of illumination, the applications in trade, industry and science (description of the colours of products and colour samples in trichromatic coordinates, employment of colour coordinates for the description of physical phenomena) and finally the problem of colour reproduction.

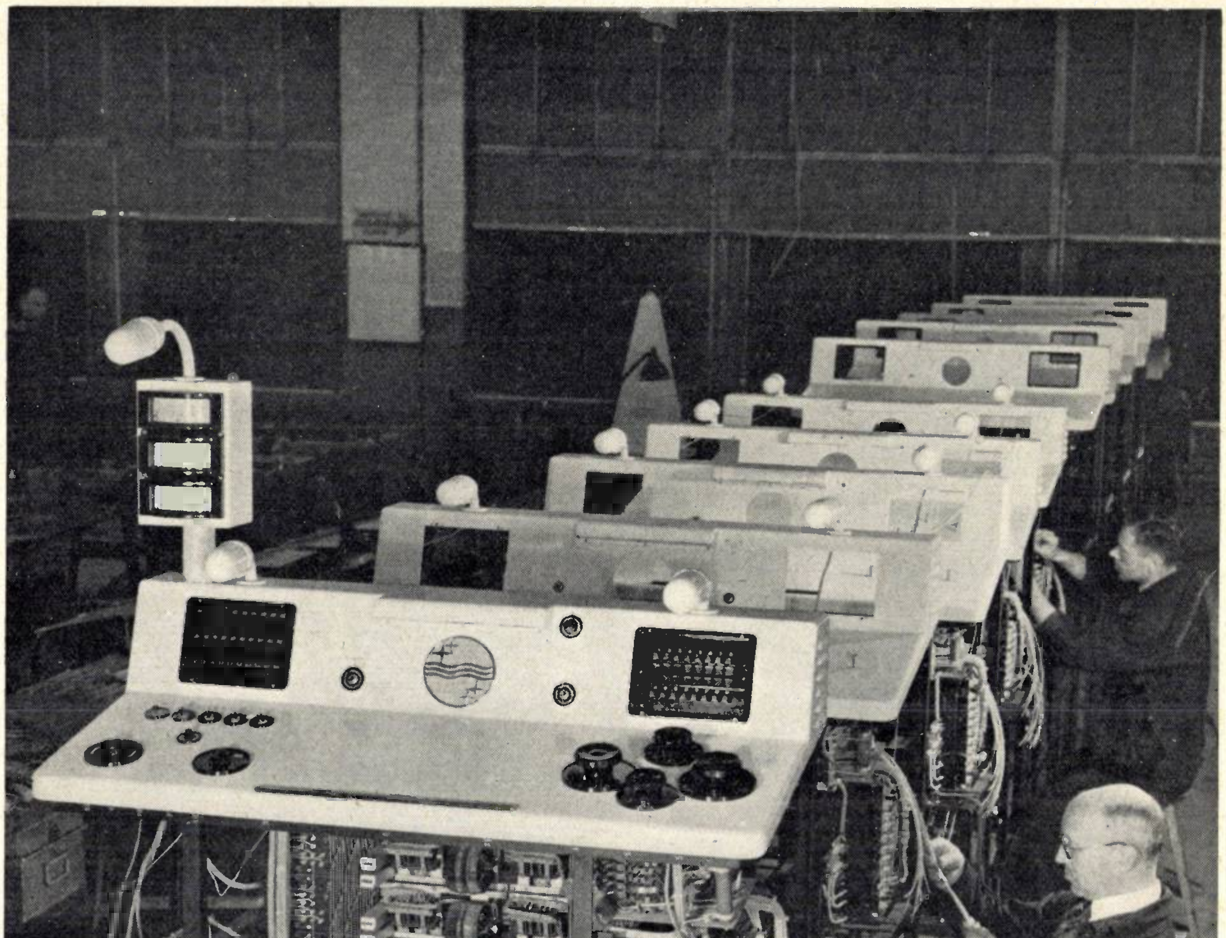
An appendix contains 15 tables, a bibliography with about 400 references to literature, a list of symbols employed and an index.

W. de Groot.

[1] The title of the forthcoming English translation will be: "Physical Aspects of Colour, an introduction to the scientific study of Colour Stimuli and Colour Sensations".

[2] The usefulness of the additive mixing laws is sometimes doubted. The number of cases in which one is concerned exclusively with these laws of mixing is indeed limited. In many cases problems arise which belong to the field of subtractive or multiplicative colour formation. The characterization of a colour by three numbers is then of little help and one is forced to have recourse to the spectral distribution function of the light source combined with the spectral transmission curve or spectral remission curve of a filter or of pigments. Moreover, in practice one is often concerned with the influence of the colour sensation, which is dependent on the surroundings. The possibility of fixing a colour by three coordinates, practically independently of the surroundings, furnishes, however, a means that can never be dispensed with.

# ASSEMBLY OF SWITCHING DESKS FOR X-RAY DIAGNOSTIC APPARATUS



50410

An X-ray apparatus of the type shown here in course of construction (the Super D) can have eight X-ray tubes connected to it to be taken into use in turn. For each tube the desk is fitted with a so-called automatic unit providing for an optimum loading of the tube. Three such units. consisting of relays, resistors and contact plates, are seen underneath the middle part of the covering plate of the first desk, and five others can be placed underneath these. The tube to be used is selected by pressing in one of the eight push-button switches at the top of the desk on the right, thereby automatically switching over all the connections for the supply, the control mechanism, the safety devices and the signalling. This greatly simplifies operation, but it makes the apparatus complicated. The fitters have to make some thousand wire connections.

# 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.

**1719/1721:** C. J. Bouwkamp: On the dissection of rectangles into squares I, II, III (Proc. Kon Ned. Akad. Wetenschappen Amsterdam **49**, 1176-1188, 1946; **50**, 58-71, 72-78, 1947).

According to Brooks, Smith, Stone and Tutte, the semi-topological problem of the dissection of a rectangle into non-overlapping, unequal, squares is reduced to a physical problem: that of the current distribution in a planar electrical network consisting of equal resistances. The networks required for squarings of an order less than 15 are drawn. There are 2, 6, 22, 67, 214 perfectly squared rectangles of the order of 9, 10, 11, 12, 13 respectively. All these squarings are codified and classified. Special attention is paid to the construction of squared squares. At present it is not known whether a simple, perfect, squared square is possible.

Number 3 of volume 2, June, 1947, of *Philips Research Reports* contains the following papers:

*R 42:* W. Elenbaas: Influence of cooling conditions on high-pressure discharges.

*R 43:* H. C. Hamaker, H. Bruining and A. H. W. Aten Jr: On the activation of oxide-coated cathodes.

*R 44:* F. A. Kröger: Photoluminescence in the quaternary system $MgWO_4$ - $ZnWO_4$ - $MgMoO_4$ - $ZnMoO_4$.

*R 45:* F. A. Kröger: Luminescence of solid solutions of the system $CaMoO_4$ - $PbMoO_4$ and of some other systems.

*R 46:* R. Loosjes and H. J. Vink: The *i*, *V* characteristic of the coating of oxide cathodes during short-time thermionic emission.

*R 47:* J. D. Fast: The reaction between carbon and oxygen in liquid iron.

*R 48:* C. J. Bouwkamp: Calculation of the input impedance of a special antenna.

Readers interested in any of the above mentioned articles may 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 please write to the publishers of Philips Technical Review.

VOL. 9 No. 6, p. 161-192 1947

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 48-CHANNEL CARRIER TELEPHONE SYSTEM

by G. H. BAST*), D. GOEDHART and J. F. SCHOUTEN.

621.395.44:
621.396.619.2

The carrier telephone cables laid in the Netherlands since 1935 have proved to be suitable, as regards attenuation and cross-talk, for the transmission of frequencies up to about 200 kc/s with a repeater spacing of 25 km. In order to take full advantage of this possibility, the Dutch P.T.T. and Philips have together developed a 48-channel carrier telephone system. Modulation is in three stages. In the first stage all channels are modulated with a "terminal" carrier of 60 kc/s, the undesired modulation products being suppressed by a universal channel band filter. In the second stage of modulation groups of 12 basic channels are assembled by channel carriers of 192, 196.... 236 kc/s to form a basic group in the frequency band 252-300 kc/s, which requires only very simple filtering. By a third modulation with four basic carriers (240, 360, 408 and 456 kc/s) the four basic groups are brought into their places in the supergroup 12-204 kc/s. The reasons for the choice of this method of modulation are fully discussed in this article. In the construction of the universal channel band filter operating at 60 kc/s use has been made of "Ferroxcube", the new magnetic core material produced by Philips, with which it has been possible not only to satisfy filter design requirements (the coils of the filter in question have a $Q$ value of 500-600) but also to reduce the volume of the coils by a factor of 5. The influence of this new filter technique on the mechanical construction of the system will be discussed in another article.

## The Development of Carrier Telephony

In long-distance telephone communications the line costs constitute a considerable portion of the total cost. Consequently it has long been the aim to obtain a number of telephone circuits from each physical circuit. The first practical applications of carrier telephony for this purpose were made possible in the period 1912-1920 with the advent of the three-electrode valve (triode). At that time open wire lines were used and were found to be capable of carrying three or more single side-band carrier channels. Between 1920 and 1930 this technique was extensively applied in various countries - particularly the U.S.A. where large networks of open wire lines were available.

Shortly after 1930 carrier telephony was also applied to multi-conductor telephone cables, it having been found possible to transmit wide frequency bands with a suitable repeater spacing. At first 12-channel systems were employed in the U.S.A., and later also in Europe. A frequency band of 4 kc/s was reserved for each channel and the highest frequency transmitted was 60 kc/s. In the U.S.A. existing cables were first adapted for this system (*i.e.* by removing the loading) but in Europe, in practically all cases, new cables were made and laid. In course of time, with wider knowledge and gradually improved technique in cable-making, it became evident that frequencies much higher than 60 kc/s and even up to 200 kc/s could be transmitted. [1]) As a consequence the carrier telephone cables laid extensively in Holland since 1935 allow nearly 50 unidirectional channels to be obtained on each pair (separate "go" and "return" cables are used).

About 1937-1938 carrier telephone systems on co-axial cables were introduced in the U.S.A. and Great Britain; these cables may be used up to at least 3 Mc/s., thus making it possible

*) Of the Netherlands P.T.T.

[1]) G. H. Bast. On the application of carrier telephony in the Netherlands Telephone System, T. Ned. Radiogenootschap, 9, 279-293, 1941/42.

to obtain several hundred channels from a single co-axial circuit.

Let us consider the influence of circuit length upon the economic application of carrier working. As already mentioned, the cost of a physical circuit is high; in carrier systems the line cost per channel is the cost of the physical circuit divided by the number of channels obtained over it. Assuming that this division of the line cost more than outweighs the cost of the extra repeater stations necessary for carrier working, it follows that there is no upper limit to the length of circuit over which carrier working is economical. However, the reduction in the line charges per channel has first to offset the cost of the carrier terminal apparatus. The relative cost of the terminal apparatus increases as the length of circuit is reduced, so that in principle there is a minimum length below which carrier telephony is not justified from the economic point of view. With the gradual advance in technique the cost of the terminal apparatus, tends to decrease, so that carrier telephony becomes justified for shorter distances; this is rather important because the distribution of traffic, as a function of distance, shows that there is a preponderance of short haul circuits, and if carrier working is economically possible at shorter distances it will lead in turn to mass production of carrier apparatus and a further reduction of cost. The reaction of the number of channels required on the basic cost of carrier terminal apparatus makes it very difficult to say definitely where the economic limit lies. It may be said, however, that generally speaking voice frequency circuits with repeaters are no longer justified, since as a rule carrier telephony proves more economical. On the other hand, for distances that can be bridged with audio-frequency telephony without repeaters this is, as a rule, less expensive.

### Carrier Telephony in the Netherlands

Just before the outbreak of the last world war the Netherlands had reached a stage where the application of carrier telephony was to be expected on an unprecedented scale. The introduction of subscriber-to-subscriber dialling on a national basis has tended to increase the number of channels required and has added impetus to the development. The ravages of war, destruction and arrears have made the need for expansion greater than ever and the case for carrier telephony even more favourable. Some figures regarding the telephone network in the Netherlands will give an idea of the position as it was at the beginning of the war. The country is divided into 20 districts, each with a district telephone exchange through which traffic with other districts is passed. The network interconnecting these 20 district exchanges is called the inter-district network. When about 1940/41 the war put an end to the expansion that was steadily taking place, there were about 1500 carrier telephone circuits in this inter-district network with an average length of about 100 km. per circuit. The cable system carrying these 1500 channels is shown in *fig. 1*.

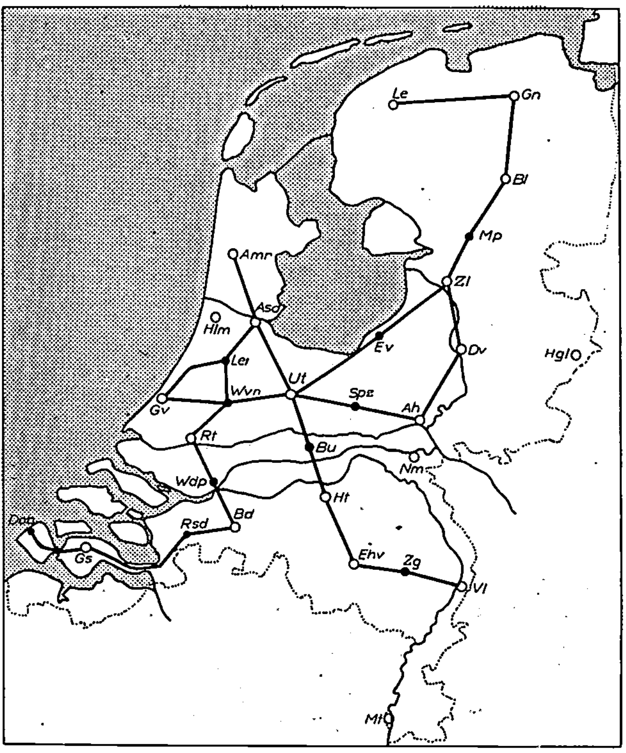


Fig. 1. The Netherlands telephone network is divided into 20 districts. The 20 district exchanges are interconnected by inter-district networks (indicated by open circles). This map shows the carrier inter-district links as used about 1940/41. The dots represent repeater stations.

The audio-frequency inter-district circuits also numbered about 1500 at that time, so that the point had just been reached where the number of carrier channels was beginning to overtake that of audio-frequency channels.

At that stage a total of about 1000 Km. of carrier cable network had already been laid. Each cable route, as already mentioned, comprised two separate cables containing "go" and "return" circuits respectively. Each cable has 24 pairs of conductors, so that by fully equipping all cables with 12 channels per pair of conductors it was then already possible to obtain (1000 ×

24 × 12) : 100 = 2880 channels of 100 km. in length. Expansion to systems with 24 or 48 channels per pair of conductors would have provided 5,760 or 11,520 channels respectively on the cable network already laid.

The prospect of such a development as this and the threat of large scale destruction during the war led to close cooperation between the Dutch P.T.T. and Philips, so as to be prepared in good time for speedy rehabilitation and further expansion. As a result a new carrier telephone system with 48 channels was developed, making full use of the possibilities of transmission over modern carrier cables. In its general set-up this system has been brought into line, on the most important points, with international standards for carrier systems.

The fundamental elements and most important details of this new carrier system will be described in a series of articles to be published in this Review. Among other things we shall deal with new constructional methods that have been made possible by using "Ferroxcube"[2]) filters. The present article will deal mainly with the basic principles of the system and in particular with the choice of the method of modulation.

## The choice of a 48-channel system

Reference has already been made to the possible influence of repeater costs on carrier systems. In the relevant frequency range the attenuation of carrier cables increases as the square root of the frequency.

As the maximum gain of a repeater has already been fixed by international agreement on the basis of the maximum and minimum transmission levels permitted, the upper limit of the frequency band transmitted determines the repeater spacing. The position of the junction centres in the Netherlands network was such that the repeater spacing could be conveniently standardized at about 25 km., a distance allowing the transmission of a frequency band up to about 200 kc/s. This introduces no difficult equalization problems[3]).

But the highest frequency transmitted is also limited by far-end cross-talk caused by mutual inductance and capacity unbalance between pairs. The effect of these unbalances increases with frequency. By careful manufacturing methods, however, and by special splicing procedure described elsewhere[4]) it was found possible to make the cable suitable for the transmission of frequencies up to about 200 kc/s. Such a frequency band provides about 50 channels each with a bandwidth of 4 kc/s. With large numbers of channels it is usual to divide them into groups of 12, so that in our case we have a system of 4 groups of 12, thus a total of 48 channels.

We shall now first describe in detail the method of modulation employed and then deal with the considerations that led to the choice of this method.

## Method of modulation used in the 48-channel system

The 48 audio-frequency channels are brought into the frequency band 12-204 kc/s by means of three stages of modulation.

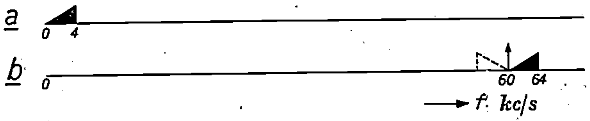


Fig. 2. Terminal modulation in the 48-chanel system. By modulating the audio-frequency channel (*a*) with a terminal carrier of 60 kc/s a basic channel (*b*) of 60-64 kc/s. is obtained which is identical for all 48 channels.

In this and all other modulation diagrams given in this article a carrier is indicated by an upright arrow, a wanted side band by a solid triangle and unwanted side band by a dotted triangle. The slope of the hypothenuse of the triangle (or of one side of the trapezium in other diagrams) represents the directions of increasing audio-frequency (or, in more general terms, of the modulation frequency).

In the first stage (*fig. 2*) each channel is modulated with a terminal carrier having the same frequency for all channels, *viz.* 60 kc/s. The carrier is suppressed in the balanced modulator, as is usual in carrier telephony, and the channel band filter (which is identical for all channels) selects the upper side band of 60-64 kc/s. In this way we now have 48 basic channels occupying the 60-64 kc/s frequency band.

In the second modulating stage (*fig. 3*) twelve of these basic channels are modulated with twelve different channel carriers of 192, 196......236 kc/s. The upper side bands thus appear on the busbars to which the channel modulators are connected as a basic group of 252-300 kc/s. The channel carrier leaks and the undesired lower side bands lying between 128 and 176 kc/s are suppressed by a common band filter. Three other groups

[2]) J. L. Snoek, Non-metallic magnetic material for high frequencies, Philips Technical Review 8, 359-360, December 1946.

[3]) See for example H. van de Weg, The equalization of telephone cables, Philips Technical Review, 7, 184-191, 1942.

[4]) See article quoted in footnote1).

of twelve basic channels are dealt with in the same way, each having twelve channel carriers of the same frequencies (192 . . . 236 kc/s) and with an identical group band filter. We have thus obtained four basic groups all in the frequency band 252-300 kc/s.

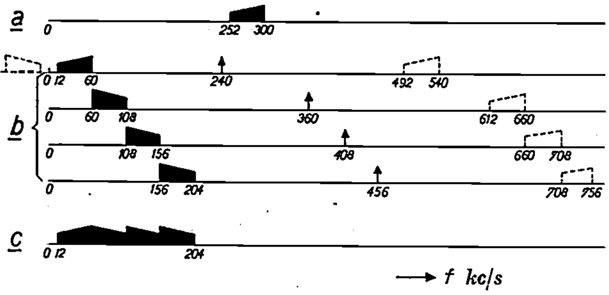


Fig. 3. Channel modulation in the 48-channel system. Twelve basic channels ($a$) ($f = 60$ - 64 kc/s) are assembled to form a basic groep ($c$) (252-300 kc/s) by modulating each basic channel with one of the twelve channel carriers of 192, 196.....236 kc/s ($b$).

In the third modulating stage (*fig. 4)* the four basic groups are modulated with four group carriers and are thus assembled to form a 12-204 kc/s super-group. These group carriers have frequencies of 240, 360, 408 and 456 kc/s. The four groups occupy the 12-60, 60-108, 108-156 and 156-204 Kc/s. bands respectively and can now be transmitted on one pair of conductors. As indicated in fig. 4, the channels in the 12-60 kc/s group are erect and the rest are inverted. This is accomplished by selecting the upper side band after group modulation in the former case and the lower side band in the latter[5]).

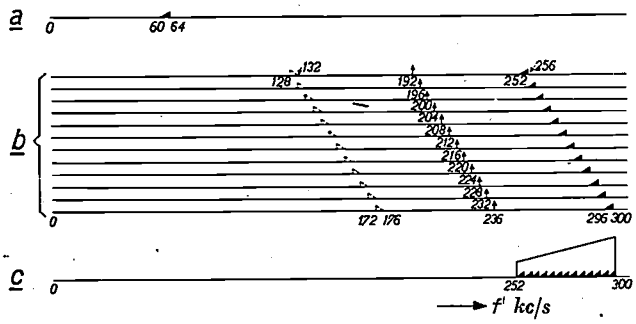


Fig. 4. Group modulation in the 48-channel system. By means of four group carriers of 240, 360, 408 and 456 kc/s. ($b$) the four basic-groups ($a$) of 252-300 kc/s are assembled in a super-group of 12-204 kc/s. ($c$). (The formation of a super-group is only partially shown.)

*Fig. 5* gives the frequency allocation of the whole of the modulation system; the block diagram in *fig. 6* shows the general arrangement of the carrier terminal equipment.

Before giving the reasons for this apparently complicated method of modulation it must be explained that the particular choice of upper or lower side bands was determined by the desire to conform with the recommendations of the C.C.I.F.(Comité Consultatif International Téléphonique)[6]). One of the objects aimed at in the design was to treat all channels as far as possible in exactly the same way in order to make the equipment interchangeable. The terminal carrier is identical for all 48 channels, as is also the channel band filter. The group band filters and the channel carriers are likewise identical for all four groups.

[5]) An erect channel is one in which the highest channel frequencies correspond to the highest frequencies $q$ of the speech transmitted, as is always the case in the upper side band in the case of single modulation; *cf.* fig. 2. An inverted channel is one in which the highest speech frequencies correspond to the lowest audio frequencies. The fact that in single modulation the upper side band $p+q$ of a modulated carrier wave $p$ is erect, and the lower side band $p-q$ inverted is directly deduced from the fact that in the former case $q$ occurs with the positive sign and in the latter case with the negative sign. In the case of multiple modulation, however, it is possible that $p < q$. Then the lower side band (formed by the difference frequencies $|p-q|$) has the frequency $q-p$, so that it is erect. This is the case in our first group in fig. 4.

[6]) The discrepancy between the initial letters of this Comittee and the abbrevitation used is due to the fact that the abbrevation C.C.I.T. was already being used for the Comité Consultatif International Télégraphique.

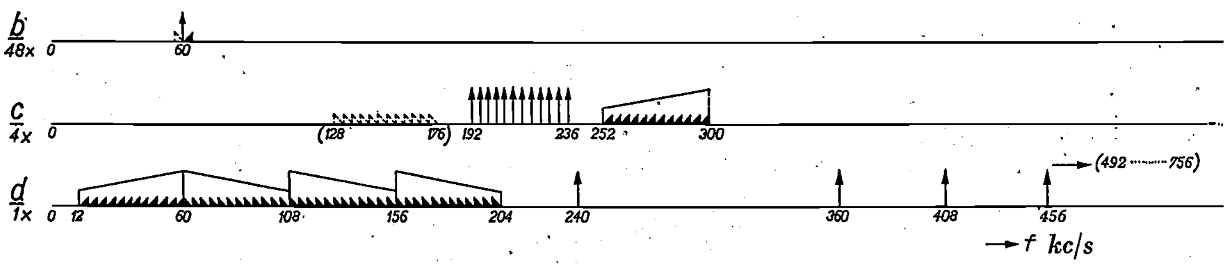


Fig. 5. Complete modulation diagram of the 48-channel system. ($a$) Audio-frequency channel. ($b$) Terminal modulation and basic channel. ($c$) Channel modulation and basic group. ($d$) Group modulation and 48 channel super-group.

### The choice of lower or upper side bands.

Upon the introduction of carrier telephony in Holland in 1938 the 8-56 kc/s band was chosen for the 12-channel system. At first, following British practice, the lower side bands of the carriers 12, 16....56 kc/s were transmitted. In the final development, however, the British Post Office adopted the American practice of using the upper side-bands of the same virtual carriers and a proposal has been made to the C.C.I.F. that this method should be standardized.

Although the channels were otherwise completely in accordance with international requirements, the situation that had arisen would inevitably have led to future difficulties in the engineering of international circuits. Consequently, as developments in cable manufacture gave the prospect of being able to use several groups of 12 channels there was every reason, when renewal of the system became necessary, not only to increase the number of channels but also to give consideration to the recommendations of the C.C.I.F.

At the last meeting of the "Commission de Rapporteurs" of the C.C.I.F. prior to the outbreak of the war (London, December 1938) a resolution was passed to standardize the allocation of the frequency bands in 12-channel systems. For 12 channels in the 12-60 kc/s range the upper side band was chosen. Without discussing systems of 24, 36 or 48 channels it was further recommended that the upper side band be taken for the 312-552 kc/s super-group and again the lower side band for 564-804, 812-1052 and 1060-1300 kc/s. This (at first sight) rather unsystematic choice arose from the solution which the British Post Office had chosen for the transmission of a large number of channels over a co-axial cable. This solution was based on the method of modulation introduced in America,[7]) where in the first stage 12 channels are combined to form a group of 60-108 kc/s and in the second stage five groups are combined into a super-group of 312-552 kc/s, the super-groups then being given their final position in a third stage of modulation.

Though the resolutions of the London meeting had not yet been translated into a general recommendation of the C.C.I.F., [8]) it was deemed advisable to adapt further development in the Netherlands to the conclusions arrived at in London.

This accounts for the choice, illogical as it may appear, of the upper and lower side bands for the four groups of the 48-channel system.

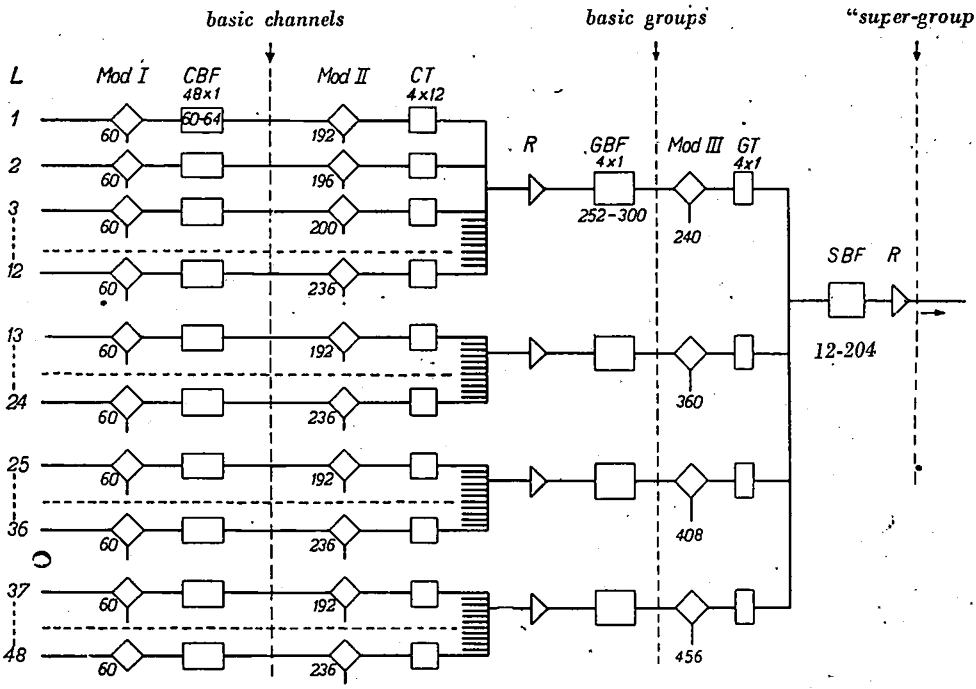


Fig. 6. Block diagram of the modulation system of the 48-channel system. The 48 audiofrequency channels enter at *L*. *Mod. I.* is the terminal modulator (60 kc/s carrier supplied); *CBF* the channel band filter; *Mod. II.* the channel modulator (channels of 193, 196.... 236 kc/s); *CT* channel transducer circuit; *R* repeaters; *GBF* group band filter; *Mod. III.* group modulator (group carriers of 240, 360, 408 and 456 kc/s); *GT* group transducer circuit; *SBF* super-group band filter. The filters *GT* are added because the outputs of the four modulators *Mod. III.* cannot be connected directly in parallel without interaction.

## Choice of number of stages of modulation

Continuing the explanation of the choice of the method of modulation for the 48-channel system, we will confine ourselves first to the case where only one system of 12 channels is required.

[7]) C. W. Green, and E. I. Green, A carrier telephone system for toll cables, Bell System Technical Journal **17**, 80-105, 1938. F. J. D. Taylor, Carrier System No. 7, Post Office Electr. Eng. J. **34**, 101-108 and 151-158 1941/42.

[8]) Meanwhile this has been done at the 14th plenary session of the C.C.I.F. at Montreux, in October 1946.

The most obvious method of placing the 12 channels in the standardized frequency band of 12-60 kc/s is to modulate the 12 audio-frequency channels directly with 12 carriers of 12, 16 . . . 56 kc/s, suppressing the lower side band and other undesired modulation products [9]) by means of 12 band-pass filters with pass-bands of 12-16, 16-20 . . . . 56-60 kc/s (single modulation [10]).

Such a system of single modulation has two drawbacks, already noticeable with only 12 channels and even more so as the number increases. In the first place a like number of dissimilar band filters have to be made, and so many different elements in the construction of a system is in itself an objection on the grounds of economy. Furthermore, the higher the frequency the more difficult it is to build a band-pass filter with a certain transmission band, in this case of 4 kc/s. One of the reasons for this is that the coils used in a filter involve losses, and these losses cause, *inter alia*, accentuated attenuation at the edges of the transmission band. This can, it is true, be compensated to a certain extent by equalisers, but it leads to undesired complication and to additional attenuation. The differences in filter performance also have the unpleasant result that the channels of the system do not all have exactly the same loss-frequency curve, even after equalisation. Another difficulty in the design of band filters for high frequencies arises from the inevitable small deviations from the nominal values of inductance and capacity and from the variations of these quantities with temperature. These variable factors lead to a certain relative deviation from the required frequency limits of the filter. The absolute deviation, which is always of importance in carrier telephony, varies directly with the frequency.

These and similar considerations have led to the adoption of multiple modulation for systems with a large number of channels. As a matter of fact various methods of multiple-channel modulation are applied also to systems having not more than 12 channels. To give an idea how far the above-mentioned disadvantages of single modulation are thereby avoided, we shall take as an example the 12-channel system as it was introduced in Holland in 1935 *(fig. 7)*.

As in the 48-channel system described above, each channel was first modulated with a terminal carrier of 60 kc/s to form a basic channel of 60-64 kc/s. The lower side-band (with other undesired modulation products) was suppressed by a channel band filter. In the second stage of modulation the channels were inverted and brought into their desired places in the final frequency band 8-56 kc/s [11]) by means of 12 channel carriers of 72, 76 . . . 116 kc/s. As may be seen from fig. 7, in the channel modulation all undesired side bands are in the region 32-180 kc/s, well beyond the range of the final frequency band. In principle, therefore, there is no need of a separate filter in each channel after channel-modulation to suppress the unwanted side bands, and this can almost be accomplished by a group filter common to all 12 channels, and having a pass-band of 8-56 kc/s. A very simple filter consisting of a single tuned circuit in each channel is sufficient to suppress any residual minor modulation products within the pass-band of the group filter. (This tuned circuit will be referred to as a "transducer", to distinguish it from a filter and to emphasize that one of its functions is to couple the channel modulators (and demodulators to the busbars without excessive interaction). The group filter is also of simple design owing to the relatively wide band-pass and because there is no functional requirement which necessitates sharp cut-offs.

Compared with the system first described in which single modulation is applied to 12 channels, the objections previously mentioned are partly

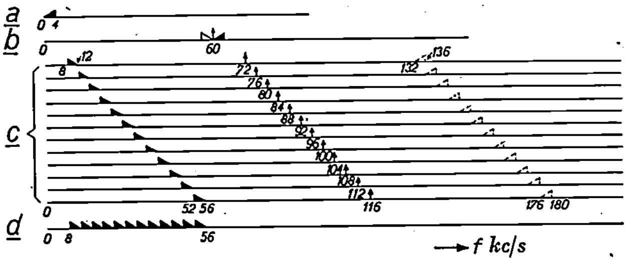


Fig. 7. Modulation diagram of the 12-channel system introduced in the Netherlands in 1935. Double modulation with a basic channel at 60-64 kc/s and final allocation of the channels in the 8-56 kc/s frequency band.

[9]) See, for example, F. A. de Groot and P. J. de Haan, Modulators for carrier telephony, Philips Technical Review 7, 83-91, 1942.

[10]) In principle the 17-channel system described in this journal was a development of this idea, but at the request of the Australian P.T.T., who were going to adopt this system, two channels were added to the original 12 on the lower side of the frequency band and three on the upper side, making a 17-channel system in the 4-72 kc/s band. See, *inter alia*, Philips Technical Review 6, 325, 1941 7, 104, 1942; 8 137 and 168, 1946.

[11]) When it was found possible and desirable to use still more channels the Netherlands 12-channel system was extended with eight more channels between 68 and 100 kc/s.

avoided; the 12 channel filters are all identical and the same frequency characteristics are obtained on all channels. The number of carriers, however, is not reduced, and the "universal" channel band filter is slightly more difficult to design than the filter for the top channel of the single modulation system.

Similar considerations hold for systems with more than 12 channels, except that the advantages of multiple modulation are then more pronounced. The channels are assembled in groups of twelve and further processes of modulation are used to form a super-group from a number of groups or (as in the co-axial system) to combine a number of super-groups each of five groups to form a hyper-group. The formation of the basic group of 12 channels in our 48-channel system is exactly analogous to the assembly of the final group in the 12-channel system described, except that the former lies in the 252-300 kc/s-band. Owing to the high-frequency of the basic channel (60-64 kc/s) the upper side bands after channel modulation are well separated from the carrier leaks and the lower side bands, so that the unwanted products can, therefore, easily be suppressed in a common group filter. In transposing the four basic groups to form the 48 channel super-group another common filter is needed, the super-group filter. But this, too, is very simple because here again the undesired side bands are well outside the required frequency range (see fig. 4). The number of carriers required is now only 17 for 48 channels and the frequency band of the universal channel filter lies much lower than those of the filters that would be needed with single modulation (up to 200 kc/s).

The result of the comparison between single and multiple modulation may be summarized as follows. The addition of one or more modulating stages, with the larger number of modulators involved, is undoubtedly a complication, but this is more than outweighed by the considerable reduction in the number of different band-pass filters and also, where there is a multiple of 12 channels, in the number of different carriers needed. An important consequence is that with the method of modulation described here, both for 12 and 48 channels, the channel band filter is identical for all channels and in the latter case operates at a low frequency compared with the frequency of the top channel. The advantage of such a universal channel filter from the constructional point of view may be judged from the fact that in a carrier telephone system these filters occupy about 40% of the volume and represent a corresponding percentage of the cost of the whole apparatus. Moreover, a very attractive and logical equipment layout is obtained, as may be seen from the fact that the apparatus consists of units absolutely identical for all channels. We shall refer to the equipment layout again in another article.

## Choice of terminal carrier frequency

If it be conceded that the basic channel is a correct engineering solution, there remains the choice of the terminal modulation frequency. We shall consider here briefly the various possible solutions, starting with the simplest case, a 12-channel system with double modulation.

One finds that the choice of the terminal carrier frequency is dependent on the band-width occupied by the group, and to a smaller extent on the frequency allocation of the basic group. For a 12-channel group a band-width of 48 kc/s is necessary and three cases arise according to whether the terminal carrier frequency and the selected side band fall in the following frequency ranges:

(a) In the frequency band to be transmitted (12-60 kc)
(b) Below the frequency band to be transmitted (<12 kc)
(c) Above the frequency band to be transmitted (>60 kc)

Dealing with these cases in the above order we choose, for example, a terminal carrier of say 24 kc/s, so that the basic channel occupies the band 24-28 kc/s. Then if we require an inverted basic group of 12-60 kc/s the frequencies of the channel carriers must be 40, 44 . . . . 84 kc/s. The unwanted upper side bands thus lie between 64 and 112 kc/s, which is just outside the frequency band of the desired group. As in the case of the 48-channel system already described, in which the basic channel occupies the band 60-64 kc/s, we have the advantage that the twelve unwanted side bands can be suppressed with a single group filter. There are, however, objections to this method. Although in theory the modulators might be built with such a high degree of symmetry that neither the input signal nor the channel carrier appears in the output[12]), in practice they do so to a slight extent owing to small unavoidable unbalances. Some of the channel carrier leaks will always fall within the basic group in case (a) and, in addition, if the basic group happens to have a frequency allocation which includes

[12]) See, for example, the article quoted in footnote 9).

the basic channel, cross-talk trouble will be experienced. Thus, in the example considered, each of the channels contributes to the energy in the 24-28 kc/s band, with the result that the channel allocated to this band suffers from troublesome cross-talk from all other channels. This disturbed (fourth) channel would, therefore, have to be left out of the group of twelve, and that would leave an odd number in the system. Irrespective of the location of the basic group, some channel carrier leaks (in the example, those at 40, 44 .... 60 kc/s) lead to operating difficulties in the seventh to twelfth channels, if the method of carrier signalling is applied for long-distance dialling, and are liable to cause possible overload of repeaters and intermodulation cross-talk [13]). Consequently we cannot contemplate the possibility of a basic channel within the basic group; also individual channel filters would be necessary to suppress carrier leaks in case (a), irrespective of where the basic group is located.

We shall now consider the choice of a low-frequency basic channel (case b). A 12-channel system with this method of modulation was, in fact, developed in Germany [14]), with an 8 kc/s terminal carrier and a universal basic channel between 4 and 8 kc/s (*fig. 8*). With this system, in the channel modulation the wanted and the unwanted side bands of each channel are separated by an interval of only 8 kc/s. Consequently most of the unwanted side bands, as well as some other modulation products not indicated in fig. 8, lie within the frequency band of the group, so that a common group filter does not suffice and another channel filter is needed.

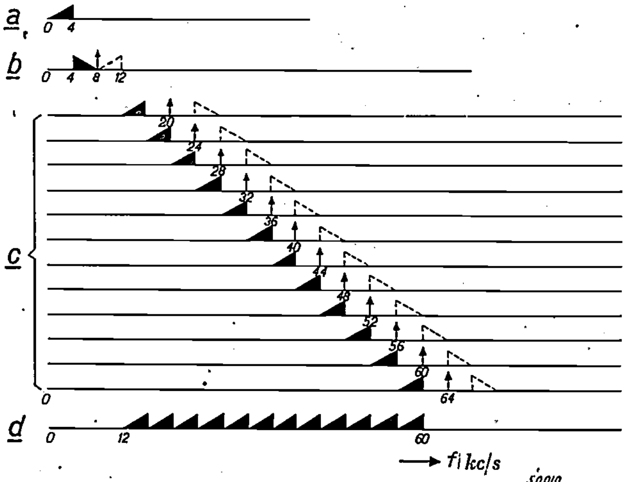


Fig. 8. Modulation diagram of a 12-channel system developed in Germany. Double modulation is used but with a basic channel at 4-8 kc/s below the final group (12-60 kc/s.)

This is a great drawback compared with the case of a basic channel located at a high frequency. On the other hand, however, a terminal carrier frequency of type (b) has the advantage that the universal channel filter is much easier to make. This is one of the reasons for the original adoption of multiple modulation.

In case (c) the wanted side bands forming the basic group are not only separated from the unwanted side bands but also from the channel leaks. 60 kc/s was chosen for the 48-channel system. It allows for the possible use of a 12-60 kc/s basic group, not containing the basic channel, in special cases, where only a double modulation system is required, without modification of standardized apparatus. The fact, however, that it is easier to make a band filter for case (b) than for case (c) with the same tolerances, is only an incidental argument against the latter, since the difficulty lies in imperfections of the component parts used. As technical development advances these imperfections will be reduced and become of minor importance compared with arguments of a fundamental nature.

The fact that filtering after channel-modulation is much simpler in case (c) is to be regarded as a fundamental advantage of this method.

It was this consideration that led to method (c) being chosen for the first carrier system in the Netherlands.

Meanwhile developments of recent years have already proved this argument to be corrrect. For the construction of the new carrier telephony system we were able to use "Ferroxcube", with which it has been possible to make coils of greatly improved properties, so that there are now no longer any insurmountable difficulties in the manufacture of a band filter for 60 kc/s. This material, indeed, more than fulfils present requirements by a margin adequate to deal also with any future demands which can be foreseen at present.

## Channel filter design

The losses in coils are usually expressed by the quality factor $Q = \omega L/R$, where $L$ is the self-inductance, $R$ the series resistance of the coil and $\omega$ the angular frequency. Over a relatively wide frequency range $Q$ is practically constant. If,

[13]) See F. A. de Groot, Signalling in Carrier Telephony, Philips Technical Review 8, 168-176, June 1946.

[14]) D. Thierbach and A. Schmid, Ein Zwölf-Kanal-Trägerfrequenzsystem für unbelastete Kabelleitungen, E.T.Z. 60, 761-768, 1939.
H. Düll, Das Deutsche Zwölfband-Trägerfrequenzsystem, Europ. Fernsprechdienst, 51, 43-49. 1939.

however, one considers the effects of dissipation in the frequency characteristic of a series of filters with a pass-band of constant width, as a function of mid-band frequency, then it is not the quality $Q$ that is decisive but rather the ratio $\varrho = R/L = \omega/Q$. With a given coil quality the attenuation distortion in the pass-band increases with the mid-band frequency and other similar difficulties are experienced.

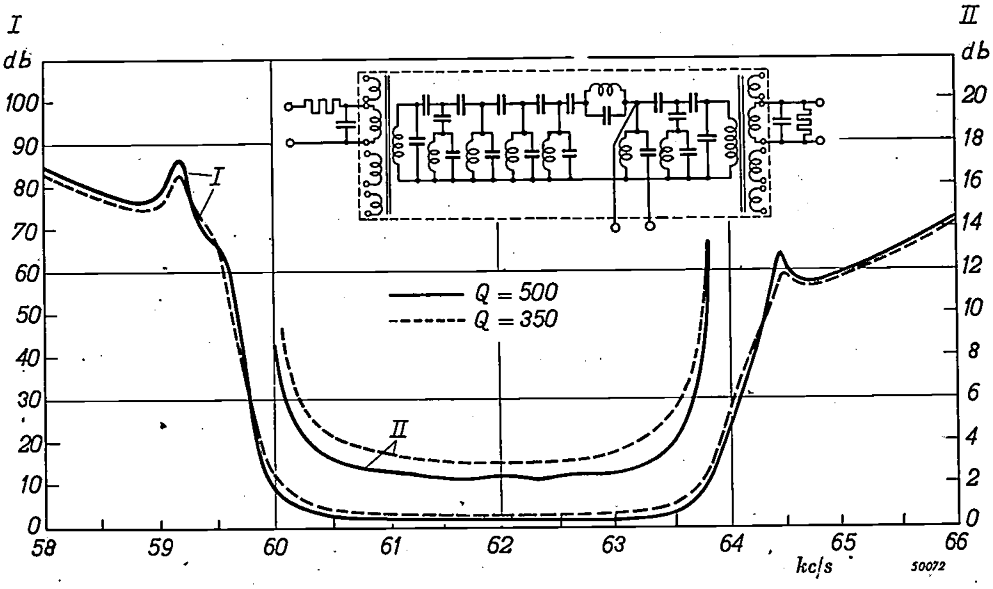


Fig. 9. Attenuation-frequency curve of the channel filter (*CBF* in fig. 6.) of the 48-channel system using coils having a "Ferroxcube" core and $Q = 500$ (fully drawn curve). The dotted curve represents the corresponding attenuation-frequency characteristic obtained when $Q = 350$. For a proper comparison of the attenuation distortion parts of both curves within the pass-band are drawn to a larger scale (see the scale on the right-hand ordinate). The diagram of the filter is shown in the inset; the terminals in the middle of the filter serve for tapping off the carrier signalling when required.

Calculations of the distortion showed that it is desirable to keep $\varrho$ below 750 $\Omega/H$, which means that a coil with $Q = 67$ is required for 8 kc/s and one with $Q = 500$ for 60 kc/s. By way of comparison it may be mentioned that in the 17-channel system previously described in this journal (see note [10])) coils were used which had a $Q$ value of about 220, whilst elsewhere dust cores have been used with which a $Q$ of 300-350 has been attained.

It is well to bear in mind that as $\varrho$ is reduced, and thus the quality $Q$ is improved, the attenuation distortion is not reduced indefinitely. Even with ideal coils and condensers there is always appreciable distortion due to the fact that a band filter can never be exactly terminated with its image impedances. According to the termination conditions, the image impedance at the edges of the pass band tends to zero or to infinity [15]). When a filter is terminated with a constant resistance, as is actually the case in practice, then at the frequencies on the edge of the transmission band reflection losses occur which result in distortion. This, therefore, is the minimum distortion that would be obtained in the case of coils and condensers with infinite $Q$. As $\varrho$ is reduced so the improvement obtained also diminishes. When $\varrho$ drops from 1500 to 750 the distortion is appreciably reduced, and upon a further drop to 500 or 400 there is still a noticeable improvement, but below that any further improvement has little effect. A simple calculation shows that with a core of given shape and material the $Q$ of a coil is proportional to the linear dimensions; in principle, therefore, any desired $Q$ could be obtained even with an inferior magnetic material if the coils can be made large — which is inadmissible from the constructional point of view. By using "Ferroxcube" as the core material for a coil of 210 cm$^3$. volume, a size which up till that time would have been considered satisfactory, it was found possible to obtain a $Q$ exceeding the design requirement by a factor of 1.7. This surplus was therefore utilised to reduce the volume of the coil by a factor of $1.7^3 = 5$, so that finally a coil was produced with a $P$ of 750-630 ($Q = 500$-600, for a filter

[15]) See, for example, Balth. van der Pol and Th. J. Weyers, Electrical Filters, II, Philips Technical Review 1, 270-276, 1936, especially p. 274. By using so-called m-sections the frequency range in which the filter impedance is practically constant can, it is true, be widened, but even then the impedance runs ultimately to zero or infinity at the edges of the pass band.

at 60 kc/s) and a volume of 44 cm$^3$. We hope to revert in a later article to the far-reaching consequences this has in the construction of the system. *Fig. 9* shows the attenuation-frequency characteristic of the channel filter built with these coils as compared with the more rounded characteristic obtained at these frequencies with a $Q$ of 350.

"Ferroxcube" also meets the requirements in respect of the temperature coefficient and stability and thus provides a very important component for the construction of the new system.

### Design of the three stages of modulation

Reviewing the method of modulation in the 48-channel system in its entirety, we see that it may be regarded as a logical development of that employed in the original 12-channel system introduced in the Netherlands. In the latter system the basic channel of 60-64 kc/s was located just above and therefore outside the final group, which in this case was also the basic group. In the 48-channel system the basic group of 252-300 kc/s is located just above and outside the final super-group, which in this case is also the basic super-group.

Whereas the object of this arrangement is qualitively the same as in the case of the 12-channel system previously discussed, the analogy is quantitively slightly different because one of the group carriers must be lower than the basic group in order to obtain the C.C.I.F. arrangement. When this is taken into account it is found that the basic group must be located above 220 kc/s in order that the lowest group carrier may be just outside the super-group; this, however, is only a small increase above 208 kc/s, which would be necessary to prevent cross-talk due to group modulator unbalances. Owing to the addition of the third stage of modulation the choice of the terminal carrier frequency could now be reconsidered, because unbalance of the terminal modulator can no longer cause cross-talk in the final super-group; but, as already explained, the choice of a frequency of 60 kc/s can be justified on other grounds. Moreover 60 kc/s was already recognised by the C.C.I.F. as a pilot frequency and must, therefore, be generated in any case. Another consideration had some influence in the final choice of the basic group. The group carriers chosen are multiples of 24 kc/s, which is of importance in the design of the carrier supply apparatus. Using a basic group of 252-300 kc/s, group carriers of 240, 360, 408 and 456 kc/s are required in order to provide the channel frequency allocation recommended by the C.C.I.F. If the basic group were located 24 kc/s lower this would be very near the minimum of 220 kc/s mentioned above and would make the filter design problems more difficult.

### Demodulation

The process of demodulation [16]) is analogous to that of modulation and also consists of three stages. The super-group is conducted to four group demodulators each supplied with a group carrier of 240, 360, 408 and 456 kc/s respectively. From each of these modulators a basic group of 252-300 kc/s is obtained, which is passed through an individual group filter to attenuate adjacent groups and remove the unwanted side band. Each group is then applied to 12 channel demodulators, each of which is fed with one of the 12 channel carriers of 192, 196 . . . 236 kc/s, so that each demodulator delivers a basic channel in the 60-64 kc/s band. This is selected by the individual channel band filters. Finally, in the 60 kc/s terminal demodulator the 60-64 kc/s band is reduced to the original audio-frequency band.

In conclusion it is to be noted that the system of modulation described here can be freely extended to a system with several super-groups, as is desired for the transmission of several hundred channels *via* a coaxial circuit. In that case a series of five basic groups is combined into a basic super-group, which is transposed to the final hyper-group by a fourth stage of super-group modulation. By selecting for the basic super-group frequencies higher than the band to be transmitted, the advantage of easy suppression of the undesired modulation products is retained in the super-group modulation, whilst retaining also the essential advantage of the system, that all channels are dealt with as uniformly as possible.

[16]) The process called demodulation is absolutely identical with modulation. One is, it is true, usually inclined to believe that in modulation a frequency spectrum is transposed from low to high frequencies, whilst in demodulation the reverse takes place; or one may well assume that in modulation the carrier wave has a higher frequency than that of the band to be modulated and that the reverse is the case in demodulation. With multiple modulation, however, neither of these two features need be present, as may be seen most clearly in the case of the 240 kc/s group carrier in the method of modulation described here. One may, nevertheless, speak of demodulation, as simply implying that the reception end is meant.

# AN IMPROVED METHOD FOR THE AIR-COOLING OF TRANSMITTING VALVES

H. de BREY and H. RINIA.

621.396.694.032.42

A method is indicated which makes it possible in principle to use air for cooling all transmitting valves which are at present cooled with water. The recognised principle is that for effective cooling it is necessary to have a large number of short air passages connected in parallel, and this has led to a system in which the air admitted is divided into a number of air currents each of which serves a definite zone (of short length) of the cooling fins. Once the dimensions of the cooling fins have been chosen, then with a given maximum anode temperature and a given ratio of the ventilator power to the power to be dissipated, the maximum specific anode loading is determined (*i.e.* the dissipation per square cm anode surface). Valves now in use already work with a specific anode load of more than 60 W/cm$^2$. In principle unlimited total powers can be dissipated. Nevertheless, the cooler may be so small that even for high powers the anode capacity and the weight are considerably less than in other air-cooling systems.

This article describes an air distributor, for distributing the air among the cooling zones, and a particularly favourable and simple construction of the cooling fins.

In a transmitting valve heat is liberated at the anode upon the conversion of D.C. energy into high-frequency A.C. energy. A smaller amount of heat is developed in the filament and the grids. As the power for which the valves were constructed increased, this lost energy also increased. Thus more and more attention has had to be paid to the question of how best to get rid of this heat. An important step in the case of powers above a certain limit was the abandonment of the construction in which a glass envelope entirely surrounds the electrode system and the adoption of a system with an external anode which forms a large part of the valve wall and is cooled with water. Until now this has been the usual construction for transmitting valves with a total dissipation greater than 4 to 10 kW.

There are, however, objections to this water cooling. The anode which is to be cooled is usually at a high potential, for instance 20 kV; the cooling water, on the other hand, comes from the mains at earth potential. An insulating connection is therefore necessary to carry the water to the anode. Along this cooling-water line a voltage gradient of not more than 1 kV/m is permissible, so that the line in question must be quite long. Furthermore, in transmitting installations the cooling-water available is often not of sufficient purity.

A cooling method with air instead of water has long been sought. Before the external anode with water-cooling became customary glass transmitting valves were sometimes air-cooled with a fan. Much more effective is the cooler already described in this periodical for the transmitting valve PA 12/15 [1])[2]), which was originally designed for water-cooling. The cooler in question consists of a cylinder of copper or aluminium into which the cylindrical anode is soldered. The cylinder is provided with a number of fins about as long as the anode itself. A fan blows air from the bottom to the top through the slits between the cooling fins. In this way a quantity of heat corresponding to about 10 kW loss in the transmitting valve can be dissipated.

When investigating whether this air-cooling system can also be realized for larger powers, various objections are encountered. These are connected with the fact that at higher powers the anode must be longer, so that with the cooling system described the air passage also becomes longer. During its passage through the slits the air rises in temperature; it therefore cools the fins at the end less than at the beginning, so that in the axial direction there is an appreciable temperature gradient in the anode. This is unfavourable, since it is a question of the maximum temperature occurring in the anode. A second disadvantage of the long air passage is that the slits must be made fairly wide, since otherwise the resistance to the air flow would become large and the fan would have to produce a fairly high pressure. It is necessary to make a compromise here because a widening of the slits, with a given circumference of the anode,

[1]) M. van de Beek, Air-cooled Transmitting Valves, Philips Techn. Rev. 4, 121-127, 1939.

[2]) In the type numbers of the Philips transmitting valves the number in front of the oblique stroke indicates the maximum anode D.C. voltage in kV, that following the stroke the delivered H.F. power (in round numbers) in kW (in the case of small valves in W), in class C adjustment.

is obtained at the expense either of the thickness or or of the number of fins. The former is unfavourable for the heat conduction through the fins, the second means a reduction in the cooling surface which must be compensated by making the fins wider in a radial direction — a measure which promotes the occurrence of a considerable temperature drop in that direction. Thus for high powers one always arrives at coolers of disproportionately large size, which are not only heavy but are also unsuitable from an electrical point of view, especially on short waves, since the anode capacity assumes an undesired large value.

### The new cooling system

A cooling system which does not possess the disadvantages mentioned and which in contrast to the older methods can also be employed for transmitting valves of very high power has been realized in the construction to be described in the following. The principle of this system was given by the late Dr. P. H. Clay with an entirely different application in view, namely for the heaters of air engines. The principle can, however, be used in many other fields.

Instead of simply passing the current of air along the whole length of the anode fins it is first divided into a number of air currents, each of which cools only a certain zone of the anode. The way in which this distribution of the air is accomplished will be explained in the following section. Because of the short length of these air channels "in parallel", it is now possible, without the necessity of too high a fan pressure, to make the slits narrow and thus increase their number. A sufficiently large cooling surface is then attained with fins which are narrow in a radial direction, and the cooler thus becomes smaller. In this way also adverse temperature differences are avoided, both in the anode, because of the short length of each cooling zone, and in the fins (radial), because they are narrow.

With the new cooling system neither the dimensions of the anode nor the total power to be dissipated are limited; for a long anode one simply needs more cooling zones than for a short one. With a given ratio of the fan power to the power to be dissipated and with the maximum permissible anode temperature, the specific anode loading (*i.e.* the power to be dissipated divided by the anode surface cooled) depends only on the dimensions of the cooling fins.

The new cooling method is distinguished not only by the above-mentioned subdivision of the air current admitted, but also by a much more economical use of the air than was the case with the older cooling methods. In the older methods large quantities of air were often passed through which increased only slightly in temperature. A much smaller amount of air and thus a much smaller fan is sufficient if the air in the cooler is made to assume a temperature of the same order as the anode temperature. In order to attain this, quite different dimensions of fins and slits are necessary than have hitherto been usual. It is here very much a question of the choice of the corrrect dimensions. By choosing the correct proportions we were successful in reducing the air consumption, which in the cooler previously described amounts to about 1.5 m$^3$ per minute and per kW of power to be dissipated, and in other models as much as 2-3 m$^3$/kW·min, to 0.8-1 m$^3$/kW·min. At the same time the dimensions and the weight of the cooler have been appreciably reduced.

We shall now deal in turn with the form of the fins and the construction of the air distributor in which the air current is distributed among the cooling zones; a new model for the cooling fins; the question of whether the ventilator should have a blowing or a suction action; the shape of the cooler housing; and how the heat accumulated in the transmitting valve can be dissipated in the event of trouble in the mains.

### *The fins and the air distributor*

In principle one may choose between two forms of fins: longitudinal fins *(fig. 1a)* in planes through the axis of the anode, and transverse fins *(*fig. 1*b)* perpendicular to the axis [3]). *Figs. 2a* and *b* show how in each case air currents can be passed along short lengths of the fins by means of partitions whose planes are perpendicular to those of the fins.

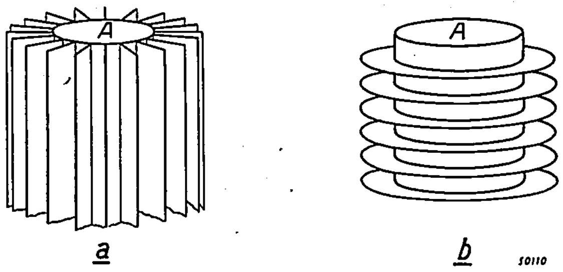


Fig. 1. Anode (*A*) of a transmitting valve, provided with (*a*) longitudinal fins, (*b*) transverse fins. The former are to be preferred.

[3]) We shall not consider here the less practical solutions such as radially directed pins or lengths of wire soldered to the anode surface.

The air currents are obtained from a fan which either blows or sucks the air through the cooler. Let us assume the latter case (we shall shortly see that a blast bas certain advantages over suction); it provides a single air current which still has to be divided into the above-mentioned smaller air currents. The way in which the fan can be connected to the air distributor, indicated very roughly in fig. 2, may be seen in *fig. 3* for longitudinal fins, in *fig. 4* for transverse fins.

A clear picture is given by *fig. 5a* of an air distributor for longitudinal fins, which, as will be seen in the following section, possesses some advantages over transverse fins. The anode provided with longitudinal fins is surrounded by piled metal boxes which are alternately open front and back and closed at the sides or closed front and back and open at the sides. As indicated by the arrows, the air is admitted from the left and right and escapes at front and back. The air entering can only escape between the fins to the adjacent higher and lower compartments. In order to obtain a uniform anode temperature "dead angles" between the cooling fins must be avoided, *i.e.* places where the air is not in motion. For that purpose the bottom planes of the compartments are provided with rims (*R* in figs. 2*a*, *b* and figs. 5*a*, *b*, *c* and *d*). The optimum dimensions of this rim and also of the compartments themselves were determined by means of enlarged models where the air was replaced by running water. It was found that the design according to fig. 5*b*, which is the simplest to construct, is adequate, and that the rim need not fit especially tightly around the fins. In agreement with this, in the actual model so much play may be allowed that the transmitting valve can easily be slid into the air-distributor, while the air leak due to this play is insignificant.

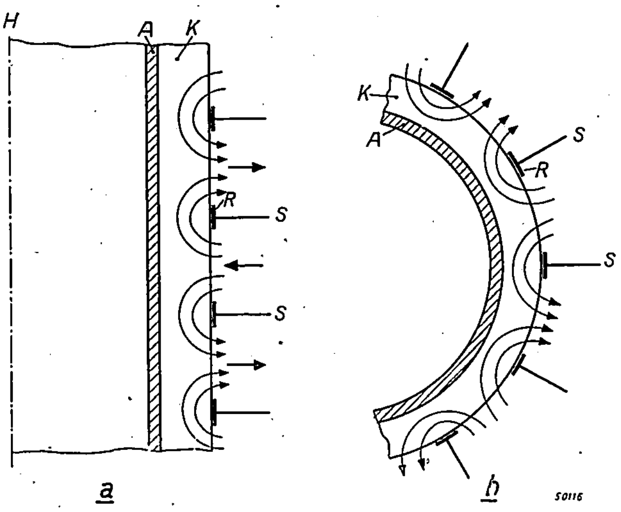


Fig. 2. Partitions divide the cooling fins into zones, each of which is served by a separate current of air; *a*) shows the case of longitudinal fins (cross-section through the axis *H* of the anode), *b*) that of transverse fins (cross-section perpendicular to the axis). *A* is the anode, *K* the cooling fins, *S* the partitions with the rim *R* serving to distribute the air better. Arrows indicate the air current.

Slightly modified models of air distributors will be discussed later.

### *Form of the fins*

This air distributor could be employed with a transmitting valve provided with coarse cooling fins as described in the article referred to in footnote [1]). Thanks to the principle of separate cooling zones this would be a great improvement. An investigation in the Philips Laboratory has shown, however, that much better results can be attained by making the slits much narrower and increasing the number of the fins. The question then arises

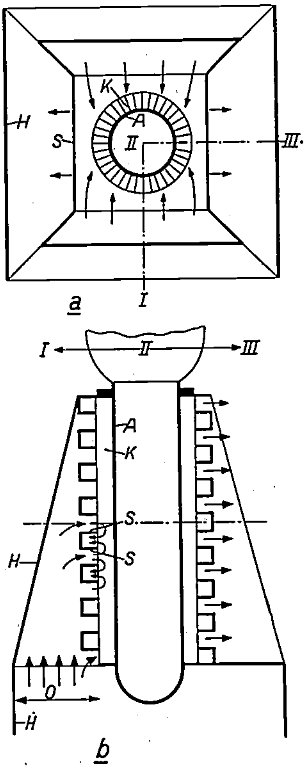


Fig. 3

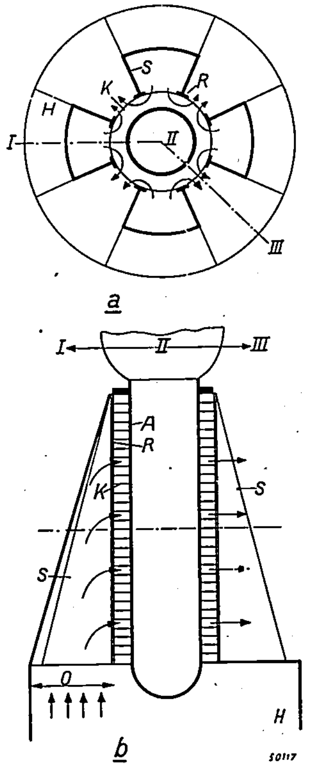


Fig. 4

Fig. 3. Transmitting valve with longitudinal fins and air distributor. *a*) is the view from above on a transverse cross section at a point at half the height of the anode, *b*) in the left-hand half is a cross section through the plane indicated in *a*) by *I-II*, the right hand half is a cross section in the plane *II-III*. *A*, *K* and *S* have the same meanings as in fig. 2. *H* is the cooler housing through which the air is admitted from below *via* the openings *O*.

Fig. 4. Transmitting valve with transverse fins and air distributor. *a*) shows the view from above of a transverse cross section at a point at half the height of the anode, in *b*) the left and right-hand halves are cross sections in the planes indicated in (*a*) by *I-II* and *II-III*. The letters have the same meanings as in fig. 3.

as to how to construct an anode with such fine fins. The problem of heat contact between the anode and the cooling fins then also becomes prominent. In this respect the ideal must be considered to be the construction of anode and fins of copper in one piece, but the casting or fraising of an object with such fine fins is very difficult. The fins must therefore be attached to the anode, and the best way is to solder them on; for good heat conduction a pure metal (for example tin or cadmium) rather than an alloy is used as solder.

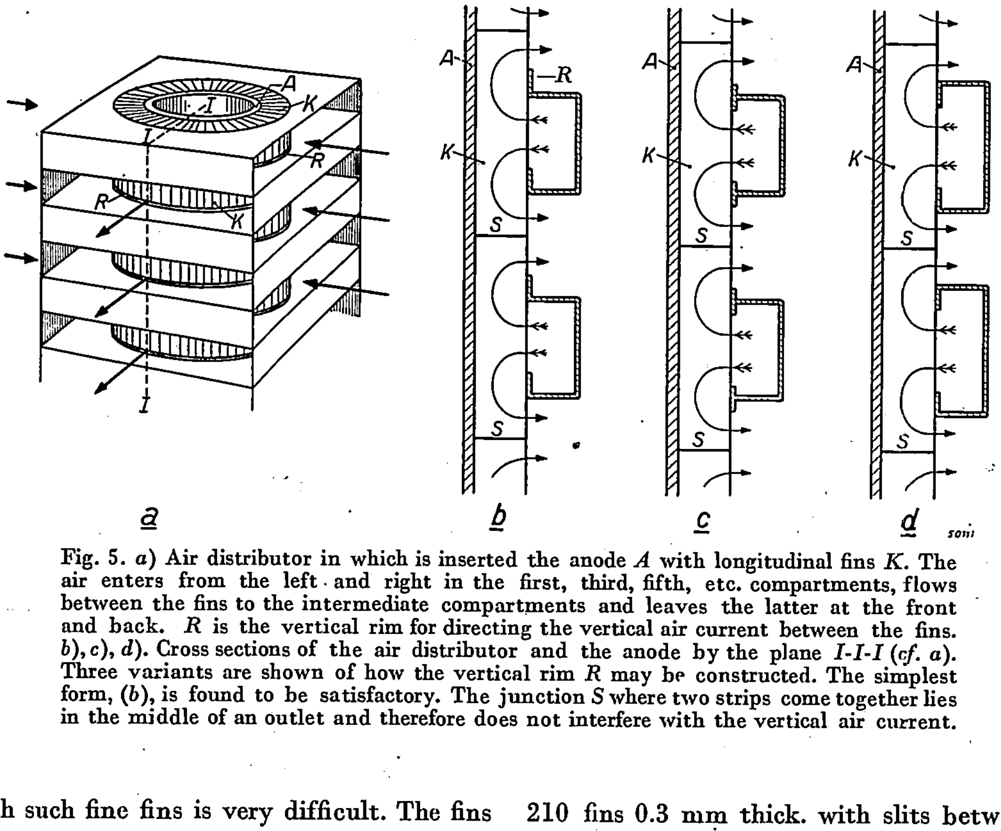


Fig. 5. *a*) Air distributor in which is inserted the anode *A* with longitudinal fins *K*. The air enters from the left and right in the first, third, fifth, etc. compartments, flows between the fins to the intermediate compartments and leaves the latter at the front and back. *R* is the vertical rim for directing the vertical air current between the fins. *b*), *c*), *d*). Cross sections of the air distributor and the anode by the plane *I-I-I* (*cf. a*). Three variants are shown of how the vertical rim *R* may be constructed. The simplest form, (*b*), is found to be satisfactory. The junction *S* where two strips come together lies in the middle of an outlet and therefore does not interfere with the vertical air current.

As already stated, there is a choice between two forms of fins: longitudinal and transverse. An objection to the latter form is that due to fluctuations of temperature the ring-shaped fins may work loose, which is not the case with longitudinal ribs. A very simple solution has been found for the problem of their construction: strips of copper, first folded as indicated in *fig. 6*, are soldered side by side on the anode until it's whole length is covered. For practical reasons a single strip as wide as the length of the anode is not used, but strips only a few cm (normally 4 cm) wide placed side by side. The heat exchange with the air could be further improved by corrugating the fins or making their surface rough. The strip with the dimensions given in fig. 6 can be made in any length desired and may thus serve for anodes of different diameters (the transverse fins of fig. 1*b* on the contrary would have to be made in as many sizes as there are anode diameters). The result may be seen in *fig. 7*. The anode of the transmitting valve TA 12/20 shown here is covered by seven strips. Around the circumference of the anode, whose diameter is 60 mm, there is room for about 210 fins 0.3 mm thick. with slits between 0.6 mm wide at the inside and 0.9 mm at the outside. The breadth of the fins is 10 mm.

Where two strips come together, as at *S* in fig. 5*b*, the fins do not in general form prolongations of each other. Such a joint might hinder the vertical

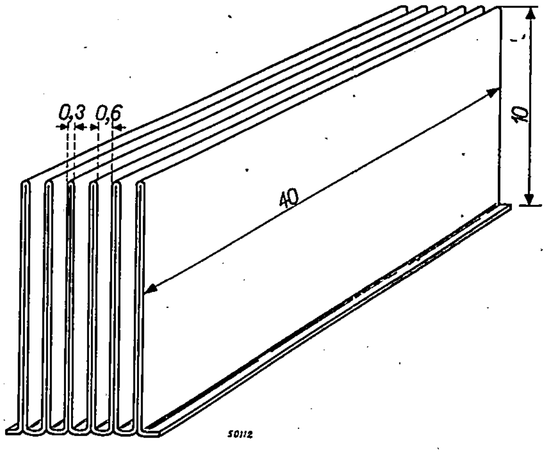


Fig. 6. A copper strip of 0.15 mm thickness is folded in the manner indicated (dimensions in mm) and then soldered to the anode. Such strips can be made in any length desired.

air current. In order to prevent this it is so arranged that the joints fall just in the middle either of an inlet or outlet opening (in fig. 5*b* an outlet). Each strip thus forms a zone which is cooled by a separate air current.

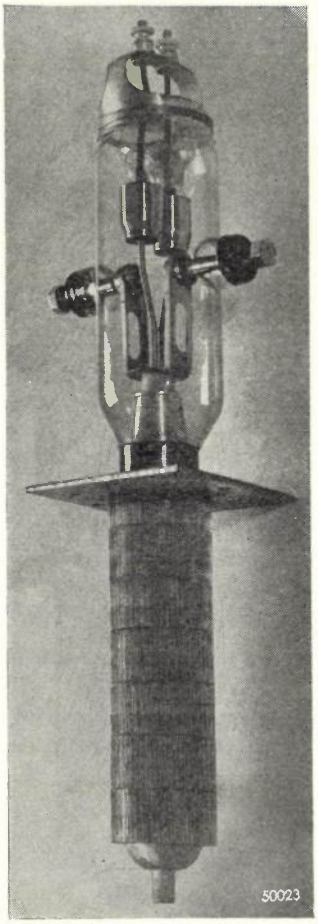


Fig. 7. Seven strips like that shown in fig. 6 are soldered onto the anode of this transmitting valve type TA 12/20.

In order to obtain a temperature distribution over the anode as uniform as possible the air must be made to penetrate deep into the slits everywhere. This is promoted by leaving a small annular opening between adjacent zones, for instance by placing the strips on the anode not side by side in contact but at a slight distance from each other. For cases with a very large dissipation per square cm surface this may be an advantage.

With the construction shown in fig. 7 it is possible to dissipate 20 kW (*i.e.* 4800 cal/sec) at a maximum anode temperature of 150° C and a temperature of the air admitted of 25° C, and with an air consumption of 20 m$^3$/min. This air current requires a difference in pressure at the fan of about 12-15 cm water column and is furnished by a fan whose motor uses about 1 kW (5 % of the power to be dissipated).

*Blast or suction?*

The question must be considered whether the fan should blow or suck; in other words whether it can best be placed in the inlet or in the outlet connection. The former is preferable on the following grounds.

Whichever of the two solutions is chosen in a given case, a certain volume of cold air must always enter the air distributor per unit of time. In order to maintain this air current the fan must in both cases provide the same difference in pressure. The two solutions differ, however, in the fact that with blowing there is cold air in the fan and with suction warm air. From the point of view of construction the latter is less desirable, so that for that reason alone it is preferable to place the fan at the inlet.

There is, however, still another argument in favour of that arrangement. The effective power *P* which the fan supplies to the air can be divided into a potential and a kinetic part:

$$P = pV + \tfrac{1}{2}Mv^2, \quad . \quad . \quad . \quad . \quad (1)$$

where $p$ is the difference of pressure caused by the fan, $V$ the volume and $M$ the mass of the air passing per unit of time, and $v$ the velocity of of the air.

If we give the quantities dependent on temperature the indices 0 and 1 respectively, as they relate to cold and warm air, the following is valid for blowing:

$$P_0 = pV_0 + \tfrac{1}{2}Mv_0{}^2,$$

and for suction:

$$P_1 = pV_1 + \tfrac{1}{2}Mv_1{}^2.$$

If $T_0$ is the absolute temperature of the cold air and $T_1$ that of the warm air, then $V_1 = V_0T_1/T_0$ (the pressure difference $p$ is small compared with atmospheric pressure) and $v_1 = v_0T_1/T_0$, so that apparently $P_0 < P_1$. Thus by having a fan with a blowing action a smaller type may be used and the energy consumption will be slightly less than when the fan works by suction.

*The cooler housing*

The cooler housing serves for the connection of the inlet and outlet lines to the air distributor. When the fan is placed in the air inlet an outlet channel may be entirely omitted if the warm air is allowed to flow freely out of the air distributor. If this should raise the temperature of the space into which this air escapes too much, the warm air must be conducted to the outside. The system can be arranged for this without much difficulty, but it is really only necessary for high powers.

The shape and the choice of material of the cooler housing are determined mainly by the following considerations:

1) the (electric) anode capacity must be kept as small as possible;
2) dielectric losses must be avoided;
3) where the material is in contact with the distributor it must be resistant to temperatures of 150 to 180 °C;
4) the fan must be insulated from the anode.

The points mentioned under (2) and (3) present objections to making the housing completely of insulating material, which objections can only be overcome by the use of ceramic material. Since, however, ceramic parts of the desired shape and dimensions were not immediately available, we at first tried a metal casing fitted over the air distributor and connected with the fan by an insulating pipe. This casing must be as small as possible in view of the capacity.

If it is a question of two valves in push-pull connection, the capacity between the anodes is kept lowest by arranging the air distributor for unilateral admission of the air, and by causing the air to enter at the sides of the cooler facing away from each other.

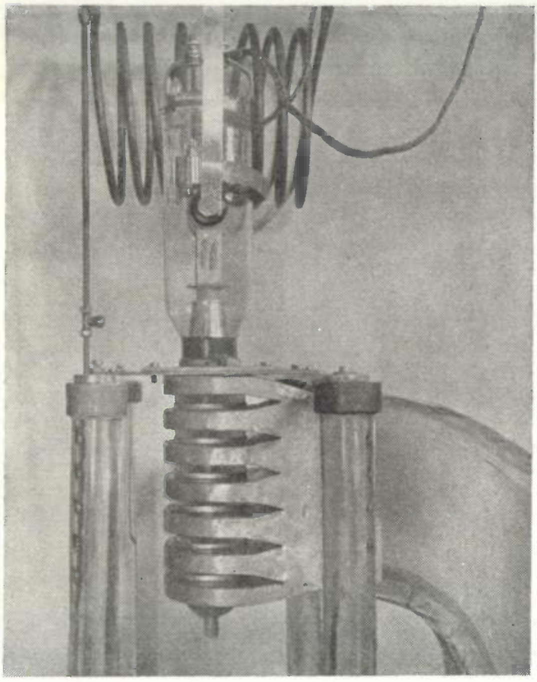


Fig. 8. Transmitting valve TA 12/20 with an air distributor into which the air is led from one side, instead of from two sides as shown in fig. 5*a*. The warm air flows away freely.

*Fig. 8* shows how in such an air distributor annular channels lead the air around to the corresponding cooling zones. Between two anodes, each provided with such a cooler and placed with their axes 40 cm apart, the capacity amounts to about 20 pF, which is not objectionable for waves longer than 10 m.

The designer of transmitters may object to this arrangement on the ground that the inlet channels obstruct access to the transmitter valve and other components of the apparatus. From his point of view it would be ideal to lead in the air from below and let it escape above. In the models sketched in figs. 3 and 4 that was the case, but the dimensions of the cooler housing and thus also the anode capacity are fairly large in those cases. This can be met to some extent by proceeding as in *fig. 9*. Here the lower half of the anode is situated inside the air inlet tube. Through the openings indicated by *O* around the cooling fins only half as much air now flows as in the design according to figs. 3 and 4; as a result in the case of fig. 9 these openings can now be made smaller without involving a higher air resistance, which results in smaller dimensions of the cooler housing and thus a smaller anode capacity.

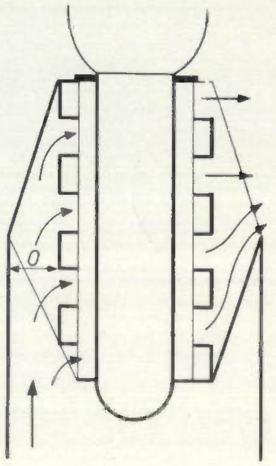


Fig. 9. Here the anode has been allowed to sink halfway into the cooler housing, so that only the air cooling the upper half of the anode flows through the openings *O*. These openings, and thus also the dimensions of the cooler housing, can therefore be made smaller than in the designs according to figs. 3 or 4. This results in a decrease of the anode capacity.

A further elaborated form of this construction is shown in *fig. 10*, where the cooler housing consists partly of ceramic material. The warm air flows out freely through openings in the cap. The air distributor consists of piled aluminium

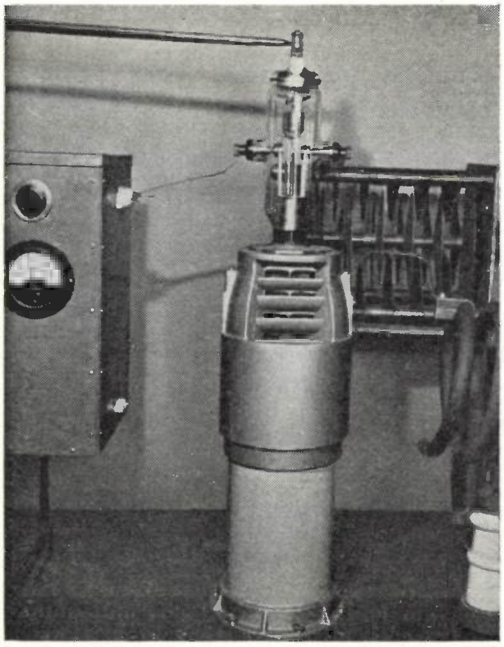


Fig. 10. Photograph of a cooler with air admitted from below. The lower part of the cooler housing consists of ceramic material.

castings which are stream-lined to keep the air resistance as low as possible. In *fig. 11* the form of these castings is shown. A separate outlet channel is also present here.

In conclusion it should be mentioned that it is desirable to allow the air, before it enters the fan, to pass through a filter which retains dust and insects.

### *Dissipation of heat upon failure of the mains*

In the event of a breakdown in the mains supplying the transmitter as well as the motor of the fan, the heating current and the anode current of the transmitting valve are both interrupted. In the filament and the grids, however, a certain amount of heat is still stored up; the heat capacity of the anode and the cooler is so small in the construction described that the anode temperature would rise higher than is permissible if the cooling air also immediately ceased to flow. Owing to the moment of inertia of the fan and the rotor, however, the fan continues to run for some time after the mains voltage has dropped out, in most cases long enough to prevent the overheating in question. If necessary, a fly-wheel may be attached to the motor shaft.

### Applications to higher powers

The above examples of the new cooling system all refer to the valve TA 12/20. The anode of this valve is loaded with 45 W/cm$^2$, at a dissipation of 20 kW. It is quite possible to go as far as 60 W/cm$^2$ with the same fins, with the quite permissible anode temperatur; of 180 °C and using a fan whose motor consumes no more than for instance 5% of the power to be dissipated. The temperature of 180 °C lies far enough below the melting point of the tin solder used (232 °C). When a solder with a higher melting point is used (for example cadmium, 321 °C) the anode temperature could be raised higher than 180 °C.

For several larger types of transmitting valves — which, like the TA 12/20, have until now been cooled with water — we find the following values of the dissipation and of the specific anode loading:

| Transmitter valve types | Dissipation kW | Specific anode loading W/cm$^2$ |
|---|---|---|
| TA 12/35 | 22.3 | 40 |
| TA 18/100 | 77 | 45 |
| TA 20/250 | 145 | 85 |

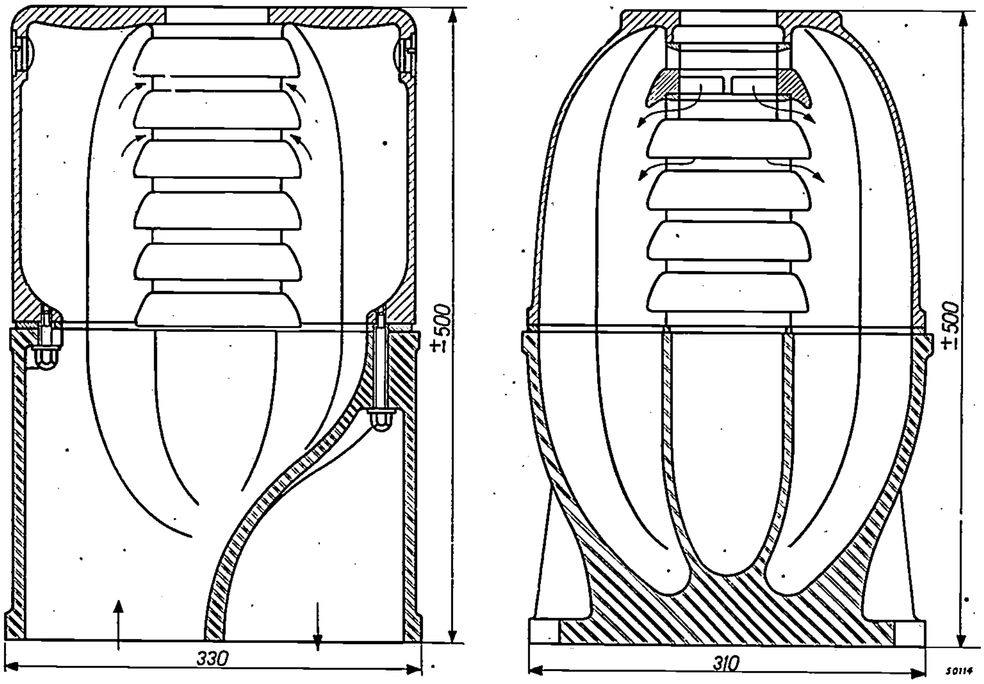


Fig. 11. Cross sections of a cooler with an inlet and outlet channel each describing 1/4 of a winding of a helix. The air distributor consists of stream-lined aluminium castings one on top of the other. The cooler housing is partly of ceramic material. The main dimensions are given in mm.

From the values of the specific loading it is evident that for the first two types in the list the new cooling system can be used unaltered in the form described. This is not, however, the case for TA 20/250, the largest transmitting valve made by Philips. By increasing the area of the cooling fins, making the slits narrower and dividing the anode into a larger number of narrower cooling zones, it is, however, possible also in this case to replace the water-cooling of this valve — which requires 130 litres of water per minute — by air-cooling.

In *fig. 12* several graphs[4]) are given which relate to a certain anode, cooled in the manner indicated by fig. 8, with a maximum temperature of 180 °C. The anode in question is that of the valve TA 12/35; when cooled with water it can dissipate 22 kW; cooled with air by the old system (described in the article referred to in footnote [1])), 10 kW. From fig. 12 it may be read off that the same anode cooled by the new method can be used for a transmitting valve which dissipates 40 kW for example, provided 31.5 m³ of air is supplied per minute. For this an air pressure of 23 cm water column is required. A fan is needed with a motor using about 2.3 kW (6% of the dissipation). With the dissipation of 40 kW mentioned the specific anode loading is about 70 W/cm².

It also follows from fig. 12 that it is no use to attempt, at constant anode temperature, further to increase the specific anode loading by increasing the amount of air $V_0$ passing through. The pressure difference required is approximately proportional to $V_0^2$, the fan power to $pV_0$, thus to $V_0^3$; the dissipation, however, is proportional to $V_0$. The "relative fan power" (*i.e.* the fan power divided by the dissipation) is thus proportional to $V_0^2$; it therefore increases rapidly upon increasing the specific anode loading. The solution must then be found, as already stated for the case of TA 20/250, in choosing different dimensions of the fins and slits and length of the zones, and/or permitting a higher anode temperature.

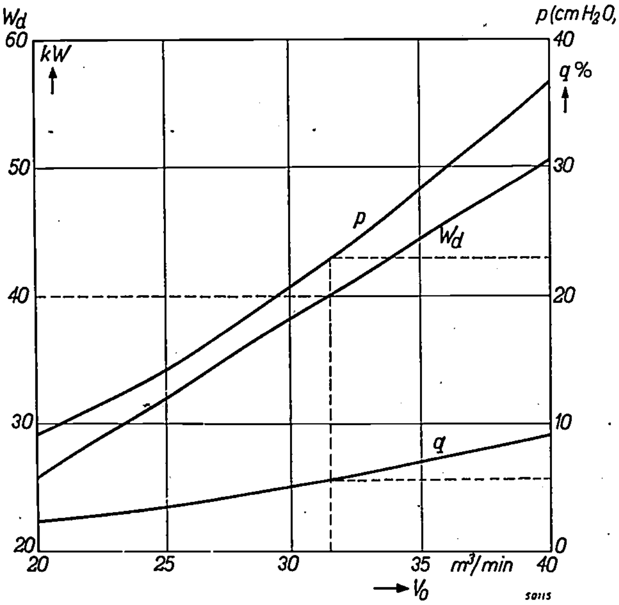
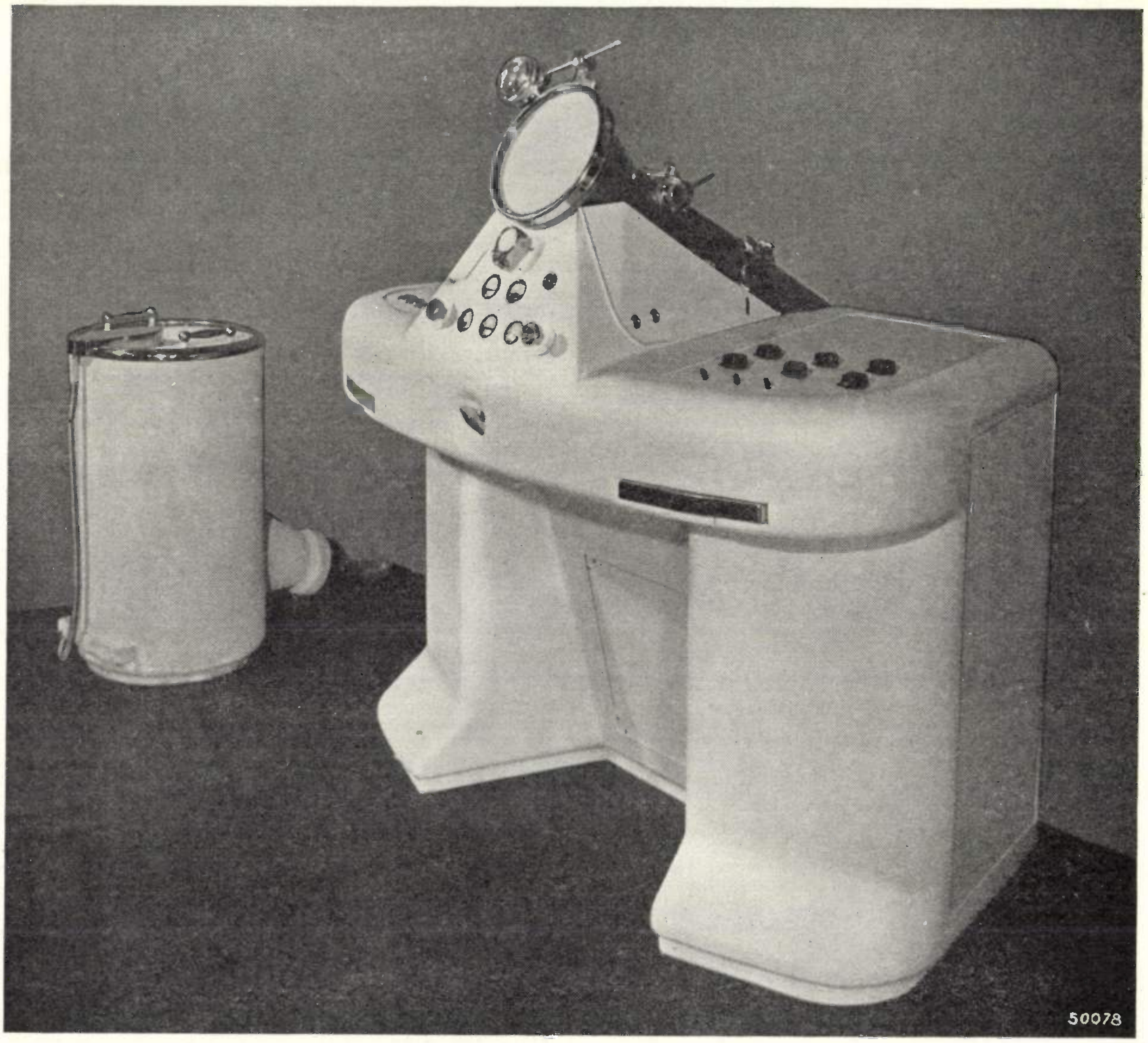


Fig. 12. Graphs relating to an anode like that of the transmitting valve TA 12/35 with a maximun temperature of 180 °C. The following are plotted as functions of the necessary amount of cold air $V_0$ in m³/min: the total dissipation $W_0$ in kW, the required pressure difference $p$ in cm water column, and the power taken up by the fan motor in percent ($q$) of the dissipation.

In conclusion it should be noted that the application of the cooling system described need not of course be restricted to transmitting valves, but may prove useful in several other fields.

[4]) These graphs are based upon measurements carried out by J. C. van Warmerdam, who has also made an important contribution towards the practical construction of the coolers described.

# A NEW ELECTRON-MICROSCOPE FOR 100 kV



In view of great interest shown in recent years in the electron-microscope for laboratory work, Philips have developed a new instrument as illustrated above. This electron-microscope is based on the pioneer work carried out in the Laboratory for Technical Physics of the Technical High School at Delft; its principles have been recently described in an article in this journal [1]): the magnification is continuously variable from 1000 × to 150 000 ×; focusing is greatly facilitated by a special method; with a few simple turns of the hand one can produce on the screen, instead of the magnified image of the object, an electron diffraction pattern of a part of the specimen previously screened and selected.

In this photograph the microscope tube is seen directed obliquely upwards. It will be noted that it has an exceptionally large viewing screen on which the image appears. Details of the image can be examined under a magnifying lens (turned away to the side in the picture). Micrographs on a 4 × reduced scale can be taken on 35 mm film with the aid of a camera built into the tube. The stabilized high voltage is supplied by a generator, on the left of the photograph, in a special housing connected with the electron-microscope by a flexible cable.

[1]) J. B. Le Poole: A new Electron-Microscope with Continuously Variable Magnifications, Philips Techn. Rev. 9, 33, 1947.

# IMPROVEMENTS IN THE CONSTRUCTION OF CATHODE-RAY TUBES

by J. de GIER and A. P. van ROOY.

621.385,832

**The use of a flat glass base with chrome iron pins has long been known in the manufacture of radio valves. By applying this construction to cathode-ray tubes more space has become available and it has thus been possible to introduce some improvements of an electron-optical nature without having to make the tube any larger. Furthermore, a new technique has been developed for the mounting of the electrodes which ensures better centering. As a result a sharper light spot is obtained, particularly at the edge of the screen. These improvements have been incorporated in a new oscillograph tube, type DG 7-3, which also has an electric screening that prevents the two pairs of deflecting plates affecting each other electrically at high frequencies.**

In the construction of cathode-ray tubes for use in an oscillograph a number of improvements have been worked out in recent years which have led to a much better quality of the image. We will discuss these improvements with reference to a new type of tube (with electrostatic deflection in both directions) in which they have already been incorporated. In the main these are improvements of an electron-optical nature, the principles of which are not new but the application of which would have involved longer tubes if the old method of construction had been maintained.

These improvements, which we shall now deal with successively, consist of:

1) changes in the leads and in the shape of the envelope;
2) a new method of mounting;
3) electron-optical improvements resulting from 1) and 2);
4) a screening between the pairs of plates.

## Changes in the leads and in the shape of the envelope

Hitherto, in the manufacture of cathode-ray tubes, a so-called "pinch" had been used for carrying the electrical leads through the glass (*fig. 1a*). Owing to the large number of leads required for these tubes (eight or nine) plus, in some cases, a number of supports to which the electrodes are affixed, cross-shaped and ring-shaped pinches have had to be employed, which were not at all satisfactory from the glass-technical point of view. Moreover — also in the simpler forms as illustrated in fig. 1*a* — the distance between the pinch and the point where it is fused in had to be several centimeters in length to prevent the pinch from being heated to too high a temperature in the fusing process; furthermore the nature of this process is such as to cause considerable variations in this length, with the result that specimens of the same type of tube are apt to show differences in length, which of course have to be allowed for in the construction of the apparatus in which the tubes are used.

Difficulties of the same nature had been experienced also with radio valves and there they were overcome by replacing the pinch by the flat base of pressed glass with a number (*e.g.* nine) of chrome iron pins. [1]) Figs. 1*b* and *c* show how the assembly of the inner parts of a cathode-ray tube can be mounted on such a standardised glass base. The saving in length compared with fig. 1*a* is already noticeable here, but it is still more apparent in the cross-sectional drawings of *figs. 2a* and *b*. This

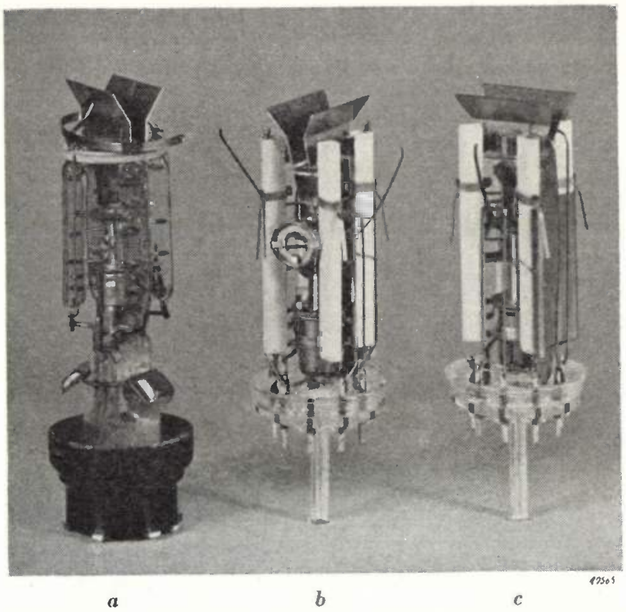


Fig. 1. *a*) The electrode system of a cathode-ray tube (DG 7-1) with the old glass construction: leads passed through a "pinch" and the electrodes fixed by means of glass "beads".

*b*) and *c*). The electrode system of the new cathode-ray tube (DG 7-3) with glass-technical improvements. Leads as in radio valves of the "Key-Valve" type: nine chrome iron pins in a base of moulded glass, the cap being dispensed with. Electrodes fixed in sintered glass contained in ceramic rods.

[1]) Philips Techn. Rev. 4, 170-175, 1939 and 6, 321-328, 1941.

saving in length is due partly to the fact that the outer ends of the lead pins serve at the same time as contact pins. The base cap seen in fig. 1*a* is thus dispensed with entirely, whilst, moreover, there are no longer any variations in length due to the cementing on of the base cap. Furthermore, the fusing of the glass base onto the accurately cut envelope can be done with much narrower tolerances than was possible with the old method.

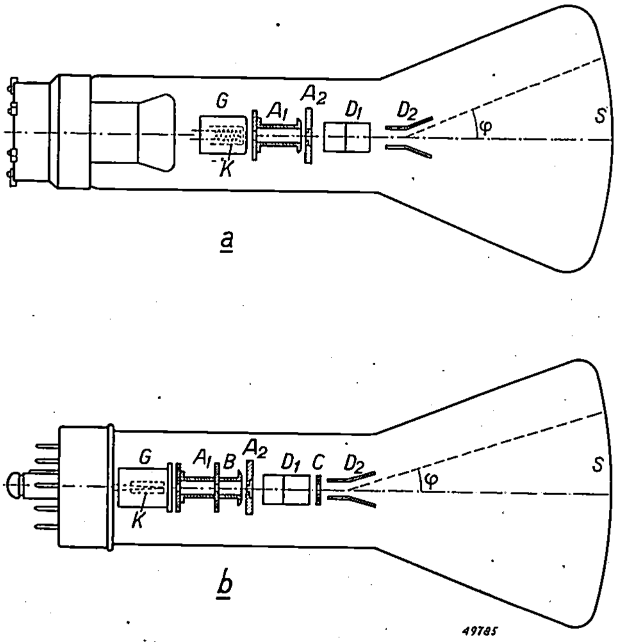


Fig. 2. Diagrammatic cross section of a cathode-ray tube, *a*) with pinch (type DG 7-1), *b*) with moulded glass base (type DG 7-3).
In both cases the electron gun consists of the indirectly heated cathode *K*, the control grid *G*, the focusing anode $A_1$ and the final anode $A_2$. $D_1$ and $D_2$ are the pairs of plates for deflecting the electron beam in two directions perpendicular to each other. In *b*) *B* is a diaphragm, *C* a part of the electric screen between the pairs of deflectors. The replacement of the pinch by the flat base gives a gain in space which is utilised for the greater part to lengthen the distance between the deflectors and the screen (*S*). In this way, for a given size of picture, the maximum deflection angle $\varphi$ of the beam is reduced, which has several advantages.

Figs. 1 and 2 both relate to an oscillograph tube with a screen diameter of 7 cm, figs. 1*a* and 2*a* being those of a tube type DG 7-1, while figs. 1*b* and *c* and 2*b* are of a new tube [2]) type DG 7-3, a photograph of which is reproduced in *fig. 3*. The saving in length previously referred to averages about 30 mm on an overall length of approx. 150 mm; how this has been utilised will be shown later on. The variation in length has been reduced from 15 mm to 6 mm, which is all to the good for the construction of the apparatus.

From the electrical point of view the flat glass base has the advantage over the pinch with base cap in that there is a smaller capacitance between two adjacent pins or wires. This we will revert to in the last part of this article.

As to the shape of the envelope of the DG 7-3 tube it is to be noted that the part which is lined with the fluorescent layer is flatter than that in the older types, whilst the curvature of the end where it bends round into the conical face has a smaller radius (compare figs. 2*a* and *b*). As a result the useful screen diameter is relatively large, which is of importance when considered in combination with the improved sharpness of the light spot at the edge of the screen, which will be discussed farther on.

## New method of mounting

Before proceeding to discuss the improvements in the assembly of the electrode system we would remind our readers that this system comprises two groups of electrodes. Those of one group form together the "gun" supplying a beam of electrons, which can be deflected in two directions perpendicular to each other by the electrodes of the other group, the deflecting plates .All these electrodes must be accurately fixed in relation to each other, and therefore in the assembling of the various component parts they are "threaded" in their proper sequence on a centering pin with spacers in between, after which the whole of the mount is secured in a gauge. The electrodes are provided with radially directed supports, or poles, which have to be fixed in some way or other to strong insulators.

In the old method of mounting three or four "beads" were used, small glass rods which were heated to the softening point and into which the supporting poles of the electrodes were then pressed in. After the last bead had cooled down the centering pin and spacers were removed, leaving a mount such as is shown in fig. 1*a*. In actual practice, however, it is not easy to get invariably good results with this "bead technique": if the bead is over-heated slightly then the glass begins to flow, whereas if the temperature is not quite high enough the glass does not adhere properly to the metal support pressed into it, with the result that after cooling the support works loose; consequently the mount is then no longer exactly centered and this ultimately has an adverse effect upon the sharpness of the spot of light. The drawback of the flowing of the glass is particularly evident when soft glass is used, whilst unsatisfactory adhesion to the wire occurs particularly with hard glass; a good compromise cannot be found,

[2]) A description of an oscillograph incorporating this tube will appear in this journal shortly.

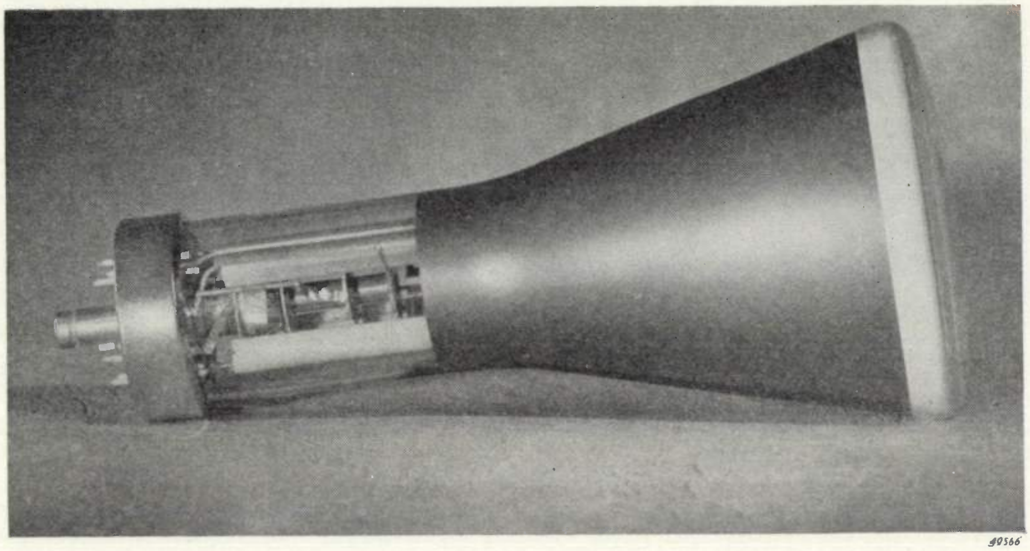


Fig. 3. The new oscillograph tube DG 7-3. Screen diameter 7 cm, overall length approximately 15 cm. The cap on the left protects the pumping stem and has a stud, so that the tube fits into the socket in only one way and the pins are automatically connected in the right way.

also because of the fact that only those kinds of glass can be used which have a coefficient of expansion not differing too much from that of the wire used for the metal supports.

During recent years a new technique has been developed whereby the glass beads have been replaced by ceramic rods (*fig. 4*) having a groove filled with sintered glass [3]). Thanks to the heat resistance of the ceramic material, for the mounting of the electrode system this rod can be heated till the sintered glass is liquefied. The glass is held in the groove by capillary action and readily flows round the electrode poles inserted in it, so that an excellent adhesion is obtained. Figs. 1*b* and *c* show the electrode system of the oscillograph tube DG 7-3 mounted in this way. This method, which thus allows of a more accurate mounting and, moreover, saves time, is already being applied also for other types of tubes.

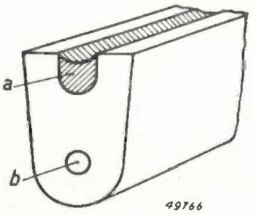


Fig. 4. Section of one of the ceramic rods used in the new mounting technique in the place of the glass beads. *a* is a channel filled with sintered glass. The opening *b* is for a supporting pole, which is afterwards welded to a lead pin.

## Electron-optical improvements

As already mentioned, the replacement of the pinch by the flat base meant a saving of about 30 mm in length. If we leave the dimensions of the tube and of the electrode system roughly unchanged, then the distance between the deflecting plates and the screen is increased by that amount. The angle $\varphi$ (fig. 2) through which the electron beam has to be deflected to describe on the screen an image of a certain maximum size decreases approximately in inverse proportion to that distance. In several respects it is advantageous to have a small angle. In the first place it means greater deflection sensitivity: less tension is required on each pair of plates to give a certain deflection on the screen. The sensitivity of the new tube (DG 7-3) is in fact about 15% greater than that of the older types DG 7-1 and DG 7-2. But a still more important result of the smaller deflection angle is that it greatly reduces the errors of deflection causing defocusing at the edge of the screen.

The manner in which one of these errors of deflection arises is shown in *fig. 5.* The electrostatic "lens", formed by the electric field between the focusing anode $A_1$ and the end anode $A_2$, focuses the electron beam in a round spot $P$ on the screen; the tension between the deflecting plates $D'$-$D''$ (the other pair of plates is disregarded here) is assumed to be still zero. When applying a positive voltage to $D'$ and a negative voltage (with respect to the final anode) to $D''$ the electrons in the beam on the side near $D'$ are accelerated whilst those on the side near $D''$ are retarded. Now, with a given tension between the deflecting plates, the deflection of an electron beam is the smaller according as the velocity of the electrons is higher. Therefore the electrons near $D'$ will undergo a smaller change in direction than those near

[3]) Various other applications of sintered glass are dealt with in an article by E. G. Dorgelo, Philips Techn. Rev. 8, 2-7, 1946.

$D''$; they strike the screen at $P'$ and $P''$ respectively. The originally circular spot of light $P$ becomes an oval spot $P'P''$. Calculations show [4]) that the magnitude of this error of deflection is proportional to the second power of the mean deflection angle, so that a relatively small reduction of the latter is sufficient to reduce the error appreciably.

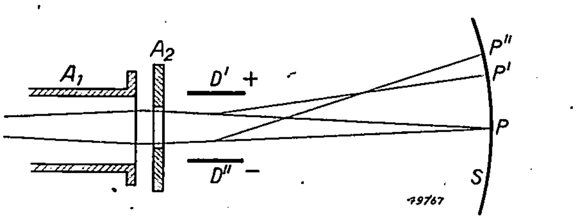


Fig. 5. Explanatory illustration of one of the errors of deflection. The "lens" formed by the electric field between the focusing anode and the anode ($A_1$ and $A_2$ respectively) concentrates the electrons of the beam into a round spot $P$ on the screen $S$, so long as the tension between the deflecting plates $D'$ and $D''$ is zero. $P'$ and $P''$ are points where the outermost rays of the beam strike the screen when there is a voltage of the indicated polarity between $D'$ and $D''$.

Another cause of unsharpness lies in the highly inhomogeneous field at the edges of a deflecting plate causing defocusing when the electron beam passes very closely to the plate. In the tube DG 7-3 the distances between the plates are the same as in the corresponding older types, but thanks to the smaller deflection angle the beam can be kept sufficiently far away from the plates to avoid any trouble on that account.

The greater distance from the lens to the screen — to which the advantages just mentioned are to be ascribed — has, however, also a less favourable effect. The magnification, as given by the ratio of the lens-screen distance to the lens-cathode distance [5]), is thereby increased and results in reduced sharpness of the light spot on the screen. In order to avoid this effect, part of the extra length available has been utilised to increase the lens-cathode distance so as to reduce the magnification and thus give a greater sharpness in the middle of the screen. The gain in sharpness at the edge of the screen due to the reduced errors of deflection is much greater.

This increase in the lens-cathode distance has been obtained by extending the focusing anode ($A_1$ in fig. 2). At the same time a diaphragm ($B$, fig. 2$b$) has been introduced, such as is usual in other types of cathode-ray tubes. By limiting the beam diameter a diaphragm contributes towards greater sharpness of the light spot. In principle the diaphragm could be placed anywhere in the beam in front of the deflecting plates, but by placing it in a field-free space — such as in the middle of the tubular focusing anode — it does not need to be so precisely centered and, moreover, it avoids undesired effects caused by secondary electron emission. The secondary electrons released from the diaphragm then all return to the wall of the field-free space without any chance of their being drawn into the beam; such would happen if the diaphragm were placed in the final anode, for owing to their very low velocity the secondary electrons would then all be driven onto the deflecting plates and constitute a troublesome current load. The secondary electrons might also be taken up in the beam if the diaphragm were placed at the outer end of the focusing anode, because after passing the final anode their velocity is lower than that of the electrons coming from the cathode (they have not passed through the cathode-focusing anode voltage difference )and consequently they would be sent off at a greater angle by the deflecting plates and not strike the screen in the same spot as the main beam. There is no sense in placing the diaphragm at the input end of the focusing anode because the beam there is already narrow. Therefore the best place for the diaphragm is about half-way along the focusing anode.

## Screening between the pairs of deflectors

Besides the mechanical and electron-optical improvements the tube DG 7-3 incorporates another new feature of an entirely different nature, a screening between the two pairs of deflecting plates. This circumvents the trouble, occurring especially at high frequencies, of a voltage on one pair of plates tending to generate a voltage on the other

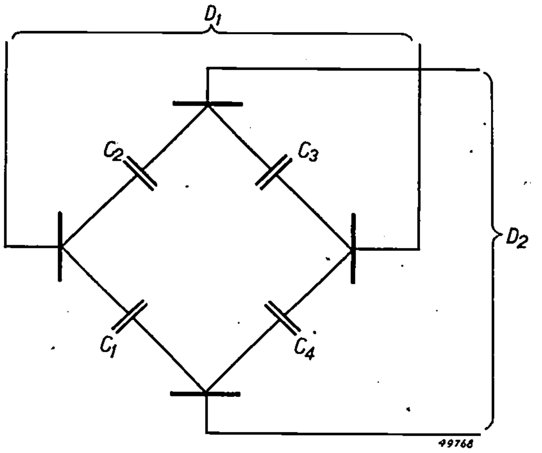


Fig. 6. $C_1$, $C_2$, $C_3$ and $C_4$ are the stray capacities between the deflecting plates (and their leads) belonging to different pairs. Through these capacities the pairs of plates $D_1$ and $D_2$ are apt to exercise an adverse electrical effect upon each other.

[4]) P. Deserno, Arch. Elektrotechn. 29, 139-148, 1935.

[5]) Strictly speaking, one should not take the lens-cathode distance but that from the lens to the smallest diameter of the beam between the cathode and the lens; it is in fact this smallest diameter that is thrown on the screen, but it is so close to the cathode (1-2 mm) that for the sake of simplicity we may roughly speak of the lens-cathode distance.

pair, which of course is undesirable: This effect — sometimes called "cross-talk" in aualogy with certain phenomena occurring in telephony — manifests itself in a distortion of the oscillogram, which in the case of a frequency of 100 000 c/s and over may be very troublesome. The cause of this lies in the stray capacities $C_1$, $C_2$, $C_3$, $C_4$ (*fig. 6*) between the plates (including their leads) which belong to different pairs, or rather in the inequality of those capacities. As may be calculated, the pair of deflectors $D_2$ would not be affected by $D_1$ if $C_1$ were equal to $C_2$ and $C_3$ equal to $C_4$; inversely $D_1$ would not be affected by $D_2$ if $C_1$ were equal to $C_4$ and $C_2$ equal to $C_3$. This effect could, therefore, be neutralised in both directions if the four capacities were made equal with the aid of correcting capacitors, but these are so small (only a few pF) that correction is impracticable. A simpler way is to apply a screening so as to reduce these capacities far enough for the differences to be so small as to render the "cross-talk" imperceptible. That it has been possible to achieve this is due partly to the fact that the capacities between the pins in the flat glass base are so much smaller than those between the lead wires in a pinch with base cap.

The screening referred to consists of two metal partitions, one ($C$ in fig. 2*b*) between the two pairs of deflecting plates, with an aperture for the passage of the beam, and another between the two pairs of lead wires. These partitions, clearly to be seen in fig. 1*c*, are connected to the final anode. By these means the capacities have been reduced from a few pF to less than 0.1 pF. Provided also the external leads are properly screened there will no longer be any trouble from mutual effect between the deflectors, not even at very high frequencies.

It goes without saying that the various improvements described here will not be confined to one type of tube but will be applied also in other types where necessary, as is in fact already being done.

# ON THE CRYSTALLINE STRUCTURE OF FERRITES AND ANALOGOUS METAL OXIDES

by E. J. W. VERWEY, P. W. HAAYMAN and E. L. HEILMAN†

546.723-3:548.73

Ferrites are binary oxides, the technically most important type of which is indicated by the general chemical formula $MFe_2O_4$ (M a bivalent metal). The ferrites of particular importance in electrotechnology are those with a crystal structure analogous to that of the mineral spinel $MgA1_2O_4$. These ferrites form an essential component of the new magnetic material for high frequencies, "Ferroxcube", and also of certain resistance materials which have a large (negative) temperature coefficient of resistance. The magnetic and electrical properties of these ferrites and of the mixed crystals of which they form a part depend very closely upon certain peculiarities of their crystal structure. The latter is described in this article. It is found that the data about the ferrites in question and their mixed crystals with substances of analogous structure can be summarized in three "rules", which considerably facilitate the preparation of materials with certain dcsired physical properties.

In recent years all kinds of new materials of importance in electrotechnology have been developed in the Philips Laboratories. Among these certain ferrites and the mixed crystals of ferrites occupy a special position. The ferrites are binary oxides with the formula $MFeO_2$ or $MFe_2O_4$, where M is respectively a monovalent or bivalent metal. In this article we shall deal exclusively with ferrites with bivalent metal. The new magnetic material for high frequencies, "Ferroxcube", is composed of mixed crystals of these ferrites; this material has recently been discussed in detail in this periodical[1]). These ferrites also constitute part of certain mixed crystals which have a practical significance because of their large negative temperature coefficient of resistance. All the ferrites used in "Ferroxcube" and in the resistance materials mentioned have this in common, that they have the same crystal structure, namely that of the mineral spinel $MgA1_2O_4$, which crystallizes in the cubic system. It has become customary to call ferrites and other related oxides having the spinel structure and corresponding to the formula $XY_2O_4$, where X and Y indicate metals, also spinels. We shall often use that term in this article and in that way avoid the confusion which might arise from the term "ferrites", because of the fact that ferrites with bivalent metal are also known, which have a different structure.

Since the magnetic and electrical properties of spinels are very closely connected with the position of the metal ions in the crystal lattice, it was desirable to study the spinel structure in detail. The results of this crystallographic investigation will be discussed in this article; in a subsequent article we shall return to the electrical properties of the spinels and their employment as resistance materials (their magnetic properties were the subject of the article already referred to).

## The spinel lattice

The structure of an ideal crystal lattice is completely given as soon as the arrangement of the atoms in the so-called elementary cell is known. The elementary cell is the smallest structural unit and in the most general case it is a parallelopiped. By placing such parallelopipeds side by side and piling them on top of each other in such a way that corresponding edges are parallel, the whole crystal is obtained.

From X-ray analysis the following is now known about the structure of an elementary cell of the spinel lattice.

The elementary cell is a cube containing 8 molecules of $XY_2O_4$, *i.e.* a total of 56 atoms. Due to the presence of such a large number of atoms per elementary cell, its structure is quite complicated. We shall first concern ourselves with the position of the oxygen ions and then with that of the metal ions [2]).

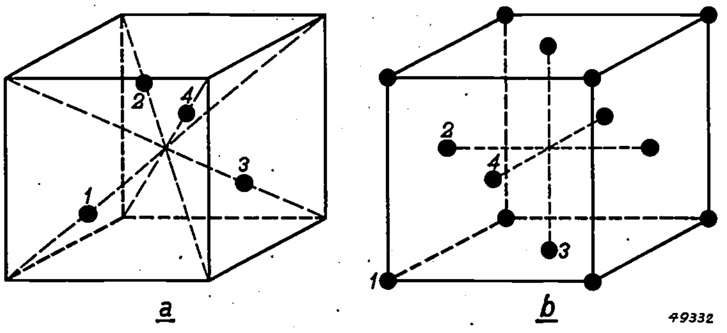
Let us imagine for the time being that all the metal ions are removed from the spinel lattice. The lattice of the remaining oxygen ions is then relatively simple. The elementary cell of this oxygen ion lattice is then found to be twice as small linearly as that

[1]) Philips Techn. Rev. 8, 359, 1946.

[2]) For the sake of convenience we speak here of oxygen ions and metal ions. This does not mean, however, that we wish to imply that one is concerned with a pure ionic binding in the spinel lattice.

of the spinel lattice, so that it contains only four oxygen atoms. In other words, if the elementary cube of the spinel lattice is divided into eight equal cubes — which we shall call "octants" — these eight "octants" are absolutely identical as far as position of the oxygen ions is concerned.

The position of the four oxygen ions in an octant is now such that on each body diagonal of the octant there is one oxygen atom, as represented in *fig. 1a*; the distance of the ion to the closest corner point of the octant is the same for all four ions and amounts to 1/4 of the length of the diagonal.



Figs. 1 *a*) anb *b*): Two possible choices of the elementary cell of the oxygen ion lattice in spinels. The large black dots indicate oxygen ions. *b*) is formed from *a*) by shiftirg all the lattice points parallel to the body diagonal on which oxygen ion *1* is siuated.

The centres of the oxygen ions in this oxygen ion lattice have the same spatial arrangement as the centres of a packing of spheres with a cubic symmetry where the empty space between the spheres is as small as possible, *i.e.* the so-called closest cubic packing of spheres.

Whether the oxygen ions in the spinels "touch" each other like the spheres in the cubic packing depends, of course, on the size of the metal ions which must be situated between the oxygen ions. Now in general the metal ions are considerably smaller than the oxygen ions. Therefore they can be placed in the interstices between the cubic packing of the oxygen ions without causing it to "swell" too much. Thus the oxygen ions nearly touch each other.

It should be mentioned here that actually the oxygen ion lattice of the spinels deviates somewhat from the arrangement of the closest cubic packing of spheres. The deviations are due to the fact that the metal ions do not push aside the oxygen ions directly surrounding them everywhere in the same way. These displacements, however, take place in such a way that the cubical symmetry is retained.

We shall now discuss the position of the metal ions in the interstices between the oxygen ions. There are two kinds of interstices:

Firstly, the interstices surrounded by four oxygen ions forming a tetrahedron; these interstices may be called tetrahedron spaces. Secondly, the interstices surrounded by six oxygen ions forming an octahedron and consequently called octahedron spaces. The tetrahedron and octahedron spaces are given in figs. *2a* and *b* respectively. The arrangement of the tetrahedra and octahedra can be made somewhat clearer by choosing the elementary cell of the oxygen ion lattice differently. If one imagines all the lattice points to be displaced in the direction of a body diagonal in such a way that the oxygen ion indicated as *1* in fig. 1*a* lies at the lower left-hand corner, the elementary cell shown in fig. 1*b* is obtained. In the same way figs. 2*a* and *b* then become figs. 2*c* and *d*. By reference to these figures it may be seen that per elementary cell of the oxygen ion lattice (*i.e.* per octant of the elementary cell of the spinel lattice) there are 4 octahedron spaces and 8 tetrahedron spaces. Thus per elementary cell of the spinel lattice we have at our disposal 96 spaces.

Which of the 96 spaces are occupied by the 24 metal ions?

As far as the position of the metal ions is concerned the 8 octants of an elementary cell are found to fall into two groups of 4 octants, in such a way that the 4 octants of the same sort always have one edge in common (see *fig. 3*). In the octants of one sort only the 4 octahedron spaces are occupied (*cf.* fig. 2*f*); in the octants of the other sort all the octahedron spaces are unoccupied and only two of the tetrahedron spaces are occupied, as indicated in fig. 2*e*.

*Fig. 4* shows the arrangement of the oxygen ions as well as of the metal ions in the elementary cell of the mineral spinel $MgAl_2O_4$, The illustration gives somewhat more information than follows from the above. So far we have been concerned with the position of the metal ions without

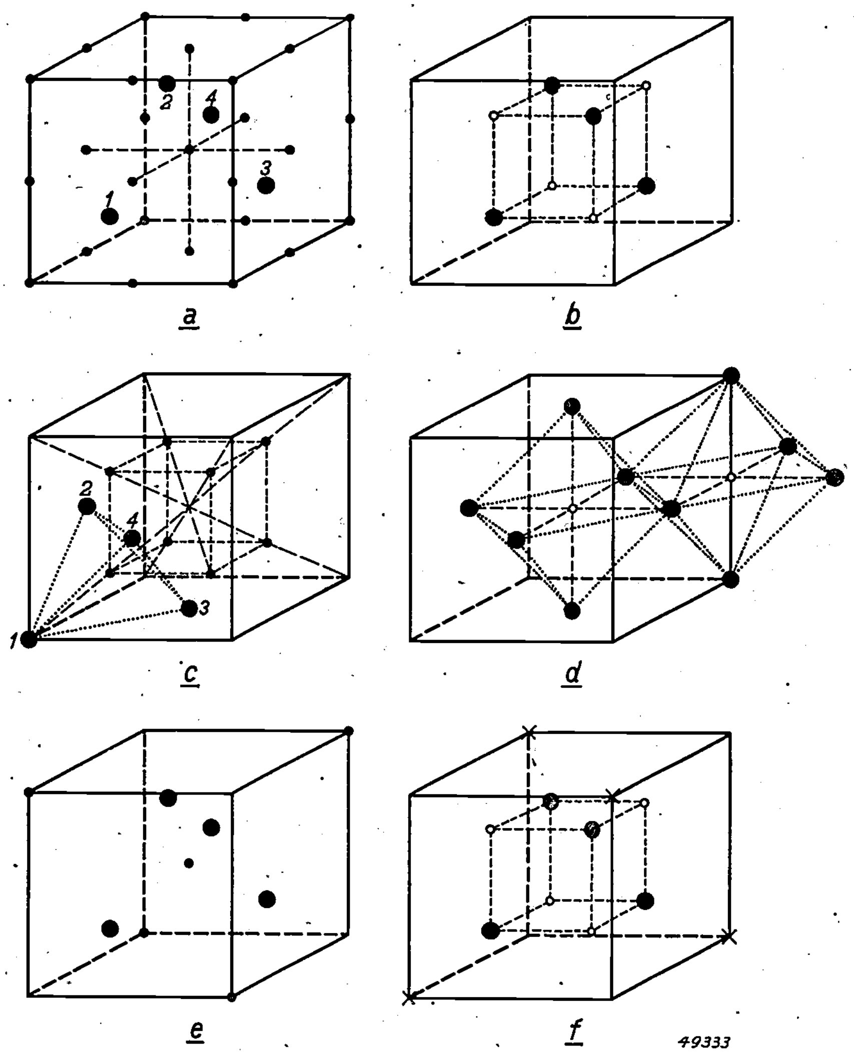


Fig. 2*a*) anb *b*): Position of the centres of the tetrahedron spaces (small black dots) and octahedron spaces (small circles) in the elementary cell of the oxygen ion lattice chosen according to fig. *1a*. *c*) and *d*): The same as in *a*) and *b*) when the elementary cell is chosen according to fig. 1*b*), in other words *c*) and *d*) are formed from *a*) and *b*) respectively by the shifting of all lattice points parallel to the body diagonal on which the oxygen ion *1* is situated. For the sake of clarity in *c*) and *d*) all the oxygen ions and the centres of all the octahedra are not indicated. *e*) and *f*) are formed from *a*) and *b*), respectively, by the omission of the centres of all those tetrahedra and octahedra which are not occupied by the metal ions; in *f*) moreover, we have indicated with crosses the centres of those occupied tetrehadra which in the text are considered as belonging to the octants of the other sort.

taking into account the fact that there are two kinds of metal ions present in every spinel, while in the illustration the positions of the two kinds of metal ions are indicated separately. We shall now discuss this latter point.

## The distribution of the different metal ions among the available spaces

The general chemical formula $XY_2O_4$ is satisfied not only by the spinels which are built up of one bivalent and two trivalent metal ions (per "molecule") (for example $MgAl_2O_4$), but also by the spinels which are built up of one tetravalent and two bivalent metal ions (per "molecule") (for instance $Mg_2TiO_4$). We shall devote our attention to the first-mentioned possibility. Analogous considerations hold for the other case.

We have seen that in the spinel structure 8 tetrahedron spaces and 16 octahedron spaces, whose position in the elementary cell was indicated, are occupied by metal ions. In the following these 24 spaces will be indicated as those "available" for the metal ions. How are the bivalent and trivalent metal ions now distributed among the available spaces? (As already stated, the answer to this question is of great importance for predetermining the physical properties of the spinels.)

At first this was not considered to be any problem at all, the 8 bivalent ions being located — it was thought — in the 8 available tretahedron spaces and the 16 trivalent ions in the 16 available octahedron spaces. In many cases this is indeed true; spinel proper, $MgAl_2O_4$, may serve as an example of this (*cf.* fig. 4).

Barth and Posnjak[4], however, pointed out that this simple assumption is by no means correct in every case. By studying X-ray diffraction photographs of a number of spinels in which the two

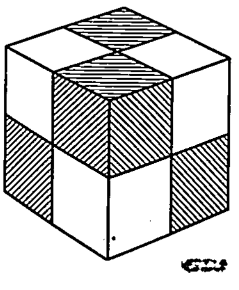


Fig. 3. The cube represents symbolically the elementary cell of the spinel lattice. The four shaded and the four non-shaded octants are occupied respectively in the same way by the metal ions, namely as in figs. 2*e*) and 2*f*) respectively.

[3] In Fig. 2*e* the centre of the cube and four corners are indicated as occupied by metal ions. It must not be forgotten, however, that each of the occupied corners must be considered as belonging to four similar octants, so that only 1/4 of it belongs to the figure in question. There are therefore, as claimed in the text, $1 + 4 \times 1/4 = 2$ occupied tetrahedron spaces per octant.

[4] T. F. W. Barth and E. Posnjak, Z. Kristallogr. **82**, 325, 1932.

kinds of metal ions have a sufficiently large difference in scattering power for X-rays, they were able to show that there are also spinels with the 8 bivalent ions in 8 of the 16 available octahedron spaces and with the 16 trivalent ions distributed equally over the remaining 8 available octahedron spaces and the 8 tetrahedron spaces. The 8 bivalent and 8 trivalent ions are at the same time distributed at random among the 16 octahedron spaces in question. In other words in the available octahedron spaces of an elementary cell one finds only on an average equal numbers of bivalent and trivalent ions. This also means that the concept of "elementary cell" has here lost its significance as far as the metal ions are concerned.

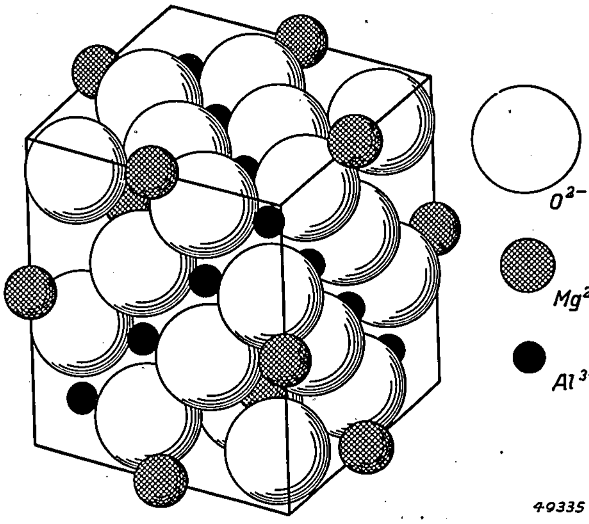


Fig. 4. Elementary cell of spinel proper, $MgAlO_{24}$. It may be seen that the oxygen ions are much larger than the metal ions.

### The electrostatically most stable configuration of a spinel lattice

When it is seen that in some spinels the equally charged metal ions are situated only in the octahedron spaces and in others over octahedron and tetrahedron spaces, the question arises as to how this distribution is determined.

It might be assumed that it depends upon the size of the ions, so that for example the smallest ions will occur as far as possible in the tetrahedron spaces, which are considerably smaller than the octahedron spaces; but this is contrary to what has been observed.

Another factor which may determine the distribution of the metal ions among the available spaces is the electrostatic energy of the spinel lattice (in the following called "lattice energy"), *i.e.* the energy gained when the ions, first considered to be at an infinite distance from each other, are joined to form the spinel lattice. Because if the chemical binding in a spinel lattice is brought about only by the electrostatic (Coulomb) forces (attraction between ions of the same sign), that distribution of metal ions will be most stable where the lattice energy is greatest. In order to judge whether this is actually the case we have calculated the lattice energy for several possible distributions of metal ions. The results of these calculations are given below; the comparison with the observations will be dealt with in the following paragraph.

We consider spinels *A*), built up of bivalent and trivalent metal ions ,and *B*) built up of bivalent and tetravalant metal ions. In case (*A*) as well as in case (*B*) there are two possibilities. In case (*A*):

*Aa*) The bivalent ions are situated only in the tetrahedron spaces and the trivalent ions only in the octahedron spaces. The lattice energy *E* per molecule of $XY_2O_4$ then amounts to

$$E_{Aa} = 150.3 \frac{e^2}{a} \text{ erg},$$

where *e* is the charge of the electron in e.s.u. and *a* is the lattice constant in cm, *i.e.* the length of the edge of the elementary cell.

*Ab*) The bivalent ions are situated only in the octahedron spaces and the trivalent ions are distributed equally over the octahedron and tetrahedron spaces:

$$E_{Ab} = 143.6 \frac{e^2}{a}.$$

In case (*B*):

*Ba*) The tetravalent ions are situated only in the tetrahedron spaces and the bivalent ions only in the octahedron spaces:

$$E_{Ba} = 142.1 \frac{e^2}{a}.$$

*Bb*) The tetravalent ions are situated only in the octahedron spaces and the bivalent ions are distributed equally over octahedron and tetrahedron spaces:

$$E_{Bb} = 150.3 \frac{e^2}{a}.$$

As to the calculations which led to these results, the following should be noted. In case (*Aa*) we obtain the same value for the lattice energy as in case (*Bb*) due to the fact that in the latter case, in which the bivalent and tetravalent ions occur in equal numbers distributed at random in the octahedron spaces, we assumed that electrostatically this presents the same picture as a distribution of trivalent ions whose number is equal to that of the sum of the bivalent and tetravalent ions. In a similar way in the calculation in case (*Ab*), where the bivalent and trivalent ions occur in the octahedron spaces, we have considered them as 2 1/2 valent ions.

A comparison of the calculated values of the lattice energy now shows that the most stable state of a spinel lattice built up of bivalent and trivalent metal ions corresponds to case (*Aa*), where the trivalent ions are located exclusively in the octahedron spaces. On the other hand for a spinel lattice built up of bivalent and tetravalent metal ions the most stable state is found to be that where the bivalent and tetravalent ions are distributed over the octahedron spaces (case *Bb*). This conclusion is valid only when the lattice constant *a* in cases (*Aa*) and *Ab*) (and in cases (*Ba*) and (*Bb*), respectively, has about the same value, which is quite plausible.

The fact that cases (*Aa*) and (*Bb*) must correspond to the electrostatically most stable states can also easily be understood qualitatively. It will be advantageous from the point of view of energy if the most highly charged metal ion is surrounded by as many negatively charged oxygen ions as possible. And that is exactly true in cases (*Aa*) and (*Bb*).

## Checking against observations

In order to compare the above theoretical results with the actual facts we determined by X-ray analysis the distribution of the metal ions in a large number of spinels. The data thus obtained combined with those already known from the investigations of Barth and Posnjak lead to the following conclusions.

For aluminates $MAl_2O_4$ and chromites $MCr_2O_4$ (M bivalent metal), where the metal ions are bivalent and trivalent, as well as for the titanates $MTi_2O_4$ and the stannates $M_2SnO_4$ (M bivalent metal), where the metal ions are bivalent and tetravalent, the actual distribution of the metal ions over the available spaces is in agreement with our calculation on the basis of the electrostatic lattice theory.

As far as the ferrites are concerned the situation is not so simple.

In $ZnFe_2O_4$ and $CdFe_2O_4$ the distribution of metal ions corresponds to the electrostatic theory; in $MgFe_2O_4$ and $CuFe_2O_4$, on the other hand, it does not: here the trivalent ions (ferric ions) are divided among octahedron and tetrahedron spaces. In the other ferrites with spinel structure, for example $CoFe_2O_4$, $MnFe_2O_4$ and $Fe_3O_4$, the difference in scattering power between the bivalent and trivalent metal ions is too small to make it possible to draw any conclusions about the location of the ions from the relative intensities of the X-ray reflections. Some conclusions may, however, be drawn in this respect if we compare the values of the lattice constant for different aluminates and ferrites with each other.

Table I.

Columns (2), (4) and (6) give respectively the values (in Å) of the lattice constant for different aluminates, chromites and ferrites, columns(3) and (5) the differences between these values for the corresponding chromites and aluminates and for the ferrites and chromites respectively. The values in column (1) are the radii (also in Å) of the bivalent metal ions. The radii of the trivalent ions are given under their chemical symbols. Cu chromite and Cd aluminate with spinel structure are unknown, The value of the lattice constant of Mn ferrite is uncertain. The figures published for the radius of the $Cu^{2}+$ ion are contradictory and this value is therefore not given here.

| (1) | | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | | $Al^{3+}$<br>0.57 | | $Cr^{3+}$<br>0.64 | | $Fe^{3+}$<br>0.67 |
| $Ni^{2+}$ | 0.78 | 8.05 | 0.25 | 8.30 | 0.06 | 8.36 |
| $Cu^{2+}$ | | 8.07 | — | — | — | 8.37 |
| $Mg^{2+}$ | 0.78 | 8.07 | 0.24 | 8.31 | 0.05 | 8.36 |
| $Co^{2+}$ | 0.82 | 8.08 | 0.24 | 8.32 | 0.04 | 8.36 |
| $Zn^{2+}$ | 0.83 | 8.07 | 0.23 | 8.30 | 0.12 | 8.42 |
| $Fe^{2+}$ | 0.83 | 8.12 | 0.22 | 8.34 | 0.05 | 8.39 |
| $Mn^{2+}$ | 0.91 | 8.26 | 0.23 | 8.49 | 0.06? | 8.55? |
| $Cd^{2+}$ | 0.97 | — | — | 8.57 | 0.12 | 8.69 |

In *table I* the values are given of the lattice constant for different aluminates, chromites and ferrites. Upon passing from an aluminate to the corresponding chromite the lattice constant increases, in agreement with the fact that the radius of the $Cr^{3+}$ ion is larger than that of the $Al^{3+}$ ion. For all pairs of corresponding chromites and aluminates this increase is approximately equal to 0.24 Å. Since the difference between the radius of the $Fe^{3+}$ ion and that of the $Cr^{3+}$ ion is about one half the difference between the radius of the $Cr^{3+}$ ion and that of the $Al^{3+}$ ion, it might be expected that the increase in the lattice constant upon passing from a chromite to the corresponding ferrite would be about $0.24 : 2 = 0.12$ Å. This is indeed true in the case of Zu and Cd ferrite, *i.e.* for the ferrites for which the distribution of the metal ions corresponds to the electrostatic theory. For all other ferrites, however, a remarkable fact is observed: the increase of the lattice constant is much smaller, and for all of them about equal to 0.05 Å.

Now among the ferrites which, as far as the lattice constant is concerned, show a deviating but mutually similar behaviour belong Mg and Cu ferrite, where the ferric ions are divided among the tetrahedron and octahedron spaces. From this we feel justified in concluding that the distribution of the metal ions in Co, Mu and $Fe^{2+}$ ferrites. is the same as in Mg and Cu ferrite, *i.e.* that the ferric ions in all these ferrites are also distributed among the tetrahedron and octahedron spaces.

There are other no less important arguments for the correctness of this conclusion.

One argument, for example, is furnished by the fact that all these ferrites are ferromagnetic with the exception of Zn and Cd ferrite. This question is briefly discussed in the article referred to in footnote [1]); we shall not go into it here. Another argument can be deduced from a consideration of the conductivity of the mixed crystals of ferrites, to which we shall revert in a subsequent article.

As a conclusion to the discussion of the probable distribution of the metal ions among the available spaces we should like to mention the following. The excellent agreement between the purely electrostatic theory of the spinel lattice and experiment in the case of aluminates, chromites and Zn and Cd ferrites need not suggest that the chemical binding of these spinels is almost purely electrostatic, although electrostatic forces undoubtedly play an important part. There are many indications that the binding in the spinels cannot be entirely approximated by the electrostatic conception of chemical valence and that homopolar forces (*i.c.* atomic binding in contrast to ionic binding) makes a significant contribution to the total picture of the binding forces. The explanation of the fact that the ferrites do not all behave in the same way, as far as the distribution of the metal ions is concerned, must be sought in certain finesses connected with these non-electrostatic forces. We shall not go further into it here because the theory is not yet able to explain the phenomenon satisfactorily.

## Rules for the distribution of metal ions in spinels

On the basis of the results discussed above we may now set out the following rules for the structure of spinels built up of bivalent and trivalent or bivalent and tetravalent metal ions.

1) The trivalent and tetravalent metal ions occupy the octahedron spaces in agreement with the electrostatic conception of the structure of spinels.
2) Exceptions are the $Fe^{3+}$ ions, which have a preference for the tetrahedron spaces. $In^{3+}$- and $Ga^{3+}$-ions, which have not been mentioned in this article, are also exceptions to rule (1).
3) $Zn^{2+}$ and $Cd^{2+}$ have a strong preference for the tetrahedron spaces and are able to drive the ions mentioned under (2) out of these spaces.

From the measurements of intensity on X-ray diffraction photographs we were able to deduce that these rules also remain valid for the formation of mixed crystals of the spinels. Several examples are given in *table II*.

Although we were unsuccessful in completely solving the problem of the structure of the spinels theoretically — the preference of $Fe^{3+}$ ions for the tetrahedron spaces remains somewhat mysterious — the investigation has had the result that with the help of the above-formulated rules we can predict the position of the metal ions in any arbitrary mixed crystal with very great probability, and we can therefore also prepare materials with a desired ion distribution. As will appear from a following article, this has not unimportant practical consequences.

Table II.

Structure of the mixed crystals of two spinels with the components taken in a mol ratio of 1 : 1. The symbols of the ions situated in the octahedron spaces are placed between parentheses. Roman numerals indicate the valence of the iron ions.

| Components of the mixed crystal | | | Structure of the mixed crystal |
|---|---|---|---|
| $Fe^{III}\,(Fe^{II},\ Fe^{III})\,O_4$ | and | $Fe^{II}\,(Al,\ Al)\,O_4$ | $Fe^{III}\,(Fe^{II}\ Al)\,O_4$ |
| $Fe^{III}\,(Cu,\ Fe^{III})\,O_4$ | and | $Zn\,(Fe^{III},\ Fe^{III})\,O_4$ | $Zn_{0,5}\,Fe^{III}_{0,5}\,(Fe^{III}_{1,5},\ Cu_{0,5})\,O_4$ |
| $Zn\,(Fe^{III},\ Fe^{III})\,O_4$ | and | $Zn\,(Cr,\ Cr)\,O_4$ | $Zn\,(Fe^{III},\ Cr)\,O_4$ |
| $Zn\,(Ti,\ Zn)\,O_4$ | and | $Fe^{III}\,(Mg,\ Fe^{III})\,O_4$ | $Zn\,(Fe^{III},\ Mg_{0,5},\ Ti_{0,5})\,O_4$ |
| $Mg\,(Ti,\ Mg)\,O_4$ | and | $Fe^{III}\,(Mg,\ Fe^{III})\,O_4$ | $Fe^{III}\,(Ti_{0,5},\ Mg_{1,5})\,O_4$ |
| $Zn\,(Fe^{III},\ Fe^{III})\,O_4$ | and | $Fe^{III}\,(Fe^{III}_{1,67})\,O_4$ | $Zn_{0,5}\,Fe^{III}_{0,5}\,(Fe^{III}_{1,83})\,O_4$ |

# 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.

**1722:** H. J. Lindenhovius: Het meten van impedanties bij hoge frequenties en toepassingen van de staande golfindicator (T. Ned. Radiogen. **12**, 60-82, 1947). (The measurement of impedance at h.f. and applications of the standing wave indicator).

In this article a survey is given of the different methods used in measuring impedances.

For frequencies below about 300 Mc/s the method used most frequently is that which employs a tuned circuit and in which the impedance is determined from its damping and detuning influence on the circuit.

For higher frequencies the lumped circuit can be replaced by a tuned transmission line but in that case some difficulties arise and it is preferable to use an untuned transmission line and to determine the impedance from the voltage distribution along the line.

The voltage distribution is characterised by the standing wave ratio and the position of the voltage minimum. A new diagram has been designed, which enables one to determine the impedance graphically and in a most comprehensible way even for very large values of the standing wave ratio. In connection with the measurement of the voltage distribution the standing-wave indicator is described.

Finally a number of other applications of the standing-wave indicator are dealt with, such as the measurement of characteristic impedance and attenuation constant of a transmission line and the measurement of net power flow along a transmission line.

**1723:** P. Cornelius: Eén eenhedenstelsel in de electriciteitsleer (Faraday **16**, 57-67, 1947). (One system of units in electro-magnetic theory).

A short survey is given of the basic formulae of electromagnetism using rationalized Giorgi units (M.K.S. units). The didactic value of using these units is stressed.

**1724:** J. L. Snoek: Zeitabhängige Erscheinungen in Eisen enthaltenden Stoffen unter dem Einfluss mechanischer und magnetischer Kräfte (Schweizer Archif angew. Wiss. und Techn. **13**, 9-14, 1947).

For the contents of this paper the reader is referred to the article by J. L. Snoek and K. F. du Pré, Philips Techn. Rev. **8**, 57, 1946 and to the book: J. L. Snoek, New developments in ferromagnetic materials, Amsterdam 1947 (see abstract No. 1729).

**1725:** J. H. van Santen and G. H. Jonker: Effect of temperature on the permittivity of Barium Titanate (Nature **159**, 333, 1947).

Above a certain transition temperature ($\vartheta$) crystals of barium titanate $BaTiO_3$ and related compounds, such as $SrTiO_3$, (Ba, Sr) $TiO_3$ and $TiO_2$, show a cubic structure. At the temperature $\vartheta$ the permittivity has a sharp maximum. It decreases monotonically with increasing temperature. This decrease is explained from the Clausius-Mosotti formula. Taking into account the thermal expansion and ignoring the tempcrature dependence of the polarisability, it follows that for higher values of $\varepsilon$

$$\frac{1}{\varepsilon} = \beta\,(\mathrm{T} - \mathrm{C}).$$

With the titanates the polarisability proves to be independent of the temperature. Then $\beta$ is the coefficient of thermal expansion.

It is believed that in the cubic region ($T < \vartheta$) there is no permanent dipole moment, whereas in the tetragonal region ($T < \vartheta$) thc assumption of dipoles seems to be quite plausible.

**1726:** M. Gevers: The relation between the power factor and the temperature coefficient of the dielectric constant of solid dielcctrics (Thesis, Delft, 1947).

In this thesis the relation between the power factor (tan $\delta$) and the temperature coefficient of the dielectric constant of solid, amorphous dielectrics is dealt with theoretically as well as experimentally. This subject has already been treated in a paper by M. Gevers and K. F. Du Pré (Philips Techn. Rev. **9**, 91, 1947). The methods of measurement are described and a number of special cases are dealt with. Finally some remarks are made on the properties of mixtures of dielectrics. (Also published in Philips Res. Rep. **1**, 197, 279, 361, 447, 1946, R 15, 20, 25, 30.)

**1727:** J. F. H. Custers: On the relation between deformation and recrystallization texture of nickel-iron with cubic orientation (Physica **13**, 97-116, 1947).

Polycrystalline nickel-iron (~ 50 weight % Ni) which has been severely cold rolled exhibits on recrystallisation at about 1000 °C a so-called cubic orientation. Aluminium does not show this texture upon recrystallisation. To trace any difference in slip mechanism between Ni-Fe and Al, the deformation and recrystallisation textures of polycrystalline Ni-Fe with cubic orientation were investigated and the findings were compared with observations of Burgers and Louwerse on single crystals of Al. On the whole the deformation textures were found to be the same. There are, however, marked differences due to the Ni-Fe specimen being not a mono-crystal. The recrystallisation textures are strongly different. The reason is to be found in the difference between the deformation textures. This can be explained by Burgers' theory. The second part contains a discussion of Barrett's criticism of this theory.

**1728:** J. M. Stevels: The effective permittivity of compressed and sintered samples of $TiO_2$ Rec. Trav. chim. Pays-Bas **66**, 71-74, 1947.

The effective permeability of compressed and sintered samples of $TiO_2$ has been measured and the results are discussed in terms of the theory developed by Polder and van Santen (abstract 1698). The experimental curves show that in loose powders the holes are more or less disc-shaped. The more the samples are sintered the more nearly spherical the holes become.
As a whole the results confirm the theory mentioned.

---

Number 4 of Volume 2, August, 1947 of *Philips Research Reports* contains the following papers:

*R 49:* F. L. H. M. Stumpers: On a non-linear noise problem.

*R 50:* J. L. Meyering and M. J. Druyvesteyn: Hardening of metals by internal oxidation, part II.

*R 51:* T. H. Oddie and J. L. Salpeter: Minimum cost chokes.

*R 52:* A. J. Dekkers and W. Ch. van Geel: On the amorphous and crystalline oxide layer of aluminium.

Readers interested in any of the above mentioned articles may 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 please write to the publishers of Philips Technical Review.

VOL. 9 No. 7, p. 193-224 1947

# 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 ELECTRON MICROSCOPE FOR 400 KILOVOLTS

by A. C. van DORSTEN, W. J. OOSTERKAMP and J. B. le POOLE. 621.385.833

M.F. 43 78.

When the acceleration voltage of an electron microscope is raised its resolving power is improved, theoretically. A more important advantage, however, is that at high voltages thick specimens can be studied, which at lower voltages would not give a satisfactory image, due, *i.a.*, to the excessive scattering of the electrons. In this article an experimental electron microscope is described which has been built in the Philips Laboratory in Eindhoven. It works with voltages up to 400 kilovolts. The instrument contains an objective aperture, adjustable from the outside, with four different openings, which provide the means for obtaining the best possible contrast in the image for any specimen. It has been found, for instance, that in a special case of yeast cells, whose internal structure is not sharply defined at 100 kV, the fine structures can be clearly observed when 350 kV is applied. Finally measures are discussed for the protection of the observer against the X-radiation excited in the microscope.

### Advantages of an electron microscope with high acceleration voltage

Most electron microscopes used until now work with an electron beam of a energy of 50 to 150 kV. In the Philips Laboratory in Eindhoven an electron microscope has been built for use with electron beams having an energy of 400 kV. Before describing this instrument we shall discuss the advantages connected with the use of a high acceleration voltage.

We shall first devote our attention to the resolving power. When the acceleration voltage is increased the resolving power is improved, at least theoretically, *i.e.* the distance between two details which can just be separately distinguished is reduced. This is explained as follows. It is known that moving electrons are to be regarded as having the nature of a wave, the wavelength $\lambda$ being given by the equation

$$\lambda = \frac{12.3}{\sqrt{V}}\,\text{Å}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where $V$ represents the acceleration voltage in volts. From this it may be seen that upon increasing the voltage the wavelength of the electrons is reduced, which means an increase in the resolving power [1]). In order to examine this more closely it must be noted that at the high voltages now under consideration formula (1) is no longer exactly valid. A correction must be introduced according to the relativity theory. It then becomes

$$\lambda = \frac{12.3}{\sqrt{V_c}}\,\text{Å}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

where

$$V_c = V(1 + 0.98{\cdot}10^{-6}\,V).$$

We shall call $V_c$ the corrected voltage. This value must also take the place of $V$ in the formula for the power of a magnetic electron lens, which formula is derived in the article referred to in footnote [1]). When the acceleration voltage of the microscope amounts to 400 kV, the corrected voltage is 560 kV. It may be seen from formula (2) that as the accelleration voltage is increased by the relativity correction so the wavelength is reduced more rapidly than would be expected from formula (1).

The limiting value which can be reached for the resolving power of an electron microscope depends not only upon $\lambda$, but also on the magnitude of certain inevitable aberrations, especially those due to diffraction and spherical aberration. It has been found that the highest resolving power is obtained when there is a certain ratio between the diffraction error and the spherical aberration. On that

[1]) See for example J. B. Le Poole. A new Electron Microscope with Continuously Variable Magnification, Philips Techn. Rev. 9, 33-45, 1947.

basis the following relation has been derived for the resolving power:

$$d_{\min} = 0.56 \sqrt[4]{\lambda^3 C} . \quad . \; . \; . \; . \quad (3)$$

It can further be shown that the smallest attainable value of the quantity $C$ occurring in formula (3) is proportional to the dimensions of the lens, which in turn are proportional to the square root of the corrected voltage. If the corrected voltage is chosen $n$ times as large, $C$ thus becomes $\sqrt{n}$ times as small. The resolving power is then improved and

$$d_{\min, n} = d_{\min} \sqrt[4]{\sqrt{n}/\sqrt{n^3}} = d_{\min} \sqrt[4]{1/n}.$$

When the acceleration voltage rises from 150 to 400 kV and thus the corrected voltage is increased from 165 to 560 kV, an improvement in the resolving power by a factor 0.73 can therefore be expected. This means that the smallest observable distance is now more than 25% smaller.

This must indeed be considered as a not unimportant improvement in the quality of the image. But such an improvement alone does not justify the construction of the expensive apparatus which is necessary when such a high acceleration voltage is to be used in electron microscopes. Before we mention the second advantage of a high acceleration voltage we shall first consider somewhat more closely the formation of the image in an electron microscope.

For the observation of a given detail of an object it is not enough that its dimensions should be larger than those corresponding to the resolving power of the microscope. There must also be sufficient contrast in the image studied. The eye cannot distinguish sharply between two parts of the image whose brightness differs by less than a certain amount. The percentage depends upon the observer, the illumination and the nature of the object.

How is contrast brought about in an electron microscope? The scattering of the electrons in the specimen is almost exclusively responsible for the contrast. The scattering is proportional to the mass through which the beam passes, so that the heavier the part of the specimen through which they have passed the more the electrons are scattered. The electrons which are scattered beyond a critical angle strike the objective aperture and do not contribute to the excitation of light on the fluorescent screen. The remaining electrons passing through the specimen and the objective are focussed to an image by the magnetic field of that lens. The diagram of *fig. 1* illustrates this.

Let us suppose that it is desired to examine with an electron microscope, working with a low acceleration voltage, a thick specimen, for instance a microtome section, yeast cells or a chromosome specimen. The specimen will then scatter the electrons so much that nearly all of them will fall on the objective aperture. The image is consequently almost completely black, so that no internal structures can be seen.

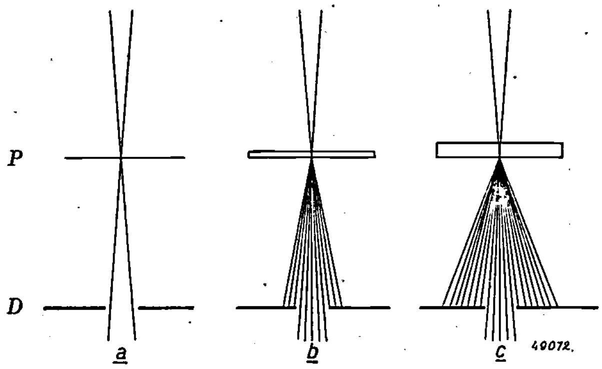


Fig. 1. The scattering of the electrons by objects of different thickness. In *a*, *b* and *c* rays are drawn which pass respectively through a very thin, a medium and a thick specimen *P*. *D* represents the objective aperture which captures part of the electrons in each case. In thick parts of the specimen the scattering is so great that only a small number of the rays reach the image. As a result the part of the image corresponding to a thick part of the specimen will be much "darker" than the rest.

By careful examination of the phenomena which occur when electrons pass through a thick object, one discovers yet another disturbing effect, namely so-called spatial scattering in the object itself. By this is meant the phenomenon that electrons are again deflected from their paths after having passed certain details in the specimen. Consequently there is finally hardly any relation between the direction in which they left the specimen and the spot whence they came. This, of course, has an unfavourable effect upon the resolving power.

Now it might be supposed that better results would be obtained with thick specimens by using a larger objective aperture. As to the first effect mentioned some improvement can indeed be expected, but in any case the error caused by spatial scattering cannot be reduced that way. The only possibility is to use a higher acceleration voltage. When the energy of the electrons is increased the average angle of scattering will decrease quite rapidly and therefore the spatial scattering will at the same time also decrease rapidly. Moreover, more electrons will be able to pass through the aperture and reach the final screen.

The second and most important advantage of a high acceleration voltage is, therefore, that thick specimens can be studied which would not give a satisfactory image at a low voltage.

It must not be concluded from this that it would be desirable to examine all specimens with electron rays of high energy. From the fact just mentioned, that when the acceleration voltage is increased the mean angle of scattering decreases quite rapidly, it follows of course that the contrast becomes smaller. When a thin specimen gives an image with sufficient contrast at a low acceleration voltage, the contrast will be seen to decrease wnen a higher acceleration voltage is used. This loss in contrast can be eliminated, however, for the greater part by choosing a smaller objective aperture, but as a rule the quality of the image will nevertheless be found to depreciate. From this it follows that a high voltage is only justified for those specimens for which, owing to their thickness, good results cannot be obtained at a lower voltage.

A third advantage of the use of an electron beam of high energy is the decrease in chromatic aberration. This aberration is caused by fluctuations in the velocity of the electrons. There are different causes of these fluctuations, of which we may mention: uncontrollable variations of the acceleration voltage, and in the case of thick specimens charges of the velocity as a result of non-elastic scattering of the electrons within the specimen. The faster the electrons the less energy they will give off to the specimen. Since chromatic aberration is proportional to the percentage of energy loss, by using electrons with a higher energy a decrease in the effect of chromatic aberration may be expected. Finally it may be mentioned that when a higher acceleration voltage is applied the heating of the specimen is less at the same current density, and the amount of ionization with the specimen, which may be found objectionable in some cases, is reduced at higher voltages.

## Description of the instruments

The electron microscope can be divided into three parts: the generator for high d.c. voltage, the acceleration tube and the microscope proper, *i.e.* the tube with the lenses, the fluorescent screen and various auxiliary apparatus. A brief description of each of these parts follows.

### *The generator for high D.C. voltage*

The apparatus for the production of the high voltage is built on the same principles as those employed in the installations which Philips have designed in the last ten years for various institutes for X-ray therapy and nuclear physics. Since this has already been described in detail in earlier numbers of this periodical [2]), a short description will suffice here.

The d.c. voltage is produced by a three-stage cascade generator with a circuit based on the principle of Greinacher and brought to practical realization by Cockcroft and Walton in Cambridge and by Bouwers in Eindhoven.

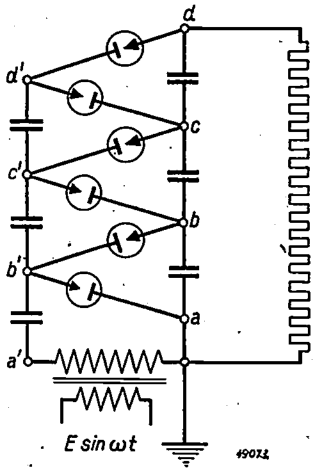


Fig. 2. The circuit of a cascade generator having six valves, which furnishes the no-load voltage $6E$ at $d$ when the amplitude of the a.c. voltage given by the transformer is $E$. In the circuits $ab'$, $b'b$, etc. are the six valves for 150 kV negative anode voltage.

The principle of this circuit is shown in *fig. 2*. If the amplitude of the a.c. voltage furnished by the transformer is $E$ at $d$, the d.c. voltage is 6 $E$. Since in the installation described here the transformer has a peak voltage of 75 kV, the non-loaded generator can produce a d.c. voltage of 450 kV.

For the rectifier valves needed in such generators, in previous cases Philips usually employed mercury vapour valves. Since these are provided with oxide cathodes they have the advantage of a low filament power (8-14 watts). It was not considered advisable, however, to use them in this installation, because these valves have an ignition voltage which is sometimes low, but which may also be of the order of several kilovolts, and since for the electron microscope the voltage on each valve must be constant to within a few tenths per thousand, *i.e.* in this case to within 50 to 100 volts, it was feared that the variable ignition voltage might have an unfavourable effect on the stability of the voltage. Vacuum rectifier valves were therefore used. At the high valve voltage these cannot, it is true, be provided with oxide cathodes [3]), but since the total direct current and thus also the

[2]) Philips Techn. Rev. 1, 6-10, 1936; 2, 161-164, 1937.
[3]) Philips Techn. Rev. 8, 199-205, 1946 (No. 7).

current through the valves is small, it was found possible to use tungsten cathodes with a filament power of about 20 watts.

A high-frequency current is used to feed the filament of the valves [2]). This has the advantage that the current source can be kept at earth potential. This heating current is furnished by a generator with a power of 200 watts and a frequency of 500 000 c/s.

In *fig. 3* the electron microscope is shown with its high voltage installation (except the transformer). For further particulars about this installation, which in certain respects differs from previous types, reference is made to the text under the illustration.

### *The acceleration tube*

The experience gained in the Philips Laboratories in the designing of X-ray installations for high voltages has been turned to account in the construction of the acceleration tube.

An X-ray tube for a million volts has been described in detail in this periodical [4]). This tube consists of three units connected in series. Each has an intermediate partition which is kept at a certain fixed potential, so that the voltage is divided into six equal stages. The electrons emitted by the tungsten cathode are accelerated between the cathode and anode of the first tube and shot into the hollow anode, then flying with a constant velocity through that narrow connection to the second tube, where they are accelerated a second time between cathode and anode, and again accelerated between cathode and anode of the third tube. At

[4]) J. H. van der Tuuk, A Million Volt X-ray Tube, Philips Techn. Rev. 4, 153-161, 1939.

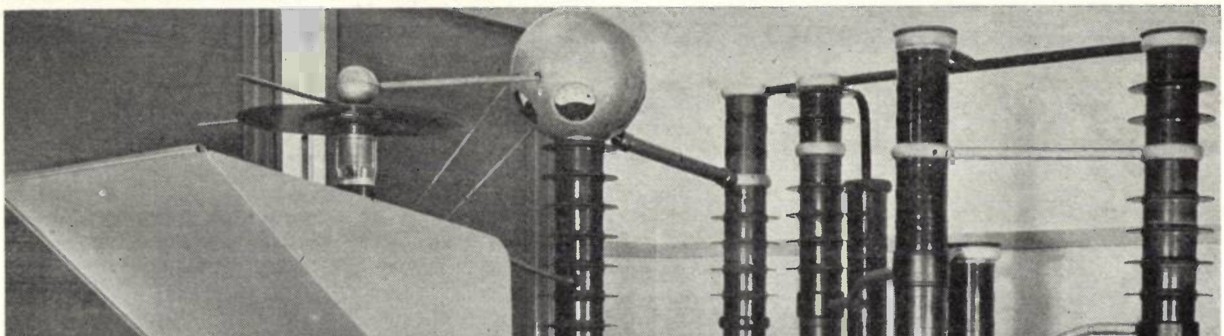


Fig. 3. The electron microscope for 400 kV with its high-voltage installation. The transformer is not visible in this photograph. The three vertical columns in the middle, the dimensions of which are approximately as 2 : 3 : 4 and which are not provided with projecting flat flanges, each contain two vacuum valves for 150 kV negative anode voltage. The other columns contain condensers and resistances. In the foreground is a control panel with various measuring instruments. Behind it is a box containing the generator for the high-frequency heating of the cathodes of the valves. On the extreme left are the variable resistances for the current control of the magnetic lenses.

the end of the hole in the third anode the electrons strike the anti-cathode. In order to reduce the length of a unit and still prevent creeping discharges there are double folds in the glass wall. To prevent direct sparking the annular-shaped hollows between the folds are filled up with an insulating body of "Philite".

The acceleration tube of our electron microscope is constructed on exactly the same principle, so that for further particulars we may refer to the article in question. The acceleration of the electrons here, however, takes place in three instead of in six stages. If it were later desired to use still higher voltages in the electron microscope it is possible to alter or to extend this component without great difficulty.

The cathode emitting the electrons is fed from a 6-volt accumulator battery. It can be displaced somewhat with respect to the other electrodes in order to make it possible to direct the beam accurately. The focussing cap has a variable voltage, negative with respect to the filament, which is important for obtaining a narrow beam of electrons.

Much care must be devoted to the accurate adjustment of the position and direction of the beam striking the specimen, because of the great influence of these factors on image quality. For that reason the accelerator tube with condenser is made adjustable, as to direction and position with respect to the microscope tube, by means of screws. In the photograph of *fig. 4* the protecting screen has been removed from the microscope in order to show the acceleration tube clearly.

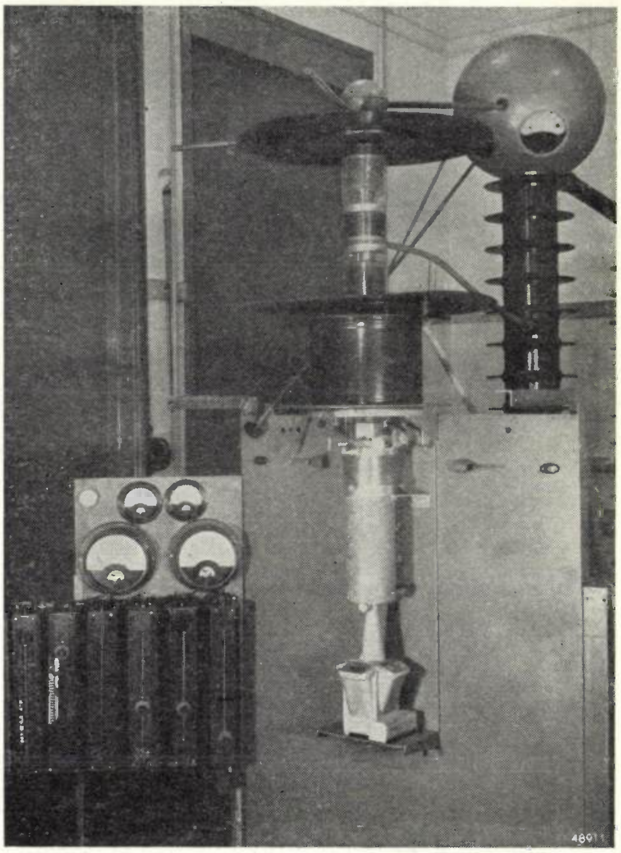


Fig. 4. The microscope proper with the acceleration tube. The protecting plate has been removed to show clearly the acceleration tube. On the right is a column containing the resistances for the potentiometer from which the current for the tube is tapped. On top of this column, in a metal sphere at a potential of —400 kV with respect to earth, is a 6-volt accumulator for the heating of the filament.

### *The microscope*

In the description of the microscope proper we can also be brief, since in various respects its construction corresponds to that of the electron microscope for 150 kV described in the article cited in footnote [1]).

*Fig.* 5 shows the microscope with the acceleration tube in cross section. The condenser lens is above the microscope at the lower end of the acceleration tube. In addition to the condenser there are four lenses: an objective, two intermediate lenses and a projector. The introduction of an intermediate lens has the advantage that the length of the tube is kept small in relation to the magnifications reached, the magnification being variable within a wide range. The distance from the condenser to the final screen is 93 cm. The magnification is continuously variable from 2000 to 100 000 diameters, the projector being left unaltered so that the whole image field with a diameter of 9 cm is always filled [5]).

At the time that this microscope was designed, from what was known about the subject it was to be expected that the resolving power of photographic emulsions and fluorescent screens would decrease to a not unimportant extent upon applying a higher acceleration voltage. It was therefore decided to design the instrument for a magnification of $10^5$ diameters. In order to attain this with a short overall length of the microscope it was considered advantageous to use four stages, *i.e.* to introduce two intermediate lenses besides objective and projector. With the photographic emulsions and fluorescent screens now available the resolving power is so good that at high voltages it will be possible to work with a lower magnification than was originally intended. Satisfactory results can then be obtained with three stages, using one intermediate lens. It is planned in that case to replace the upper intermediate lens by a

[5]) This is explained more fully in the article referred to in footnote [1]).

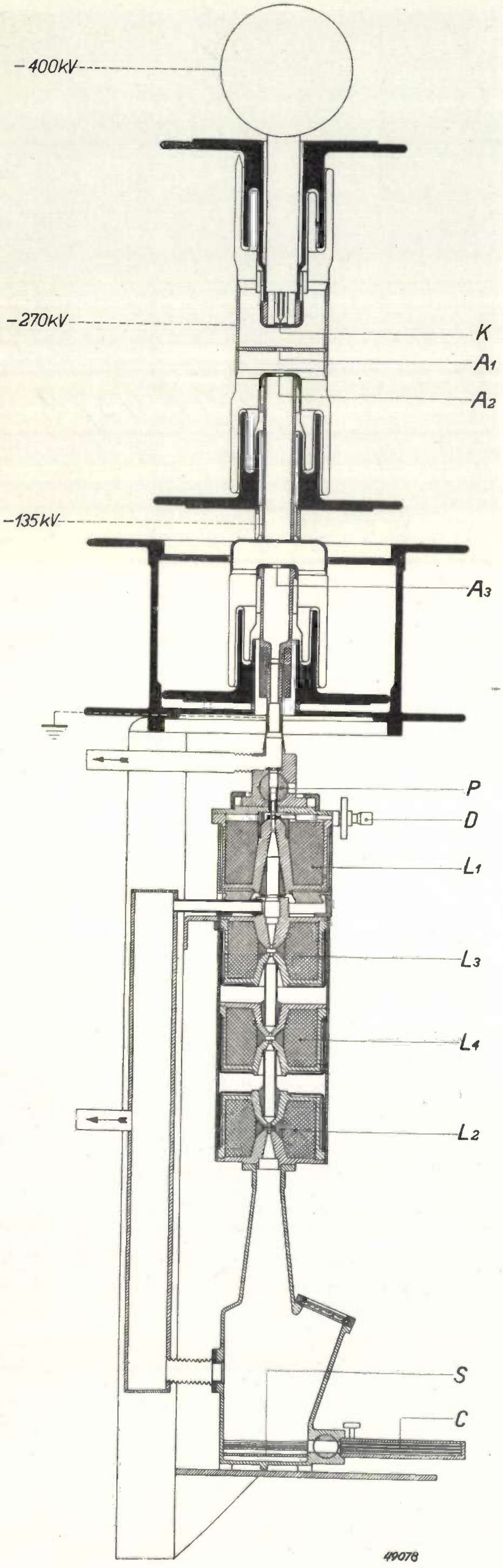


Fig. 5. Sectional view of the microscope with the acceleration tube (scale about 1 : 8). $K$ is the cathode, $A_1$, $A_2$ and $A_3$ are the three anodes, $P$ the lock for the specimen, $D$ the objective aperture. $L_1$ is the objective, $L_2$ the projector, while $L_3$ and $L_4$ are the two intermediate lenses. $S$ is the fluorescent screen and $C$ the plate holder of the photographic camera. At the left of the figure it is indicated what part of the acceleration tube is connected with earth and where the potential difference with respect to earth amounts to —135, —270 and —400 kV respectively.

diffraction lens as used in the above-mentioned 150 kV microscope.

As regards the construction of the lenses it may be noted that the shape of the iron cores is so chosen that saturation cannot occur. It should also be mentioned that, in view of the possible future use of higher acceleration voltages the microscope is so constructed that electron rays with an energy of a million volts can be employed.

We have already called attention to the fact that when using an electron microscope with a high acceleration voltage it is essential to take care to use an objective aperture best suited to the nature of the object to be studied. In order to make this possible four tungsten objective diaphragms, of different sizes and adjustable from the outside, are provided, the diameters of the openings being respectively 0.8, 0.08, 0.045 and 0.03 mm.

Finally a few remarks concerning the vacuum system. When the electron microscope is in use the pressure in the acceleration tube may not be higher than about 0.001 mm Hg. In order to attain this vacuum two independent high-vacuum pump systems are used. The vacuum is measured with a Philips vacuum meter [6]) with a glowdischarge tube, to be used in the interval from $10^{-5}$ to $10^{-3}$ mm of mercury.

When a new specimen is to be examined it is introduced into the tube by means of an air lock in the object stage. Only about 2 cm³ of air enters the microscope when this is done. Twenty seconds after closing the lock the vacuum has again reached a value permitting the high voltage to be switched on.

Five reproductions on the following pages give some idea of the results which have been obtained with the electron microscope for 400 kV. *In figs. 6, 7* and *8* yeast cells are shown; it should be mentioned that these pictures have no biological importance in themselves but merely illustrate the effect of high voltage in connection with specimens.. The exposures were made with acceleration voltages of 112, 225 and 350 kV respectively. The appearance of these pictures confirms the

[6]) F. M. Penning, Philips Techn. Rev. 2, 201-208, 1937.

statement at the beginning of this article about the advantages to be expected from the use of a high acceleration voltage with thick specimens. *Figs. 9* and *10* demonstrate the influence of the objective diaphragm on the contrast obtained.

The electron microscope for 400 kV described here is of an experimental type. Improvements and refinements are being constantly made and further investigations are being continued.

We shall conclude this article with a discussion of the measures which have to be taken to protect the observer against X-radiation.

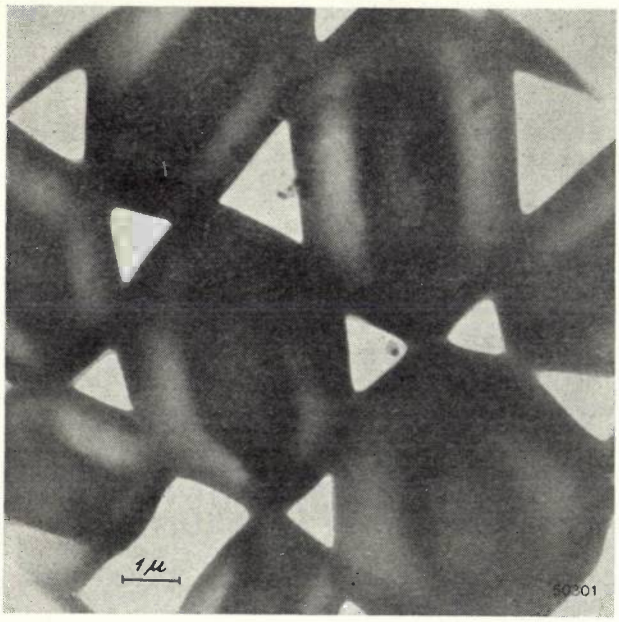


Fig. 6. Yeast cells photographed with an acceleration voltage of 100 kV, diameter of objective diaphragm 0.08mm. Magnified electron-optically 2000 ×, total magnification 5000 ×. The cells are transparent, but the internal structure is blurred. In order to indicate the true magnification of the reproduction the length 1μ is shown in this and the following figures.

### Protection against X-rays

Wherever electrons collide with matter X-rays are excited. It is sufficiently well known that exposure to X-rays for too long a period or a too intense radiation may have very harmful results on the human body. Precautions must therefore be taken that the amount of X-radiation reaching the operator of an electron microscope is kept below a certain safe limit. This is usually done by introducing between the source of X-radiation and the observer protective material which absorbs most of the X-rays. Complete absorption is impossible, but also unnecessary.

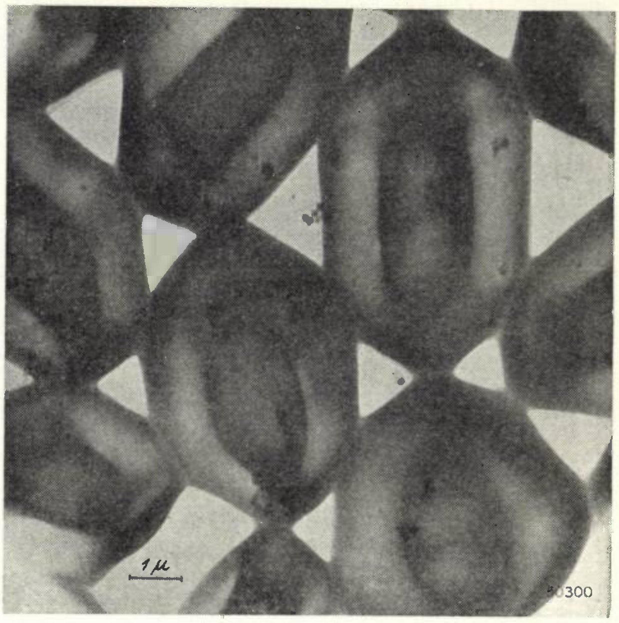


Fig. 7. Yeast cells, the same specimen as in fig. 6, with the same aperture and the same magnification, but with an acceleration voltage of 225 kV. The internal structure is now appreciably sharper.

In the case of instruments with a very low voltage, for instance up to 20 kV, the X-rays are very soft, and thus easily absorbed. The wall of the vacuum tube and the surrounding air therefore fulfil this function adequately. In the case of microscopes working with high voltages special measures must be taken to protect the observer against X-radiation. This is especially so when electrons of such high energy are used as in the case of the instrument in question.

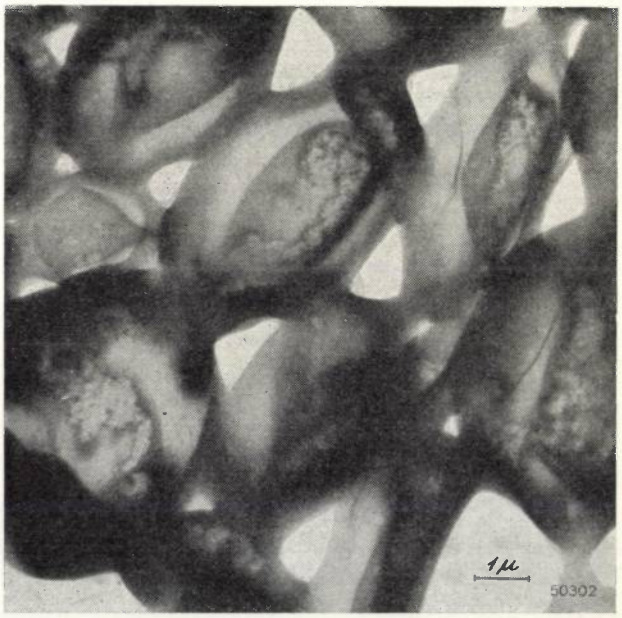


Fig. 8. Yeast cells taken with an acceleration voltage of 350 kV. The objective diaphragm with a diameter of 0.08 mm was again used and the electron-optical magnification was again 2000 diam. Thanks to the use of the very small objective diaphragm the contrast is still sufficient with this energy, while finer structures are now also visible. Some thin dark lines to be seen on the right of the picture are images of folds in the collodium film on which the specimen is fixed.

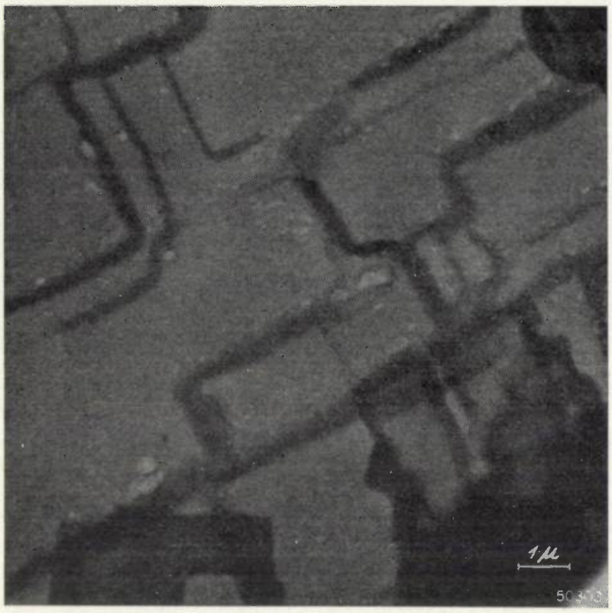


Fig. 9. Oxide film of an etched aluminium surface, taken with an acceleration of 250 kV and a diaphragm of 0.5 mm diameter. Electron-optical magnification 3000 diam.

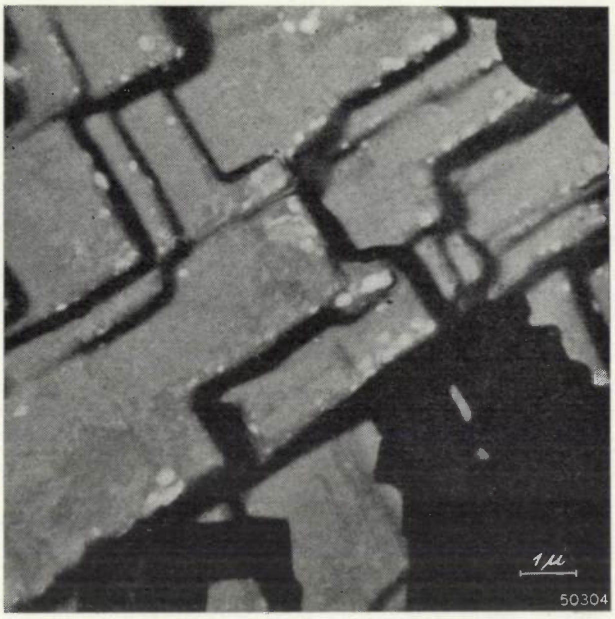


Fig. 10. The same specimen as in fig. 9 taken with the same acceleration voltage with the same magnification but with a diaphragm of 0.045 mm. The effect of the size of the diaphragm on the contrast is clearly visible.

The permissible daily tolerance is usually considered to be 0.2 r [7]), which for an eight-hour working day amounts to an X-ray intensity of about $10^{-5}$ r/sec.

The X-ray dose increases linearly with the current of the electron ray and with the atomic number of the material struck by the electrons, and, moreover, with the square of the energy of the electrons. For a source of radiation of small dimensions the radiation is inversely proportional to the source.

It is important to note that in an electron microscope the main source of X-radiation is at the diaphragms. When the electron beam has passed the condenser aperture the current is already considerably smaller, and this is even more so after the objective aperture has been passed. The current to the screen is so small — in our case only $10^{-8}$ to $10^{-9}$ A — that the X-radiation emitted there may be practically ignored. Therefore the protective screen, usually of lead, must be placed in front of the upper end of the microscope.

It is desirable, however, that the necessary protective measures to lower the intensity of the X-radiation should be taken in the construction of the instrument. In the electron microscope described here this has been done in two ways.

In the first place care has been taken that the current shall be as small as possible, *i.e.* not larger than what is necessary for the "illumination" of the part of the object to be projected. This current is fixed because of the fact a certain current density is necessary to obtain a clear image. Reduction of the current to this minimum amounts to working with a narrow electron beam which is carefully directed. It is always desirable to work with a narrow, well-directed beam, but it is an essential requirement when the acceleration voltage is high. Owing to the acceleration tube being long, the beam has a greater spread, and as a result of the high energy of the electrons an intense X-radation is thereby excited.

The second method employed for decreasing the intensity of the X-radiation consists in making the most important diaphragms of materials having a low atomic number. This is obtained by introducing a diaphragm made of beryllium both in the condenser and directly over the specimen. To promote the dissipation of heat the diaphragms are brassed into a copper fitting.

It is desirable to determine the thickness of the protective plate necessary to reduce the dose received by the observer to below the limit previously stated. In this we are helped by the investigations already made in connection with the protection of the operators of X-ray therapy equipment. An electron microscope can be compared to an X-ray installation having an anode voltage as high as the acceleration voltage of the microscope. On the basis of the investigations previously carried out the

[7]) A röntgen (r) is the dose of X-radiation which will free an e.s.u. of charge ($3.3 \times 10^{-10}$ coulomb) by ionization upon passing through 1 cm³ of air at a temperature of 20 °C and a pressure of 76 cm Hg. See also the article by B. van Dijk in Philips Techn. Rev. 4, 114-117. 1937.

thickness of the protective plate in the X-ray installation can be found, and from that it can be calculated how thick the plate must be in the case of the electron microscope, taking into account that the current of the electron beam in the microscope is much smaller than in the X-ray tube, and if necessary also allowing for a difference in atomic number between the materials struck by the electrons in the two instruments, and the distances at which the observers are situated from the X-ray source in both cases.

The graph of *fig. 11* relates to an X-ray tube for 400 kV, where a tungsten surface is struck by a current of 10 mA. These are the conditions prevailing in a certain installation for irradiation with X-rays [8]. It is found that the lead covering must be 24 mm thick to keep the dose received by the operator at a distance of 1 m below the value of $10^{-5}$ r/sec.

A current of 10 mA is normal in X-ray installations. With this electron microscope, under the above-mentioned conditions, a current of not much more than 0.02 mA is sufficient. In view of this the thickness of lead can be reduced from 24 to 8 mm. By using beryllium instead of tungsten for the diaphgrams a further reduction from 8 to 2.5 mm of lead is possible.

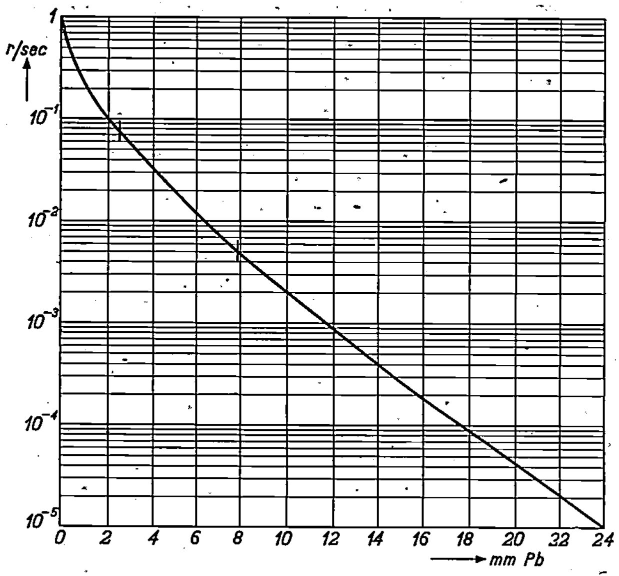


Fig. 11. The thickness of the lead protecting screen in millimeters necessary to keep the X-ray dose below a certain value when the anode voltage is 400 kV and the current is 10 mA, while the electrons impinge on tungsten and the observer is situated at a distance of 1 m from the source of radiation. When the current is reduced to 0.02 mA a lead thickness of about 8 mm is sufficient to keep the X-ray intensity below $10^{-5}$ r/sec; if, in addition, the diaphragms are made of beryllium a protection of slightly more than 2 mm of lead is sufficient.

On the the other hand in an electron microscope the distance from the source of X-radiation to the observer is usually less than 1 m. In determining the thickness of the lead protecting plate this must be taken into account.

[8] Philips Techn. Rev. 8, 105 — 110, 1946 (No. 4).

# A CATHODE-RAY OSCILLOGRAPH WITH TWO PUSH-PULL AMPLIFIERS

by E. E. CARPENTIER.

621.317.755

A description is given of the new cathode-ray oscillograph for universal use, type GM 3159, incorporating the new oscillograph tube DG 7-3. In this tube there is a push-pull amplifier for each pair of deflecting plates. A correction circuit ensures that within a frequency range of 10 - 460 000 c/s and with the maximum sensitivity the amplification is kept constant within 3 db; with reduced sensitivity this range is extended to over $10^6$ c/s. With the introduction of an amplifier for the horizontal deflection it has been possible to arrange a very simple circuit for the time-base voltage. The saw-tooth voltage is also available for other purposes extraneous to the oscillograph, for instance for measuring amplifiers; the frequency of this voltage can be regulated between 10 and 150 000 c/s. When the time-base is used the light spot can be blanked out during the retrace, thus making the picture clearer. There is a considerable improvement in the magnetic screening of the oscillograph tube. The new oscillograph is smaller and much lighter in weight than the older types.

Cathode-ray oscillographs have been described several times already in this journal [1][2][3]. The type with which this article deals (the GM 3159) is distinguished from the others, *i.a.* by the fact that it has two amplifiers, one for the vertical deflection, as is usual, and another for the horizontal deflection. With this second amplifier the oscillograph is of more universal use. A second point of difference is the better quality of the picture, due for the greater part to a new type of oscillograph tube being used. In the third place it may be said that every endeavour has been made to keep the weight and dimensions of the whole apparatus as low as possible, with the result that a handy and inexpensive instrument has been produced.

This aim at small dimensions tends to come into conflict with the desire for high sensivitity and a wide frequency band, to give the instrument the widest possible field of application. For instance, to avoid an undesirable high temperature in an apparatus of small dimensions the energy dissipation has to be limited. This makes it necessary to use the smallest possible number of amplifiers, which in turn means on the one hand a compromise with the requirement of sensitivity and on the other hand leads to a high amplification per stage, at the cost of bandwidth. Compared with the oscillograph GM 3152 [2]) the compromise referred to has necessitated some concession in respect to sensitivity, especially at the very high frequencies. The new type is not intended for the field of very low frequencies (1-10 c/s) such as occur in mechanical applications; for this purpose the oscillograph GM 3156, specially designed for these low frequencies, is still indicated (see note [3]).

## General aspects

Hitherto cathode-ray oscillographs have been designed mainly on the system represented in the block diagram of *fig. 1a*: the voltage to be oscillographed is connected at *I* and conducted *via* a variable attenuator (potentiometer) $Z_1$ to the amplifier $A_v$, which is connected to the plates $D_v$ for the vertical deflection. The horizontal deflection is obtained by applying to the plates $D_h$ a linear saw-tooth voltage induced by the time-base circuit *TB*. This is synchronised by an auxiliary voltage drawn, *via* an attenuator $Z_2$ and a switch *S*, either from the signal to be oscillographed (with the switch *S* in position *A*) or from an auxiliary signal connected at *II* (position *B*), or from the mains (position *C*), *via* the supply unit *P* furnishing the anode, grid and filament voltages for the various valves.

The new oscillograph has been designed somewhat differently. As already stated, the pair of horizontal deflecting plates also have a separate amplifier ($A_h$ in fig. 1*b*). The switch *S* now has five positions, the third, fourth and fifth corresponding to the positions *A*, *B*, and *C* in the old system, except that the amplifier $A_h$ is always in between the pair of plates $D_h$ and the time-base unit. The latter unit now has to supply only a saw-tooth voltage of small amplitude, so that it can be made much simpler. This, however, is only an incidental advantage of the introduction of the second amplifier and finds expression particularly when the switch is in position *1*, when $Z_2$ is connected direct to the input of the amplifier $A_h$; the time-base unit is then cut off. As a consequence one is no longer restric-

[1] Philips Techn. R. 1, 147 — 151, 1936 (type GM 3150).
[2] Philips Techn. R. 4, 198 — 204, 1939 (type GM 3152).
[3] Philips Techn. R. 5, 277 — 285, 1940 (type GM 3156).

ted to the observation of actual oscillograms (with the voltages as a function of time), for now also Lissajous figures (*fig. 2*) can be oscillographed, a voltage being pictured as a function of any other voltage; thanks to the presence of two amplifiers both input voltages may be small. This considerably expands the scope of the instrument, for instance for measuring frequencies of phase angles. Finally, with the switch $S$ in position *2* a voltage with the mains frequency is conducted to the amplifier $A_h$ *via* a separate attenuator $Z_3$; in this position, too, the time-base unit is switched off.

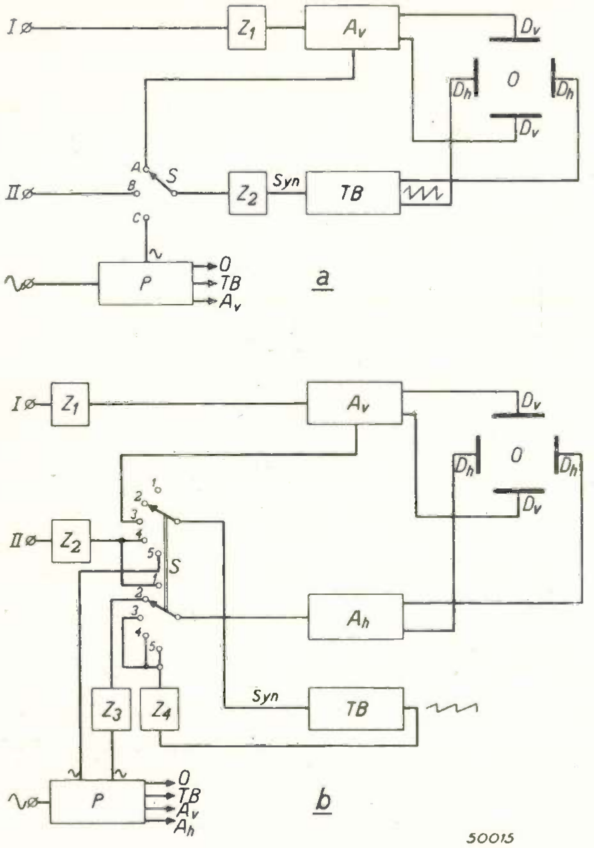


Fig. 1. *a*). Block diagram of cathode-ray oscillographs hitherto commonly used. $I$ — input connection for the voltage to be oscillographed. $Z_1$ — variable attenuator (potentiometer). $A_v$ — amplifier. $O$ — oscillograph tube with plates $D_v$ for vertical deflection and $D_h$ for horizontal deflection. $TB$ — time-base unit, the frequency of which can be synchronised with an auxiliary voltage carried in at $Syn$ *via* the attenuator $Z_2$ and the switch $S$. This voltage is either drawn from the voltage to be oscillographed ($S$ in position $A$) or is carried in at $II$ (position $B$), or is taken from the mains (position $C$). $P$ — supply unit connected to the mains.

*b*): Block diagram of the new oscillograph (GM 3159). The plates $D_h$ are connected to an amplifier $A_h$ (identical with $A_v$) whose input, when $S$ is in position *1*, is connected to the terminals $II$, so that a small voltage can be oscillographed as a function of another small voltage (both of the order of 10 mV). When $S$ is in position *2* a sinusoidal voltage of the mains frequency is fed to the amplifier $A_h$. Positions *3*, *4* and *5* of $S$ correspond to the positions $A$, $B$ and $C$ in fig. 1*a*. $Z_3$ and $Z_4$ are attenuators. The rest of the symbols have the same meanings as in fig. 1*a*.

In both systems (*a*) and (*b*) the plates $D_h$ can be connected direct to a pair of terminals.

So much then for the general layout of the new oscillograph. We shall now examine some of the principal components, very briefly the cathode-ray tube, then in more detail the amplifiers and the circuiting of the time-basc voltage. Finally we shall discuss some features of the mechanical construction.

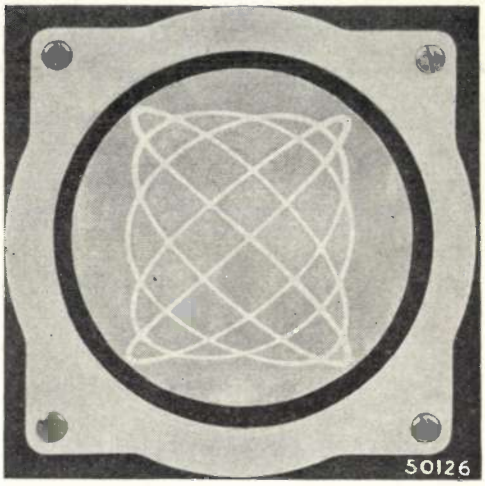


Fig. 2. Lissajous figure demonstrating 1) that the light spot retains its sharpness over the whole area of the screen and 2) that the picture fits in a rectangle. The latter is due to the symmetrical control of the two pairs of deflecting plates, avoiding trapezoidal distortion (see the article referred to in note [2]).

## The cathode-ray tube

The new oscillograph is fitted with the DG 7-3 cathode-ray tube, which incorporates the various improvements described in the last number of this journal [4]). This tube has a sharper light spot, especially at the edge of the screen, with the result that although the screen is only 7 cm in diameter a sharp picture is obtained of practically the same size as that produced with the older tubes having a screen diameter of 9 cm. The Lissajous figure in fig. 2 shows the uniform sharpness of the line over the whole screen.

Then there is the electrical screening between the two pairs of deflecting plates and their leads, thanks to which the parasitic capacitors between the pairs of plates are reduced to such an extent that even at very high frequencies there is no noticeable influence of one pair of plates upon the other, so that there is no tendency of a voltage between one pair to induce a voltage between the other pair.

[4]) J. de Gier and A. P. van Rooy, Improvements in the Construction of Oscillograph Tubes, Philips Techn. R., 9, 181 — 185, 1947 (No. 6).

## The amplifiers

### *Push-pull amplificitation*

Since the amplifiers for both pairs of deflecting plates are identical, a description of one of them suffices.

Asymmetric control, where one of the plates is earthed, causes distortion of the picture (see for instance the article quoted in footnote [2]). In order to avoid this distortion the amplifier has been built on the push-pull principle. A pre-stage has been dispensed with, because the number of valves and the dissipated energy had to be kept as low as possible in order to meet the desire for an apparatus of small dimensions. The question, then, was how to get a symmetrical output voltage with a push-pull circuit to which an asymmetric input voltage is applied (earthed on one side). *Figures 3a* and *b* show two ways of solving this problem. In the first method the amplifier $B_2$ is fed from a voltage-divider, formed by two resistors $R_b$ and $R_{g2}$, in series with a separating capacitor $C_3$ applied across the output of the valve $B_1$. The ratio of these resistances should be as $(g_0 - 1) : 1$, where $g_0$ is the voltage amplification of one half of the push-pull.

In the method according to fig. 3*b* absolute symmetry is not obtained, it is true, but it is very closely approximated. The principle underlying this method is that the difference between the two anode currents (these being unequal in the case of asymmetry) is caused to flow through an impedance $Z_k$ in the common cathode lead, thereby inducing in that impedance an alternating voltage which acts on the two grids in the same sense. A simple calculation shows that symmetry is very closely approximated if the condition $SZ_k \gg 1$ (where $S$ = the slope of the valves) is satisfied.

The following equations may be written for the alternating voltages and currents occurring in the system of fig. 3*b*, with the symbols and choice of positive directions given there:

$$\left.\begin{array}{ll} \text{Tube } B_1 & \text{Tube } B_2 \\ V_{g1} = V_1 - Z_k I_k, & V_{g2} = Z_k I_k, \\ I_{a1} = S V_{g1} & I_{a2} = S V_{g2}, \\ I_{k1} = \beta\, I_{a1}, & I_{k2} = \beta I_{a2}. \end{array}\right\} \quad . \; . \; (1)$$

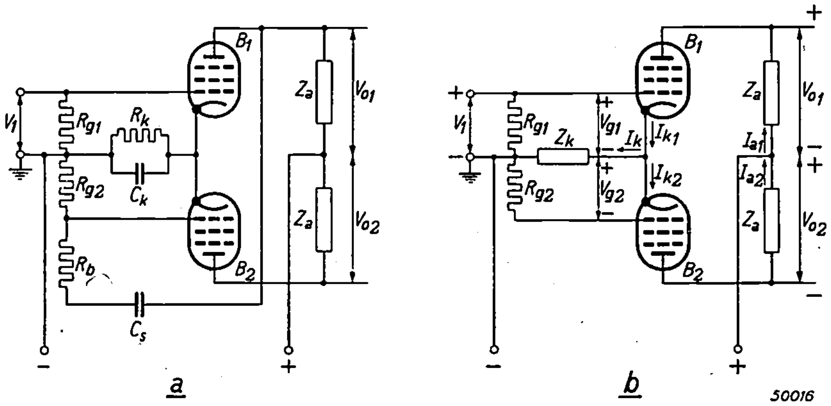


*Fig. 3.* Two methods of obtaining from a push-pull amplifier a symmetric output voltage ($V_{o1} = V_{o2}$) )notwithstanding an asymetric input voltage ($V_1$).
*a*) A voltage divider formed by the resistors $R_b$ and $R_{g2}$ conducts a part of the output voltage $V_{o1}$ from the amplifying valve $B_1$ to the control grid of the valve $B_2$. $C_s$ is a separating capacitor, $R_k$ a resistor shunted by a capacitor $C_k$ for supplying the negative grid voltage.
*b*) With this method symmetry is obtained approximately when $SZ_k \gg 1$ ($S$ = slope of the valves $B_1$ and $B_2$, $Z_k$ = impedance of the common cathode lead). The currents indicated in the diagram as $I_{k1}$, $I_{a1}$, etc. are alternating currents, and the voltages $V_1$, $V_{g1}$, $V_{o1}$, etc. are alternating voltages; they are reckoned to be positive in the direction of the arrow or in the direction marked + and —.

Here the factor $\beta$, which generally has a value of 1.2 — 1.3, indicates that the cathode alternating current is higher than the anode current; the difference between these two flows across the screen grid.

Further the equation holds:

$$I_k = I_{k1} - I_{k2}.$$

Hence:

$$\left.\begin{array}{l} I_{a1} = \dfrac{1 + \beta\, SZ_k}{1 + 2\beta\, SZ_k} SV_1, \\[2ex] I_{a2} = \dfrac{\beta\, SZ_k}{1 + 2\beta\, SZ_k} SV_1. \end{array}\right\} \quad . \; . \; . \; . \; . \; . \; . \; (2)$$

From these equations it is seen that the anode currents differ less according as $\beta\ SZ_k$ is greater than unity. If $I_{a1} \approx I_{a2}$ then, given equal anode impedances $Z_a$, naturally also $V_{o1} \approx V_{o2}$.

Which of the two methods is the more suitable for our purpose will be left unanswered for the moment; we shall revert to it after discussing the frequency characteristic.

*Amplification required; value of the anode resistance*

Between the plates of the more sensitive of the two deflecting systems (*i.e.* the pair of plates closest to the anode) a peak voltage of 20 V is needed for a total deflection of 1 cm. Reckoning on a sensitivity of 1/25 cm picture height or width per mV (r.m.s.) being required at the input of the amplifier, then an amplification of $20(/25\sqrt{2} \times 10^{-3}) = 560$ is needed. Therefore each half of the push-pull must yield an amplification of 560/2 = 280.

When a pentode is used the amplification equals the product of the impedance $Z_a$ in the anode circuit and the slope $S$. In the pentode used here (the EF 50) the slope is adjusted to approx. 4 mA/V, so that to get an amplification of 280 $Z_a$ must be 70 000 ohms.

$Z_a$ consists of a resistance $R_a$ and a parasitic capacity $C_p$ connected in parallel, the latter being taken at about 28 pF. So long as the frequency is not too high the effect of $C_p$ may be ignored, so that then $R_a \approx Z_a = 70\,000$ Ohms. At a frequency, however, of say $10^6$ c/s — which we shall take, for the present, as the upper limit of the frequency band — the impedance of the parasitic capacity has dropped to 5700 Ohms, which is small compared with $R_a$, so that then $|Z_a| \approx 1/\omega C_p$ and the amplification has dropped accordingly. How this problem is met will be shown farther on; first we shall see how the choice of the anode direct current and of the direct voltage for feeding the anode circuit is mainly determined by the value of $1/\omega C_p$ at high frequencies and by the size of picture required.

*Anode direct current; feed voltage*

If at high frequencies a picture height of say 3.5 cm suffices, then the amplifier has to give a voltage with a peak value of $3.5 \times 20 = 70$ V, thus 35 V for each half. At the highest frequency of $10^6$ c/s, when $|Z_a| \approx 5700$ Ohms, there is an anode alternating current with a peak of $(35/5700) \times 10^3 = 6.1$ mA. The anode direct current must therefore be at least equal to that value; to leave some reserve 6.5 mA has been chosen.

The voltage required for feeding the anode circuits is found in the following way; this feed voltage is roughly equal to the minimum admissible anode voltage of the pentode EF 50 (approx. 100 V) plus the maximum voltage loss in the anode resistance. This latter resistance consists of the sum of the direct voltage loss — amounting to $6.5 \times 10^{-3} \times 70\,000 = 455$ V — and the peak value of the maximum alternating voltage that each half of the amplifier is required to yield, *viz.* about 35 V. Thus we arrive at a sum of 100 + 455 + 35 = 590 V. A certain allowance has to be made, however, for mains voltage fluctations, leakage of anode current and leakance of the anode resistance. In this way we reach a figure of 675 V for the feed voltage. This voltage together with the above mentioned anode current of 6.4 mA for each of the four amplifying valves gives for a large part the total dissipation of energy. The feed voltage of 675 V is also used for the cathode-ray tube.

*The frequency characteristic*

The shape of the frequency characteristic of the amplifier depends upon the variation of $|Z_a| = R_a/\sqrt{1 + (\omega C_p R_a)^2}$ with the frequency. In *fig.* 4 curve *a* represents $|Z_a|/R_a$ where the maximum value is taken as 100%. As is seen, this curve is not very satisfactory. In order to improve this, a new method has been followed which will be explained with reference to *fig.* 5, where the push-pull circuit is temporarily dispensed with.

The anode alternating current $I_a$ is split up into two components, $I_R$ and $I_C$, which flow respectively through $R_a$ and $C_p$; $I_C$ is 90° in advance of $I_R$. When the control grid voltage $V_g$ has a constant amplitude and increasing frequency $I_a$ remains constant at the value $SV_g$, but $I_C$ increases at the cost of $I_R$, so that the voltage supplied $|V_o| = I_R R_a$ drops. This voltage would be independent of the frequency if the valve could also supply, in addition, the capacitive current $I_C$. This can be achieved by feeding the valve with an additional

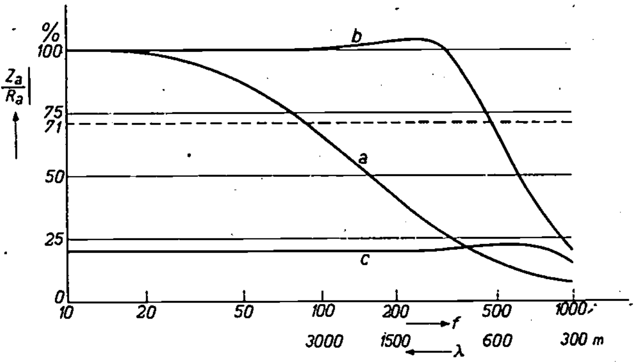


Fig. 4. The ratio $|Z_a|/R_a$, to which the amplification is proportional, as a function of the frequency $f$ (in kc/s). Curve *a*: resistance amplifier without correction. The drop in amplification as the frequency increases is due the decrease of the anode impedances $Z_a$ resulting from the presence of parasitic capacity. Amplification drops to 71% already at 85 kc/s. *b*) By applying a correction the frequency range within which the amplification lies between 100% and 71% is extended to about 460 kc/s. *c*) By connecting a resistor ($R_p$ in fig. 7) parallel across the output terminals the upper frequency limit is raised to over 1000 kc/s, at some sacrifice of sensitivity.

On the ordinate axis 100% corresponds to an amplification of about 560. On the axis of abscissae several wavelengths (λ) are shown.

input voltage of a suitably chosen value and phase, as shown in fig. 5: in (*a*) $V_1$ is the original voltage that is to be amplified and $V_b$ the extra input voltage (which has to satisfy a certain condition to be specified later), so that the total grid alternating voltage $V_g = V_1 + V_b$. As is known, in so far as the anode circuit is concerned one may imagine a pentode as being replaced by a source of current of the strength $I_a = SV_g = S\,(V_1 + V_b)$ — see fig. 5*b*. When the positive directions are chosen as indicated in the illustration the current $I_R$ in the resistance is then:

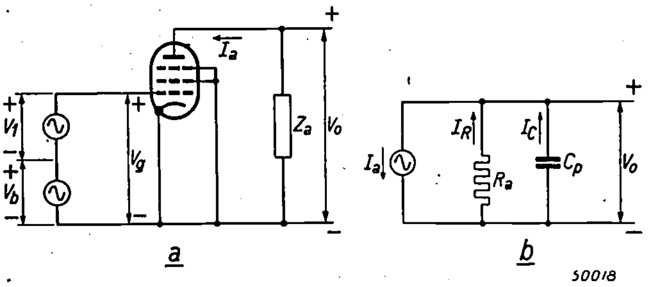


Fig. 5 *a*) A pentode, to the control grid of which an auxiliary voltage $V_b$ is applied (in addition to the voltage $V_1$ to be amplified) with the object of compensating the loss of amplification at high frequencies due to parasitic capacity. Direct voltage sources are omitted in this and the following diagrams. The + and — signs indicate the directions in which the alternating voltages $V_1$, $V_b$, $V_g$ and $V_o$ are considered to be positive; the positive current direction is indicated by a single arrow.

*b*) Replacement diagram for the anode circuit of fig. 5*a*: the current source $I_a$ feeds the parallel circuit of $R_a$ and $C_p$.

$$I_R = \frac{-V_o}{R_a} = \frac{Z_a I_a}{R_a} =$$

$$= \frac{R_a/(1+j\omega C_p R_a)}{R_a} \cdot S(V_1+V_b) = \frac{S(V_1+V_b)}{1+j\omega C_p R_a}.$$

To give $I_R$ the value that is really desired, *i.e.* $SV_1$, it is therefore necessary to chose $V_b$ such that

$$\frac{S(V_1+V_b)}{1+j\omega C_p R_a} = SV_1,$$

or

$$V_b = j\omega C_p R_a V_1 = -\frac{j\omega C_p}{S} \cdot V_o. \quad . \quad (3)$$

The extra voltage $V_b$ must therefore be 90° advanced in phase with respect to the output voltage and increase proportionately with the frequency.

This desired extra voltage can be obtained approximately by following the scheme of *fig. 6*, where a voltage divider $C_b$-$R_g$ is applied across the output. At $R_g$ a voltage $V_b' = V_o \cdot R_g/(R_g + 1/j\omega C_b)$ is obtained, which expression at $\omega C_b R_g \ll 1$ can be approximated by $j\omega C_b R_g \cdot V_o$. From a comparison with (3) we see that the desired effect can be obtained by choosing $C_b R_g = C_p/S$ and turning the voltage $V_b'$ another 180° in phase.

Let us now return to fig. 3*a*, where it is shown how, with the aid of a voltage divider $R_b$-$R_{g2}$, a push-pull circuit can be supplied with a symmetrical voltage notwithstanding the fact that the input voltage is asymmetric. It is obvious to combine this voltage divider with the voltage divider $C_b$-$R_g$ in fig. 6. The 180° phase shift required in the scheme of fig. 6 can be dispensed with if the voltage to be amplified, $V_1$, and the voltage tapped from $R_g$ are conducted to different valves, as illustrated in *fig. 7*. $R_b$ of fig. 3*a* and $C_b$ of fig. 6 come to lie parallel; $R_g$ of fig. 5 is called $R_{g2}$ in agreement with fig. 3*a*. As regards the working of the system according to fig. 7 it may be roughly said that at "low" frequencies (*i.e.* where the influence of $C_b$ is negligible) it changes over to that of fig. 3*a* and that at "high" frequencies (where $R_b$ can be ignored) it corresponds to the system of fig. 6 (where the 180° phase shift is brought about by the push-pull circuit).

This system, however, still has an important shortcoming. Though $R_b$ and $R_{g2}$ may be chosen of such values that at low frequencies the voltage

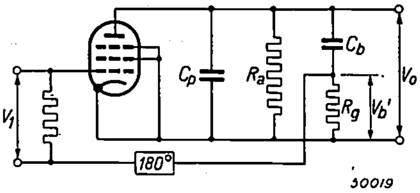


Fig. 6. The voltage $V_b'$ tapped from the voltage divider $C_b$-$R_b$ needs turning 180° in phase to serve as the compensating voltage $V_b$ of fig. 5*a*.

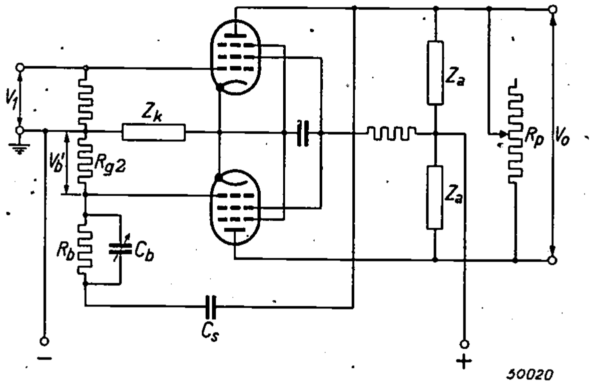


Fig. 7. Here the voltage dividers $R_b$-$R_{g2}$ of fig. 3*a* and $C_b$-$R_g$ of fig. 6 are combined into one ($R_b$, $C_b$) - $R_{g2}$ for improving the frequency characteristic at high frequencies. The impedance $Z_k$ has the same function, at high frequencies, as in the scheme of fig. 3*b*. $C_s$ is a separating capacitor. The smaller the resistance $R_p$ the less is the amplification but it is more constant at high frequencies; the maximum and minimum values of $R_p$ correspond respectively to the characteristics *b* and *C* of fig. 4.

is symmetrical, at higher frequencies this will no longer be the case, for the valve $B_2$ (fig. 7) has to supplement the deficit in amplification of $B_1$ caused by the drop in $Z_{a1}$. In fig. 3*b*, however, we have another means of making a push-pull amplifier symmetric, by choosing for $SZ_k$ a value greater than unity. This is in fact what has been applied in the amplifiers of the oscillograph GM 3159. Briefly it may be said that use has been made both of the method according to fig. 3*b* as well as that of fig. 3*a*, except that in the latter system instead of a potentiometer ratio of $R_{g2}/R_b$ (which is independent of frequency) a ratio is applied which is dependent upon the frequency, to the extent that the amplification at high frequencies deviates less from the nominal value; the danger of the symmetry being thereby disturbed is obviated by choosing a value for $SZ_k$ greater than unity.

Since $S$ is a fixed quantity, this means that $Z_k$ has to be chosen sufficiently large. And this applies only for high frequencies, since at low frequencies the system is identical with that of fig. 3*a*, where no account need be taken of $Z_k$. Consequently $Z_k$ consists mainly of the reactance of a choke coil, and for the rest a resistance required to induce the desired negative grid voltage.

The extent to which the frequency characteristic is improved by this method may be seen from fig. 4, curve *b*: the frequency at which the amplification of 3 db drops to $^1/_2\sqrt{2}$ times the nominal value (thus 71%) has been raised from 85 c/s to approx. 460 c/s. Fig. 7 shows another special feature, the variable resistance $R_p$ between the output terminals. The smaller $R_p$, the smaller the amplification but the more constant at high frequencies. With $R_p$ at its minimum the characteristic *c* of fig. 4 is obtained where the amplification drops to 71% of the normal value around 1000 kc/s. Thanks to this resistance it is therefore possible to work also in the 500-1000 kc range with a fairly constant amplification, though a larger input signal is required for the same picture height.

### Analysis of the push-pull circuit with correction.

Needless to say, the above more or less approximative statements have to be more carefully analysed to reach the best results. Here we shall show how some of the calculations work out.

For the lowest frequency to be considered the quantities $\omega C_b R_b$ and $2\omega\ C_p R_a$ should be smaller than unity, whilst $R_b \approx (g_0 - 1)\ R_{g2}$, in which $g_0 = SR_a \mp$ half the total nominal amplification.

At high frequencies $\omega C_b R_b$ must be greater than unity. Between $V_b$ and $V_o$, according to (3), a phase difference of 90° is needed, and this is approximately obtained when $g_0 \gg 1$ and $C_b = 2C_b/SR_{g2}$. The absolute value $g$ of the total amplification is then:

$$g = \left|\frac{V_o}{V_1}\right| = 2g_0 \sqrt{\frac{g_0{}^2 + (\omega_p C_p R_a)^2}{(g_0 - \omega^2 C_p{}^2 R_a{}^2)^2 + (\omega C_p R_a)^2}}\,. \quad (4)$$

For a non-corrected amplifier (not having $C_b$ and $R_b$), provided $\beta SZ_k \gg 1$:

$$g = \frac{g_0}{\sqrt{1 + (\omega C_p R_a)^2}}\,.$$

For $\omega = 1/C_p R_a$ in this case the amplification is reduced by a factor $\sqrt{2}$; if we call this angle frequency $\omega'$ then (4) may be written as follows:

$$g = 2g_0 \sqrt{\frac{1 + \left(g_0 \cdot \frac{\omega'}{\omega}\right)^2}{1 + \left(g_0 \cdot \frac{\omega'}{\omega} - \frac{\omega}{\omega'}\right)^2}}\,.$$

This function is practically constant, *i.e.* $= 2g_o$, within a wide frequency range, but as the frequency rises it ultimately begins to drop. It becomes a factor $\sqrt{2}$ smaller at an angular frequency $\omega''$, which amounts to approximately $1.5\ \sqrt{g_0} \times \omega'$. The bandwidth within which the amplification is greater than $2g_0/\sqrt{2}$ is therefore eularged by the correction by approximately a factor $1.5\ \sqrt{g_0}$. Such a high factor, however, will not be attained in practice, due to several causes, the explanation of which would lead us too far afield here. Nevertheless, the factor that is attainable, in the present case amounting to 460 kc/s : 85 kc/s = 5.4, is appreciably greater than what can be reached with the method described in the article quoted in footnote [2]); there the drop of $Z_a$ with increasing frequency was counteracted by introducing additional elements (*e.g.* one or two coils and a capacitor) in each anode impedance, by which means the bandwidth can only be increased by a factor no greater than about 3 without involving further difficulties, which we cannot enter into here.

## The generator for the time-base voltage

In the article quoted in footnote [1]) a system is described for generating the time-base voltage by employing three valves. In the new oscillograph it has been possible to simplify this arrangement considerably, generating the time-base voltage with only one valve, though amplification is needed to get sufficient amplitude, which is achieved by means of the amplifier $A_h$ (fig. 1*b*). Really, therefore, use is again made of three valves, but two of these (those of the amplifier) serve for other purposes too, which is not the case in the old system.

The principle of the system is as shown in *fig. 8*.

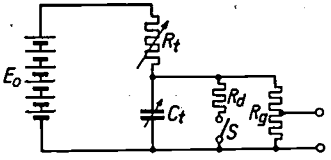


Fig. 8. Diagram of the time-base circuit. The capacitor $C_t$ is charged from a voltage source $E_0$ *via* a variable resistance $R_t$ and discharged (when the switch $S$ is closed) *via* the much smaller resistance $R_d$. Actually $S$ is a valve ( see fig. 9*a*). $R_g$ = input potentiometer of an amplifier.

When the switch S is opened then a capacitor $C_t$ is charged *via* a resistance $R_t$ from a direct voltage source $E_0$, and when the switch is closed this capacitor is discharged *via* the resistance $R_d$. Actually the function of the switch is performed by a valve, which will be referred to presently.

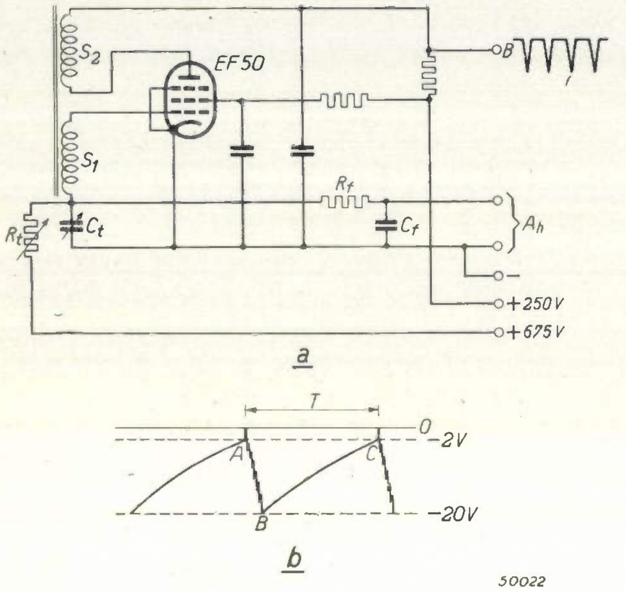


Fig. 9. *a*) Diagram of the generator for the time-base voltage. *b*) Behaviour of the voltage on the grid capacitor $C_t$. The pentode EF 50 is brought into the squegging state by the back-coupling $S_2$-$S_1$. The grid current then soon charges $C_t$ sufficiently for the resultant negative grid voltage (20 V) to block the anode current (curve *AB* in *b*). $C_t$ discharges itself *via* the resistance $R_t$ (curve *BC* in *b*) until at about —2 V oscillation starts again. Owing to $R_t$ being connected to a point with a high positive potential, a practically linear voltage *BC* is obtained. The voltage on $C_t$ is conduced *via* a filter $R_f$-$C_f$ — which cuts out the high oscillating frequency — to the amplifier $A_h$ for the horizontal deflection (*cf.* fig. 1*b*). At *B* voltage impulses can be drawn off for suppression of the electron beam during the retrace.

Both in the charging and in the discharging of $C_t$ the voltage $e_c$ behaves as an exponential function of time, so that in essence it is not linear. Consequently the horizontal velocity of the light spot on the screen will not be constant when $e_c$ is conducted to the plates for horizontal deflection. The degree of this non-linearity, however, is insignificant when the amplitude $E_c$ of the voltage $e_c$ is kept small in comparison with $E_0$.

A simple calculation shows that the percentage of velocity variation *D* while the light spot is travelling from left to right may be written as:

$$D = \frac{E_c}{E_0} \cdot R_t \cdot \left(\frac{1}{R_t} + \frac{1}{R_g}\right).$$

where $D = (v - v')/v$ with $v$ = initial velocity, $v'$ = the velocity at the end of the time-base, and $R_g$ = the input resistance of the amplifier, which is continually parallel to $C_t$.

It is seen that for $R_g \geqq R_t$ the velocity variation *D* does not exceed twice the ratio $E_c/E_0$. If for $E_0$ we use the voltage with which the amplifiers are fed (the value of which has been deduced above as 675 V) then with $E_c$ = 20 V the velocity variation is limited to less than 6%, which is not found to be at all troublesome in practice.

As remarked some way back, the function of *S* in fig. 8 is performed by a valve. This is a pentode of the type EF 50 (see *fig. 9a*). It is brought into a squegging condition [5]), that is to say it is caused to oscillate during a short period when the negative grid voltage induced in the capacitor $C_t$ is rapidly rising from about —2 to —20V (*AB* in fig. 9*b*); when the latter value is reached the slope has become so small as to stop the oscillation. A source with high voltage (675 V) supplies an opposite charge *via* the resistance $R_t$, causing the grid voltage to drop again to —2V, when the oscillation begins anew. The relaxation time *T* (fig. 9*b*) can be adjusted within wide limits by varying $C_t$ and $R_t$; in the oscillograph GM 3159 the corresponding

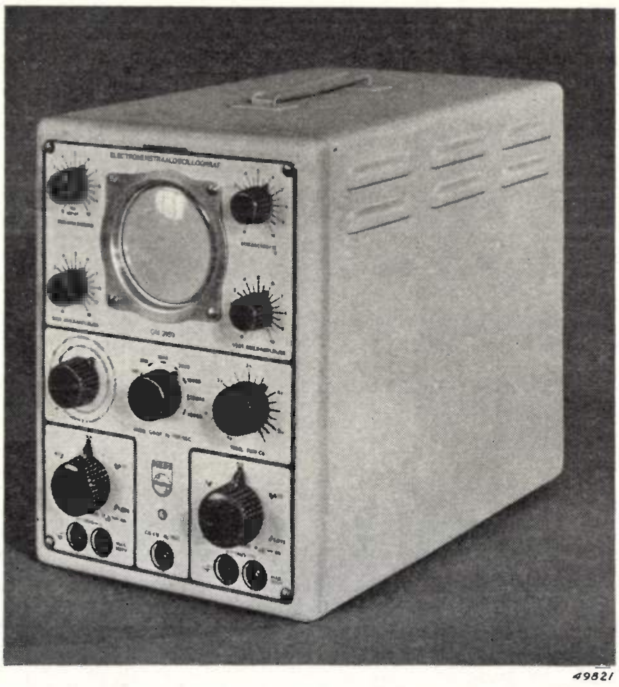


Fig. 10. External view of the cathode-ray oscillograph GM 3159. On either side of the screen, at the top, are the controls for brightness and sharpness of the light spot, and below those the controls for horizontal and vertical picture amplitude (resistor $R_p$ of fig. 7). In the middle row from left to right: the switch *S* of fig. 1*b*, the coarse and the fine frequency regulators for the time-base voltage. The two controls at the bottom regulate the input potentiometers of the amplifiers. At the very bottom on the left and right are the plug sockets of the inputs *II* and *I* (fig. 1*b*); in the middle is the socket from which the saw-tooth voltage can be taken off for other measurements (*e.g.* of amplifiers). A cap can be fitted over the front panel to protect the screen and controls during transport.

[5]) See, *e.g.* J. Van Slooten. The working of triode oscillators with grid condenser and grid resistance, Philips Techn. R., 7, 40 — 45, 1942, and Stability and instability of triode oscillators, Philips Techn. R., 7, 171 — 177, 1942.

frequency is variable between 10 and $150 \times 10^3$ c/s. If $R_t$ and $C_t$ were connected directly in parallel the voltage on $C_t$ would follow a pronounced curve *BC*, but with the arrangement of fig. 9*a*, where a source of high voltage is connected in series with $R_t$, a practically straight line is obtained (fig. 9*b*).

In the amplifier $A_h$ the saw-tooth voltage from $C_t$ is amplified sufficiently to give a deflection across the whole width of the screen; this deflection can be regulated with the potentiometer $Z_4$ (fig. 1*b*).

Furthermore, the saw-tooth voltage is carried to a separate terminal so that it can be used for other purposes outside the oscillograph, *e.g.* for measuring amplifiers [6]), without any need for a separate signal generator.

The moment at which the slope again becomes large enough to start the pentode oscillating again can be controlled by applying an additional alternating voltage to the control grid. In this way only a small voltage suffices to synchronise the frequency of the time-base voltage with that of the extra voltage referred to.

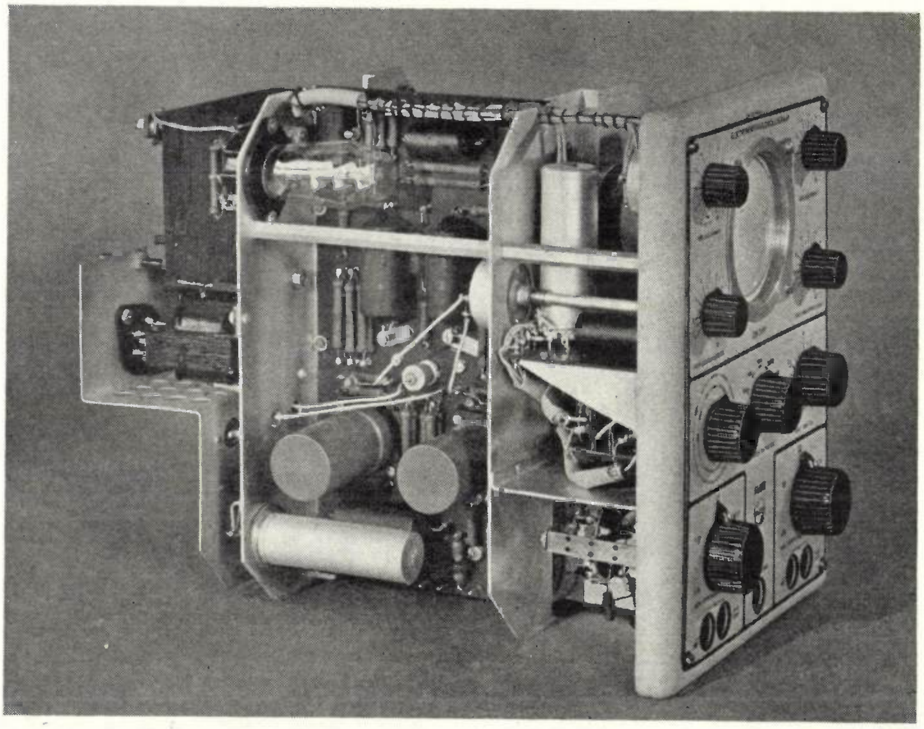


Fig. 11. An internal view of the oscillograph GM 3159. The compartment on the left contains the supply unit, the one in the middle the two amplifiers. The empty space on the left can be reached through a trap in the back of the casing and can be used for storing away the flexes.

Since the valve only takes up anode current during the short interval of oscillation, negative voltage impulses arise at the end *B* (fig.9*a*) of the resistance introduced in the anode circuit. By conducting these impulses to the control grid of the cathode-ray tube they can be used for suppression of the retrace of the beam. This often makes the picture clearer, but a small part of the oscillogram is then lost, and if this cannot be dispensed with it is necessary to switch off the beam retrace suppression, which can be done quite easily.

## Construction of the apparatus

In its outward appearance the oscillograph GM 3159 (*fig. 10*) does not differ essentially from its predecessors. There is more to be said about its internal construction, in particular about an improved magnetic screening and a new technique of assembly.

When a cathode-ray oscillograph is used in the vicinity of a strong magnetic field — for instance near transformers, electromotors and the like — that field is apt to cause troublesome deflections of the electron beam, unless the cathode-ray tube is adequately screened off against magnetic influences. That is why it is usual to insulate the tube

[6]) The use, for such purposes, of voltages whose curves deviate strongly from a sine is described by J. Haantjes in: The judging of an amplifier by means of the jump characteristic, Philips Techn. R., 6, 193 — 201, 1942, which article deals with the use of a block-shaped voltage.

in a cylinder that has good magnetic conducting properties. This effect is expressed as the screening factor, which is defined as the ratio of the magnetic field strength outside the cylinder to that inside it. The thicker the wall of the cylinder and the higher the permeability of the material of which it is made, the greater is the screening factor. By using an alloy of very high permeability for the cylinder in the new oscillograph the screening factor has been raised from 12 to 500, whilst at the same time the weight of the cylinder has been reduced from 600 to 130 grammes. Thanks to the improvement in the magnetic screening the new oscillograph can be used where there are fairly strong magnetic stray fields.

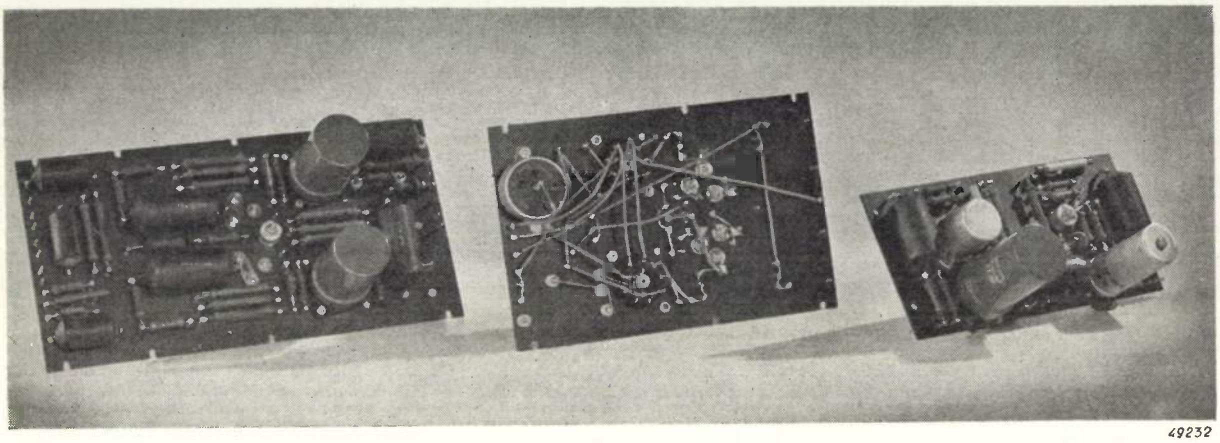


Fig. 12. Units mounted on insulating panels: on the left one of the push-pull amplifiers (rear view), in the middle the other amplifier (rear view), on the right the generator for the time-base voltage. (The two amplifier panels are not equal in size because on one of them some other parts are mounted which belong to the supply part.)

To carry off the heat generated in the oscillograph horizontal partitions tending to obstruct the natural flow of air have been avoided wherever possible. The inside of the oscillograph is divided into three compartments by vertical metal partitions (*fig.11*) acting as screens between the component parts liable to influence each other. In the middle compartment are the two amplifiers placed back to back, each mounted on an insulating panel (*fig. 12*). Pressed into these panels are brass pins, to which the electrical parts are soldered at one end and the the wiring at the other, so that the wiring can be kept very short and at the same time strongly fixed, as are also the electrical parts, thus precluding troublesome capacity variations due to shifting about. The advantage of this method from the manufacturing point of view is that the units can be pre-mounted and tested before being built into the casing (see fig. 12).

Finally it is to be mentioned that the dimensions of this oscillograph are 21 × 27 × 37 cm and that its weight is 13 kg. Compared with the GM 3152 for instance this means a saving of about 75% in volume and over 30% in weight, so that the new oscillograph gains much in handiness.

# TESTING FOR NYCTALOPIA (NIGHT-BLINDNESS)

by W. S. FREDERIK.

612.845.6-073

For certain occupations or jobs inadequate adaptation for the dark (night-blindness) may be awkward. In some cases, for instance where it means that the person affected by nyctalopia has to be exempted from night work, the possibility of simulation has to be taken into account. When such a situation arose at Philips a simple method was developed by the Medical Department of this company for testing persons for night-blindness with no possibility of simulation. With this method, which has yielded good results in practice, during the adaptation to the dark the person being tested has to look at a specially composed letter card and read words from it which differ according to the state of adaptation.

Under favourable conditions the human eye is able to observe levels of brightness as low as 0.000 003 candle/m$^2$, (1 c/m$^2$ = 0.0929 c/sq.ft) while on the side of high brightness vision is only restricted by the limit of pain, which lies at 200 000 candles/m$^2$. For covering this extremely wide range of brightness the retina contains two kinds of light-sensitive elements: the cones, which act at high levels of brightness such as occur in daylight, and the rods, which gradually take over this visual function when the brightness of the field of vision falls below about 3 candes/m$^2$ (the level at which motorists usually switch on their headlights).

The light-sensitive substance contained in the rods is called rhodopsin or visual purple, which is composed of protein and vitamin A. The cones are assumed to contain three different light-sensitive substances, but it has not yet been found possible to determine them. It is further assumed that these light-sensitive substances are continually being formed as long as the eye can see, while at the same time they are destroyed by the light striking the eye. Between this formation and destruction of the light-sensitive substances an equilibrium is supposed to be established which determines their concentration, there being more light-sensitive substance present according as the brightness is lower, because the less light that strikes the eye the more slowly the substance is destroyed. In this way the sensitivity of the eye adapts itself to the level of brightness.

As regards the still unknown light-sensitive substances in the cones the mechanism sketched above is of course purely hypothetical. For the visual purple of the rods, however, experiments would appear to confirm that this is what actually takes place. The taking over of the function of vision by the rods, mentioned above, can now be explained simply as follows: whereas at a brightness higher than about 3 c/m$^2$ the concentration of visual purple is practically nil owing to its rapid decomposition, at lower levels of brightness the concentration gradually increases [1]).

The formation of the visual purple takes some time. This can easily be observed when, coming out of a brilliantly lighted room, one steps outside on a dark night: it takes several minutes before one can see well enough in the dark. The eyes do not become completely adapted to the dark until half an hour or more later, though as a rule the increase in light-sensitivity is greatest during the first 10 minutes, after which the increase is only relatively slight.

This adaptation to the dark does not proceed in the same way for all persons. When a large number of persons are tested the light-sensitivity of some of them is found to increase slower than normal though ultimately reaching a normal value, whereas with others the normal light-sensitivity is not reached even after a long time of adaptation. Such an aberration in the latter group, when the ultimate light-sensitivity remains below a certain limit, is called night-blindness (nyctalopia). In a number of cases this defect is found to be caused by a deficiency of vitamin A, which is understandable after what has been said above about the composition of visual purple (rhodopsin).

Persons with such defective adaptation have difficulties in perception in the dark, as in a poorly lighted street at night. There are several occupations for which such people are therefore unsuited, for instance as chauffeur and in certain positions on railways, in shipping and aviation and in the police. The testing of applicants for such positions should therefore include a test for night-blindness. The usual methods are practically all based on the

[1]) Why the cones, which according to our primitive conception should then contain a good quantity of light-sensitive substance, gradually cease to function just at these levels of brightness is not explained. It may be ascribed to the existence of an absolute threshold for the light-sensitivity of those elements lying in the range of brightness between about 0.03 and 3 c/m$^2$.

same principle: after some time in a brightly lighted room the examinee is taken into a very poorly lighted room where he has to recognize certain objects within a given time.

Provided the examinee is desirous of passing the test successfully there is not much to be said against such a test, but it may well happen that he desires to be rejected. When, for example, nyctalopia necessitates exemption from night duty, the examinee may try to simulate night-blindness and if tested on the above lines need only maintain that he cannot see any of the objects he is asked to recognize. It is then not easy to expose such a simulation and it is impossible to determine his true degree of adaptation to the dark.

A case occurred at Philips where account had to be taken of the possibility of night-blindness among a group of employees, and also of the possibility of simulation. When it appeared that no simple method was known which precluded simulation, the Medical Department of Philips devised a simple apparatus which meets this problem and has proved satisfactory in practice.

The principle of this apparatus is shown diagrammatically in *fig. 1.* A letter card, which will be described later, is placed in a light-proof box and illuminated by a small incandescent lamp *via* a diffusely reflecting screen, with a certain low luminous intensity which can by varied by means of an adjustable diaphragm placed in front of the lamp. After the person being tested has been kept for 10 minutes under standardized conditions to allow his eyes to get adapted to a certain high level of brightness, the room is completely darkened and he is made to look through a slit in the cabinet at the letter card at set times, say at intervals of 1 minute.

The letter card is shown in *fig. 2.* It bears the word BUURT on a black background. Each letter consists of different parts painted in tints of grey with different reflection factors. The letters are divided into parts in such a way that when the darkest part (the part with the smallest reflection factor) is omitted a different letter results, and upon omitting the darkest part of that in turn yet another letter is formed. For example, from the U an L is formed, and from the L an I. Taking the painted word as whole and successively omitting the darkest of the five tints we read in turn the following words: BUURT - BUUR - BULT - BUI - EL,

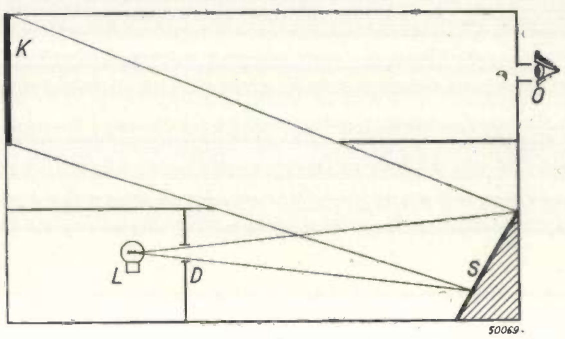


Fig. 1. Apparatus for testing night-blindness. The test person looks through the slit *O* in a light-proof box at a letter card placed at *K*. This is illuminated by one or more lamps *L via* the adjustable diaphragm *D* and the diffusely reflecting screen *S*. The horizontal partitions prevent any direct light from *D* falling upon the eye of the observer.

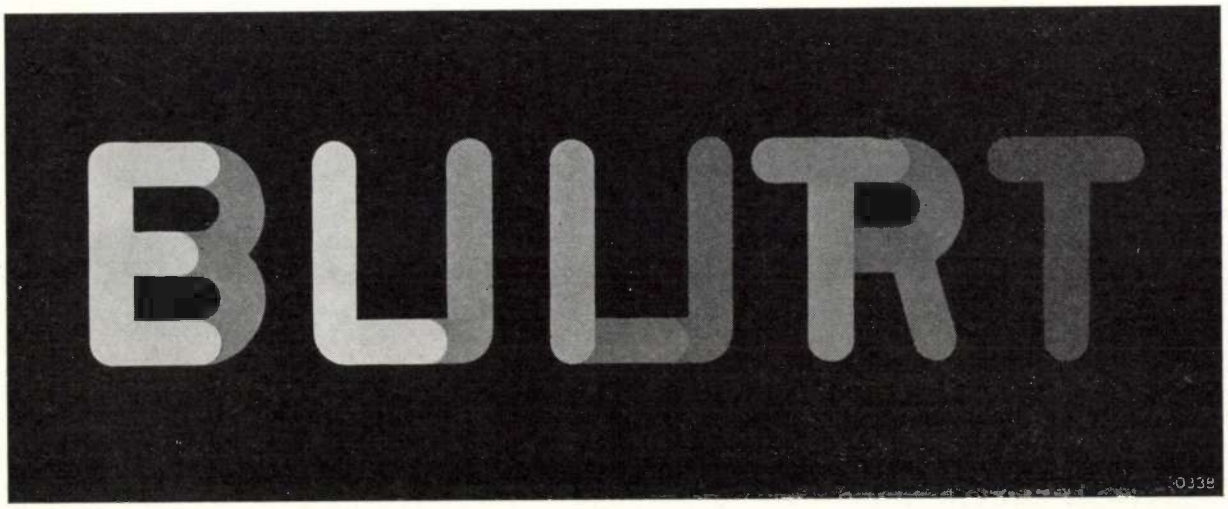


Fig. 2. Letter card used for the text. The letters, painted on a black background, are composed of five different shades of grey in such a way that upon successive omission of the darkest shades a different word results. The better the adaptation of the eye, the darker the shade that can just be observed, and thus the more complete the word that can be read. With progressing adaptation, therefore, the subject reads successively the words, EL - BUI - BULT - BUUR - BUURT.

which are all words commonly occurring in Dutch.

The intensity of illumination of the letter card is so chosen that a fully adapted normal test person can read the whole word BUURT. When he first looks through the slit, however, he will not, as a rule, see anything at all, because his eyes have just been adapted to a rather high level of brightness, thus possessing a low sensitivity to light. After a while his sensitivity has risen so far that he can see the brightest tint on the letter card and thus perceive the word EL. A little later his vision is sufficiently adapted to be able to read the word BUI, and so on. A remarkable feature about this is that in a certain state of adaptation upon a brief glance at the letter card only one word is perceived, and thus in the case last mentioned no need is felt to choose between BUI and EL. The reason for this is that at the low level of brightness of the letter card the contrast sensitivity of the human eye (*i.e.* the power of observing small differences in brightness) is much lower than in daylight, and in the five shades of grey used on the card the contrast between two successive shades is made so small (ratio of the coefficients of reflection 1 : 1.7) that it passes practically unnoticed at the low brightness mentioned, especially when only a brief glance is taken. (The series of words are built up in such a way that two adjacent parts of a letter always differ by one stage in reflection factor.) As a result, test persons who are unacquainted with the structure of the letter card do not immediately notice how the different words are formed from each other [2]).

The further procedure is as follows. Each time the subject looks through the slit of the instrument he is asked what word he sees. The numbers 1 to 5 are assigned to the five stages of sensitivity of the eye corresponding to the ability to read the five different words. The sensitivity value for the word EL is thus 1, that for BUI 2 and so on. When the sensitivity values found from the test are plotted as a function of the time, a curve is obtained for each subject representing the progress of his adaptation. Several of such curves are drawn in *fig. 3*.

Experiments were undertaken to ascertain the adaptation curves with a given letter card and a given intensity of illumination for normal persons and for those affected with night-blindness. "Normally adapting" persons were taken to be those who after 10 minutes adaptation could place themselves without difficulty in a dark street (a not very precise definition, but one appropriate to the case). The curves for night-blind people are found to be much flatter than those for normal persons.

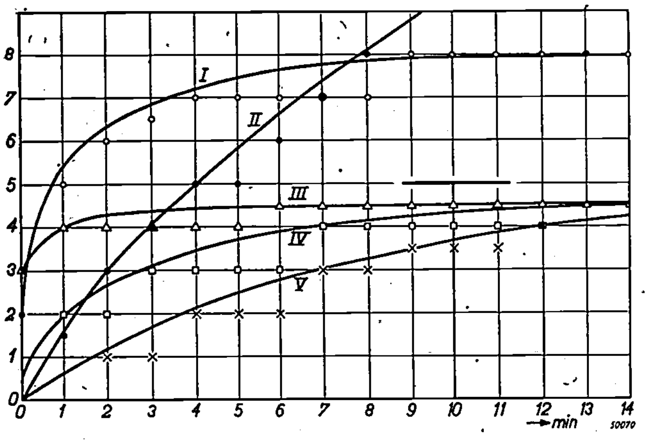


Fig. 3. Progress of adaptation to the dark for five different test persons. The values from 1 to 5 are assigned to the eye sensitivity corresponding to the ability to read the five successive words of the letter card. The sensitivity value is plotted as a function of the time for adaptation. The continuation of the curves above the value 5 was obtained as described in the text.

The three flat curves (*III*, *IV* and *V*) which at 10 minutes on the abscissa still lie below the limit indicated (sensitivity 5) represent three cases of night-blindness. There are large individual differences (as in fact also in the "normal" cases *I* and *II*). In the case of the subject *III* there is a very rapid initial adaptation, but considering the shape of the curve it is improbable that the eye sensitivity will ever rise above the value 5. On the other hand the test person *V* belongs to the group of people whose adaptation takes place very slowly but perhaps in time finally reaches a satisfactory sensitivity.

On the basis of the tests described we have taken as criterion for the existence of night-blindness that after 10 minutes the sensitivity of the eye has not yet reached the value 5. In fig. 3 three examples of this can be seen.

If desired, the progress of adaptation can be followed still farther, with the same apparatus, than to the increase of sensitivity corresponding to the value 5. When the sensitivity 5 has been reached (thus when the word BUURT has been read), the intensity of illumination of the letter card is reduced so that only the word EL can be seen. That word is then given the sensitivity value 5, the numbers 6, 7 etc. being assigned to the higher values of sensitivity reached upon further adaptation and causing the subject to read successively again the words BUI, BULT, etc. Where the curves in fig. 3 extend above the ordinate

[2]) For those who would like to make the test themselves with the letter card reproduced here, in order to convince themselves of the succesive appearance of the five words and of the fact that the contrasts between the parts of the letters are scarcely noticeable, it must be noted that the card has to be illuminated with about 0.0005 lux. This corresponds, for example, to the level of illumination on a clear, moonless, starry night. If the level of illumination is very much lower there is a chance that the whole word can never be seen, while if it is much higher the contrasts between the parts of the letters become too noticeable.

5 they have been determined in this way [3].

Conversely, at every state of adaptation the complete word BUURT can immediately be made legible — also for night-blind persons — by increasing the intensity of illumination on the letter card to a certain level. If, for example, the subject was so far adapted that he could read the word BUI, the increase in sensitivity which would then be necessary for him to be able to see the three darkest shades on the card and thus to read the word BUURT becomes unnecessary, because at the higher illumination intensity these three shades become so much brighter. If a subject should then maintain that he can only read the word BUI, it is certainly a case of simulation. Such a case, however, has not yet occurred in practice, since with the method described here the subjects are quite unable to figure out how they can impress the examining doctor with the suggestion of night-blindness.

In conclusion we have to mention some details in the application of the method. Since at low levels of brightness visual acuity is very much reduced, the letters must be made fairly large. In our apparatus the word BUURT covered an area of 10 × 50 cm, while it was viewed from a distance of 40 cm. The variation required in the intensity of illumination of the letter card was obtained by using a number of small lamps as light source (instead of one) and/or changing the opening of the diaphragm (*D* in fig. 1). The more obvious method of using a sliding resistance in series with the lamp was expressly not employed because then when the light is reduced its colour is shifted towards the red, and as the eye sensitivity depends closely on the colour of the light this would cause errors in the estimation of the state of adaptation.

A small error, of little consequence for our purpose, may occur due to the fact that the "black" background of the letter card reflected a small part of the incident light. This error can be avoided by constructing the instrument in such a way that the letter card is observed under transmitted light.

In that case the letter card is also somewhat easier to make: the letters are sawn out of a metal plate and their different parts covered with translucent paper in different numbers of layers.

There would be no fundamental difficulty in replacing the relative calibration of the apparatus with the arbitrary scale of sensitivity values from 1 to 5 by an absolute calibration. It would then be possible to obtain absolute data about the adaptation power of the persons tested. This, however, was outside the scope of our object.

[3] By slightly varying the illumination intensity to a certain degree also intermediate stages of the eye sensitivity can be measured, thus values of 1 1/2, 2 1/2, etc. There are many such intermediate values in the curves of fig. 3.

# APPLICATIONS OF LUMINESCENT SUBSTANCES

by F. A. KRÖGER.

535.376:661.4

**Luminescent substances are used when rays which are invisible to the human eye are applied and it is desired to make them visible. Such invisible rays may be beams of electrons (in cathode-ray tubes for oscillographs, television and radar apparatus), X-rays, utra-violet rays (in gas-discharge lamps), or infra-red rays. Here it is discussed what luminescent substances are most suitable for each of these cases.**

## Introduction

In technology it is often required to convert rays which are invisible to the human eye into visible rays. The originally invisible rays may be of a widely different nature. One may have to do with electro-magnetic waves differing considerably in length, or with very rapid electrons.

In the case of cathode-ray tubes, which are mainly applied in oscillographs, television sets and radar installations, it is a beam of electrons that has to be made visible. Medical practitioners meet the same problem in the screening of their patients with X-rays. The conversion of ultra-violet rays into visible light is applied in modern gas-discharge lamps. Finally there is to be mentioned the conversion of infra-red into visible rays, as was particularly applied during the last war.

It is not always that the ultimate object is to obtain visible light. Sometimes it is required to convert some irradiation or other into, for instance, ultra-violet of a suitable wavelength. This is the case, for instance, in irradiation lamps emitting ultra-violet of a certain wavelength which has a specific effect upon living organisms; applications of this are to be found, *inter alia*, in anti-rachitis and germicide lamps and so-called sun-tan lamps.

In all such cases as these use is made of substances which have the property of being able to bring about the desired transformation, and as in most cases it is a matter of producing visible light they are called luminescent substances or luminophores.

The transformation in question is called luminescence, Two effects are to be distinguished. Fluorescence is a process that predominates during the radiation, being the remission of the energy first absorbed, and after the radiation has stopped this usually disappears within a short time $10^{-5}$—$10^{-1}$ sec.). Phosphorescence on the other hand is particularly of importance after the radiation has ceased; the time during which this phosphorescence persists differs considerably according to the prevailing conditions, especially of temperature, persistences of some hours or even days being possible.

This rather rough definition of the two effects suffices for the purpose of the present article. If we were to go more closely into the theory of the phenomena of luminescence it would be necessary to define them more precisely, having regard to the mechanism of the atomic processes brought about by the phenomena observed. This, however, has already been dealt with more than once in this journal. We would only recall here Stokes' rule, which says that the wavelength of fluorescence is generally longer than that of the primarily absorbed exciting radiation [1]).

The luminescent substances used vary from one case to another. In this article a survey will be given of their most important applications and of the substances that can be used in the present stage of developments.

The conditions which these substances have to satisfy can be placed, broadly speaking, in three groups. In the first group we have those requirements that are related to the mechanism of transformation; the incident radiation has to be suitably absorbed, whilst the energy absorbed must not be converted into heat but into the desired luminescent radiation. To the second group belong those conditions that are determined by the properties to be expected of the fluorescent radiation excited, such as its spectral distribution and duration of persistence. Thirdly, the fluorescent substances, which, with a few exceptions, are obtained in powder form, must be capable of being applied in thin even layers; this depends upon the absolute as well as the relative size of the particles, whilst it is also necessary that the substances should not undergo any chemical change during manufacture (*e.g.* due to the use of adhesives).

Several of these points will be considered in more detail when dealing with particular cases, where it will also be investigated in how far the

[1]) For a closer study of the relation between absorption, irradiation and fluorcscence see the articles byW. de Groot, Philips Techn. R., **3**, 125-132, 1938, J. H. Gisolf and W. de Groot, Philips Techn. R., **3**, 241-247, 1938 and F. A. Kröger, Philips Techn. R., **6**, 353-362, 1941.

development of luminophores has progressed to be able to satisfy these requirements.

### Luminescent substances for cathode-ray tubes

Cathode-ray tubes are used for several purposes, the foremost being:

1) oscillographs,
2) television receivers,
3) radar installations.

#### *Oscillographs*

There are various forms of oscillograph tubes according to the purpose for which they are intended. For visual use a green fluorescence is desired, the persistence required being in some cases short and in others long. Short-persistent green is obtained with the aid of ZnS-Cu-Ni or $(Zn,Mn)_2$ $Sio_4$; with the former the luminescence arises from the traces of copper taken up in the crystal structure of ZnS. Nickel causes the short after-glow. Since the position of Cu and Ni in the lattice is uncertain, this system is indicated by writing ZnS-Cu-Ni; in the second case we have a mixed crystal, the manganese replacing the zinc; this is indicated by writing (Zn,Mn). Long-persistent green is obtained with ZnS-Cu or $(Zn,Mn)_2SiO_4$-As. ZnS-Cu is known to give a much longer persistence under the action of long-wave ultra-violet or blue than under the action of cathode-ray electrons. This makes it possible to extend considerably the duration of the after-glow by introducing a second luminescent substance in which cathode rays induce a violet light, which in turn brings about the green fluorescence and after-glow (the phosphorescence). The same applies for yellow luminescent (Zn,Cd)S-Cu, also under the action of blue light. The most favourable type is a double screen consisting of a yellow fluorescing layer (Zn,Cd)S-Cu on the glass and on top of that a second layer of blue fluorescing ZnS-Ag,$CaWO_4$, $Zn_2(Si,Ti)O_4$ or (Ca,Mg) $(SiO_3$-Ti).

If photographic recording is desired then a blue fluorescing screen is of advantage, and in that case a single screen with one of the last mentioned substances is sufficient.

#### *Television receivers*

In television receivers an electric signal is transformed into a visual picture by means of a cathode-ray tube. This picture may be formed on the screen of the tube itself (direct vision) or it may be projected by a system of lenses or mirrors onto a transparent screen on a magnified scale (projection television).

We have to differentiate between black and white and colour television.

For black-and-white television the fluorescent substances forming the screen have to emit a reasonably white light under the action of the cathode rays. This white colour must as far as possible be independent of the intensity of the cathode ray. Furthermore, the luminescent substances must not leave an after-glow, because otherwise rapidly moving objects would leave troublesome smudges of light.

The white colour is best obtained with a mixture of two substances, one emitting blue light and the other yellow. This is shown graphically in *fig.1* by means of a colour triangle. For yellow (Zn,Cd)S-Ag, Zn(S,Se)-Ag and $(Zn,Be,Mn)_2SiO_4$ are used, and for blue ZnS-Ag, $Zn_2(Si,Ti)O_4$ and (Ca,Mg) $(Si,Ti)O_3$, none of which substances leaves any after-glow as to be troublesome.

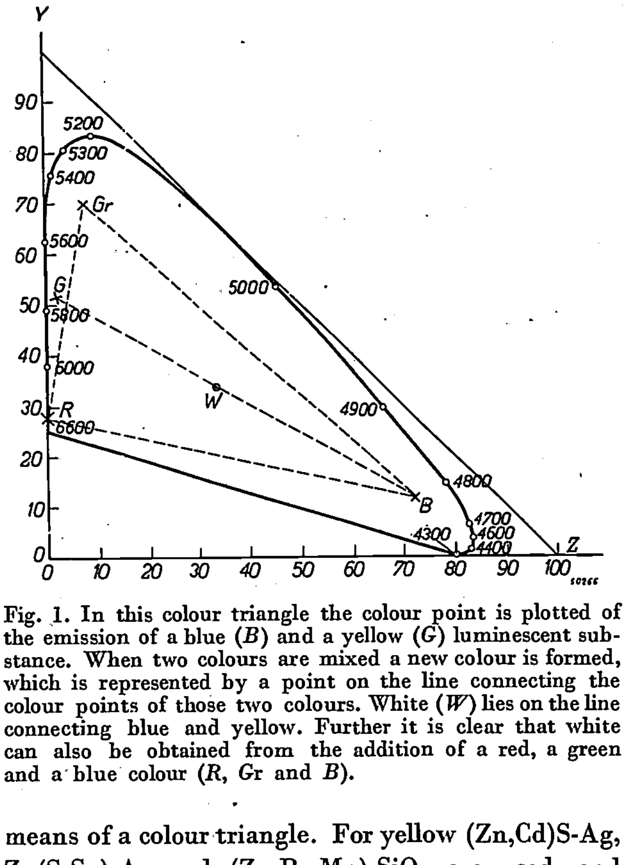


Fig. 1. In this colour triangle the colour point is plotted of the emission of a blue (*B*) and a yellow (*G*) luminescent substance. When two colours are mixed a new colour is formed, which is represented by a point on the line connecting the colour points of those two colours. White (*W*) lies on the line connecting blue and yellow. Further it is clear that white can also be obtained from the addition of a red, a green and a blue colour (*R*, *Gr* and *B*).

Where low voltages are applied, as in direct-viewing tubes (*e.g.* 5 kV), the sulphides are to be preferred on account of their high intensity of radiation. In the case of higher voltages as used in projection tubes (*e.g.* 25 kV) the sulphides and the silicates are about equal in their intensities of radiation; in this article we shall not go into the question which of these two is to be preferred in the long run.

With some mixtures applied in practice small variations in colour occur between parts of different brilliancy, due to the intensity of fluorescence of the components not depending in exactly the same degree upon the intensity of the electron beam. In the first place this arises from a divergence from the linear relation between those two quantities yielding the so-called current saturation. In the case of the sulphides, moreover, there may be a second effect: the emission from these substances consists sometimes of two bands lying in different parts of the spectrum, the one with the shorter wavelength increasing slightly at the cost of the other when the radiation density is increased.

For colour television there are various types of receivers. One familiar type is that commonly used when in the transmitter a quickly rotating disc having three sectors (red, green and blue) is employed, and in the receiver three filters, one for each of these primary colours. In that case not only is the fluorescence of the screen required to give a white colour but at the same time that white has to be of such a composition as to permit all possible colours to be obtained from it with the the aid of filters. This can be approximated fairly well by making up the white with the help of three fluorescent substances emitting light in the red, green and blue (see fig. 1). For blue we may again use ZnS-Ag, $Zn_2(Si,Ti)O_4$ or $(Ca,Mg)(Si,Ti)O_3$, for green ZnS-Cu-Ni or $(Zn,Mn)_2SiO_4$ and for red $(Zn,Mn)_3P_2O_8$, $(Ca,Mn)_3P_2O_8$, $(Zn,Be,Mn)_2SiO_4$ or the recently developed $Ca_2P_2O_7$-Bi.

*Radar installations*

It is well known that the name radar is an abbreviation for radio detection and ranging, which was highly developed during the last war. Short, high-frequency, radio impulses are broadcast by a transmitter, and their reflection back from metal objects in the air went a long way to detecting aircraft. Radar has also proved to be a valuable asset for the navigation of aircraft and ships. These are only two examples, for there are many applications of radar.

In a radar receiver is a cathode-ray tube with its screen coated with a fluorescent substance, on which the signal received is made visible. In the simplest forms of radar these fluorescent screens do not have to answer any very special requirements, so that the normal substances can be used as mentioned above when dealing with oscillograph tubes.

For the more complex forms, such as the plan-position-indicator (P.P.I.), it is a different matter. This is an instrument which - when used from the ground - indicates the position of aircraft in its surroundings, or inversely, when used from an airplane gives a picture of the landscape below. An electron beam emitted from a cathode-ray tube passes fan-like across the screen, being synchronised with the rotating antenna which broadcasts the signals and picks them up again. The movement of the beam over the screen is illustrated in *fig. 2*. The period of the rotation is from 1 to 6 seconds. If persistent fluorescent substances were not used only the rotating line would be observed, but with an after-glow on the screen the field continues to yield light after the line has passed on and gives a picture corresponding to the degrees of intensity of the cathode ray at the various points on the screen. It is therefore solely due to the after-glow that a picture is obtained over the whole screen.

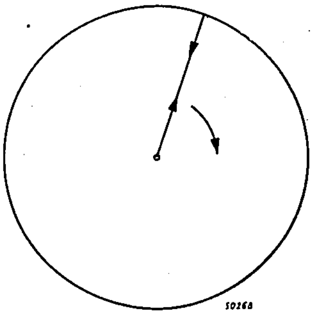


Fig. 2. A cathode-ray beam causes a narrow linear part of the screen of the plan-position-indicator to light up. This ray travels round the screen in about 1 second while the cathode-ray beam moves up and down that ray at a high velocity.

For the screen of the P.P.I. the best results are to be obtained with a substance which is strongly persistent exactly during the period of rotation, say 1 second, and after that emits no light at all. If the after-glow is shorter then the entire picture can never be illuminated, whilst if it is longer there will be an overlapping of the pictures in situations of one-second intervals, so that moving objects become blurred. Hitherto it has not been possible to comply with this requirement of 1 second constant, strong after-glow and then extinction. An acceptable compromise, however, has been reached, as illustrated in *fig.3,* where the time is plotted on the $x$-axis and the intensity of the luminescence on the $y$-axis; the broken line $ABC$ represents the ideal intensity change and the curve the approximation so far reached. Since the fluorescence under the action of the ray is much stronger than the phosphorescence there is a danger that owing to blinding nothing of the after-glow will be seen. Consequently it is necessary that the ratio of fluorescence to phosphorescence should be as small as possible.

In practice a double screen as desribed above for oscillograph tubes gives useful results. Provision is made in the first place for a strong phosphorescence. Since the fluorescence is mainly in the blue and the phosphorescence exclusively in the yellow, by using a filter which absorbs blue and allows yellow to pass through it is possible to suppress the intensity of fluorescence while maintaining the phosphorescence, By this means the ratio of fluorescence to phosphorescence can be appreciably reduced.

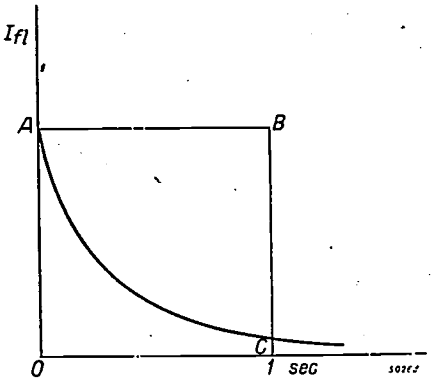


Fig. 3. Reduction in the intensity of fluorescence in a plan-position-indicator after the cathode-ray beam has passed. On the $x$-axis is the time and on the $y$-axis the intensity of the fluorescence. The broken line *ABC* represents the ideal intensity change and the curve the approximation hitherto reached.

Single screens of $(Zn,Mg,Mn)F_2$ have been suggested, but these have proved to be unsuitable on account of their chemical instability and high ratio of fluorescence to phosphorescence.

## Luminescent substances in Roentgenology

Since X-rays have a great power of penetration only a small part is absorbed by the fluorescent screens used in X-ray apparatus. However, the larger the fraction absorbed, the more energy can be converted into visible light. Therefore the first requirement to be met by luminescent substances for X-ray screens is that they should have the largest possible absorption coefficient for X-rays. This power of absorbing X-rays depends directly upon the number of electrons contained in the ions that go to make up the substance. Consequently we have to use fluorescent substances consisting for a large part of heavy ions. This is the reason why such a substance as barium-platinum-cyanide, which is otherwise of little importance as regards fluorescence, has continued to be used such a long time for X-ray screens. Further, the heavy substances $CdWO_4$, $CaWO_4$ and the recently developed $(Ba,Pb)SO_4$ and $Ba(F,Ce)_2$ are used. Of course another factor of great importance is the yield of visible light to be obtained from the conversion of the energy absorbed. Since the zinc and zinc-cadmium sulphides are favourable in this respect also these substances can be used to advantage for X-ray screens notwithstanding their relatively low power of absorption.

With these fluorescent screens long after-glow must always be avoided. If tungstates are used they can be suitably prepared so as not to cause any troublesome phosphorescence. In the case of sulphides, which under the action of X-rays glow for too long a time, the same result is obtained by incorporating extremely small quantities of an element like nickel or manganese (so-called killers) which specifically suppresses phosphorescence.

Finally it has to be considered whether the screen is to be used for visual observations or for photographic recording (with magnification). In the former case substances have to be chosen which give the optimum fluorescence with respect to the spectral sensitivity curve of the human eye *i.e.* those with a green to yellow fluorescence. In the other case account has to be taken of che properties of the photographic plate, and there substances with a particularly blue fluorescence are indicated. $BaPt(CN)_4$-aq, ZnS-Cu(Ni) and also the corresponding (Zn,Cd)S-Cu(Ni) are used as green or yellow fluorescing substances, whilst $(Ba,Pb)SO_4$ and $CaWO_4$ belong to the blue-fluorescing category. In between we have $CdWO_4$ with its bluish-green emission over a wide range.

## Luminescent substances for gas-discharge lamps

### *Discharge in mercury vapour of low pressure*

A discharge in mercury vapour of low pressure yields some visible light, but the greater part of the energy is emitted in the ultra-violet part of the spectrum, mainly on the wavelength of 2537 Å. Nevertheless, this discharge can be used for lighting purposes because it has been found possible to convert this emission into approximately white light[2]). This is done with the aid of a mixture of substances which absorb the ultra-violet radiation and each emit in a particular part of the visible spectrum. The substances used for this purpose are listed in *table I*.

Practically any colours desired can be obtained with mixtures of these substances. The last one listed is of particular importance, because with this alone a sufficiently white light can be obtained, though with small deficits in the green and the red.

[2]) See the articles in Philips Techn. R., **3**, 272-278, 1938; **4**, 342-350, 1939 and **6**, 65-73, 1941

All these substances answer the requirement of being able to absorb the radiation to be converted while at the same time giving a high yield from the conversion of the energy absorbed. The yield varies from 80 to 95 quanta of fluorescent light per hundred quanta of ultra-violet.

Table I. Some luminescent substances used for mercury discharge lamps with low pressure. The last column gives the time taken for the fluoresence to drop to about one-third.

| Luminescent substance | Colour of fluorescence | Dimming time |
|---|---|---|
| $CaWO_4$ | blue | $\sim 10^{-5}$ sec |
| $MgWO_4$ | bluish-green | ,, |
| $(Zn,Mn)_2\ SiO_4$ | green | $\sim 10^{-2}$ sec |
| $(Zn, Be, Mn)_2\ SiO_4$ | yellow to orange | ,, |
| $(Cd. Mn)_2\ B_2\ O_5$ | orange-red | ,, |
| $(Ca, Mn, Cc)_3\ P_2O_8$ | red | ,, |
| $(Ca, Mn, Sb)_5\ P_3\ O_{12}\ (Cl, F)$ | blue | $\sim 10^{-4}$ sec |
| | yellow | $\sim 10^{-2}$ sec |

Gas-discharge lamps normally work on alternating current, so that the intensity of the primary ultra-violet emission varies sinusoidally 50 times per second, which means that a dark interval occurs 100 times per second. Of course these dead moments occur also with an incandescent lamp connected to A.C. mains, but in such a lamp the slow cooling of the tungsten filament ensures that the light does not drop during these short intervals. When using luminescent substances the dark periods can be bridged over, in principle, by choosing substances which fluoresce for some time after the interruption of the excitation, so that the fluorescence starts to increase again before it is entirely extinguished; in other words, the ripple in the fluorescent light is smaller than in the ultra-violet emission of the discharge.

The duration of this after-glow differs as between one substance and another. With substances comprising manganese as the element responsible for the fluorescence the radiation drops to one-third in about 0.01 second, which is roughly sufficient for a reasonable smoothing out. For the green and red fluorescence this problem is therefore satisfactorily solved when using one of the last five substances of table I. The decay time of the blue and bluish-green fluorescing components hitherto used is much shorter ($10^{-4}$—$10^{-5}$ sec), with the result that the lamps at present on the market, which emit white light during the live periods, fluoresce with an orange-coloured light during the dead periods. This is apt to have unpleasant consequences for anyone working under the light of these lamps: there is still some flickering, and with rapidly moving objects there is a certain amount of stroboscopic effect. This difficulty can be overcome, in principle, either by modifying the known substances so as to get a longer after-glow or by developing entirely new substances having a sufficiently long after-glow. The first problem can be solved by incorporating suitable foreign ions in the crystals so as to make the substances phosphorescent. As a new substance with a longer after-glow, a blue luminescent cadmium phosphate activated with lead has recently been mentioned by English investigators, but apparently it has not yet been used in practice.

The discharge tube with mercury of low pressure can be applied for various purposes other than for illumination, as for instance in light-printing processes and for medical, cosmetic and biological purposes.

These applications require violet or ultra-violet emissions of different wavelengths. In the light-printing process the range is from 3300 to 4300 Å. Radiations of a slightly shorter wavelength (3000-4000 Å) promote pigmentation of the human skin. For light-printing lamps as well as for sun-tan lamps $CaWO_4$, $(Ca,Ce)_3P_2O_8$ or $(La,Ce)_2S_3O_{12}$ can be used. Rachitis can be checked with a source of radiation between 2700 and 3000 Å, which promotes the formation of vitamin D in the skin; for this purpose $(La,Ce)F_3$ or $Ca_3P_2O_8$-Tl have been proposed as fluorescent substances. For the promotion of plant growth blue and red light is of importance, and for that purpose tubes can be used which are lined with $CaWO_4$ (blue) and/or $Cd_2B_2O_5$-Mn (orange-red). Finally it may be added that ultra-violet radiation of very short wavelength causes on the one hand erythema in the human skin while on the other hand it kills bacteria. The 2537 Å radiation from the discharge itself serves this purpose, without any need of conversion and thus not requiring any fluorescent substance; the wall of the tube, however, has to be made of a suitable material, because ordinary glass does not let these rays pass through.

A property that is of importance in luminescent substances and therefore desired for all the applications mentioned above is stability of the intensity of the radiation, also when the substance is exposed to the discharge for any length of time. This requirement is never fully satisfied, for after a time the emission always drops a little. The cause of this reduction in fluorescence is not quite clear. Probably it is an intricate effect in which both photochemical decomposition and absorption of mercury play a part. Of the substances used tungstate is the most stable, whilst cadmium borate is the least stable. The other substances can be obtained in a fairly stable form by careful preparation.

*Discharge in mercury vapour of high pressure*

The high-pressure mercury discharge tube emits short-wave ultra-voilet ($\lambda$ = 3200 Å) as well as long-wave ultra-violet ($\lambda$ = 3650 Å) in addition to the rather large quantity of visible light (lines at 4047, 4358, 5461 and 5780 Å). With this irradiation the tube can be used directly as a source of light; but the colour of the light is of an unpleasant greenish tint, due to the very uneven distribution of the intensity over the spectrum, with light deficiencies particularly in the red and blue. As regards the red this deficiency has been successfully compensated by using the mercury discharge lamp in combination with an incandescent lamp, which has an excess of red [3]). This solution of the problem has the drawback, however, that all the ultra-violet emission is lost. A more efficient solution is to use fluorescent substances which convert the ultra-violet into the desired red and blue, which means to say that substances have to be used which can be excited by a radiation of a wavelength $\lambda < 3700$ Å; moreover, the fluorescence of these substances must not be affected by a rise in temperature.

If the fluorescent substance were to be applied direct on or in the discharge tube the rise in temperature would amount to about 400 °C. Since at such high temperatures most of the luminescent substances lose their power of fluorescence, it is desired to limit the elevation of temperature as far as possible. This can be achieved well enough by applying the fluorescent substance on the bulb enveloping the quartz discharge tube. Even so, in the case of a bulb 10 cm in diameter the temperature rise is still 100 - 150 °C, for a power of about 200 watts.

*Table II* gives a list of some red-fluorescing substances which can be excited with long-wave ultra-violet. This table also shows the temperature at which the intensity of the fluorescence has dropped to 80% of the maximum. In *fig. 4* graphs are given for some of these substances showing how the intensity of their fluorescence is governed by the temperature.

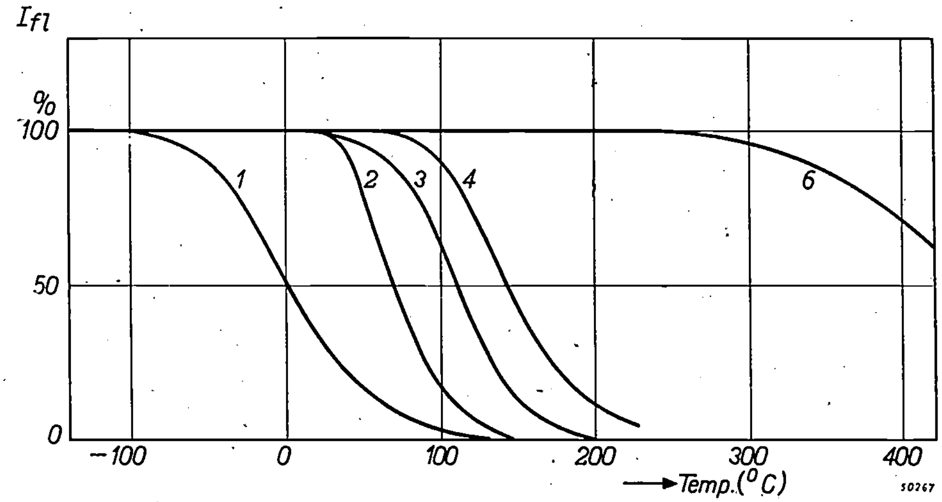


Fig. 4. Graphic representation of the manner in which the intensity of fluorescence in some luminescent substances varies with the temperature. The numbers beside the curves refer to the substances in table II.

Table II. Red fluorescing substances that can be activated with long-wave ultra-violet.

| | Fluorescent substances | Temp. at which $I_{fl}$ = 80% $I_{\max}$ |
|---|---|---|
| 1 | Cd (W,U) $O_4$ | —35 °C |
| 2 | (Ca, Mg) O.6 $Al_2O_3$-$Mn^{4+}$ | 45 |
| 3 | $SrAl_2O_4$-$Mn^{4+}$ | 84 |
| 4 | $Mg_2TiO_4$-$Mn^{4+}$ | 112 |
| 5 | (Zn, Cd) S-Cu | 200 |
| 6 | $(Al, Cr)_2 O_3$ | 372 |

The first three of the substances listed in table II cannot be considered because their fluorescence is quenched at too low a temperature.

Of the others only (Zn,Cd)S-Cu has a sufficiently high maximum yield, and that is why this substance is used in practice. But then there is the

[3]) Philips Techn. R., 5, 353-359, 1940.

drawback that this yellow substance absorbs blue and violet, so that the light from the lamp gets a greenish tint.

*Discharge lamps with rare gas*

For illuminated advertisements discharge lamps are often used which are filled with a rare gas. When neon is employed the discharge itself gives a red light. The ultra-violet radiation of short wavelength contained in the emission spectrum of rare gases can be transformed into visible light with the aid of luminescent substances, the most important of which for this purpose are $Zn_2SiO_4$-Mn, $(Zn,Be)_2SiO_4$-Mn and $MgWO_4$.

**Detection of infra-red**

In the beginning of this article we mentioned Stokes' rule, which says that luminescent substances convert the absorbed electromagnetic energy into a radiation of longer wavelength. It therefore seems strange that these substances can be used for detecting infra-red rays, but that is indeed so.

The use of luminescent substances for detecting infra-red rays became of particular importance during the last war, though it had been known a long time already that some of them, possessing the property of being able to store a large quantity of energy, can be influenced by infra-red rays. The presence of infra-red radiation can therefore be detected with the aid of such substances. This can occur in two entirely different ways, because irradiation of an activated system with infrared may have two opposite effects:

1) extinction of the phosphorescent light, and
2) accelerated emission of the stored energy (increased phosphorescence).

Several systems show the two effects to a different degree. Sometimes a system reacts in only one of these two ways, while in other cases both effects occur. Either of these effects can be used equally well for the detection of infra-red, and as a matter of fact substances have been prepared for both of them.

Any luminescent substances that can be considered suitable for this purpose must be capable of storing as large a quantity of energy as possible.

Substances used for the extinction effect must have a weak but persistent phosphorescence. As such ZnS-Cu-Mn has been used. When infra-red rays strike this substance a short flash of fluorescence is observed, but after that the phosphorence is completely quenched.

Substances used for the increased phosphorescence effect must have the least possible spontaneous phosphorescence, as otherwise the infra-red effect would be too little noticeable and, moreover, the store of energy would be soon exhausted.

In America and Germany, about the same time, substances were developed which after having once been irradiated retain the stored energy for months on end. These are Sr(S,Se)-(Sm,Eu) and Sr(S,Se)-(Sm,Ce). Also ZnS-Cu-Pb has favourable properties in this respect. To give an idea of the activity of these substances it may be noted that 0.1 mm$^3$ of such a substance viewed through a lens with a focal distance of 2.5 cm can produce for one hour a light of an intensity equal to 100 times the threshold value of the eye.

At the risk of superfluity we would once more point out that in any case the infra-red radiation only has its effect when the system has previously been irradiated, either with short-wave rays or with corpuscular rays. In case 2) it is not, therefore, a matter of visible light being excited by infra-red radiation. Thus the phenomena are by no means contradictory to Stokes' rule.

In conclusion we would mention in a few words other systems of infra-red detectors where use is likewise made of fluorescent substances [4]). Here the sensitive layer is a substance from which electrons are released by exposure to light. After having been accelerated these released electrons are concentrated on a fluorescent screen and made visible that way. For these systems the same luminescent substances can be used as described above for oscillograph tubes.

[4]) G. Holst, J. H. de Boer, M. C. Teves and C. F. Veenemans, Physica 1, 297-305, 1934.

# 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 Rescarch Laboratory, Kastanjelaan, Eindhoven, Netherlands. Those papers of which no reprints are available in sufficient number, are marked with an asterisk.

**1729:** J. L. Snoek: New developments in ferromagnetic materials; Elsevier publishing Company, Inc. New York, Amsterdam 1947, 136 pages, 52 fig.

This book is not a textbook on the subject but it gives a collection of publications on research work, carried out in 1940-1945. It is divided into three chapters dealing with the "statics" and "dynamics" of ferromagnetism and the development of magnetic materials respectively. The first chapter includes such topics as the general theory of hysteresis, hysteresis at low values of the induction, crystal anisotropy and magnetostriction in ternary systems, permeability and coercive force of cubic ferromagnetic oxides, effect of cold rolling on the alloys of Ni and Fe. The second chapter chiefly deals with the magnetic after effect and dis-accomodation in alpha iron as well as in ferromagnetic non-metals. Other topics referred to are: eddy current problems, magnetic skin-effect, large Barkhausen discontinuities. The third chapter gives extensive information on new magnetic ferrites developed by the author and co-workers ("ferroxcube" 1,2,3,4) and in addition gives some information on magnet steels.

**1730:** E. J. W. Verwey: Theory of the stability of lyophobic colloids (J. phys. and colloid Chem. **51**, 631-636, 1947.

The theory of the interaction of the double layers, surrounding colloidal particles in suspension in the presence of an electrolyte, has been reviewed in relation to the stability of lyophobic colloids. It is concluded that the interaction of the double layers must be associated with an increase of free energy, leading to a repulsion between the particles. The calculated repulsive potential for special cases has been combined with the van der Waals-London attractive potential calculated by Hamaker, to obtain curves of potential vs. distance. Predictions based on these curves appear to agree well with various experimental data. For example, the influence of electrolytic concentration and of the valencies of the ions on flocculation is satisfactorily explained in terms of the theory, although many complicated phenomena remain to be correlated with it.

**1731:** Th. P. J. Botden and F. A. Kröger Fluorescence of cadmium borates activated by manganese (Physica **13**, 216-224, 1947).

In the system CdO. $B_2O_3$ four compounds exist, each of which show fluorescence when activated by manganese. $Cd_2B_2O_5$-Mn is excited by $\lambda < 3200$ Å and cathode rays and than shows an orange luminescence, whereas $CdB_2O_4$, $Cd_2B_6O_{11}$ and $Cd_3B_2O_6$ show a green cathodo-luminescence only.

The luminescence of the $Cd_2B_2O_5$-$Mn_2B_2O_5$ phase has been studied quantitatively both as a function of the manganese content and as a function of the exciting radiation and temperature. It is shown that an ultra-violet emission appearing at low temperatures in some of the products cannot be responsible for the anomalous decrease of the red luminescence towards low temperatures.

**1732:** F. A. Kröger: Fluorescence of tungstates and molybdates (Nature **159**, 674, 1947).

A number of molybdates and tungstates which are non-luminescent at ordinary temperatures prove to show luminescense at low temperature. The temperature-quenching of the luminescence of a number of tungstates and molybdates in the interval —200 °C to + 200 C° is given in a graph.

**1733:** F. A. Kröger: Tetravalent manganese as an activator in luminescence (Nature **159**, 705, 1947).

In a number of systems, such as zinc aluminate, magnesium aluminate, $\alpha$- and $\beta$-aluminium oxide and magnesium titanate, all activated with manganese, the luminescence is markedly different according as the products are prepared under oxidizing or reducing conditions. The reduced products (except the titanate) show a green cathodo-luminescence, which is due to divalent manganese. The oxidized products show a deep red luminescence upon excitation by cathode rays or ultra-violet (3650 Å). The valency of Mn in magnesium titanate was found by titration to be between 3 and 4. Analogy between the emission spectra of $Mg_2TiO_4$-$Cr^{3+}$ and $Mg_2TiO_4$-Mn shows that Mn is present in the form $Mn^{4+}$.

**1734:** E. J. W. Verwey: Nieuwe onderzoekingen over de atoomrangschikking in spinellen in verband met hun physische eigenschappen (New investigations on the atomic arrangement in spinels in relation to their physical properties) (Chem. Weekblad **43**, 229-232, 1947).

For the contents of this paper the reader is referred to Philips Techn. Rev. **9**, 186-191 and 239-248, 1947.

**1735:** A. van der Ziel and A. Versnel: Total emission noise in diodes, Nature **159**, 640, 1947.

Measurements of the input damping $1/R$ and the equivalent saturated diode current $I_e$ were carried out at 7.25 m wavelengths on a diode (cathode area 10 cm$^2$, anode-cathode distance 0.1 cm). Both quantities are approximately proportional to $V_a^{-2}$. The total emission may be described by assuming the "equivalent noise temperature" $T_e$ of the conductance $1/R$ to be equal to the cathode temperature. It is expected that this result will hold for a very wide frequency range, because $1/R$ and $I_e$ are both proportional to the square of the frequency.

**1736:** A. van der Ziel: Total emission damping in diodes, Nature **159**, 675, 1947.

Measurements of the "total emission damping", which is a transit-time effect due to space charge and occuring at ultra high frequencies,were carried out at 5.8 m wavelength on a diode (cathode area 10 cm$^2$, cathode-anode distance 0.1 cm). The conductive part $1/R$ and the reactive part $\omega\Delta C$ are plotted against $V_a$ (heater voltage 6, 8, 10 and 12 V.) It is proved that $\omega\Delta C$ and $1/R$ are proportional to $V_a^{-1}$ and $V_a^{-2}$ respectively. It is to be expected that $\Delta C$ will scarcely depend on the frequency and that $1/R$ will be proportional to $\omega^2$, except for the highest frequencies.

**1737:** J. L. Snoek: Gyromagnetic resonance in ferrites, Nature **160**, 90, 1947.

In polycrystalline magnetic ferrites of great homogeneity prepared by the author the tangent of the loss angle between $B$ and $H$ is found to rise at about 10 Mc/sec. from values less than 0.01 to values exceeding 1. At the same time the permeability goes down to very low values. The results are compared with the Landau and Lifshitz theory of gyromagnetic resonance expected to occur in a ferromagnetic dielectric if the frequency of the applied field equals the precession frequency of the spins around the direction of the internal field. (In Landau's model the specimen is assumed to be a single crystal and the applied field is at right angles to the crystal field). According to Landau

$$\frac{\chi_t \cdot f_h}{I_{\max}} = \frac{\sqrt{2}}{2\pi}\,\frac{e}{mc} = 4 \times 10^6 \text{ cycles/Oe.sec.}$$

($\chi_t$ = transverse susceptibility, $f_h$ = frequency for which the real part of the susceptibility is halved, $I_{\max}$ = saturation magnetisation) whereas for the polycrystalline sample this expression equals 1.75 $10^6$ cycles/Oe.sec.

The approximate agreement warrants the conclusion that the rapid decrease of the permeability in ferrites at high frequencies is probably due to gyromagnetic resonance around directions prescribed by the internal field.

**1738:** E. J. W. Verwey and E. L. Heilmann: Physical properties and cation arrangement of oxides with spinel structures. I. Cation arrangement in spinels, J. Chem. Phys. **15**, 174-180, 1947.

For the contents of this paper see the article by E. J. W. Verwey, P. W. Haayman and E. L. Heilmann, Philips Techn. Rev. **9**, 186-191, 1947 (No. 6).

**1739:** E. J. W. Verwey, P. W. Haayman and F. C. Romeyn: Physical properties and cation arrangement of oxides with spinel structure. II. Electronic conductivity, J. Chem. Phys. **15**, 181-187, 1947.

For the contents of this paper see the article by E. J. W. Verwey, P. W. Haayman and F. C. Romeyn, Philips Techn. Rev. **9**, 239, 1947 (No. 8).

**1740:** A. H. W. Aten: Activation of hafnium with neutrons, Science **105**, 386, 1947.

With slow neutrons acting on Hf a period of 20 sec. is found; with fast neutrons periods of 20 sec., 10 min. and 6 hours. As the stable Hf isotopes are 174, 176, 177, 178, 179 and 180 the second period is supposed to be due to Hf 175. The results are compared with those of Flammersfeld. The periods found are useful for the determination of Hf as a contamination of Zr whereas the long periods of Zr may be used for the determination of Zr as a contamination of Hf.

**1741:** J. A. Keverling Buisman, W. Stevens and J. van der Vliet: Investigations on Sterols. I. A new synthesis of 7-dehydrocholesterol (provitamin D), Rec. Trav. Chim. Pays Bas **66**, 83-92. 1947,

A new synthesis of 7-dehydro-cholesterol (provitamin $D_3$) from cholesterol is described, which gives a higher yield than the well-known Windaus synthesis. The synthesis is carried out via a 7-bromo cholesteryl ester, from which hydrogon bromide is eliminated to give an ester of 7-dehydro cholesterol.

**1742:** A. H. W. Aten: High energy ions in crystal lattices, Phys. Rev. **71**, 641-642, 1947.

The valency is determined of $P^{32}$ ions formed by (n,p) and (n,$\alpha$) reactions due to recoil in crystals containing S and Cl respectively. It is an open question whether the phosphate and phosphite ions are formed inside the lattice or whether they adopt their final state at the moment the crystal is dissolved.

**1743:** W. P. van den Blink: Note on the influence of the water content of an electrode-coating on the hydrogen content of weld metal, Welding J. Res. Supp. July 1947.

The question is discussed whether an increase in the moisture content of the coating of a welding electrode increases the hydrogen content of the weld metal. It is shown that both an increase and a lowering of the hydrogen content may result, dependent on the partical pressures of $H_2O$, CO, $H_2$ and $CO_2$ in the welding arc.

**1744:** K. W. de Langen: The manifold of physical quantities, Physica **13**, 349-352,1947.

In electrodynamics all operations are carried out in a manifold of quantities each element of which may be written in the form

$$a \text{ cm}^k \text{ g}^l \text{ sec}^m \text{el}^n$$

(el = electrostatic unit of charge). The numbers $k$, $l$, $m$, $n$ determine the class of the quantity (dimension). The author advocates the use of quantity equations ("Grössengleichungen").

**1745:** J. H. F. Custers: The texture of copper wire drawn with backpull: Physica **13**, 366-378, 1947.

The preferred orientations of copper wire drawn in the normal way and with backpull respectively have been investigated. In the outer zones these orientations show distinct differences. In wire drawn with backpull the so-called conical fibre texture which predominates especially in the outer zones has developed less intensively than in wire drawn in the normal way. It is, however, still an open question whether properties like tensile strength are inproved in such a degree as to lead one to expect special advantages by the application of backpull.

**1746:** J. F. H. Custers: Note on the measurement of specular and diffuse photographic density, Photogr. J. Sect. B **87**, 59-63, 1947.

In measuring the density of photographic films, the smallest value or so-called diffuse density is measured when the measuring instrument catches the whole transmitted flux and the largest value or so-called specular density when only the flux transmitted normally to the layer is recorded.

The author has made some measurements in order to test a method recommended by Pitt (1938) for separating specular and diffuse density. He arrives at the conclusion that Pitts method can be used only for films or layers which scatter light according to Lambert's cosine law whereas it is of little value for all ordinary photographic films or plates because these show a scattering mechanism which obeys quite different laws.

**1747:** P. M. van Alphen and C. J. Dippel: New technical possibilities in the microreproduction and multiplication of documents. Rapports de la 17me conférence de la F.I.D. Berne, 25-30 août 1947.

An analysis is made of the possibilities and limitations in connection with the resolving power and optical conditions for recording a micro-picture and making it legible.

The contents of this paper cover partly the articles by C. J. Dippel and coworkers in Philips Techn. Rev. 9, 65-72, 1947 (No. 3) and Philips Techn. Rev. 9, 289, 1947 (No. 10).

VOL. 9 No. 8, pp. 225-256 1947/1948

# 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

## FIFTY YEARS OF ELECTRONS

by W. de GROOT.

537.122:537.533

In the publication "On Cathode Rays" by J. J. Thomson (Oct. 1879) the existence of free electrons was first irrefutably established. This important article, which appeared just 50 years ago, is discussed in detail. A short account is then given of the investigations which preceded Thomson's publication and of the results of the further theoretical study of electrons. In conclusion a survey is given of a number of technical applications in the field of "electronics".

### Introduction

We can assign no more suitable date to the discovery of the electron than the date of publication, October 1897, of J. J. Thomson's paper "On Cathode Rays" [1]). The object of Thomson's investigations was to acquire more knowledge about the nature of cathode rays, about which at that time the most divergent opinions were held.

Some people considered these rays to be charged "particles", others denied their corpuscular nature and were of the opinion that it was mainly a question of a phenomenon in the "ether".

Thomson remarked that an explanation which starts from the corpuscular nature as a working hypothesis is more likely to be successful and can more easily be tested by known laws (namely those of mechanics) than an explanation on the basis of an hypothesis about the ether, because the properties of the "ether" itself are too little known to base conclusions upon them.

The experiments described were thus devised "to test some of the consequences of the electrified-particle theory".

### Thomson's deflection experiments

The objects of Thomson's experiments were:

a) to verify that the cathode rays carry a charge and that this charge accompanies the rays when they are deflected by a magnetic field;

b) to investigate quantitatively the qualitatively known deflection in an electric field, which also indicates the presence of a charge;

c) to measure the energy of the rays and, by a combination of this measurement with that of the magnetic deflection, to determine the velocity and the ratio of charge to mass;

d) to determine the same quantities by a combination of the magnetic and electric deflections.

In all these experiments the cathode rays were obtained by a discharge in a tube filled with a rarefied gas, in which electrodes were fused or cemented in. An induction coil or a galvanic battery served as source of voltage.

The experiment *sub* a), an improved version of an experiment by Perrin [2]), was performed with the tube of *fig. 1.* Cathode rays leave the cathode *K* and through an opening in the anode *A* reach the space *R*, where they are deflected by a magnetic field and, through an opening in an earthed cylinder *B*, reach a second conductor *D* mounted inside that cylinder. An increase in the charge, registered by an electrometer connected with *D*, is only shown when the rays strike the opening. This proves that a charge is indivisibly connected with the cathode rays. The charge is a negative one.

[1]) J. J. Thomson, On Cathode Rays, Phil. Mag. 44, 293-316, 1897 (October, article dated Aug. 7, 1897). Some of the experiments described in this article had already been published elsewhere by Thomson (Proc. Cambr. 9, 1897) and demonstrated (Royal Institution, Friday evening lecture Apr. 30, 1897, see Electrician, May 21, 1897).

[2]) In Perrin's experiment a Faraday cage was placed directly opposite the cathode, when it was found that no charge was conducted to the cage when the cathode rays were deflected by a magnet.

The magnetic deflection was further investigated quantitatively by means of the tube of *fig. 2*, which is nothing else but an air-pump bell jar fitted with electrodes. It is found that the rays are not all equally deflected (magnetic spectrum) and that the components making up the largest part of the beam, causing a strong fluorescence of the glass wall, are not the same as those which cause the strongest luminescence of the gas. The most striking characteristic observed in this tube, however, was that, with a given voltage between anode and cathode, the appearance of the deflected beam is always the same whatever the nature of the gas filling.

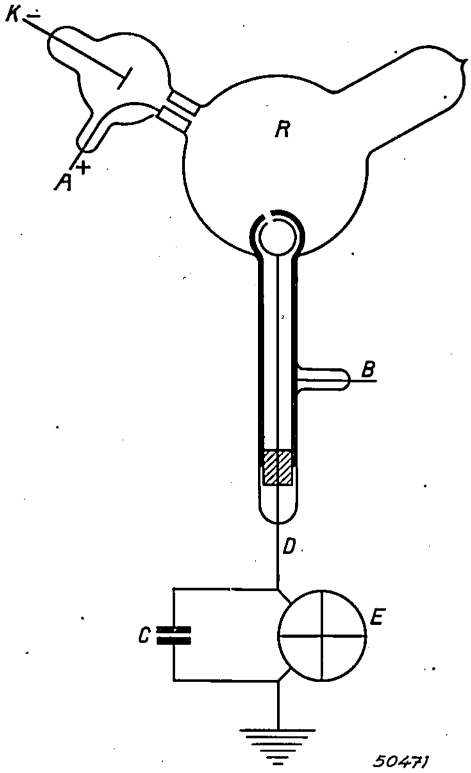


Fig. 1. Discharge tube according to J. J. Thomson, with which the (negative) charge of magnetically deflected cathode rays was demonstrated. *K* cathode, *A* anode, *B* earthed cylinder, *D* electrode for capturing the cathode rays, *E* electrometer, *C* condenser. The magnetic field is excited in the space *R* by wire coils placed outside the tube.

The deflection *sub* b) by an electric field was investigated in the tube of *fig. 3*, the precursor of our modern cathode-ray oscillograph.

Experiment c) was performed in a similar tube but without deflection plates and provided with a screened electrode (as in fig. 1). The innermost insulated electrode contained a thermo-element which was struck by the cathode rays and whose increase in temperature during a given time was measured. At the same time, with the aid of a quadrant electrometer, the charge taken up by the element was measured.

If, in the time $t$, $N$ particles strike the thermo-element, each bearing a charge $e$, the total charge is

$$Q = Ne \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad (1)$$

From the temperature increase the total energy $W$ is also known:

$$W = N \cdot \frac{1}{2} mv^2 \, , \quad . \; . \; . \; . \; . \quad (2)$$

where $m$ represents the mass and $v$ the volocity of the particle.

Finally the radius $r$ was measured of the orbit described by the particles in a field $H$ [3]).

$$\frac{mv^2}{r} = evH \quad . \; . \; . \; . \; . \; . \; . \; . \quad (3)$$

When $rH = I$ it follows from (1), (2) and (3) that

$$v = \frac{mv^2}{erH} = \frac{Nmv^2}{NerH} = \frac{2W}{QI} \quad \text{and} \quad \frac{m}{e} = \frac{rH}{v} = \frac{QI^2}{2W} \, .$$

In this way the following results were obtained in tubes with different gas fillings.

| | air | $H_2$ | $CO_2$ |
|---|---|---|---|
| $m/e$ | 0.44 | 0.47 | $0.45 \cdot 10^{-7}$ g/emu |
| $v$ | 0.28 ex $1.2 \cdot 10^{10}$ cm/sec | | |

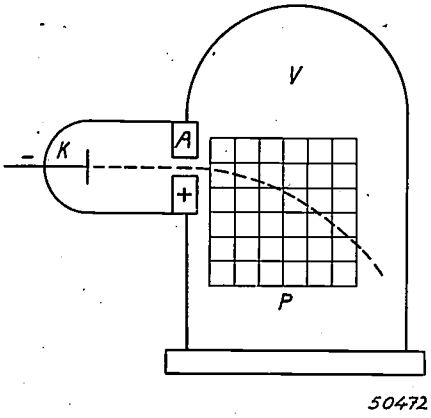


Fig. 2. Discharge vessel in which the magnetic deflection was investigated by Thomson. The rays from the cathode *K* pass through an opening in the anode *A* into the bell jar *V*. The curvature in a magnetic field (excited by coils of wire placed outside the vessel) is measured by causing the rays to pass along a glass plate *P*, upon which there is a network of lines which is photographed together with the cathode rays.

The deflection experiments *sub* d) are based on the measurement of the changes in direction taken by the ray in an electric or magnetic field. If $F$ is

[3]) It would be better to write $evB$. In electromagnetic units as used by Thomson, however, $B = H$.

the electrical field strength the lateral acceleration is $Fe/m$ and the lateral velocity after travelling a distance $l$ between the plates is $(Fe/m)\ l/v$, so that the following holds for the angle of deflection $\vartheta$:

$$\vartheta = \frac{Fel}{mv^2}.$$

The deflection in the case of a magnetic field is $\varphi = l'/r$, where $l'$ represents the distance travelled in the field, so that

$$\varphi = \frac{He\ l'}{mv}.$$

The magnetic field was generated by two coils whose diameter $l'$ was equal to the length $l$ of the plates in the electric deflection experiment. Therefore $l'$ may be taken equal to $l$ and one then finds that

$$v = \frac{\varphi}{\vartheta}\frac{F}{H}$$

and

$$\frac{m}{e} = \frac{\vartheta}{\varphi^2}\frac{H^2}{F}\ l.$$

These experiments gave the following results:

| | air | $H_2$ | $CO_2$ |
|---|---|---|---|
| $m/e$ | 0.12 | 0.15 | $0.15 \cdot 10^{-7}$ g/emu |
| $v$ | 0.22 ex $0.36 \cdot 10^{10}$ cm/sec | | |

The differences in the values of $m/e$ obtained by the two methods must be ascribed to systematic errors. The agreement between the values obtained by the same method with different gases indicates that the cathode-ray particles are independent of the nature of the gas. The value of $m/e$ at present accepted as the correct one is $0.5658 \times 10^{-7}$ g/e.m.u. ($e/m = 1.76 \times 10^7$ e.m.u./g $= 1.76 \times 10^8$ coul./g.). Thomson was at first at a loss to interpret the value found for $m/e$. For certain reasons he was of the opinion that the value of $e$ for the "corpuscles" was many times larger than the charge of a monovalent electrolytic ion. He realized that the corpuscles formed an important component of matter and assumed that each of the atoms contained a large number of these components, the larger the higher the atomic weight. The phenomenon of electric polarization giving rise to a dielectric constant (specific inductive capacity) differing from unity was considered by Thomson to be connected with the corpuscles.

Thomson's publication ends with a discussion of the construction of the atom and describes a model of an atom built up from the results of an experiment with floating magnetic needles [4]). By means of Thomson's model the first attempt was made to explain the periodic system.

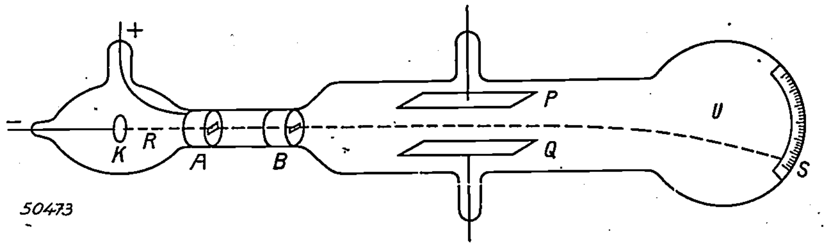


Fig. 3. Thomson's tube for the study of the electrical deflection. The cathode rays formed in the space $R$ pass through the openings in the electrodes $A$ and $B$. Thus a narrow beam reaches the space $M$ containing a pair of plates $PQ$. $Q$ is given a positive and $P$ a negative potential with respect to $B$. The ray is deflected in the electric field between $P$ and $Q$ and the deflection is observed by the displacement of the spot of fluorescence on the glass wall at $S$. $S$ is a paper scale on which the displacement is read off.

## Previous history

While conceding all honour due to Thomson, we must not forget his predecessors. In 1820 already Faraday studied discharges in rarefied gases. He distinguished two light phenomena: the negative glow at the cathode and the positive column at the anode, separated by the Faraday dark space. Plücker (1859) and Hittorf (1869) studied the phenomena at the cathode at lower gas pressures. Hittorf described cathode rays and their deflection by a magnetic field, while Goldstein (1876) observed the deflection in an electric field. To

[4]) Thomson mentions as author a certain Professor Mayer. The experiment consists in floating a number of short rod magnets held in corks, for instance with the south pole uppermost. They repel each other but are attracted by a strong magnetic north pole placed above the surface of the liquid. Certain configurations are formed according to the number of magnets.

Hittorf's mind the current flowed from the positive to the negative pole. In his opinion the phenomena at the cathode were the final phase of the process of the discharge. It was too much at that time to think of reversing the order and ascribing the phenomena of discharge to negative particles leaving the cathode.

In 1897 Crookes repeated Hittorf's experiments with an impıoved technique, especially as regards the evacuation. His lecture entitled "Radiating Matter, or the Fourth State of Aggregation" did much to popularize cathode rays, which by then had definitely come to be recognised as having their origin in the cathode.

In 1883 Hertz declared himself to be an adherent of the corpuscular hypothesis and expressed the desirability of combining the magnetic and electric deviations and thus determining the velocity of the cathode-ray particles, while he had already indicated the possibility of determining the charge electrically or magnetically.

In 1887 he discovered the photo-electric effect, the releasing of a negative charge from a metal conductor by irradiation with ultra-violet light, a phenomenon which was further investigated by Hallwachs (1888). In 1892 Hertz found that thin layers of metals allow the passage of cathode rays. Encouraged by Hertz, in 1894 Lenard succeeded in making cathode rays leave the discharge tube through a window and studying them at will in a very high vacuum. In this way he studied the magnetic deflection. In 1895 he investigated the absorption and scattering in thin metal foils. He found those properties to increase with increasing deflection (decreasing velocity) and, where the rays are of constant velocity, to be proportional to the mass passed through, regardless of the nature of the medium. In the same year, 1895, Röntgen discovered X-rays (or Röntgen rays as they are often called, after their discoverer), which are formed when fast electrons strike a wall or a metal conductor.

Simultaneously with the development sketched above, the idea grew that an electric charge in atomic distribution also occurs *in* matter. Already around 1870 Weber had put forward the hypothesis that conduction in metals takes place by means of discrete charges which move from atom to atom and are also responsible for the "circulating currents of Ampère", thus for magnetism.

In 1873 Maxwell reluctantly gave as his opinion that in order to explain the phenomena of electrolysis it might be necessary to assume an atomic distribution of electricity, and in 1874 Stoney made an estimation of this elementary charge ($3 \times 10^{-11}$ esu) on the basis of the data then available [5]). In 1891 Stoney called this charge the electron. Originally the term meant the charge itself and not a definite particle bearing this charge. In this sense the atom contained positive as well as negative "electrons".

Later on the name electron came to be used for the "corpuscles" discovered by Thomson in the cathode rays, which we now call "free electrons" in contrast to the "bound electrons" occurring in atoms.

The hypothesis that the charges in the atom can vibrate about an equilibrium and thus govern the optical properties of matter was proposed by Lorentz (1875), among others, and worked out by him into a theory (1892).

In 1896 Zeeman discovered the magnetic resolution of the spectral lines. Together with Lorentz he decided, from the sign of the circular polarization exhibited by the lines when viewed along the lines of force, that the particles in the atoms responsible for the emission of light carry a negative charge. From the magnitude of the splitting, morever, he was able to make a rough estimation of the quotient $e/m$, which, expressed in emu/g, was found to have the value $10^7$, the same value which Thomson shortly afterwards found for the cathode-ray particles [6]). Though Zeeman himself did not at that time venture to draw conclusions from his results, it may be claimed that it was this discovery that lent strength to the idea of "bound" electrons in the atom.

## Further development

We have seen that after the quotient $m/e$ of the cathode rays had been determined Thomson had some doubts about the charge $e$ and at first thought it to be much larger than the charge of a monovalent ion. Since 1896 Thomson and Rutherford had been studying the ionization of gases by X-rays. C. T. R. Wilson had found that water vapour can condense on the gaseous ions, a phenomenon which he subsequently utilised to show the paths of ionizing particles in a gas (Wilson Chamber, 1912). From an estimation of the number of water droplets, combined with a determination of the total charge, Thomson (1898) was able to determine the charge of one droplet, which he considered equal to the ion charge. He found it to be of the same order as the charge of electrolytic ions. A

[5]) The value of $e$ at present accepted as correct is $4.80 \times 10^{-10}$ esu = $1.6 \times 10^{-19}$ coulomb.

[6]) Thomson points this out in his lecture (see footnote [1]).

similar result had shortly before been found by Townsend for the charged water droplets foimed above the surface of an electrolyte when gas bubbles escape from it upon current being passed through. Wilson (1903) improved upon Thomson's method by studying the effect of a vertical electric field on the rate of settling.

Since in the process of ionization of X-rays a molecule is not divided into two ions, as in electrolytic dissociation, but the atom itself is split into charged components (monatomic gases like argon are also ionized by X-rays), it was concluded that the charge found in the gaseous ions, being of the same order as the ionic charge in electrolysis,was the charge $e$ of the electron.

Wilson's method was later (1909) worked out in Millikan's laboratory to a precision method for the determination of $e$. The accuracy of this method is even now scarcely surpassed by the indirect determination of the electrolytic unit charge, which, as will be known, is derived from the electrochemical equivalent in connection with the determination of Avogadro's number from X-ray interferences in crystals.

From the values of $e$ ($1.6 \times 10^{-19}$ coulomb) and $e/m$ ($1.76 \times 10^{8}$ coul/g) it follows that the mass of the electron is $9.1 \times 10^{-28}$ gram, *i.e.* 1/1873 of the mass of a hydrogen atom.

The discovery of bound electrons led gradually to certain conceptions concerning the structure of the atom. As we have seen, Stoney already assumed that positive and negative charges are present in the atom.

Lenard (1903) imagined these charges as occurring in pairs (dynamids), while Thomson, in view of his experiments with the model described above, assumed that the atom consisted of a sphere homogeneously filled with positive charge, within which, under the influence of the positive attraction and the mutual repulsion, the electrons take up certain equilibrium positions.

To Rutherford (1911) is due the conception that the positive charges in the atom are collected in a positive nucleus, which also carries the larger part of the atomic mass [7]).

Rutherford, who had been studying radioactive phenomena since 1897, was led to this conclusion by experiments on the scattering of alpha particles by thin metal foils. Rutherford's conception of a positive nucleus surrounded by a swarm of negative electrons formed the basis of Bohr's (1913) model of the atom and his ideas about the structure of the periodic system, which ideas were supported especially by Moseley's experiments (1912) on the X-ray spectra of atoms.

In 1924 De Broglie predicted theoretically the wave nature of the electron, which was experimentally confirmed in 1927 by the experiments of Davisson, Germer and G. P. Thomson.

In 1925 Uhlenbeck and Goudsmit explained a number of unaccountable phenomena in the spectra by ascribing to the electron the properties of a top. This so-called spin of the electron is in fact the chief source of magnetism.

Finally, in 1932, Anderson discovered the positive electron in cosmic rays, possessing the same ratio $e/m$ but having a positive charge.

Soon afterwards it was found that also many artificial radio-active isotopes emit positive electrons, forming the counterpart of the negative electronic rays (beta rays) familiar since the first days of radio-activity. The existence and properties of the positive electron were more or less foretold by Dirac. It has the tendency to combine with a negative electron under emission of radiation (annihilation). Conversely, from a quantum of radiation of sufficiently high energy (> 1 Me V) an electron pair (+ and —) can be created.

## The rôle of free electrons in technology

It must not be considered as mere chance that the electron was discovered at a time when important progress was being made in vacuum technique owing to the development of the incandescent electric lamp.

Edison had already discovered that when a third electrode is introduced into the bulb of a carbon filament lamp and connected with the positive end of the filament a negative current passes through the vacuum to that electrode. This thermionic emission of incandescent bodies was carefully investigated by Richardson (1901). The negative particles emitted by a filament are nothing else but free electrons. In this way they can move in an evacuated space without any "gas discharge" being present.

Richardson's experiments led to the invention of the diode as a rectifier and detector of electrical oscillations (Fleming 1904). By the addition of a third electrode or grid, the triode was formed (Lee de Forest 1909), which is the ancestor of radio receiving and transmitting valves in all their various forms. Besides the possibility of generating and detecting electrical waves the radio valve is also of importance as an amplifier of weak A.C. voltages (radio receiving and transmitting installations, line telephony, sound amplification, physical methods of measuring).

[7]) See for example W. de Groot, Nuclear Physics, Philips Techn. Rev. 2, 97-102, 1937.

From the diode came the rectifier valves with their various applications (charging of batteries, welding technique, electrolysis).

An important factor in the development of all these applications is the discovery of the electron emission of metal oxides (Wehnelt 1903).

The release of electrons from metals by irradiation with light led to the construction of photocells. In modern photocells use is made of the amplification by secondary electron emission, a phenomenon which is also used in radio valves and which consists in the fact that a surface struck by electrons itself emits electrons.

The cathode-ray tube originated directly from Thomson's experiments; it is used in the form of the cathode-ray oscillograph in technology and in the laboratory to visualize and measure alternating currents and all kinds of phenomena of short duration. From it in turn are derived the applications in radar and television on the one hand, and on the other in the electron microscope.

When a very high voltage is applied between cathode and anode of a discharge tube part of the energy with which the electrons strike the anode is converted into X-rays. The vacuum X-ray tube with heated cathode (Coolidge 1913) has been developed into a modern apparatus in which voltages of 25 to 2000 kV are employed. These X-ray tubes are used extensively in medical diagnostics and therapy and in the testing of materials.

Finally the study of discharges in rarefied gases, together with the application of the oxide cathode, has led to a number of new light sources (sodium and mercury lamps, fluorescent lamps), which are employed not only for illumination but also for other purposes (ultra-violet irradiation, the analytical lamp).

Thus from the discovery of the electron in the physical laboratory a large number of industrial applications have followed which have been useful in many ways, and which in turn have had a stimulating effect on scientific research. It is not without pride and satisfaction that physicists and technologists may look back upon the accomplishments of the last fifty years, for which so much is due to the work of J. J. Thomson.

# EMERGENCY SUPPLY SYSTEMS WITH ACCUMULATOR BATTERIES

by H. A. W. KLINKHAMER.

621.316.261

Emergency supply systems are for serving electric mains-fed plant, such as telephone or lighting installations, in the event of a breakdown in the mains supply. A very suitable system is one where the plant is fed from a rectifier with a battery of lead-cell accumulators connected to it in parallel. The rectifier has to be of such a construction as to ensure a constant output voltage practically independent of mains voltage fluctuations and variations of load. This is necessary to maintain a permanent charge of 2.1 / 2.2 V per cell in the battery, which modern experience proves to be the condition ensuring the longest life for a battery. A rectifier answering these requirements is therefore called a "preserving rectifier". This article explains the advantages of such an emergency supply system compared with a dynamo and the obsolete two-battery system (with two batteries feeding the plant alternately). Finally it is shown that a type of rectifier already described in this journal, with a highly saturated transformer, is quite suitable as a "preserving rectifier", two special applications of which for emergency supply systems are described.

Electricity mains are nowadays the most commonly used source of power for all sorts of plant and installations. The advantages are sufficiently well known. This source of power, however, has one drawback, or perhaps it is better to say that it has a drawback in common with almost all other sources of energy, in that it is liable to a breakdown. Though for many plants and, for instance, for domestic supplies this may not constitute any serious objection provided the interruptions are not too frequent or of too long a duration, there are cases where an interruption in the working of the plant for more than a few seconds — or even not that short — cannot be allowed. To give a striking example, under no conditions may the illumination of an operating table in a hospital fail while an operation is being performed, for that would jeopardize the life of the patient [1]). Less fatal but equally intolerable is the failure of the lighting in subways, roadtraffic tunnels, cinemas or halls, where a panic is likely to be caused or a situation may arise favourable for pick-pockets. Another important case is that of the telephone, for the slightest interruption in the power supply may have very serious consequences, resulting, for instance, in the mutilation of a telex message that is just being transmitted.

Where in such and similar cases it is desired to keep the plant going in the event of a breakdown in the mains one must have an emergency supply system immediately available. As such there are to be considered in the first place a dynamo driven by a combustion engine or else a battery of lead-cell accumulators [2]).

Here we will deal with supply systems employing accumulator batteries. These possess several obvious advantages, such as simplicity of the installations, noiseless working and the fact that they can be started up without any delay. To judge properly the value of the last mentioned advantage let us consider the steps taken in the case of some P.T.T. plants working with dynamos in order to avoid interruption in the event of a failure of the mains. To bridge over the starting time of the combustion engine, which may be a matter of 30 seconds or so, a compressed air engine is mounted on the dynamo shaft and starts running immediately the mains power fails. Still it would take a few seconds, however, before the dynamo gets up speed, and therefore, to avoid even such a short interruption, a small electro-motor is also set up which keeps the dynamo turning (without load) during normal working, whilst the inertia of a coupled flywheel helps to bridge over the starting period of the compressed air engine. Finally a compressor is needed to recharge the compressed air cylinder after an interruption in the mains supply. There is no denying that a battery system is much simpler.

Nevertheless, designers of power supply systems often show an aversion to the battery solution. From what follows it will be realized, we hope, that this aversion is unwarranted, provided one applies the developments of the last decade in working with batteries. The advantages of the emergency supply system with battery to be described here are such as to merit its application

[1]) This is so important that a safeguard has to be provided not only against interruptions of the mains supply but also against fusing of the incandescent lamp filament. Special lamps have been developed incorporating a reserve filament that comes into action automatically.

[2]) We will disregard here the nickel-iron storage batteries that are used in some cases.

on a much wider scale, both to replace dynamo units and to be installed in cases where hitherto one has had to manage without an emergency supply system on account of the complications involded.

## Hygienics of the lead-cell battery

The aversion to an accumulator battery is due to the idea that it necessarily involves heavy expenditure, a great deal of upkeep and considerable attention. As regards upkeep and attention, it is true that a battery cannot just be stored away in a corner somewhere, as a sort of can of preserved electrical energy, to be used only in case of need. Until fairly recently it was the common belief that, in order to keep a battery in tip-top condition and to make sure that it lasts as long as possible, it must be used; unless it is periodically discharged and fully recharged the plates deteriorate through sulphation ("hardening") and thus reduce the capacity of the battery.

For a number of years already this idea has been refuted by carefully checked practical experience over a great length of time,[3]) and it is now known that a lead-cell battery lasts longest when used in an entirely different way, being kept under what is called a preservation charge. Before proceeding to explain what this means we would point out that the method that used to be commonly practised with batteries for emergency supply systems was based on the obsolete idea mentioned above. The plant was connected to a battery instead of the mains, with a second battery being charged from the a.c. mains *via* a rectifier while the plant was being run off the first one, so that as soon as the first battery was discharged its function was taken over by the second one, and so on alternately. This was the so-called two-battery system, as illustrated diagrammatically in *fig. 1.* It is relatively expensive, because two batteries have to be maintained, while, in a manner of speaking, at any moment one has available as reserve only the power capacity of one battery. It calls for a great deal of maintenance and attention, for the acid strength and meter readings have to be repeatedly checked to see whether it is not time to change over the batteries. If this should be forgotten only once then accidents will happen.

The prejudice that arose against batteries as a medium for emergency supply is thus understandable.

Against the old battery treatment we now have a new method, that of keeping the battery charged continuously with a low charging current, just sufficient to compensate the interal losses. This is done by maintaining a permanent voltage of 2.1 / 2.2. V per cell on the battery.

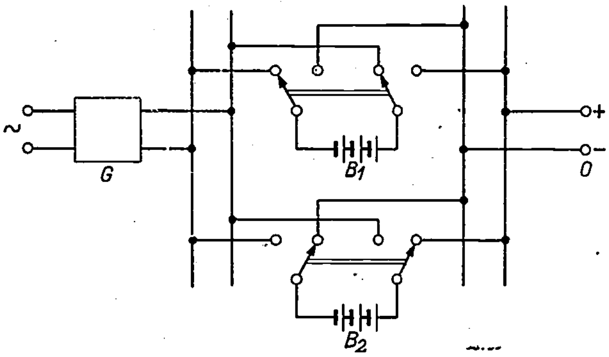


Fig. 1. Two-battery system. While the battery $B_2$ is feeding the plant connected at $O$ the battery $B_1$ is being charged by the rectifier $G$. When battery $B_2$ becomes exhausted two switches are thrown over and the functioning of the batteries is interchanged until $B_1$ is exhausted, and so on.

The fact that under these working conditions the battery will last longest may be made clear from the following somewhat simplified considerations. In the process of discharging sulphate of lead is formed in a fine crystalline structure, and after a time, about a week, this begins to change into a coarse crystalline formation. This is the phenomenon of the "hardening" of the plates already mentioned. Provided not too long a time is allowed to elapse the sulphate of lead can, it is true, be reduced again to a fine crystalline structure by repeated deep discharging followed by overcharging, but then it is inevitable that in the overcharging of the battery it will "boil" or "gas" for quite a time, and this "gasing" is most harmful for the life of the battery. It destroys mechanically the skin of $PbO_2$ right down in the depth of the pores, thus reducing the active surface. Later on a fresh skin of $PbO_2$ is formed on the exposed lead, thereby increasing again the active surface. Every time this takes place, however, the acid burns deeper and deeper into the metal, causing the plate to become spongy and shortening its life. Now, neither of these two causes of ageing — hardening through discharge and becoming spongy through the gasing — occur when a battery is kept under a preservation charge, because then it does not lose any charge and does not gas.

[3]) See, *i.a.*, E. Anderfuhren. Techn. Mitt. Schweiz T.T. **19**, 146-151, 1941; H. Grau, Die Stromversorgung von Fernsprechwähleranlagen, Verl. R. Oldenburg 1943, page 14; also U. Lamm ASEA J. **20**, 3-9, 1947 (No. 1)., in which article the line of thought followed is similar to that here.

It may be said, therefore, that a battery (not subjected to shocks) permanently kept to a cell voltage of 2.1 — 2.2 V has practically eternal life, whereas if it is left to perform its natural function of giving off and taking up charges a battery must inevitably age.

### Single-battery system versus two-battery system

In practice the method of preservation charging for an emergency supply system could be applied according to the diagram in *fig. 2*. It is assumed that the plant is fed from the a.c. mains *via* a rectifier. If the plant can be run equally well on a.c. or d.c. voltage this rectifier can be dispensed with. There is only one battery, permanently connected to its own rectifier, which must be capable of keeping the battery up to a preservation charge or charging it properly after a discharge. In the event of the mains failing the battery is then switched over, maybe automatically, from its charger to the plant, and when the mains voltage comes on again it is switched back to the rectifier.

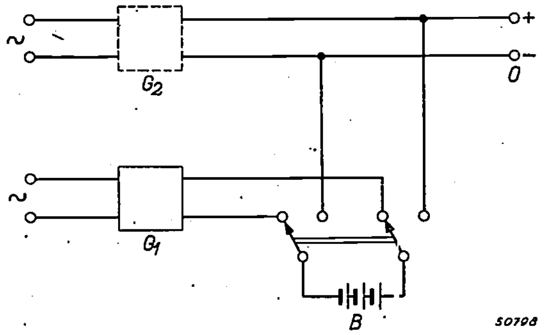


Fig. 2. Emergency supply system with only one battery and two rectifiers. While the plant connected at $O$ is being fed from the rectifier $G_2$ (sometimes direct supply from the mains is also possible) the battery $B$ is being kept under a preserving charge from its own rectifier $G_1$. In the event of the mains failing $B$ is switched over to the plant.

A still simpler circuit is to connect the plant and the battery in parallel to the same rectifier. This system, which is the most important from a practical point of view, is represented in *fig. 3*. In case of a mains breakdown the permanently connected battery takes over the feeding of the plant without any switching and without any interruption.

A "single battery system" as outlined here has important advantages over the two-battery system previously described.

In the first place the full capacity of the battery is always available for keeping the plant running in case of need. For a certain duration of "reserve running" roughly only half the battery capacity is required for a single-battery system compared with the two-battery system, thus halving the initial cost and the space required. Furthermore, one can then dispense with the switchboard with bus-bars, switches, separate mains switch, voltmeter and ammeter for each battery.

A second advantage is one of efficiency. In a two-battery system every kWh of the power output is first accumulated in the battery, which is not the case with a single-battery set. Since, as proved by practical measurements, the mean useful effect of a battery for this kind of plant is only about 65%, the efficiency of the normal plant with the single-battery system is much higher than that with a two-battery system.

Another advantage of a single-battery system is that there are never any gases or caustic vapours hanging about in the vicinity of the accumulators, because a battery maintained by a preservation charge does not gas.

Finally we come to what are possibly the most important advantages: the life of a battery maintained by a preservation charge is very much longer, as explained above, and might be said to be almost unlimited if the battery did not now and then have to come into action when the mains fail. Upkeep and attendance can be reduced to practically nil: with a single-battery system on the principle of fig. 3 the battery functions and ceases to function without any manual aid or automatic switching, and if the rectifier is of a suitable construction (about which more will be said later) the installation can be left entirely to itself; apart from a simple manipulation after a mains breakdown, all that is required is a monthly overhaul.

### The preservation rectifier

As a characteristic feature of preservation charging it has been said above that a voltage of 2.1 to 2.2 V per cell has to be maintained on the battery. It is the satisfying of this condition that constitutes the main problem in a single-battery system.

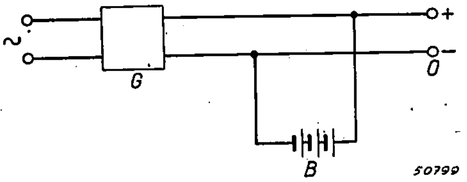


Fig. 3. A single-battery system where the plant $O$ and the battery $B$ are permanently connected in parallel to a preserving rectifier $G$.

With the arrangement according to fig. 2 it is in the first place necessary that the output voltage of the rectifier should vary as little as possible with the inevitable fluctuations of the mains voltage, which may amount to ± 5% and even sometimes ± 10%. With a system according to fig. 3, which is to be preferred because it dispenses with all switching operations and which we shall, therefore, consider exclusively from now on, something more has to be demanded of the rectifier: its output voltage must also be independent of fluctuations in the load.

Ordinary rectifiers do not by any means possess these properties. Their output voltage drops considerably the higher the power consumed, with the result that with every increase of the load the battery, connected parallel to the rectifier, takes part in supplying current to the plant and thus loses some of its charge. If the rectifier voltage drops 10% — which may well happen if both the mains voltage and the load change simultaneously — this means a dissipation of about 80% of the battery's capacity [4]). Needless to say, very little then remains for the functioning of the single-battery system.

Nevertheless, for the lack of anything better the ordinary rectifier has in the past been used for this purpose. By continuous readjustment one endeavoured to make the rectifier current at any moment as far as possible equal to the current consumption plus a small excess to compensate the battery losses. Of course, this readjusting by hand is only a makeshift and only partly helps to achieve the advantages of the single-battery system. As a matter of fact in the beginning the single-battery system was mainly applied for the sake of its simplicity and the cutting out of switching operations; one was then scarcely aware of the importance of the principle of preservation charging, so that it was not fully realized what advantages were lost owing to the rectifier voltage not being kept sufficiently constant.

Once impressed with the importance of the new treatment of batteries, one had to look for a battery charger which in spite of mains voltage fluctuations and load variations continues to supply a constant d.c. voltage. Such an apparatus we will call a preservation rectifier, in view of the object of keeping a battery under a preservation charge [5]). As a matter of fact a similar apparatus had already been developed for another purpose, for cases where a constant voltage was required not for the charging of a battery but for the power plant itself. Such is particularly the case with telephone exchanges. In a previous article [6]) it has been explained that the supply voltage of a telephone exchange is confined to very narrow limits for a proper functioning of the automatic selectors, while on the other hand the load varies considerably owing to the varying number of telephone conversations being carried at a time. (A battery connected also as a reserve likewise benefits when for these reasons the voltage is kept constant.)

Various kinds of supply rectifiers have been worked out for this purpose. In the first place there is a group of constructions where the voltage is regulated by a resistance in the direct current circuit which automatically increases as the voltage rises. This regulation of the resistance is effected for instance by the sliding of a contact by a servomotor governed by a contact voltmeter or by a magnet coil connected to the voltage to be regulated, or by some other means, in any case mechanically. Mostly this necessitates the special measures to prevent oscillations due to the inertia of the masses to be moved. Furthermore, wear and tear is inevitable. If such an apparatus were to be used as a preserving rectifier in a supply system according to fig. 3, then a difficulty would arise in the event of mains failures of long duration, for the battery would then have given off much of its charge and its voltage will have dropped considerably. Since the regulating mechanism adjusts the output voltage of the apparatus to the normal value after the mains voltage has returned, far too high currents may result. This means that special measures have again to be taken, which make the apparatus complicated, or else it must be continuously watched and operators must be ready to take action immediately after every breakdown of the mains [7]).

[4]) The relation between the voltage and the state of charge of a battery is not unequivocal. Only after very small quantities of the charge have been dissipated can it be said that the battery after recharging to its original voltages has again reached practically its original state of charge. If, however, the battery is discharged to a higher degree then it must be recharged for some time to an overvoltage before it gets back its full state of charge.

[5]) It is pointed out that this term is used here in a narrower meaning than that in which it is used in literature; generally it is taken to mean a rectifier capable of keeping an unloaded battery under a preserving charge, thus compensating the leak losses and maintaining the battery voltage at 2.1 to 2.2 volt per cell. We are using the term preserving rectifier here for an apparatus that keeps the battery voltage independent of both the mains voltage and the load within the limits mentioned.

[6]) H. A. W. Klinkhamer, A rectifier for small telephone exchanges, Philips Techn. Rev. 6, pp. 39-45, 1941.

[7]) The fact that notwithstanding these complications this kind of apparatus has frequently been used as a preserving rectifier for the single-battery system right up to the present time clearly shows what value is attached to this system. B. Stange, for instance, in E.T.Z. 64, 341-344 and 372-377, 1943, maintains that the complications with such an apparatus are more than compensated by the advantages of the single-battery system.

It has been endeavoured to overcome the drawback of mass inertia and of wear by governing the reactance of the a.c. side of the rectifier by electrical or magnetic means *via* the output voltage, instead of by mechanical means. We would mention here as an example the connecting in series of a choke pre-magnetised with the direct current.

## A preserving rectifier with highly saturated transformer

Some years ago Philips constructed a rectifier (see the article quoted in footnote [6]) of such a design that the voltage is kept constant not by a regulating reactance of the output voltage but by means of a highly saturated transformer in a special circuit as represented in fig. 4*b*. (Fig. 4*a* shows the circuit of a normal rectifier by way of comparison.)

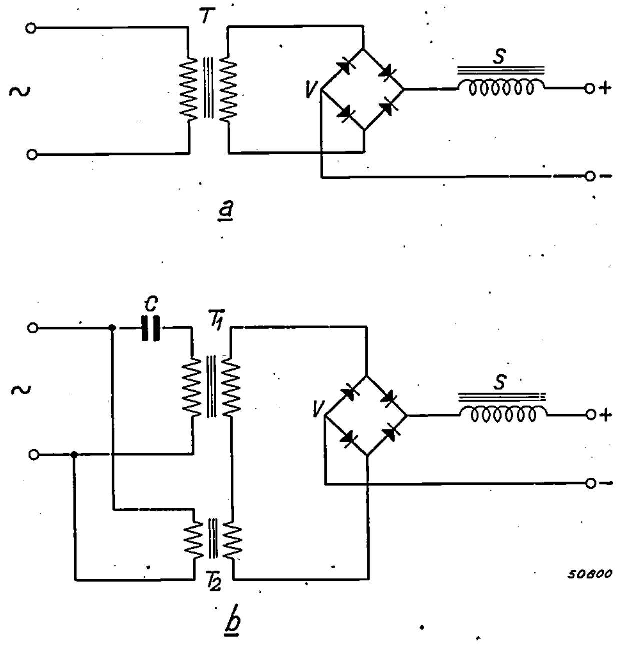


Fig. 4. Arrangement of a normal rectifier with transformer *T*, four selenium valves *V* with Graetz circuit, and smoothing choke *S*.

*b*) Arrangement with a preserving rectifier according to a principle described earlier [9]. $T_1$ is a transformer with highly saturated iron core, *C* a condenser, $T_2$ a small normal transformer, *V* selenium valves, *S* choke.

The working of this circuit is not immediately apparent. For an explanation and for the deduction of the characteristic of such a rectifier from the data of the elements and the magnetisation curve of the iron core see the article referred to in footnote [6]). Here we will deal only with those properties of this kind of rectifier that make them specially suitable for use as a preserving rectifier.

The direct voltage is to a very high degree independent of load and mains voltage fluctuations, as may be seen from *figs. 5* and *6*. Fig. 5 gives the characteristic, *i.e.* the direct voltage as a function of the load current, for the nominal mains voltage. It will be seen that within a wide range of current intensities — in this case between approximately 0.3 and 3 amperes — the voltage varies by only a few percent (from 63-60 volt). Fig. 6 gives the characteristic for three different values of the mains voltage and shows that even with a varying mains voltage the direct voltage remains fairly constant, within about 3% for 10% mains voltage difference[8]).

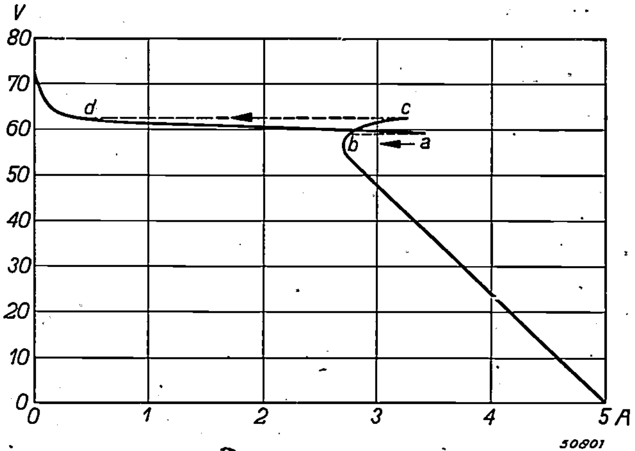


Fig. 5. Characteristic of a rectifier according to fig. 4*b*. The voltage varies but little within a wide range of load currents (approx. 0.3—3 Amp.). The characteristic has two branches. When the load rises to point *a* the working point jumps over to the steep branch (*a* → *b*), so that the current taken off continues to be limited. When the load drops the working point does not jump back to the flat branch until the voltage of the permanently connected battery has risen to point *c* (*c* → *d*).

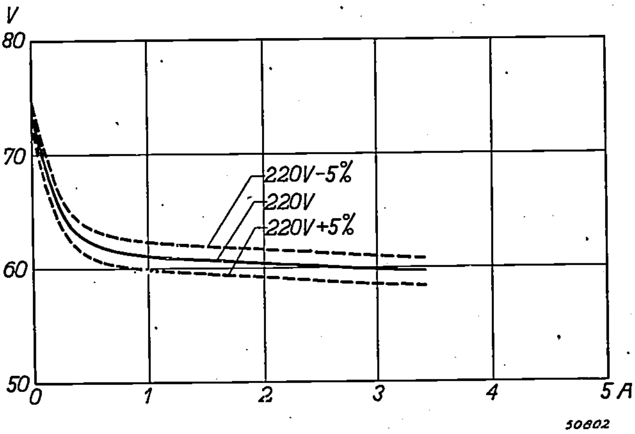


Fig. 6. Characteristics as in fig. 5 for three different mains voltages (nominal 220 V). Only the flat branch is drawn. It will be noted that the voltage rises with falling mains voltage and *vice versa*. The variations are small: about 3% for 10% mains voltage variation.

[8]) It should be noted that these characteristics are obtained at the nominal frequency of the mains. If the mains frequency shows considerable fluctuations, as is the case in many countries in the present post-war period, then the properties of the rectifier system described here are less favourable, in that the output voltage is not sufficiently constant.

If only on this account the apparatus is therefore already capable of answering the requirements of a preserving rectifier, whilst, moreover, it has the advantage of the absence of inertia and wear, since there are no moving parts. But from a closer investigation it will be found to have yet another important property which avoids in a simple manner the last of the drawbacks mentioned above as attaching to former constructions, *i.e.* the necessity of manual manipulations immediately after a mains breakdown. As may be seen from fig. 5, the characteristic of the rectifier consists of two separate intersecting branches, one flat and one steep. The rectifier is so designed that for the current intensities normally required for the plant the working point always lies on the flat branch. Variations in the state of charge of the battery are thereby extremely limited. Should a load peak arise which is too great for the highest current of the flat branch then the rectifier adjusts itself, the working point jumping over to the steep branch from *a* to *b* along the lower dotted line (see fig. 5). The battery then supplies that part of the total current that is lacking. When the peak has passed the working point it remains for a moment on the steep branch of the characteristic, the rectifier thus supplying more current than is being consumed by the plant, and this surplus serves to charge up the battery again. As the battery voltage rises so the working point climbs along the steep branch until the point *c* is reached, where the rectifier readjusts itself, the working point jumping back along the upper dotted line to the flat branch, towards point *d*.

In the event of the mains failing then the battery comes into full operation and is discharged more or less according to the duration of the mains breakdown. Even though the mains interruption may last so long as to cause the battery voltage to drop considerably, when the mains voltage comes on again there need be no fear of very high currents arising. As a matter of fact the working point of the rectifier then immediately comes to lie on the steep branch of the characteristic, at the level corresponding to the battery voltage at that moment. Consequently the current continues to be limited and can never reach a dangerous level, while this does not call for any manual action. During the normal working the battery is then charged up again until the battery voltage reaches the top of the steep branch (point *c*).

The battery is not then fully charged, it is true, because charging would have to continue for a time at a still higher voltage. For that purpose a small "booster" is built into the rectifier, which can be connected, automatically or otherwise, in series with the rectifier proper to get quick charging, dropping out of action as soon as the desired voltage of the battery is reached. Even though the booster may not have automatic action it involves no extra work worth mentioning: since it is only a matter of a supplementary charging it is not neccesary to boost up immediately after a mains failure, for this can be done within a day or two by sending someone to switch on the booster. If another small preserving rectifier is used as booster then no supervision at all is required for rapid charging.

In fig. 5 it is seen that the characteristic rises steeply on the extreme left. Consequently the voltage on the battery would there be higher than is normally desired. In most cases of normal working, however, this point is never reached, for the working point can only reach such a position if the installation remains entirely unloaded for one or two days. But even if the working point should reach this area there would still be no harm done to the battery, because the current intensity with which the battery is then further charged is extremely low. If necessary that rising part of the characteristic can, as a matter of fact, be removed by applying a small rest-load.

All these properties of this preserving rectifier make it possible to leave the accumulator battery to look after itself, apart from an inspection say once a month. We thus have the full benefit of all the advantages of the single-battery system: lasting qualities of the battery, practically its full capacity being available in case of need, simplicity of the circuit, continued working of the plant in case of mains failure without switching operations and without any interruption, functioning without attendance and practically no supervision.

### Some special applications of preserving rectifiers

It should by now be sufficiently clear why we regard the single-battery system with a preserving rectifier of the construction described here as the ideal emergency supply system for all sorts of cases. We will now mention a few special cases where the preserving rectifier is used in somewhat different ways.

One application for telephony has already been fully dealt with in the article quoted in footnote[6]). In that case the preserving rectifier with battery served for the feeding of terminal exchanges in the telephone network. In these small exchanges, where there are no permanent attendants, the properties of these rectifiers show to their fullest advantage.

But also in larger telephone exchanges where personnel are always in attendance the advantages of this system are still of importance. Such has proved to be the case with a trial installation taken into use some years ago in a so-called junction exchange of the Dutch P.T.T. In these exchanges the daily variation of load differs from that in the terminal exchanges; there is also a certain permanent (rest) load due to the counting apparatus for the trunk calls. This does not, however, make any essential difference for the action of the rectifier and the battery.

For this trial installation the capacity required was about five times that needed for terminal exchanges. The designing of a preserving rectifier for high ratings is a somewhat more difficult problem than that of increasing the capacity in the case of normal rectifiers. Owing to the high saturation of the transformer core the iron losses in a preserving rectifier of the type described are greater than normal, that is to say the heating that arises becomes troublesome already with small transformer volumes. For this reason it is more economical in the case of higher capacities not to rely solely upon a preserving rectifier for the supply of the total power but to have such a rectifier connected in series with an ordinary battery charger. This also has the advantage that the influence of mains voltage fluctations can be still further restricted, for, as shown in fig. 6, the output voltage of the type of preserving rectifier described rises as the main voltage of a normal rectifier drops with falling mains voltage, and when we have the two connected in series these effects partly neutralize each other. This was utilised in the installation for a junction exchange. The power required (15 Amp., 60 V) is supplied for 3/4 by a preserving rectifier and for 1/4 by an ordinary battery charger. *Fig. 7* is a diagram of the set-up that has now already been applied for several junction exchanges. With a mains voltage variation of 10% the direct voltage varies by only 1%. *Fig. 8* gives the characteristics of the two rectifiers when each is used separately, as also the resulting characteristic when the two are connected in series. It is remarkable that in spite of the steep slope of the characteristic $G_2$ the resultant characteristic is not noticeably steeper than $G_1$. Consequently the effect of load variation on the output voltage is not greater with the series connection than with the preserving rectifier alone. This is quite understandable when it is borne in mind that the moments of opening and closing

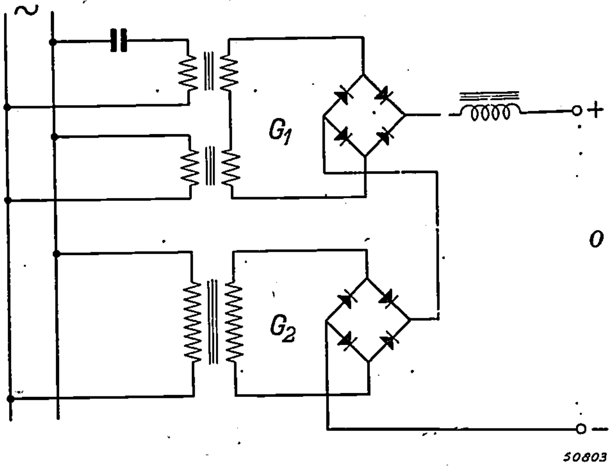


Fig. 7. Series connection of a preserving rectifier $G_1$ and an ordinary rectifier $G_2$. A similar arrangement is being used for the feeding of a number of junction exchanges of the Dutch P.T.T.

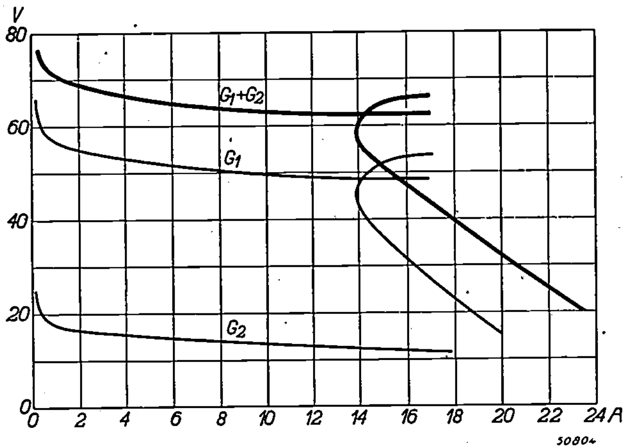


Fig. 8. Charactcristics of the two rectifiers of fig. 7, each taken separately: $G_1$, $G_2$; the resulting characteristic of the series connection of the two rectifiers is $G_1 + G_2$.

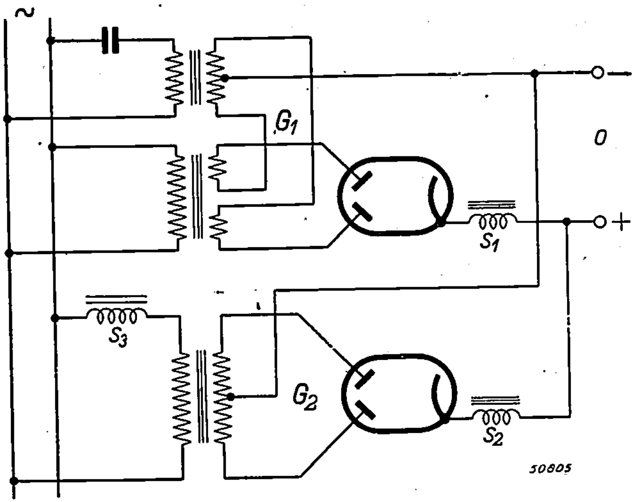


Fig. 9. Parallel circuiting of a preserving rectifier $G_1$ and an ordinary rectifier $G_2$. In the special case where we applied this arrangement it was desirable to use rectifying valves instead of selenium valves. $S_1$ and $S_2$ are chokes. $S_3$ is a variable choke by means of which the flat part of the resulting characteristic (see $G_1 + G_2$ in fig. 10) can be shifted to the desired current intensity range.

of the valves for the two rectifiers connected in series are not mutually independent. One may not therefore simply add together the ordinates of the two characteristics for each abscissa (current intensity). (The theoretical deduction of the actual course of the resultant characteristic is very difficult.)

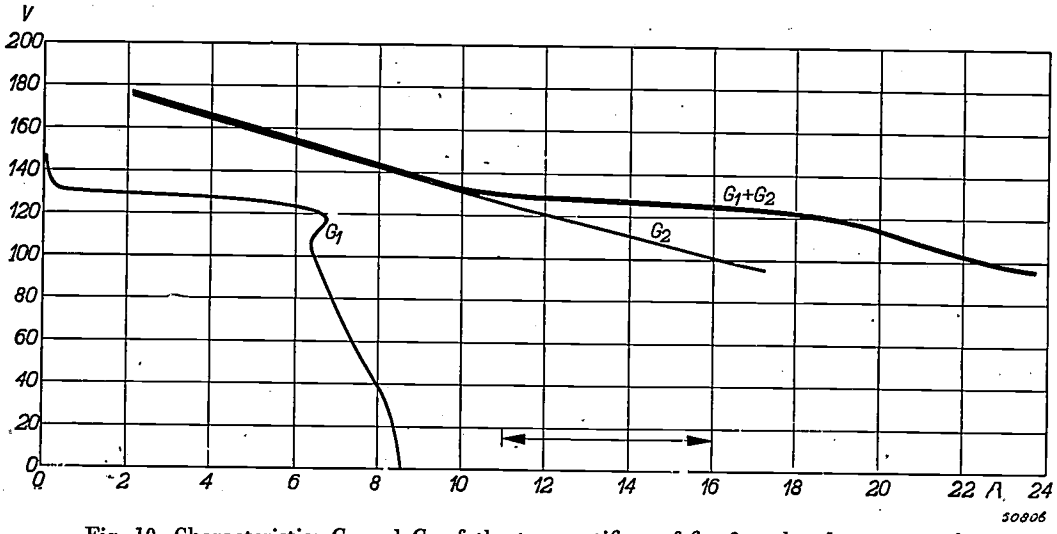


Fig. 10. Characteristics $G_1$ and $G_2$ of the two rectifiers of fig. 9 each taken separately. Since the internal resistance of a rectifying valve varies with the current in a manner different from that of a selenium valve, the characteristic $G_1$ differs somewhat from that in fig. 5; though there is a flat and a steep branch these flow into each other without intersecting. $G_1 + G_2$ is the resulting characteristic of the parallel circuiting of the two rectifiers. Within a limited current intensity range (indicated on the abscissa) the voltage is fairly constant, whilst only a relatively small part of the power has to be supplied by the preserving rectifier.

A somewhat different manner of coordinated action between an ordinary rectifier and a preserving rectifier is obtained when these are connected in parallel. Such an arrangement has been applied, for instance, for the feeding of a set of magnetic mains switches in a distributing station of the electrical network of the Philips works. The set-up is shown in *fig. 9*, whilst *fig. 10* gives the separate characteristic of the two rectifiers and the resultant characteristic of the parallel connection. Since the output voltages in this case are identical, the two rectifiers here do indeed behave independently of each other and the resultant characteristic is obtained simply by adding the two abscissae for each ordinate. It is seen that the resultant characteristic has a fairly steep slope on the whole but that in a small range it is practically horizontal. In the case in question it was in fact only a matter of a small range of current intensities, so that the arrangement described had the advantages that the horizontal part of the characteristic could be attained in that small range with a quite small and thus inexpensive preserving rectifier.

# SEMI-CONDUCTORS WITH LARGE NEGATIVE TEMPERATURE COEFFICIENT OF RESISTANCE

621.315.59:
537.312.6:548.7

by E. J. W. VERWEY, P. W. HAAYMAN and F. C. ROMEYN.

In view of the increasing importance in electrotechnology of resistor materials having a large negative temperature coefficient of resistance, materials having this property have been developed in the Philips laboratories. These materials are mixed crystals of $Fe_3O_4$ with certain substances having the same crystal structure as $Fe_3O_4$, the so-called spinel structure. They offer considerable advantages compared with the materials hitherto used for this purpose. In the first place they are much more constant in manufacture: their resistivity is exclusively governed by the mixing proportions of $Fe_3O_4$ and the other component. In the second place the electrical properties (the values of the resistance and its temperature coefficient) of the resistors made with these materials — here called *h.t.c. resistors* — are much more constant in use: for most applications a h.t.c. resistor can be used in air without requiring any special precautions, even at a temperature of some hundred degrees centigrade.

This article deals with the practical possibilities of these resistors and the physical-chemical background underlying the development of these new materials.

## Introduction

Semi-conductors are solids whose specific resistivity is much greater than that of metals. The conduction mechanism in semi-conductors may be based either on the movement of ions (electrolytic conduction) or on that of electrons (electronic conduction). Only the electronic semi-conductors are employed in practice as materials for resistors, because electrolytic conduction is accompanied by chemical changes and polarization phenomena which are found to be troublesome. In this article, therefore, we are dealing only with the electronic semi-conductors.

Although the conduction in these semi-conductors is due to the same movement of electrons as occurs in metals, the mechanism of that phenomenon is essentially different from that in the metals. This manifests itself in the fact, among others, that with increasing temperature the specific resistivity of these semi-conductors diminishes, whereas in the case of the metals it is the other way round. It appears that the relation between the specific resistivity $\varrho$ of semi-conductors and the absolute temperature $T$ can be expressed in good approximation by the formula

$$\varrho = \varrho_\infty e^{b/T} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where $\varrho_\infty$ and $b$ are positive temperature-independent constants. The temperature coefficient $a$ of the resistance is therefore

$$a = \frac{1}{\varrho}\frac{d\varrho}{dT} = -\frac{b}{T^2}, \quad . \quad . \quad . \quad . \quad (2)$$

and thus decidedly negative.

For electric circuits a resistance with a negative temperature coefficient (abbreviated: t.c.) obviously offers certain advantages. It is for this reason too that semi-conductors are so highly important in electrotechnology. But in their application semi-conductors have often turned out to be less satisfactory than was to be expected. This is due to the fact that for practically all semi-conductors hitherto employed the electrical properties — the value of the resistance and of the t.c. — are subject to considerable changes in use; moreover, in the manufacture of semi-conductors it has generally been very difficult to get absolutely invariable results.

In the course of the last few years, however, the Philips laboratories have been developing semi-conductors which have a highly negative t.c. and exhibit these two disadvantages to a very much less extent. These semi-conductors are mixed crystals and compounded of $Fe_3O_4$ with certain substances satisfying the general chemical formula $XY_2O_4$. An essntial requirement is that these substances must have the same crystal structure as $Fe_3O_4$ (this chemical formula may also be written as $FeFe_2O_4$), the so-called spinel structure. The resistors made from these semi-conductors are here called h.t.c. resistors. The properties of these h.t.c. resistors are fairly stable, so that they can be used in air, even at a temperature of some hundred degrees centigrade, without any special precautions. They are also very well reproducible: in the manufacture of the material for h.t.c. resistors the desired value of resistivity can easily be obtained by a suitable choice of the mixing proportions of $Fe_3O_4$ and the other component.

It is to be pointed out that the use of h.t.c.

resistors is not confined to those cases where materials with a negative t.c. are needed. It has been found that the absolute value of the t.c. of h. t. c. resistors at room temperature is larger than that of metals by a factor 10. Consequently if the action of any particular instrument is based upon the variability of the resistance with temperature and it makes no essential difference whether the t.c. is positive or negative, it will often be easy to increase the sensitivity of the instrument by using a h. t. c. resistor instead of a metallic one.

The possibilities of the h. t. c. resistors will be discussed more fully in the first section of this paper. There will then follow an explanation of the physical-chemical background underlying their development. This will be done with reference to two articles recently published in this journal, one on the electronic conduction in semi-conductors [1]) and the other on materials having the spinel structure [2]). It will be found that the good stability of h. t. c. resistors and constancy of their electrical properties are easily understandable from the conduction mechanism of the semi-conductors (which becomes evident, *inter alia*, from equation (1)) and from the characteristics of the spinel structure.

## Applications of h. t. c. resistors

The uses to which h. t. c. resistors can be put may be placed under two headings, one where the t.c. must be essentially negative and the other where only the high absolute value of the t.c. is of importance.

### *Applications where the t.c. must be essentially negative*

Here two cases may be distinguished. In the first case good use is made of the fact that it takes some time for the current to heat up the h. t. c. resistor and for equilibrium to be reached between the electrical energy supplied (Joule heat) and the heat carried off through convection, conduction and radiation. In the second case the h. t. c. resistor is used in a state of equilibrium; the inertia of the phenomenon is then taken into account or else attempts are made to reduce it by special measures.

A typical example of the first case is to be found in radio technique. In radio valves with an indirectly heated cathode it takes some time before the cathode reaches its ultimate temperature. The t.c. of the heater filaments (generally of tungsten) being positive, their resistance is lower when the current is first switched on than it is in the ultimate state. Where the heater filaments are connected in series with other parts of the circuit it may therefore happen that when the current is switched on those other parts have to withstand a much greater voltage than prevails during the normal working. The consequences for the parts in question may be fatal. This difficulty can now easily be circumvented by connecting a h. t. c. resistor in series with the heater filaments. The resistivity and the dimensions of the h. t. c. resistor can easily be chosen of such a value that at the initial temperature the dangerous over-voltage is practically entirely taken up by the h. t. c. resistor, while thanks to its highly negative t.c. this resistance cuts itself out almost completely at the higher final temperature of the resistance. The principle of this application of h. t. c. resistors is of course not limited to radio sets but can also be applied, for instance, for lowering switch-on peaks such as occur with motors.

As regards the second kind of application we have to consider further the current-voltage characteristic of the h. t. c. resistors in the stationary state. This is represented diagrammatically in *fig. 1*, where the voltage $V$ across the resistance is plotted as a function of the current $I$ flowing through the resistor. For a certain current the value $R$ of the resistance is given by the tangent of the angle $\Theta$ between the horizontal axis and the line drawn from the origin through the "working point" $A$. The higher the intensity of the current, the higher is the tempera-

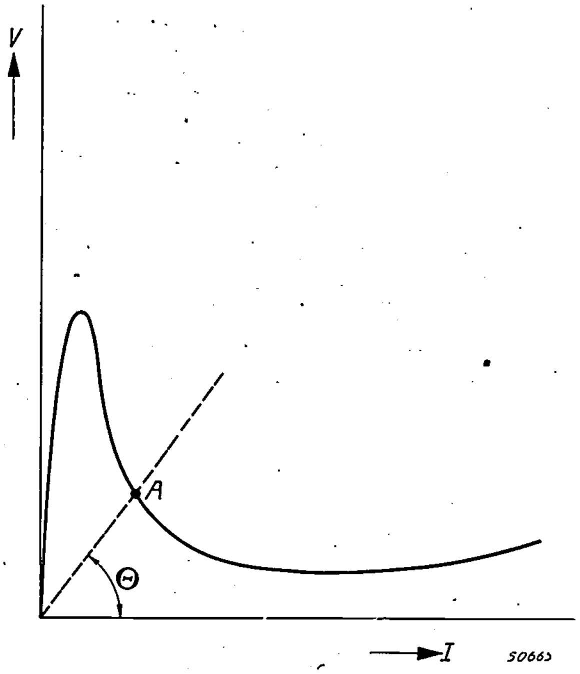


Fig. 1. The voltage $V$ across a h.t.c. resistor as function of the current intensity $I$ (qualitative). At a certain value of $I$ the resistivity is given by tan $\Theta$.

[1]) Philips Techn. Rev. 9, 46-54, 1947 (No. 2).
[2]) Philips Techn. Rev. 9, 186-192, 1947 (No. 6).

ture of the resistor and, owing to the negative sign of the t.c., the smaller the value of the resistance. The reduction of the resistance with increasing current intensity becomes more gradual as the current intensity becomes greater.

The remarkable shape of the characteristic can be understood when it is considered that in the state of eqµilibrium the Joule heat generated in the resistor must be equal to the heat carried off to the outside. The precise form of the characteristics will therefore depend not only upon the electrical properties of the h. t. c: resistor, but also upon the manner in which heat is exchanged with the surroundings. Consequently in a vacuum a h.t.c. resistor will behave differently than in a gas.

Now several parts of this characteristic can be used in electric circuits. First of all it is seen that over a large part of the curve the voltage varies but little with the current intensity, namely in that part corresponding to high loads. This kind of resistor can therefore be used as a voltage stabilizer.

Secondly, there is a part of the curve having a negative slope, thus corresponding to a negative differential resistance d$V$/d$I$. Since negative differential resistances play a part in the excitation of electric vibrations, we have here another field open for the use of h.t.c. resistors, at least for low frequencies.

The combination of an ordinary "temperature-independent" resistor and a h. t. c. resistor in series again produces a horizontal straight line within a certain range of the $V$-$I$ characteristic, as may be seen from *fig. 2*. Such a combination of resistors can therefore also be used for the stabilization of voltage. In many cases this has advantages not obtained when using the flat part of the characteristic of the h. t. c. resistor itself: the horizontal part of the characteristic lies at a much lower current intensity; there is no maximum having to be exceeded before the condition is reached corresponding to the horizontal part of the characteristic.

In all these applications a h.t.c. resistor will reach the state of equilibrium more quickly and the fluctuations of the voltage or of the current follow accordingly when (a) the thermal capacity of the resistance is lower and (b) the heat exchange with the surroundings is more rapid. In connection with (a) these resistors can be applied in the form of very thin wires. This is made possible by employing the known ceramic methods, as for instance by spraying the ceramic mass with a suitable binder, and then baking the wires in a hanging position. As regards (b) this can be brought about by placing the resistor in a small tube filled with a gas consisting of very light molecules (*e.g.* helium).

Among the group of applications being dealt with here there is also the case where a resistance has to be provided which is independent of the ambient temperature within a certain temperature range. This is again attained by connecting a suitably dimensioned h. t. c. resistor in series with a metallic resistor.

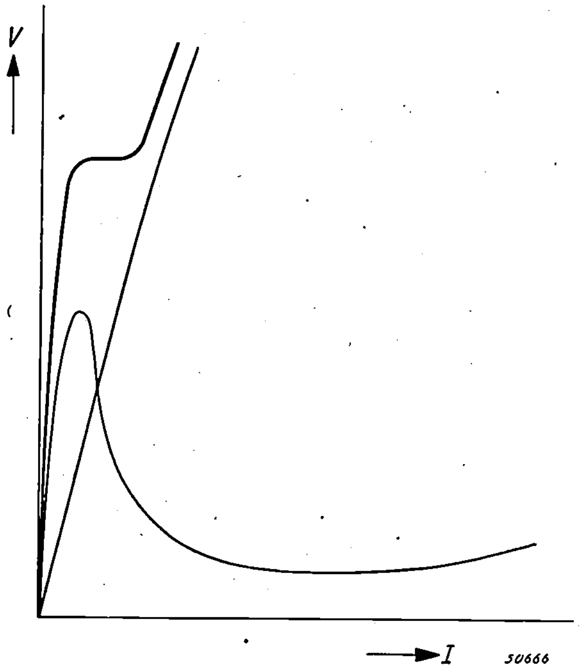


Fig. 2. The voltage $V$ across a resistor comprising a h. t. c. resistor in series with a metallic resistor, as function of the current intensity $I$ (qualitative).

### *Applications where the high absolute value of the t.c. is of importance.*

As already indicated in the introduction, here we have in mind the application of h. t. c. resistors in all kinds of instruments whose action is based on a t.c. differing from zero and in which it is in principle immaterial whether the t.c. is positive or negative. In the past metallic temperature-dependent resistors have usually been used for this purpose — for instance resistance thermometers or bolometers. The sensitivity of such instruments is greater the higher the absolute value of the t.c.

Now the specific resistivity $\varrho_{\mathrm{met}}$ of metals at not too low a temperature is approximately proportional to the absolute temperature $T$:

$$\varrho_{\mathrm{met}} = \text{const. } T.$$

From this it follows that the t.c. of $\varrho_{\mathrm{met}}$ is practically independent of the value of $\varrho_{\mathrm{met}}$. In other words the t.c. has approximately the same value for all metals:

$$\alpha_{\text{met}} = \frac{1}{\varrho_{\text{met}}} \frac{d\varrho_{\text{met}}}{dT} = \frac{1}{T},$$

that is to say at room temperature $T = 300°$ and $\alpha_{\text{met}} = 0.33\%$.

In the case of semi-conductors the situation is quite different. Here $\alpha$ is indeed dependent upon the value of the specific resistivity, for from equations (1) and (2) it follows that

$$\alpha = \frac{\log \varrho_{\infty} - \log \varrho}{T}.$$

From the point of view of measuring technique a specific resistivity of $10^5$ Ω cm is quite suitable. Assuming further that $\varrho_{\infty} = 10^{-2}$ Ωcm (a still lower value of $\varrho_{\infty}$ is as a matter of fact not difficult to attain) then with the aid of the last formula we find:

$$\alpha = \frac{-2-5}{T} (\log 10) \approx -\frac{16}{T}.$$

For a h.t.c. resistor for which $\varrho = 10^5$ Ω cm is at room temperature it follows that $\alpha \approx 5\%$.

The superiority of h.t.c. resistors compared with metallic resistors is therefore obvious, apart from the further consideration that the high value of the specific resistivity is in itself an advantage in resistance thermometry.

Long-wave infra-red rays cannot be investigated with the aid of photo-conduction, photo-electrical emission or photographic methods, since the energy quanta are too small, so that bolometers have to be employed. For these bolometers h.t.c. resistors can be made in the form of thin membranes provided on either side with a metallic contact. The area of the surface chosen will generally be such that the beam just radiates the whole membrane. The membranes have to be very thin to keep their thermal capacity very low, this being favourable for the starting time of the instrument.

H.t.c. resistors are also suitable for measuring energy at still longer wavelengths, in the range of radio waves of very high frequency (cm waves). They are then used in the form of very small balls (size of a pin's head) with two very thin metal filaments baked in to serve as contacts.

In all these applications the measuring current sent through the resistor has to be sufficiently weak as not to cause any measurable change in the temperature of the resistor. Thus the value of the resistance is practically exclusively governed by the energy supplied (in the form of heat or radiation), as opposed to the case of the first group of applications, where the equilibrium temperature of the resistor and thus also the value of the resistance are dependent upon the Joule heat of the current.

Another possibility coming under this group of applications is the case where the temperature and in consequence the resistivity of a h.t.c. resistor is regulated by an external heating current. The current in one circuit can be continuously controlled with the aid of the current in another circuit, with the two circuits completely separated electrically. In such a case a h.t.c. resistor can be used, for instance, in the form of a slender rod surrounded by a coiled metal wire in such a way that a small space is maintained between the two conductors over the entire length, this assembly, with leads, then being mounted in an evacuated envelope or a tube filled with an inert gas.

To conclude this summary of the possibilities of h.t.c. resistors we would mention that apart from the shapes already spoken of (thin filaments, membranes, minute balls) these resistors can also be made in the form of small rods, plates, tubes, etc.

## The physical-chemical background underlying the development of h.t.c. resistors

### *The temperature-dependance of the resistance in semiconductors*

In a good conductor, for instance a metal, the conduction electrons are more or less free. When the temperature rises the number of free electrons remains unchanged. The electrical resistance, however, increases, owing to the fact that at a higher temperature the electrons become more scattered and are checked in their motion by the increased thermal movement of the atoms.

In semi-conductors the situation is quite different. At low temperature the material contains no free electrons. In the article quoted in footnote [1]) we saw, however, that by introducing either optical energy or thermal energy the electrons can be given more or less freedom of movement in the crystal lattice. If only little energy is required for this, that is to say if the electrons are only "weakly bound", it may happen that the material will already have considerable conductivity at room temperature. Anyhow, the number of electrons released by the thermal energy increases rapidly with the temperature. This effect as a rule exceeds by far the other effect of increased scattering and checking of the electrons. As a result the electric resistance diminishes rapidly with rising temperature; in other words the temperature coefficient of resistance is negative.

The relation (1) between the value of the resistance and the temperature is obtained roughly as follows. Let us suppose that in the crystal lattice there are $N$ places occupied by weakly bound electrons and that an energy $\varepsilon$, called the activation energy, is required to give such an electron more oı less freedom of movement. At the absolute temperature $T$ there will be on an average $n$ of the $N$ electrons in a more or less free state, that is to say there will be interstices in $n$ of the said $N$ places. In the thermo-dynamic equilibrium the relation then holds:

$$\frac{n \cdot n}{N-n} = \text{const. } e^{-\varepsilon/kT},$$

in which $k$ is Boltzmann's constant [3]). Since $n$ is much smaller than $N$ we may also write this formula as follows:

$$n = \text{const. } \sqrt{N}\, e^{-\varepsilon/2kT}.$$

The specific conductivity is now equal to the product of the mobility of the electron in the more or less free state and $n$, if the latter number is taken per cubic centimeter. Since the mobility of electrons is practically independent of temperature, ultimately we actually get for the temperature-dependence of the specific resistivity a formula of the form (1), where $b = \varepsilon/2k$:

$$\varrho = \text{const. } e^{\varepsilon/2kT} \quad . \; . \; . \; . \; . \; . \quad (3)$$

*Temperature-dependence of the resistance of $Fe_3O_4$*

After this introduction we will now proceed to discuss the semi-conductors from which h.t.c. resistors are made. The basic material, as stated in the introduction, is the compound $Fe_3O_4$, known in its mineral form under the name of magnetite and sometimes also called ferro-ferrite because it can be regarded as a ferritic $MFe_2O_4$, with bivalent iron playing the part of the metal M. This substance is a fairly good electron conductor, even a very good conductor compared with other oxidic semi-conductors: the specific resistivity at room temperature amounts to $5 \times 10^{-3}$ Ω cm, thus only a few hundred times the specific resistivity of metals. Yet $Fe_3O_4$ is a semi-conductor, in the sense that within a large temperature range it satisfies approximately the law (1). This is demonstrated in *fig. 3*, giving the logarithm of the specific resistivity $\varrho$ of $Fe_3O_4$ as a function of $1/T$ (in this graph the temperature increases from right to left). We will first consider that part of the curve lying between approximately 130°K and 200°K. Here the resistivity increases with falling temperature, the curve approaching a straight line. This is in agreement with equation (1), for from this equation it follows that:

$$\log \varrho = -\frac{b}{T} + \log \varrho_\infty, \quad . \; . \; . \quad (4)$$

so that log $\varrho$ is indeed a linear function of $1/T$. Further, we see that this practically linear part of the curve has only a small slope. In the temperature range for which $T > 200$°K the curve runs almost horizontal; for $T > 300$°K (this part of the curve is not drawn in the graph) the t.c. becomes even positive. Presumably here the metallic effect of increased scattering and checking of the electrons becomes predominant; moreover, $b$ is presumably no longer small compared with $N$ (*cf.* the deduction of equation (3)).

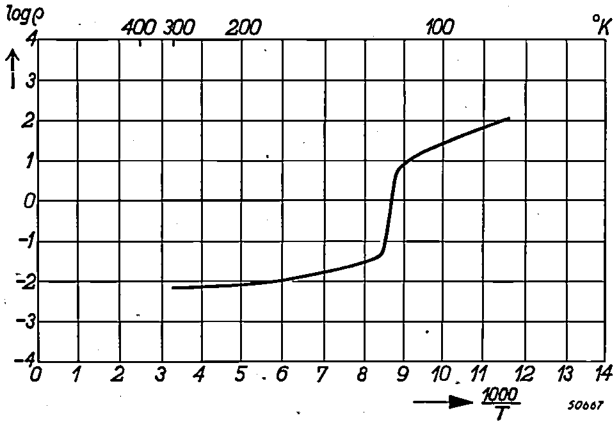


Fig. 3. Logarithm of the specific resistivity $\sigma$ (in Ω cm) as function of $1/T$, where $T$ is the absolute temperature, for $Fe_3O_4$.

In the right-hand part of the graph, *i.e.* at very low temperature, we see a peculiar phenomenon which is only observed with this substance [4]) $Fe_3O_4$, in contrast to its behaviour previously discussed. At about 130°K the resistivity jumps suddenly and as the temperature is further reduced the slope of the curve, *i.e.* the t.c., increases discontinuously. It would lead us too far afield here to enter upon a closer discussion of the nature of this transition point (see the article quoted in footnote [4]). Suffice it to add that this point disappears as soon as the $Fe_3O_4$ contains a small quantity (less than 1%) of another spinel — an abbrevation for a substance having the spinel structure.

From the foregoing it is evident that the compound $Fe_3O_4$ alone is not usually [5]) suitable for the

[3]) A more exact resolution of this formula shows that the factor by which the exponential function is multiplied is slightly dependent upon temperature.

[4]) See E. J. W. Verwey and P. W. Haayman, Physica, The Hague, 8, 979, 1941.

[5]) In a very special case use can be made of the great change in resistance at the point of transition.

applications described in this article. Both its specific resistivity and its t.c. are much too small [6]). Nevertheless, $Fe_3O_4$ is quite a suitable basis for the preparation of suitable resistance materials. Now in order to make the development of h.t.c. resistors understandable we must discuss somewhat more closely the reason for the larger specific conductivity of $Fe_3O_4$ which is extraordinarily high for semiconductors.

### *Relation between the electron conductivity and the spinel structure of $Fe_3O_4$*

The explanation of the high specific conductivity of $Fe_3O_4$ is closely related to a typical property of the spinel structure of this substance. This spinel structure has already been fully dealt with in a previous article in this journal (see note 2). There it was shown that the metal atoms or ions [7]) in the spinel structure may be arranged in at least two ways, there being two kinds of holes available for the metal ions in this crystal structure, namely per molecule one so-called tetrahedral hole and two octahedral holes. In the case, for instance, of a spinel with the chemical formula $M^{2+}M_2^{3+}O_4$ where $M^{2+}$ and $M^{3+}$ are respectively bivalent and trivalent metals, the two ways in which the metal ions may occupy the available holes are as follows. In the one type of structure the tetrahedral holes are occupied exclusively by $M^{2+}$ and the octrahedral holes exclusively by $M^{3+}$. In the second type of structure all the $M^{2+}$ions are in the octehedral holes, whilst half of the $M^{3+}$ions are in octahedral and the other half in tetrahedral holes; the distribution of the $M^{2+}$ and $M^{3+}$ions among the octahedral holes is quite irregular. The first type of structure is indicated by the formula $M^{2+}(M_2^{3+})O_4$ and the second by $M^{3+}(M^{2+}M^{3+})O_4$, the symbols of the ions situated in the octahedral holes being bracketed in each case. There are also spinels with the chemical formula $M_2^{2+}M^{4+}O_4$ which likewise have two possible types of structure similar to these.

In the article referred to it has been seen further that it is not possible to determine directly by X-ray analysis to which type of structure $Fe_3O_4$ belongs. A number of "rules" have, however, been used making it possible to predict with a great degree of probability the arrangement of the metal ions in a certain substance with spinel structure. In a somewhat abbreviated form the rules in question are:

1) The trivalent and quadrivalent metal ions occupy the octahedral holes.
2) Exceptions are the $Fe^{3+}$-ions, which preferably occupy the tetrahedral holes.
3) $Zn^{2+}$- and $Cd^{2+}$-ions are capable of driving the $Fe^{3+}$-ions out of the tetrahedral holes.

According to these rules, therefore, $Fe_3O_4$ should have a structure of the second type, $Fe^{3+}(Fe^{2+}Fe^{3+})O_4$. In other words, in the octahedral holes $Fe_3O_4$ will have a "mixture" of bivalent and trivalent ferri ions. It is to this fact that we turn for an explanation of the exceptionally high conductivity of this compound, because if $Fe^{2+}$ and $Fe^{3+}$ are indeed irregularly distributed among crystallographically equal sites the transition of an electron from an $Fe^{2+}$-ion to an $Fe^{3+}$ion (the former becoming an $Fe^{3+}$-ion and the latter an $Fe^{2+}$-ion) will make little difference in this arbitrary distribution. Since, moreover, the electric field from the surrounding ions at two crystallographically equal sites is equal, the energy of the electron before and after the transition will likewise be practically equal. Furtheımore this transition may take place simultaneously in a large number of places. This is a rather extraordinary situation and the exceptionally low value of the activation energy for $Fe_3O_4$ will no doubt be connected with it. Thanks to this low activation energy the additional electrons contained in the $Fe^{2+}$-ions (compared with the $Fe^{3+}$-ions) are more or less free (since in the octahedral holes in $Fe_3O_4$ there are precisely just as many $Fe^{2+}$-ions as $Fe^{3+}$-ions, there are half as many electrons as there are octahedral holes available.) Just as is the case with the conduction electrons in a metal, these electroms can hardly be localised and are to be regarded rather as belonging to the whole of the ferri ions in the octahedral holes. They are continuously moving about in the lattice and thus give rise to the high conductivity of $Fe_3O_4$.

If the basic point (arbitrary distribution of the $Fe^{2+}$- and $Fe^{3+}$-ions among the octahedral holes) of this argument and the argument itself are correct, then when the $Fe^{3+}$-ions in the octahedral holes are replaced by other trivalent ions the conductivity should drop considerably. The fact that this is so has been confirmed by making mixed crystals of $Fe_3O_4$ with other spinels. An exceptionally simple case is that of the mixed crystal $Fe_3O_4$ and $FeAl_2O_4$. When these two compounds are mixed in the molecular proportion of 1:1 precisely one of every two

[6]) Large (negative) values of the t.c. are often met with in semi-conductors with a high specific resistivity. The fact that these two propertics are often coupled together can be understood from equations (2) and (4); in so far as the term log $\varrho_\infty$ in equation (4) may be ignored both the logarithm of the specific resistivity and the t.c. are proportional to the same quantity *b*.

[7]) Here one may indeed speak of metal ions, since the oxidic compounds referred to here in any case somewhat approximate the ionic bond type and may be included among the so-called polar compounds.

$Fe^{3+}$-ions is replaced by an $Al^{3+}$-ion. With this mixed crystal we have, on the one hand, been able to establish by X-ray analysis that $Al^{3+}$-ions have been substituted for the $Fe^{3+}$-ions in the octahedral holes, from which it was deduced with the aid of our rules that the structure of this mixed crystal is of the type $Fe^{3+}(Fe^{2+}Al^{3+})O_4$. On the other hand we found that the specific conductivity of this mixed crystal is only 0.004 $\Omega^{-1}$ $cm^{-1}$, whereas that of $Fe_3O_4$ is 200 $\Omega^{-1}$ $cm^{-1}$.

*Temperature-dependence of the resistance of mixed crystals of $Fe_3O_4$ and a non-conducting spinel*

In the light of what has been said above it will not be so very surprising that useful resistance materials can be obtained by making mixed crystals from $Fe_3O_4$ and another, non-conducting, spinel. Since as a rule any two spinels easily form mixed crystals we have in principle a very wide choice. For instance one can take another ferrite, or one may use spinels built up from entirely different ions, such as $MgAl_2O_4$, $MgCr_2O_4$, $ZnCr_2O_4$, $Zn_2TiO_4$, etc., or maybe a combination of these spinels. Speaking generally, one may say that — as is to be expected from the foregoing — the higher the content of foreign spinel the higher the specific resistivity and likewise the t.c.

An example is the system $Fe_3O_4$-$ZnCr_2O_4$, for which the relation between log $\varrho$ and $1/T$ is represented in *fig. 4*. Looking first at the right-hand part (room temperature and lower) we see that as the content of $ZnCr_2O_4$ increases the curves gradually assume a steeper slope, which means to say that the t.c. increases. Moreover, it is to be seen that at a certain temperature also the resistance increases with increasing $ZnCr_2O_4$ content. This latter behaviour is easily understood, for with increasing $ZnCr_2O_4$ content — according to our "rules" formulated above — the numbers of $Fe^{2+}$- and $Fe^{3+}$-ions (which continue to be mutually equal) decrease in the octahedral holes (see *table I*). Since, as we have seen, the presence of the "mixture" of these ions in the crystallographically equivalant sites is an essential condition for the low resistivity, the latter must therefore increase with the $ZnCr_2O_4$ content.

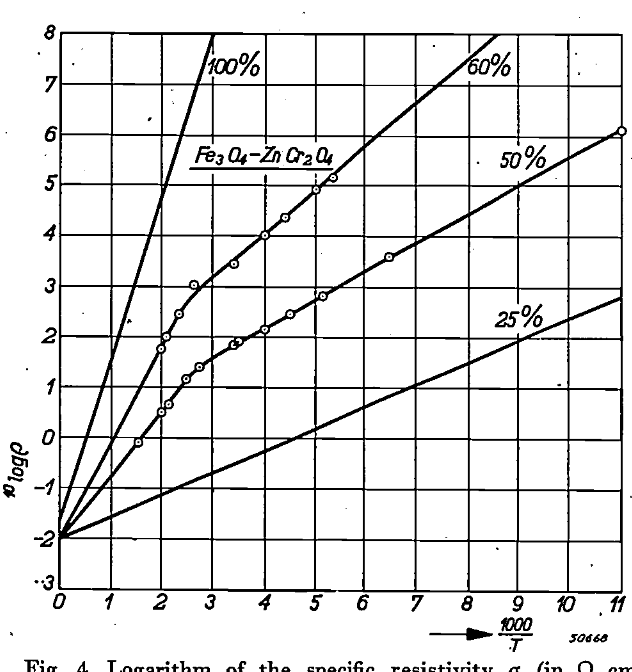


Fig. 4. Logarithm of the specific resistivity $\sigma$ (in $\Omega$ cm) as function of $1/T$ ($T$ absolute temperature) for the mixed crystal $Fe_3O_4$-$ZnCr_2O_4$ for different percentages of $ZnCr_2O_4$.

Table I. Distribution of the metal ions among the available tetrahedral and octahedral holes in the mixed crystal $Fe_3O_4$-$ZnCr_2O_4$ at various mixing ratios.

| Mixing ratio $ZnCr_2O_4 : Fe_3O_4$ | Tetrahedral holes | | Octahedral holes | | |
|---|---|---|---|---|---|
| | $Zn^{2+}$ | $Fe^{3+}$ | $Cr^{3+}$ | $Fe^{2+}$ | $Fe^{3+}$ |
| 1/4 : 3/4 | 1/4 | 3/4 | 1/2 | 3/4 | 3/4 |
| 1/2 : 1/2 | 1/2 | 1/2 | 1 | 1/2 | 1/2 |
| 3/4 : 1/4 | 3/4 | 1/4 | 3/2 | 1/4 | 1/4 |

The left-hand part of fig. 4, partly an extrapolation to $T = \infty$, shows a sudden bend in the curves for the highest $ZnCr_2O_4$ content; we shall not go into an explanation of this here.

Since the slope of the log $\varrho$-$1/T$ curves determines the t.c., it is important to note that the "fan" of these curves may have a very different appearance according to the variety of spinels used as the non-conducting component of the mixed crystal. We will therefore go more closely into a second system, that of $Fe_3O_4$-$MgCr_2O_4$. *Fig. 5* gives the log $\varrho$-$/1T$ curves for various compositions of this system. It is seen that close to the mixing ratio 1 : 1 there is a sudden increase in the slope of the curves, greater than the difference between neighbouring curves for other mixing ratios. The fan of curves is thereby clearly divided into two groups. The same appears also in *fig. 6*, where the "activation energy" $\varepsilon$ — which according to equations (1), (3) and (4) is proportional to the slope of these curves — has been plotted as a function of the composition of the mixture both for the system $Fe_3O_4$-$MgCr_2O_4$ and for the system $Fe_3O_4$-$ZnCr_2O_4$. Just before the mixing ratio 1 : 1 is reached the activation energy for the mixed crystal $Fe_3O_4$-$MgCr_2O_4$ appears to rise suddenly, contrary to the case of $Fe_3O_4$-$ZnCr_2O_4$, where the activation energy increases very gradually with the $ZnCr_2O_4$ percentage.

This behaviour of the $Fe_3O_4$-$MgCr_2O_4$ system may again be explained by means of our "rules" combined with what has been said above about the mechanism of conduction in $Fe_3O_4$. Contrary to the case of $Fe_3O_4$-$ZnCr_2O_4$, where a certain number of $Fe^{2+}$- and $Fe^{3+}$-ions are found in the octahedral holes for any mixing ratio, here the number of $Fe^{3+}$-ions in the octahedral holes for a $MgCr_2O_4$ content of 50% or higher is zero (see *table II*). The conduction mechanism which was so favourable for a low resistance can then no longer function and is apparently replaced by another less favourable mechanism; hence the jump observed in the activation energy.

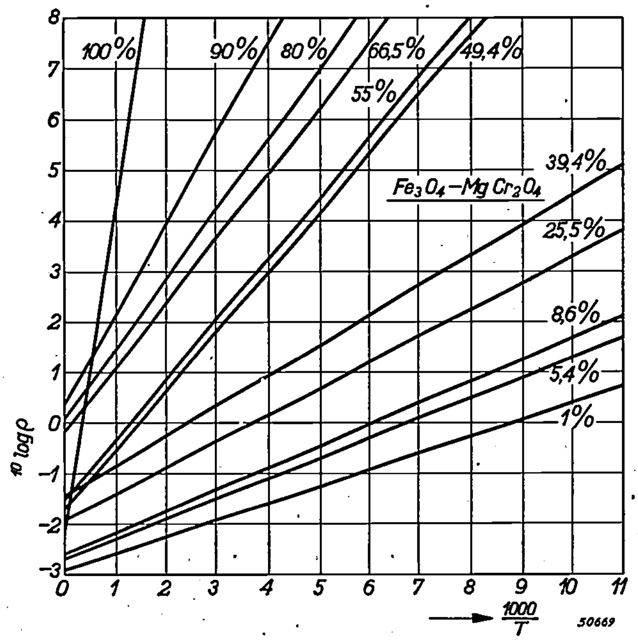


Fig. 5. Logarithm of the specific resistivity $\sigma$ (in Ω cm) as function of $1/T$ ($T$ absolute temperature) for the mixed crystal $Fe_3O_4$-$MgCr_2O_4$ for different percentages of $MgCr_2O_4$.

**Table II.** Distribution of the metal ions among the available tetrahedral holes in the mixed crystal $Fe_3O_4$-$MgCr_2O_4$ at various mixing ratios.

| Mixing ratio $MgCr_2O_4$ : $Fe_3O_4$ | Tetrahedral holes | | | Octahedral holes | | | |
|---|---|---|---|---|---|---|---|
| | $Fe^{3+}$ | $Fe^{2+}$ | $Mg^{2+}$ | $Mg^{2+}$ | $Cr^{3+}$ | $Fe^{2+}$ | $Fe^{3+}$ |
| 1/4 : 3/4 | 1 | 0 | 0 | 1/4 | 1/2 | 3/4 | 1/2 |
| 1/2 : 1/2 | 1 | 0 | 0 | 1/4 | 1 | 1/2 | 0 |
| 3/4 : 1/4 | 1/2 | 0 | 1/2 | 1/4 | 3/2 | 1/4 | 0 |
| | 1/2 | 1/4 | 1/4 | 1/2 | 3/2 | 0 | 0 |

For a content of $MgCr_2O_4$ greater than 50% the distribution of the metal ions among the available holes is not unambiguouly determined by our rules, as is evident also in table *II*. It is not impossible that there may then be a "mixture" of $Fe^{2+}$- and $Fe^{3+}$-ions in the tetrahedral holes; the latter, however, will be farther apart, so that the situation will presumably indeed be less favourable for good conduction than in the case where the octahedral holes contain this mixture.

This abnormal behaviour of the $Fe_3O_4$-$MgCr_2O_4$ system can be turned to good account if, for instance, a material is desired which has a specific resistivity of 100-1000 Ω cm at room temperature and at the same time a high t.c. for such a resistance value. Mixed crystals containing $MgCr_2O_4$ satisfy this requirement much better than those with $ZnCr_2O_4$, since the resistance value quoted corresponds to a non-conducting spinel content of 50-60%, for which content the slope of the curves for mixed crystals with $MgCr_2O_4$ is relatively great.

As regards the choice of $MgCr_2O_4$ or $ZnCr_2O_4$ as non-conducting spinel we would make the following observation. In order to regulate easily the resistivity of the mixed crystal it is desirable that the conductivity should originate exclusively from the presence of $Fe^{2+}$- and $Fe^{3+}$-ions in the crystal. Therefore in principle we should not use metals which, like iron, are apt to give ions of different valencies, such as Mn, Co,Ni, for then the conductivity might arise from the presence of, say, $Co^{2+}$- and $Co^{3+}$-ions, which is just what we want to avoid. From this point of view $MgAl_2O_4$ is therefore the most suitable non-conducting spinel, because Mg and Al can only be bivalent and trivalent respectively. It appears, however, that a mixture of $MgAl_2O_4$ and $Fe_3O_4$ is difficult to sinter into a homogeneous phase. For this reason we have used the more easily sintering chromites. Although chromium may occur in various valencies, under the experimental conditions in question we have not found any change in the valency of chromium in mixed crystals of $MgCr_2O_4$ with $Fe_3O_4$. $Zn_2TiO_4$ allows of a still lower sintering

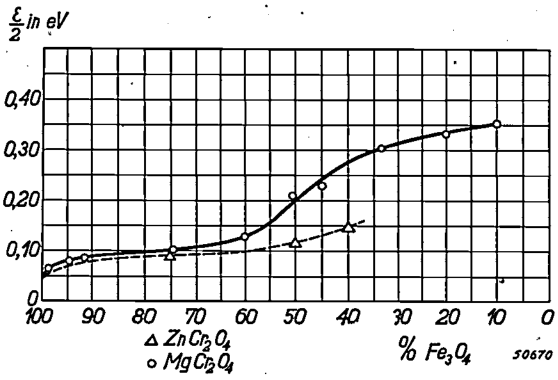


Fig. 6. The activation energy $\varepsilon$ (in eV) as function of the percentage of $Fe_3O_4$ for mixed crystals of $Fe_3O_4$-$ZnCr_2O_4$ (broken line) and $Fe_3O_4$-$MgCr_2O_4$ (fully-drawn line).

temperature but the t.c. is not as high as with $MgCr_2O_4$.

*Stability and reproducibility of h.t.c. resistors*

The principle upon which h.t.c. resistors are based — producing semi-conductors with a certain specific resistivity and temperature coefficient by preparing mixed crystals of conductive $Fe_3O_4$ and a non-conducting spinel — has various advantages over other already known means.

Resistors having the desired properties have previously been made from conducting and non-conducting grains homogeneously mixed and then sintered to a cohesive mass. As an example may be mentioned the mixture of Si powder and clay [8]). With this mixture the resistivity is governed by the transitory resistance from one Si grain to the other. The number of contact points naturally depends upon the mixing ratio, the resistivity increasing with the concentration of clay. This kind of resistor has all sorts of drawbacks: the resistivity is extremely sensitive to small fluctuations in the mixing ratio and is difficult to reproduce; moreover, there is always the danger of the resistance mass not being sufficiently homogeneous, so that the current may easily concentrate along certain paths, which may even lead to short-circuiting due to local heating and resultant reduction of resistivity [9]). Since silicon readily oxidizes in the atmosphere these resistors cannot be used without special precautions being taken.

A second method of making resistors suitable for the applications described is based upon the preparation of substances with a small deviation from the stoichiometric composition (*cf.* the article quoted in footnote [1]). In this kind of substance the resistivity usually drops with increasing deviation from the stoichiometric composition, particularly as long as the deviation is still small. In the case of an oxidic material, for instance NiO, in order to get a certain deviation from the stoichiometric state in the direction of an excess of oxygen, the oxygen pressure and the temperature have to be very accurately adjusted in the preparation of the mateiial. In practice this proves to be difficult and a reproducible adjustment to a certain resistivity is faced with insurmountable obstacles. Furthermore the slightest change in the oxygen content of the material during use will result in a relatively large change in the resistivity.

The question now arises whether the deviation from the stoichiometric state in the resistors made with $Fe_3O_4$ as basic component may not lead to difficulties. $Fe_3O_4$, too, can be prepared with a certain deviation from the stoichiometric state. In particular it is possible to get a deviation in the direction of an excess of oxygen. It is even possible to prepare any composition between $Fe_3O_4$ and $Fe_2O_3$. All these substances are homogeneous and have the spinel structure. The relative excess of oxygen, compared with $Fe_3O_4$, is obtained because the lattice contains too few $Fe^{2+}$-ions. The limiting case, $Fe_2O_3$ (the so-called $\gamma$-modification of ferric oxide) should therefore really be written in our notations as $Fe^{3+}(Fe^{3+}_{1.67})O_4$. The intermediate cases can be regarded as mixed crystals of $Fe_3O_4$ and $\gamma$-$Fe_2O_3$. All these mixed crystals, however, can only be prepared by some artifice or other and they are not stable; only those containing little $Fe_2O_3$, less than 10 %, can be prepared in the usual way by heating ferric oxide directly in an atmosphere containing more or less oxygen.

It is to be expected that the resistivity of these mixed crystals will increase with the $Fe_2O_3$ content, for according to our "rules" the higher the $Fe_2O_3$ content the fewer will be the number of $Fe^{2+}$-ions in the octahedral holes. And, if our explanation of the good conductivity in $Fe_3O_4$ is correct, this should be manifested in an increase of the specific resistivity. It is known that $\gamma$-$Fe_2O_3$ is almost an insulator. The values found for the specific resistivity of $\gamma$-$Fe_2O_3$ however are not all reproducible, firstly because this material cannot be sintered, so that measurements have to be taken with the powder, and secondly because traces of $Fe^{2+}$-ions considerably reduce the resistivity; the situation here is thus comparable to the abovementioned case of NiO, which also becomes conductive through small deviations from the stoichiometric state. With 100% $Fe_3O_4$ on the other hand we find that small deviations from the stoichiometric state cause but little change in the specific resistivity.

This insensitivity of $Fe_3O_4$ to small deviations from the stoichiometric composition usually prevails also in mixed crystals of $Fe_3O_4$ with non-conducting spinels. If, therefore, the sintering temperature, the oxygen content of the atmosphere and the duration of heating are so chosen that the mixed crystal has approximately the spinel composition, then any small variation in the process of manufacture will have little effect upon the resistivity. This explains for a large part the constancy and

[8]) These so-called "Starto" resistors have been dealt with in this journal (Philips Techn. Rev. 1, 205-210, 1936).

[9]) Of course such difficulties may arise with all inhomogeneous resistors, so that it is quite evident that there is an advantage in using mixed crystals, these being homogeneous.

also the great stability of the electrical properties of h.t.c. resistors. As a matter of fact the composition itself is quite stable: the densely sintered materials show relatively little reaction; moreover, the reactivity of $Fe_3O_4$ towards oxygen seems to be retarded by the presence of the other spinel. Consequently in many applications where the ambient temperature is low enough or the load sufficiently weak for the temperature not to rise above a few hundred degrees Centrigrade, these h. t. c. resistors can be used without any screening. (At higher temperatures they have to be used in a protective gas, *e.g.* nitrogen, or in a vacuum.) When it is considered, moreover, that any desired resistivity of the h. t. c. resistors can easily be obtained by a suitable choice of the mixing ratio of $Fe_3O_4$ and a non-conducting spinel in the mixed crystal, the great advantages that these new materials possess over those hitherto used will be still more apparent.

# THE DRYING LAMP AND ITS MOST IMPORTANT APPLICATIONS

by Th. J. J. A. MANDERS.

621.384.3:66.047

Nowadays electric drying lamps are being used for numerous drying processes in industry. These lamps are electric sources of infra-red of the incandescent type specially designed for the purpose. In this article first some properties of infra-red rays are brought to mind, it then being shown how infra-red rays are absorbed in a layer of water and how they then behave with respect to lacquers. The conditions are considered which have to be satisfied by a drying lamp. In the designing of the Philips drying lamp it has been endeavoured to make this lamp answer the requirements as closely as possible. All sorts of factories have already been equipped with large numbers of drying lamps for drying semi-manufactured goods and final products. Among the many possibilities offered by this lamp the drying of lacquers and of textiles and paper is particularly discussed.

## Introduction

Drying is a process applied in many industries. Investigations into the best methods of production have in many cases recently led to a conversion from the old drying methods (*e.g.* with a coal-fired oven) to the use of drying lamps. These are electric infra-red sources specially designed for the purpose.

The object of drying is usually to cause water or some other liquid to evaporate. In some cases this is attended by a chemical conversion, for instance oxidation in the hardening of enamels and polymerisation in the drying of synthetic lacquers.

As a rule a drying process is speeded up by heating. This heating is a question of the transmission of heat, and as is known this transmission can be brought about in three ways, by:

a) conduction,
b) convection,
c) radiation.

In the drying methods used in industry conduction plays a minor part. By convection heat is transferred from the source to the object to be heated through a medium (a liquid or a gas). The heat transfer coefficient between air and metal is low and between air and lacquer still lower. The heat then has to penetrate deeper into the object through conduction. The heat conduction coefficient of most substances that have to be dried is likewise low. In heating through convection these are two adverse factors which do not arise in the case of radiation, for in the latter case the heat is transferred entirely without the aid of a medium. It is for this reason that the method of radiation usually yields the best results. Under properly balanced working the energy is absorbed in the different layers of the object and at the same time evenly converted into heat down to a certain depth. In the layers in which this takes place the temperature rises rapidly and uniformly. This accounts for the fact that drying by radiation as a rule takes much less time than is the case by convection.

With the drying lamp the heat transfer takes place for the greater part through infra-red radiation. It is therefore advisable, before dealing with our subject proper, to give a brief account of some of the properties of infra-red radiation, as we shall have to refer to these when discussing this lamp and its applications.

## The properties of infra-red radiation

The wavelength of infra-red rays is longer than that of red light. The boundary between visible and invisible radiation lies at about 7600 Å. Infra-red radiation of a wavelength of $4.2 \times 10^6$ Å has, however, also been observed. Light waves are often compared with sound waves, and if we do so here we may say that visible light covers one octave and infra-red radiation nine octaves.

To judge the nature of the radiation of a solid body a graph may be drawn of the energy emitted as a function of the wavelength. It will then be seen that for any temperature of the body the spectral emission curve has a maximum for one certain wavelength.

In *fig. 1* spectral emission curves are given for a black body radiator with a temperature of 6000 °K, for a photo lamp ("Photolita"), a normal incandescent lamp ("Bi-Arlita") and the drying lamp discussed in this article. For a temperature of 6000 °K the maximum of the emission lies at a wavelength of 4800 Å. In the case of the "Photolita" lamp the filament temperature is 3400 °K and $\lambda_{\max} = 8150$ Å, whilst these figures for the 100 W "Bi-Arlita" lamp are 2850 °K and 9500 Å. The temperature of the tungsten coil in the drying lamp is 2200 °K and the wavelength of the maximum emission is 12 000 Å.

For black body radiation there is a simple relation

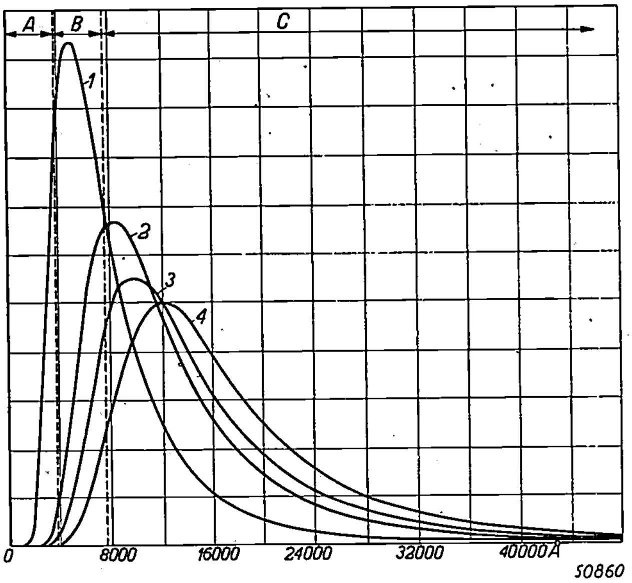


Fig. 1. Relative spectral energy distribution for some radiators at an equal quantity of total energy emitted:
*1)* for a black body at a temperature of 6000 °K;
*2)* for the coil of a "Photolita" lamp (temperature 3400 °K);
*3)* for the coil of a 100 W "Bi-Arlita" lamp (2850 °K);
*4)* for the 250 W drying lamp with a temperature of 2200 °K.
*A* is the range of ultra-violet radiation, *B* that of visible light and *C* that of infra-red radiation.

between the absolute temperature $T$ of the incandescent body and the wavelength of the maximum emission. This is Wien's displacement law:

$$\lambda_{\max} T = 288 \cdot 10^5,$$

where $\lambda_{\max}$ is expressed in Angström and $T$ in degrees Kelvin. Since all known incandescent bodies differ more or less from a black body radiator, their displacement laws differ from Wien's. According to Geiss, the displacement law for a disc of tungsten in the temperature range from 2200 to 3400 °K is [1]):

$$\lambda_{\max} T \doteq 263 \cdot 10^5 + (T - 2500) \cdot 10^3$$

and for a coiled tungsten wire, where the radiation from the part between the windings of the coil may be regarded as being practically black, for the same interval:

$$\lambda_{\max} T = 268 \cdot 10^5 + (T - 2500) \cdot 10^3.$$

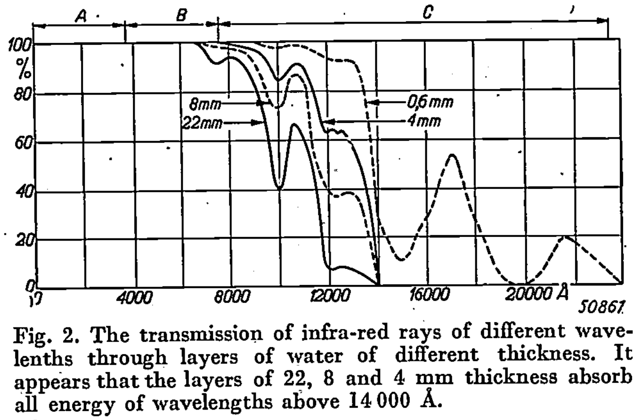


Fig. 2. The transmission of infra-red rays of different wavelenths through layers of water of different thickness. It appears that the layers of 22, 8 and 4 mm thickness absorb all energy of wavelengths above 14 000 Å.

With the help of this equation it is possible to design a lamp having a certain value for the wavelength of its maximum emission.

Since the evaporation of water is an important task of the drying lamp it is worth while investigating what happens when infra-red rays are directed upon a layer of water.

Water readily lets visible light pass through it but not so infra-red rays. Investigations with shallow depths of water have shown that the most important absorption bands in the infra-red part of the spectrum lie in the wavelengths of 15 000, 20 000, 30 000, 47 500 and 60 000 Å. *Fig. 2* gives a graphic representation of the transmission factor of infra-red radiation of different wavelengths through layers of water of different thickness. It is to be seen that the radiation with wavelengths above 14 000 Å is practicaly entirely absorbed by the layer of water unless the latter has very little thickness, *viz.* less than 0.5 mm.

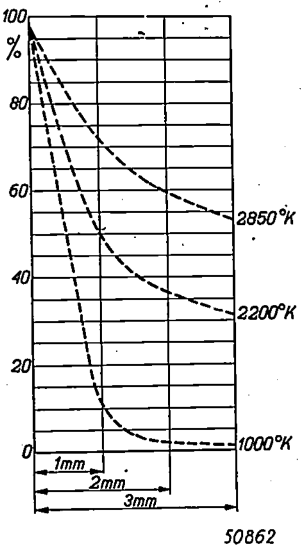


Fig. 3. Transmission of the energy emitted by three different infra-red sources (normal incandescent lamp, drying lamp, heating element of 1000 °K) through successive depths of water. Of the radiation from a drying lamp 50%, 13% and 5% is absorbed in the first, second and third millimeter respectively.

Fig. 1, however, shows that the lamps commonly used radiate a continuous spectrum, thus emitting energy of quite different wavelengths. It is therefore worth investigating at what depths of water the rays emitted from various infra-red sources are mainly absorbed. The graphs in *fig. 3* indicate the

[1]) W. Geiss, Trocknen mit infraroten Strahlen, Electrical Service 20, 287-293, 1945/46.

transmission as a function of the thickness of the layer when irradiating with a heating element (temperature 1000 °K), a tungsten filament lamp (2850 °K) and the Philips drying lamp (2200 °K).

Drying lamps are also being widely used for the drying of lacquers, and we therefore also have to find out how infra-red rays behave towards the most commonly used kinds of lacquers. This is shown in *fig. 4*, where curve *1* gives an idea of the transmission of infra-red energy through lacquers as a function of the wavelength. It is to be borne in mind, however, that the shape of this curve depends upon the thickness of the layer and the kind of lacquer. It appears that with lacquers the infra-red energy with a wavelength < 14 000 Å is almost entirely absorbed, but if a longer wavelength is used the amount of energy allowed to pass through increases. By way of comparison the spectral energy distribution of the Philips drying lamp is also drawn in the same diagram (curve *2*).

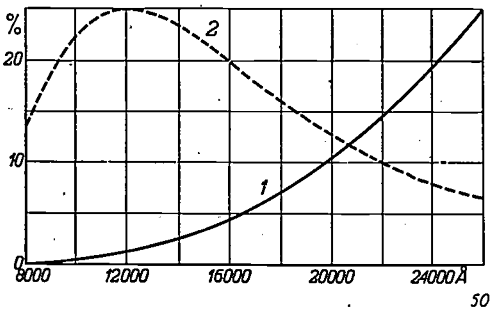


Fig. 4. Transmission of infra-red rays when drying lamps are used for lacquers. Curve *1* gives an idea of the transmission of a certain kind of lacquer for infra-red energy of different wavelengths. For comparison curve *2* gives the spectral energy distribution of the Philips drying lamp.

Having now considered the behaviour and properties of infra-red radiation, it is possible to indicate how a drying lamp can best be constructed.

## The requirement to be fulfilled by a drying lamp

When one is about to design a drying lamp the first question to be considered is what the wavelength of the maximum emission is to be. We have seen that this depends upon the absolute temperature of the filament of the lamp.

It is of primary importance that the radiation should be converted into heat in the object to be dried as completely as possible and in exactly the right place. If the absorption takes place only on the surface it means that the underlying layers have to be heated through conduction. Since most substances that have to be dried are bad heat conductors this is not desirable. With lacquers we have the further difficulty that if the surface is dried too intensively the skin becomes tough while the underlying layers are still soft. On the other hand too high a transmission means that energy will be lost or the sub-layers heated too much. It is therefore a matter of great importance to balance carefully the ratio of absorption to transmission. The ideal solution is to have sufficient energy penetrating close down to the sub-layer, so that the whole of the layer to be dried absorbs the radiation while the sub-layer is moderately heated.

When the drying lamp has to be used for evaporating water we can deduce from fig. *2* the wavelength required for the maximum emission and thus also the temperature of the lamp filament. The graph shows that the energy of infra-red rays having wavelengths longer than 14 000 Å is practically entirely absorbed in the top layer, except in the case of a very thin layer of water, *e.g.* < 1 mm. Therefore, to ensure that the rays penetrate deep enough into the layer of water, $\lambda_{\max}$ should be chosen lower, but the wavelength may not be so short that too much radiation is allowed to pass through.

From fig. 3 it is to be seen that a radiator with a temperature of say 2850 °K would not be suitable for our purpose, because the water would allow too much of the energy to pass through it [2]).

A second objection attaching to such a high temperature is the unfavourable effect it has upon the life of the lamp, which must be economically justified.

A $\lambda_{\max}$ of 12 000 Å appears to be a good choice for evaporating water. The filament temperature corresponding to this wavelength is 2200 °K. Fig 3 shows to what extent the energy from such a lamp is absorbed in the successive layers. The first millimeter of water absorbs 50%, the second 13% and the third 5%.

Now that we have seen what is the most suitable filament temperature for evaporating water, the next question to be considered is how a radiator should be constructed for the drying of lacquers. Fig. 4 supplies the answer. It is seen that the top of the distribution curve should anyhow lie below 15 000 Å, because otherwise the transmission would be too high. If this condition is satisfied it appears further that for this purpose the value of $\lambda_{\max}$ is not very critical. Closer investigation has shown that it makes practically no difference in the drying time whether $\lambda_{\max}$ is 12 000 Å or slightly lower or

[2]) It is in connection with this that the "Infraphil", an apparatus for medical therapy with infra-red rays, is fitted with a lamp that has a filament temperature of 2800 °K. The radiation from this lamp has to penetrate the water-containing epidermis to impart heat to the underlying tissues. See Philips Techn. Rev. 8, 177-182, 1946 (No. 6).

higher. Apparently, therefore, the same lamp can be used for lacquers as for water.

Now that we have found the optimum wavelength for the maximum emission we shall consider the further requirements that the source of radiation has to satisfy.

The temperature found is so high as to make it necessary to envelop the filament in a glass bulb. In order to reduce evaporation this bulb has to be filled with an inert gas, just as is the case with normal incandescent lamps. Further, the lamp should be constructed in such a way that the largest possible part of the radiation is directed upon the object to be dried. The part of the radiation that is not emitted in the desired direction must therefore be reflected. In the construction of a reflector, however, it must be taken into account that the reflection coefficient of different materials depends upon the wavelength of the radiation. *Fig. 5* gives a general idea of the reflection coefficients of some suitable materials for visible light and for the most important part of the infra-red radiation (7600 up to 20 000 Å). It appears that within the range in question silver has a great reflectivity (92-98%), as is also the case with gold (84-98%), whilst nickel (65-80%) and chromium (58-70%) are less suitable. These graphs also show that evaporated aluminium (unlike polished aluminium) is an exceptionally good reflector.

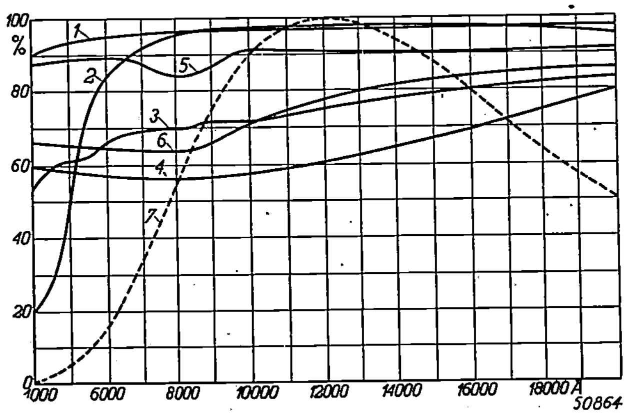


Fig. 5. The reflection coefficient, expressed as a percentage of the incident energy, of some metals that could be used as mirror of a drying lamp, *1* is silver, *2* gold, *3* nickel, *4* chromium, *5* evaporated aluminium and *6* polished aluminium. For comparison curve *7* gives the spectral energy distribution of the radiator.

The choice between silver, gold and aluminium as material for the reflector depends also upon the question whether the mirror is to be applied inside the bulb or outside. In America an external reflector is very much used. When an external reflector is applied in the drying, for instance, of lacquers usually a gold mirror is recommended because of the influence of the lacquer vapours upon the reflector. In Europe an internal reflector is most commonly used, because of the following advantages:

1) it retains its high reflection coefficient much better than an external reflector, because it is not subject to chemical or mechanical influences and is free from dust;
2) it takes up less space and requires no special means of attachment;
3) the whole radiation passes through the wall of the bulb only once. This means less absorption in the glass wall; in the case of an external reflector part of the radiation passes through the wall of the bulb three times.

For a radiator with an internal reflector gold, silver and aluminium vaporised in vacuum are all equally useful as regards reflection. Since gold, however, is expensive and silver is in this case less satisfactory from the point of view of manufacture, aluminium is used. The aluminium is evaporated and condensed on the inside of the glass bulb.

For a proper bundling of the rays it is advisable to shape the reflector like a paraboloid and to place the radiating body as far as possible in its focus.

If the rays pass through only once the question of absorption in the glass wall is not very important. A graph of the transmission factor of normal calcium glass of a thickness of 1 mm for various wavelengths [3]) shows that appreciable absorption begins at a wavelength of 25 000 Å. Within the range that counts most for our purpose the transmission of normal glass is good.

## Properties of the Philips drying lamp

We will now give a brief description of the latest type of Philips drying lamp. On account of the considerations mentioned above and also on specifically lamp-technical grounds which cannot be analysed here, this lamp has been given a shape as outlined in *fig. 6* and shown in the photograph in *fig. 7*. The glass bulb has a parabolic part and a spherical widening at the level of the filament.

The incandescent body is a tungsten coil which takes up 250 W. The temperature of 2200 °K determines — as already remarked — also the life of the lamp. The life of an ordinary incandescent lamp (for instance the 100 W "Bi-Arlita", temperature 2850 °K) is 1000 burning hours, whereas that of the Philips drying lamp is 5000. The bottom of the bulb is relatively strongly convex, giving it a robust form. To get a uniform irradiation the wall of the bulb is slightly satin-frosted and the incandescent

[3]) Philips Techn. Rev. 8, 180, 1946 (No. 6).

element is of the smallest possible dimensions. This drying lamp may be provided with either a screw cap or a bayonet cap. At temperatures above 200 °C the usual method of cementing the lamp in the cap is not reliable and for that reason a lamp base has been developed which allows of this joint being made without the use of cement.

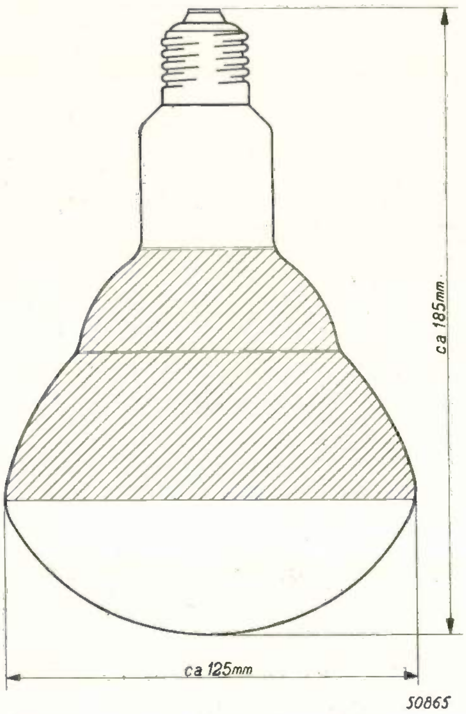


Fig. 6. Dimensions and shape of a Philips drying lamp.

With the aid of a thermopile the intensity of radiation has been measured at various points of planes perpendicular to the axis of the lamp, the planes being at distances of 10 to 100 cm away from the front of the lamp. *Table I* gives the results of these measurements expressed in mW/cm². As a comparison, the radiation intensity of the sun at its zenith under favourable conditions is 100 mW/cm². *Fig. 8* gives a graph of the distribution of radiation intensities. The table and the graph show there is uniform radiation intensity on planes 20-60 cm away from the bottom of the bulb within a cone having an apex angle of 30 degrees. With the aid of these data it is possible to determine the optimum arrangement of a number of drying lamps in relation one to the other and to the object.

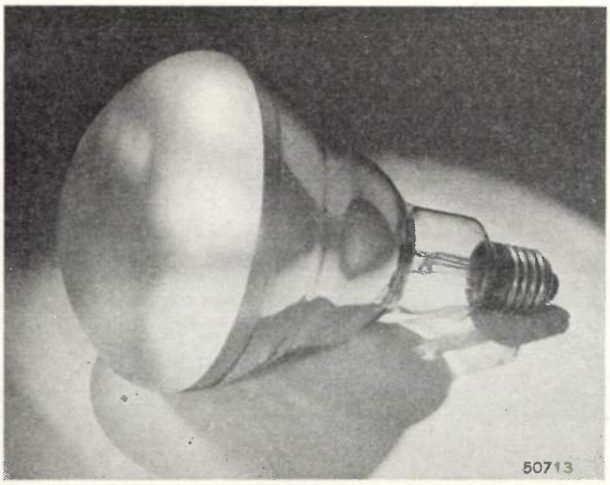


Fig. 7. The 250 W Philips drying lamp.

It will seldom occur that only one drying lamp is used for a certain purpose. In industries it is usual to work with a drying plant containing a number of lamps.

The drying time depends upon various factors:

a) the nature of the substance to be dried,
b) the nature of the sub-surface (its thickness and heat conductivity),
c) the distance between the lamps and the object and the number of lamps used, or in other words the energy supplied per cm².

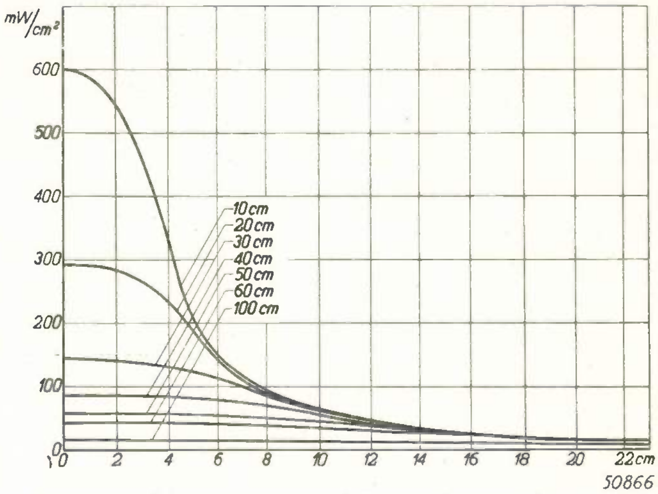


Fig. 8. Graphic representation of the distribution of the radiation intensities (in mW/cm²) in planes perpendicular to the lamp axis at different distances away from the lamp varying from 10 to 100 cm. The horizontal axis gives the distance between the point in question and the lamp axis.

Table I. Radiation intensity at different distances from the lamp expressed in mW/cm². The distance from the front of the lamp to the plane of measurement (in cm) is represented by *h* and the distance from the point of measurement to the axis of the lamp (in cm) by *r*.

| *h* \ *r* | 0 | $2^1/_2$ | 5 | 10 | 15 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 600 | 500 | 250 | 65 | 32 | 19 | 10 | 5 | 3 | 2 |
| 20 | 290 | 275 | 190 | 65 | 31 | 19 | 10 | 5 | 3 | 2 |
| 30 | 145 | 140 | 125 | 65 | 30 | 18 | 10 | 5 | 3 | 2 |
| 40 | 85 | 85 | 80 | 56 | 29 | 18 | 9 | 5 | 3 | 2 |
| 50 | 56 | 56 | 53 | 43 | 26 | 17 | 9 | 5 | 3 | 2 |
| 60 | 42 | 42 | 40 | 33 | 23 | 15 | 9 | 5 | 3 | 2 |
| 100 | 14 | 14 | 14 | 14 | 13 | 12 | 9 | 5 | 3 | 2 |

In factories the objects to be dried are usually placed on a conveyor belt which passes through a tunnel in which a number of lamps are mounted in such a way that their beams are directed upon the belt. The number of lamps, the distances away from the object and the speed of the conveyor belt should be so chosen that the drying is completed by the time the object(s) have reached the end of the tunnel, taking into account the maximum tempcrature permissible for the object(s) to be dried.

*Fig. 9* shows such a drying tunnel.

Such an installation can be built up from building elements consisting of the actual assembly unit and a lampholder (see fig. 9). These building elements can be combined in a simple manner to form larger units in flat or curved planes, an example of which is given in *fig. 10*.

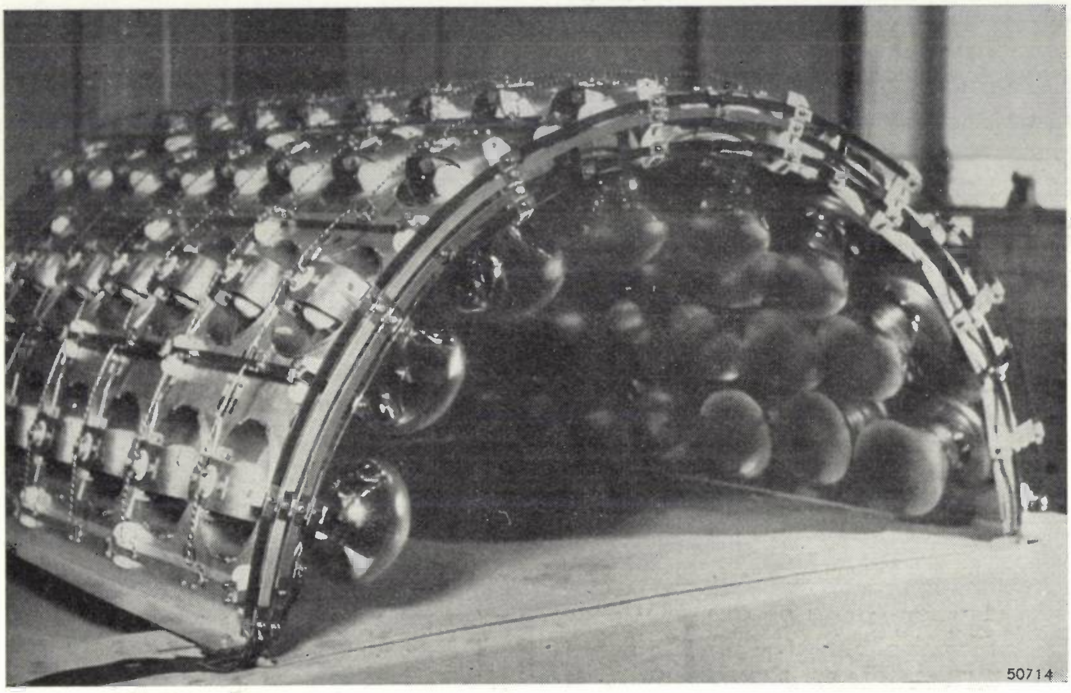


Fig. 9. A drying tunnel with Philips drying lamps, built up from "building elements" each containing one drying lamp.

## Applications

### *The drying of lacquers*

The drying of lacquered objects is an important application of drying lamps.

Lacquers contain a non-volatile binding agent and a volatile solvent. Furthermore most lacquers contain a pigment, generally an anorganic compound that does not dissolve but whose particles are mixed in the lacquer and give it a certain colour.

The most important kinds of lacquers have an oil, synthetic or cellulose base. The drying and hardening processes of each of these kinds are based on different principles. In all three cases, it is true, the solvent has to be evaporated, but oil-base lacquers harden through oxidation of the binder (the oil), whilst in the case of synthetic-base lacquers (so named because they contain synthetic resins) although the binder is also hardened to a certain extent through oxidation it takes place mainly through polymerisation. In the case of cellulose-base lacquers it is only a matter of evaporating the solvent.

Both the evaporating of the solvent and the oxidising or polymerising of the binder are processes which can be accelerated by raising the temperature, though only to within certain limits, because too rapid vaporisation causes blisters in the lacquer surface and too high temperatures may also change the colour.

According to the temperature at which the drying takes place, lacquers can be divided into those which dry in the atmosphere, those which dry fairly well already at room temperature, and baking enamels which only dry at a certain temperature. For the last mentioned category the drying process has to take place between that specific temperature and the maximum temperature permissible for the ob-

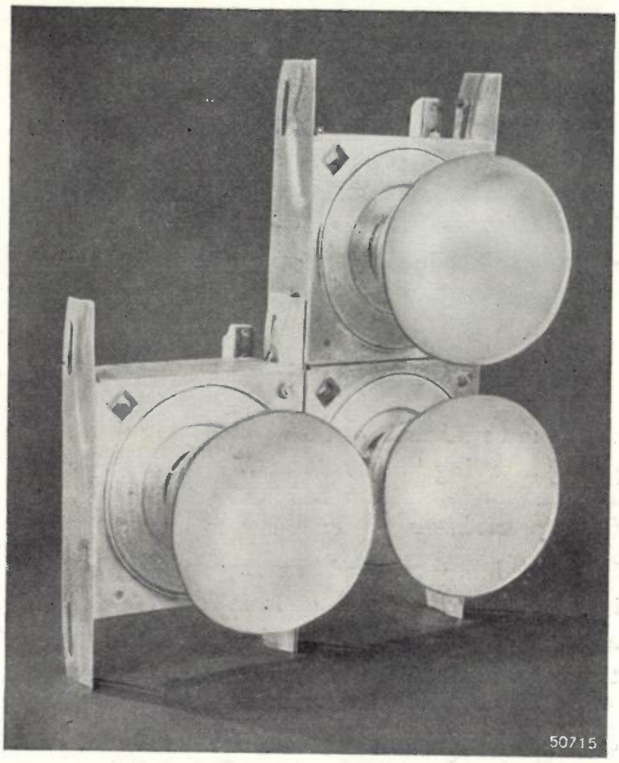


Fig. 10. Combination of three building elements (new type) each fitted with a drying lamp.

ject. It is recommendable that the temperature actually applied approaches the maximum as near as possible. Experiments have shown that during the drying process the temperature of the layer of lacquer remains fairly constant. Baking enamels of an alkyd-resin base of different colours have the same hardness after 10 minutes irradiation as after 60 minutes processing in a baking oven. This opens great possibilities for increasing production or for the saving of space.

One of the first manufacturing firms in America to apply drying lamps on a large scale was the Ford Motor Co. They use drying lamps for drying lacquered motorcar bodies, which was formerly done with steam. *Fig. 11* shows a drying tunnel installed by the Ford Motor Co. at Detroit for car bodies carried through it continuously on a conveyor.

About 1939 the Ford factories were already using some 35 000 drying lamps. For the drying of the ground lacquer an oven is used, built in two halves placed around the car body suspended from a conveyor. The oven moves along with the conveyor for 7 minutes, after which time the ground lacquer is dry, whereas formerly this took 30 minutes. The drying time of the top coat has even been reduced from 80 to 14 minutes.

Apart from motorcar factories drying lamps are nowadays being very widely used in numbers of other industries where the drying of lacquers plays an important part (*e.g.* in aircraft factories).

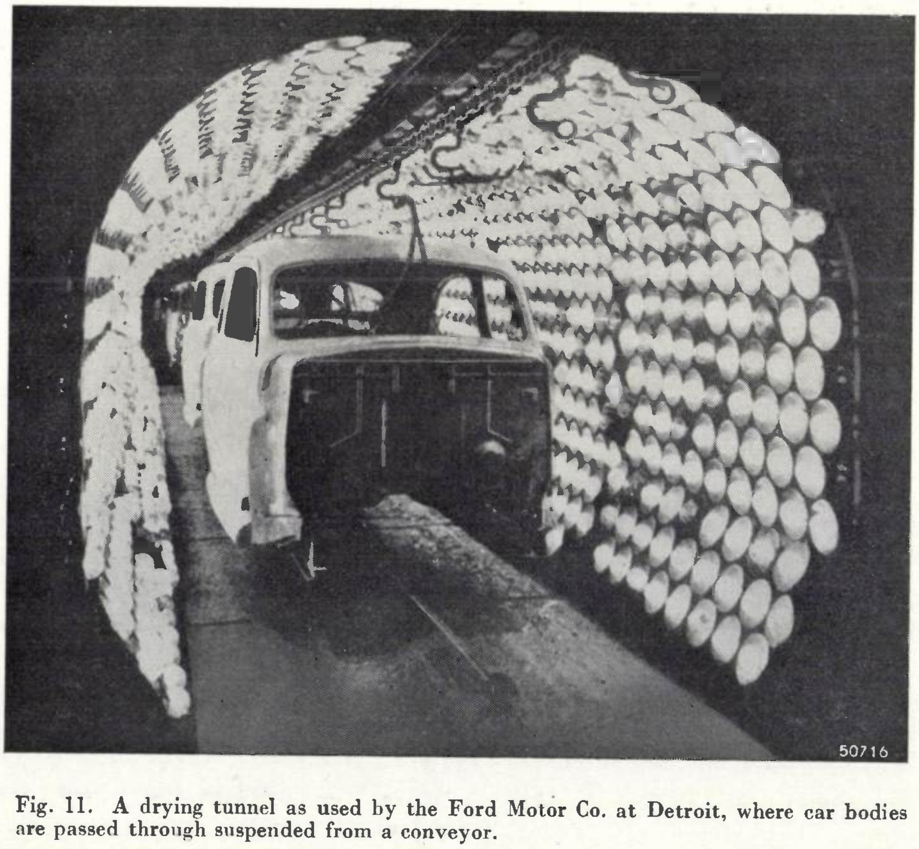


Fig. 11. A drying tunnel as used by the Ford Motor Co. at Detroit, where car bodies are passed through suspended from a conveyor.

### *The drying of textiles and paper*

In the textile and paper-making industries drying lamps are used to evaporate superflous water from semi-manufactured goods or the finished products.

The "Vezelinstituut" at Delft, a laboratory for testing textile yarns, have conducted a series of experiments in order to find out in how far the drying of textiles by means of drying lamps had an adverse effect upon the breaking strength, the elongation upon fracture and the tint of dyed and undyed woollen, cotton, linen and rayon fabrics. As a result of a large number of practical tests it was established that irradiation with drying lamps has no adverse effect whatever upon these properties of those four materials. Further it was investigated how fabrics can be most economically dried with the aid of drying lamps.

Drying lamps are now being used in several textile mills. To mention an example: in one factory the yarns wound on spools used to be dried in an oven for anything from 4 to 16 hours, 45 kg being dried in 4 hours, during which time 7% of moisture was extracted. Now the yarns are rewound at a rate of 27.5 m/min. and passed through in a broad band side by side underneath a wall of drying lamps having a total capacity of 11 kW. Underneath the yarns is a reflecting plate which throws back onto the yarns the rays that have passed through. In this manner it is possible to dry in 1 hour the same quantity of yarns as it used to take 4 hours to dry.

Drying lamps are also being applied with great success in paper making, as may be illustrated by two examples.

When the temperature of the paper track is raised to 55-75 °C immediately before the last stage of wet pressing the viscosity of the water is reduced and thus the water can be more easily pressed out. This has been known for a long time but it was very difficult to bring about this increase of temperature in a practicable manner. Drying lamps afforded an excellent solution. It was found that with these lamps and under the same compression 25-36% more water could be pressed out of the paper,which means a considerable increase in efficiency.

In another factory the after-drying of the paper used to be done with electric resistance elements, the plant having a capacity of 18.5 kW. After drying lamps had been installed a capacity of 4.2 kW was found sufficient for the same drying time, thus giving a saving in current of 75%.

### *Other possibilities*

Drying lamps are used in many other fields, as for instance for the drying of leather in tanneries and in boot and shoe factories, for the drying of tobacco and also of developed films.

One important application may be that of grass drying, which is still in an experimental stage. If it comes to be applied on a large scale it will mean a revolution in agriculture, for whereas there can only be one or two hay crops in a season young grass can be mown and dried five to seven times. Preliminary experiments have shown that drying lamp installations give about the same yield as drying machines fired with coke. When drying lamps are used, however, the product is of a better quality and, moreover, the cost of the plant is estimated to be much lower.

Drying lamps have also been used for drying vegetables and preserving fruit.

The future will undoubtedly reveal many more fields of application for these drying lamps.

VOL. 9. No. 9, pp. 257-288 1947/1948

Philips Technical Review

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

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

# CHROMATIC ADAPTATION OF THE EYE

by P. J. BOUMA † and A. A. KRUITHOF.

535.733:612.843.31

When two kinds of light are compared successively one finds that the sensations of colour brought about by coloured objects do not as a rule differ much, although, as a simultaneous comparison teaches, there may indeed be a very great difference in the colour stimulus of the object when under the two kinds of light. This difference between the colour sensation and the colour stimulus is due for the greater part to the chromatic adaptation of the eye. Experiments for determining the hues of the colours shown by a series of pigments (Ostwald cards) under different kinds of light have furnished new data concerning chromatic adaptation. Some conceptions of chromatic adaptation known from literature have been checked with the data collected from the experiments. It appears that large deviations arise when predicting according to the laws derived from these conceptions the hue changes accompanying a change of the kind of light. From the experimental data a new and much better law is obtained, which can be formulated by means of a hyperbola in the mixture diagram. The hyperbola passes through the two colour points representing the kinds of light compared and, according to an old theory of Helmholtz and von Kries regarding chromatic adaptation, should further pass through three fixed points (the colour points of the hypothetical physiological primary colours.) The experiments, carried out with eight sources of light, prove that such is the case with a reasonable approximation, though the three fixed points found here deviated considerably from the primary points determined by other methods. With the aid of the three fixed points, the hyperbola required for the comparison can be constructed without any further empirical data for any pair of kinds of light.

## Colour stimulus and colour sensation

In a previous article in this journal [1]) it was explained that a distinction has to be made between the conceptions of "colour stimulus" and "colour sensation". A spot of light having a given spectral distribution, *i.e.* a given colour stimulus, evokes a colour sensation depending upon many other conditions as well, especially upon the colour of the surroundings. This fact is connected with the phenomenon of chromatic adaptation, *i.e.* the adaptation of the eye to the colour of the surroundings.

The influence of the surroundings is only fully felt when spots of light are compared successively. This manner of comparison, where the surroundings are changed and the eye is given time to adapt itself to the new surroundings, is therefore characteristic for the investigation of colour sensations. The colour stimulus associated with a spot of light, on the other hand, is determined by the usual methods of colorimetry, which are based upon simultaneous comparisons under standardised conditions.

A colour stimulus is usually indicated by means of the trichromatic coefficients in a mixture diagram completed by its luminosity [2]). One may also use, however, as characteristic features the dominating wavelength, colorimetric purity and luminosity. A colour sensation, on the other hand, is characterized by hue (indicated by the name of the colour,) saturation and lightness.

[1]) P. J. Bouma and A. A. Kruithof, Colour stimulus and colour sensation, Philips Techn. Rev. 9, 2-7 1947, (No. 1).

[2]) See *e.g.* the article by P. J. Bouma in Philips Techn. Review 2, 39-46, 1937; more comprehensively in P. J. Bouma, Physical aspects of colour, Philips Techn. Library, publ. Meulenhoff, Amsterdam, 1947, especially chapters V and VI.

An important practical question is: what changes do the colour sensations brought about by surrounding objects undergo when a given source of light is replaced by another, for instance when passing from daylight to incandescent lamplight? Clearly we have here a case of successive comparison. In the preceding article it was already pointed out that in this case the changes observed in colour sensation prove to be surprisingly small. At the same time it was stated, however, that so far very little research has been done in regard to these changes which go to show the influence of chromatic adaptation. A series of experiments will now be described which have been carried out in order to investigate how the most important characteristic of the colour sensation brought about by an object — the hue — varies when the source of light is changed. The results of the experiments will then be systematically discussed and an attempt will be made to build up on these considerations a general picture of chromatic adaptation [3]).

The ultimate object is to be able to predict colour sensations as observed in practice from the purely physical data of the coloured objects and the sources of light.

### Experimental determination of hues under different kinds of light

The experiments were conducted in the following manner: A large sheet of white paper, almost as large as the whole field of vision, was illuminated with a certain kind of light. The illumination was approximately 140 lux. After the observer's eye had become adapted both to the level of brightness and to the colour of his field of vision, various coloured objects were placed in the middle of the paper in succession and the observer was asked to name the hue of the colour which he thought each object had under the given conditions.

The coloured objects used were the 100 cards of a complete circle of Ostwald's colour atlas, namely those of the nc-circle [4]). This is a series of rather saturated colours, which give a sensation of saturation that is fairly constant and the hue of which (in daylight) within the range of the numbers 0-99 passes from yellowish green through yellow, orange, red, purple, blue and green and back again to yellowish green. The order in which the cards were shown was quite arbitrary. Since the observer did not know what cards were being shown to him there was no question of suggestion ("that is a green card, so I shall have to call it green").

When naming the hues the observer had to make a choice out of the 36 numbered hue names of the following system:

Principal hues: pure yellow (1), orange (7), red (13), purple (19), blue (25), and green (31).

Intermediate hues: for instance between blue and green: blue with a little green (26), blue with a considerable amount of green (27), bluish-green where neither the blue nor the green predominates (28), green with a considerable amount of blue (29), green with a little blue (30).

In this way each card, which in the atlas has one of the numbers $j$ = 0--99, could be given a hue number $N$ = 1—36.

After the whole series had been gone through the test was repeated with a different kind of light. The lights used were in the first instance artificial daylight ($K$) from a tubular fluorescent lamp of so-called daylight colour, and incandescent lamplight ($G$) from 150 W "Bi-Arlita" lamps (colour temperature 2850 °K). Later on other kinds of light were used.

The experiments with the lights $G$ and $K$ were carried out with three observers $A$, $B$ and $C$ having normal colour vision and another observer $D$ having abnormal colour vision (deuteranomaly). They were repeated a number of times under each kind of light. The average of $N$ was taken for each observer under a given light over the various series of tests. A mathematical process was then applied in order to eliminate as far as possible the change errors in the relation between $j$ and $N$. *Table I* gives in abbreviated form the relation thus found between $j$ and $N$ for the three observers $A$, $B$ and $C$ under artificial daylight $K$ and under incandescent lamplight $G$.

Referring to this table we observe the following:

1) Changing over from one kind of light to the other causes the hue number $N$, for most cards, to differ by only one unit or less. This is a quantitative expression of the fact already mentioned that thanks to chromatic adaptation a change of light mostly causes only small changes in the colour sensations, Differences larger than $N_k - N_g = 1$ occur only in the range of the colours between red, purple and blue. We know from our every-day life that colours in this range always show great differences

[3]) Some of the experiments, as also the fundamentals of the new method of representing chromatic adaptation have already been described in P. J. Bouma and A. A. Kruithof, Hue estimation of surface colours as influenced by the colours of the surroundings, Physica, The Hague, 9, 957-966, 1942 and 10, 36-45, 1943. Similar investigations have recently been made by H. Helson and J. Grove: Changes in hue, lightness, and saturation of surface colours in passing from daylight to incandescent lamp light, J. Opt. Soc. Amer. 37, 387-395, 1947, (No. 5).

[4]) This series is also used for the experiments described in the previous article [1]).

between daylight and incandescent lamplight.

2) There is a fairly wide discrepancy between the names given by the three observers for the hue of the same card. For the greater part such differences are not due to actual differences in colour vision but simply to the fact that each one of us has learnt to connect certain sensations of colour with the usual hue names in a somewhat different way. This effect can be eliminated by starting with a given hue sensation (*i.e.* a given value of $N$) and asking oneself the question which card (number $j_g$) gives this hue sensation under incandescent lamplight and which card (number $j_k$) gives the same hue sensation under daylight. The difference $j_g$—$j_k$ then serves as a measure for the change of hue sensation when changing from daylight to incandescent lamplight. In *fig. 1* the difference $j_g - j_k$ (as an average of the three observers $A$, $B$ and $C$) derived from table I is plotted as a function of $j_g$.

Table I.

Hue numbers $N$ (1 to 36) given by the observers $A$, $B$ and $C$ to the 100 Ostwald cards of the nc-series numbered $j = 0$ to 99 when observed under artificial daylight ($N_K$) and under incandescent lamplight ($N_G$).

| $j$ *) | Observer $A$ | | Observer $B$ | | Observer $C$ | |
|---|---|---|---|---|---|---|
| | $N_K$ | $N_G$ | $N_K$ | $N_G$ | $N_K$ | $N_G$ |
| 0 | 33.92 | 33.86 | 35.35 | 34.75 | 34.93 | 34.69 |
| 4 | 35.48 | 0.05 | 1.08 | 1.07 | 1.36 | 1.51 |
| 8 | 2.28 | 2.50 | 3.31 | 3.36 | 3.20 | 3.26 |
| 13 | 4.95 | 5.06 | 6.84 | 7.05 | 6.98 | 6.42 |
| 17 | 7.81 | 6.88 | 8.89 | 8.59 | 9.11 | 8.40 |
| 21 | 11.84 | 10.19 | 12.19 | 11.54 | 11.37 | 10.26 |
| 25 | 15.91 | 13.56 | 14.12 | 13.42 | 14.82 | 13.16 |
| 29 | 18.82 | 14.75 | 14.90 | 14.42 | 16.09 | 14.36 |
| 33 | 19.86 | 17.91 | 15.62 | 15.14 | 18.19 | 16.22 |
| 38 | 21.69 | 20.31 | 18.09 | 16.20 | 19.99 | 18.82 |
| 42 | 22.56 | 21.20 | 20.87 | 18.30 | 21.44 | 19.98 |
| 46 | 23.11 | 22.11 | 22.60 | 20.85 | 22.62 | 21.50 |
| 50 | 23.84 | 23.05 | 24.06 | 23.52 | 23.64 | 23.09 |
| 54 | 24.30 | 24.20 | 24.80 | 24.51 | 24.17 | 23.98 |
| 58 | 25.27 | 24.91 | 25.00 | 24.93 | 24.89 | 24.85 |
| 63 | 25.91 | 25.91 | 25.31 | 25.14 | 25.38 | 25.83 |
| 67 | 26.30 | 26.47 | 25.97 | 25.69 | 25.78 | 26.34 |
| 71 | 27.02 | 27.19 | 27.25 | 26.78 | 27.43 | 27.35 |
| 75 | 28.25 | 27.92 | 28.87 | 28.20 | 28.32 | 28.18 |
| 79 | 29.25 | 28.98 | 30.48 | 29.75 | 29.20 | 28.99 |
| 83 | 29.69 | 29.42 | 31.08 | 30.85 | 29.71 | 29.31 |
| 88 | 30.80 | 30.50 | 31.88 | 31.57 | 30.46 | 30.31 |
| 92 | 32.30 | 31.78 | 32.51 | 32.50 | 31.19 | 31.11 |
| 96 | 33.03 | 32.75 | 33.56 | 33.07 | 32.17 | 32.24 |

*) Old numbering according to Ostwald. More recent editions of the colour atlas contain only the cards in the nc-circle used here, which are numbered 1 to 24.

3) The method of representing these results as applied in fig. 1 is most suitable to demonstrate the difference between the results of a successive comparison and those of a simultaneous comparison. For simultaneous comparison two cards are placed side by side, one of which ($j_k$) is illuminated with the light $K$ while the other ($j_g$) is illuminated with the light $G$. Pairs of cards ($j_k$, $j_g$) are then selected which, when compared in this way, show the same hue (no denomination is then necessary!).

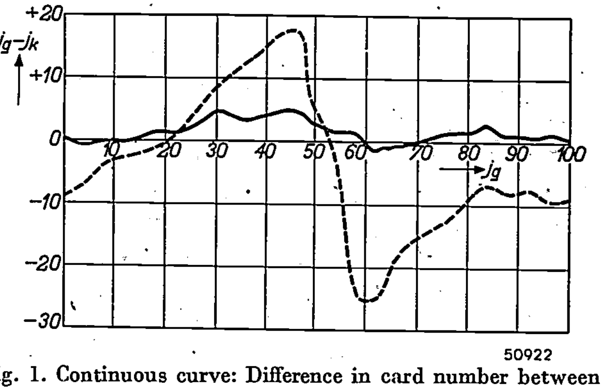


Fig. 1. Continuous curve: Difference in card number between the Ostwald cards $j_k$ and $j_g$ which, observed successively under artificial daylight and under incandescent lamplight, have the same hue (*i.e.* an observer ascribes to them the same hue name, or hue number $N$). This difference here is the average for three observers $A$, $B$ and $C$ (see table I) and is plotted as a function of the number $j_g$.
Broken-line curve: The differences $j_g$ — $j_k$ calculated for simultaneous comparison of the cards under illuminations $G$ and $K$.

The fact that the continuous curve shows practically everywhere much smaller ordinates than the broken-line curve gives expression to the influence of the chromatic adaptation of the eye.

For this case one can also find by calculation, according to known methods, which cards go to make up a pair [5]). The calculated differences $j_k — j_g$ then occurring are likewise plotted in fig. 1 (broken-line curve). Here we have a clear demonstration of the phenomenon that due to chromatic adaptation the differences in colour stimulus as a rule find expression in the colour sensations on a reduced scale.

This general phenomenon was confirmed in this way not only in the cases of the three observers $A$, $B$ and $C$ but also in that of the deuteranomalic fourth observer $D$. The differences in the individual curves $j_k — j_g$ for the three observers $A$, $B$, $C$ were found to be relatively small. Consequently from now on we will consider the average of the perceptions of $A$, $B$ and $C$ for the comparison between artificial daylight and incandescent lamplight.

## Corresponding specimens

When the cards used in the experiments are illuminated with one kind of light (*e.g.* artificial daylight) we get a certain collection of colour sensations, and when the same cards are illuminated with another kind of light (*e.g.* incandescent lamplight) we

[5]) See the literature quoted in footnote [2]).

get another collection of colour sensations. This fact is expressed in table I and is due to two causes: in the first place the colour stimuli are changed, each card under the changed lighting being represented by a different point in the mixture diagram; in the second place the chromatic adaptation of the eye is changed.

Both these collections, however, are associated with the same complete series of hues (yellow, orange, red, etc. up to yellow again) but a given hue in the first collection generally belongs to a card different from that in the second collection. The cards $(j_k, j_g)$ having corresponding hues in the two collections of colour sensations are called "corresponding specimens".

The question now arises as to which cards become "corresponding specimens" as a consequence of chromatic adaptation when two different kinds of lights are used.

In order to formulate this question graphically we will suppose that we have calculated the points representing the cards in the mixture diagram for each kind of light. In this way we get for incandescent lamplight the 24 points (see the note under table I) of the broken-line polygon in *fig. 2* and for the artificial daylight the 24 points of the full-line polygon. For each card number on the first polygon, for example $j_g = 50$, we can determine from fig. 1 the specimen corresponding to it experimentally — in the case in question $j_k = 47.2$ — and find its position on the second polygon (of course as a rule we have to interpolate between the 34 plotted points). When we now come to consider each pair of corresponding specimens the question arises whether there is a general rule defining the correspondence between the points of the first polygon and the points of the other polygon.

If a theoretical rule can be found for this correspondence it will enable us to predict the hue of any given object under any kind of light. To explain this we will first say something about the "white" colour sensation.

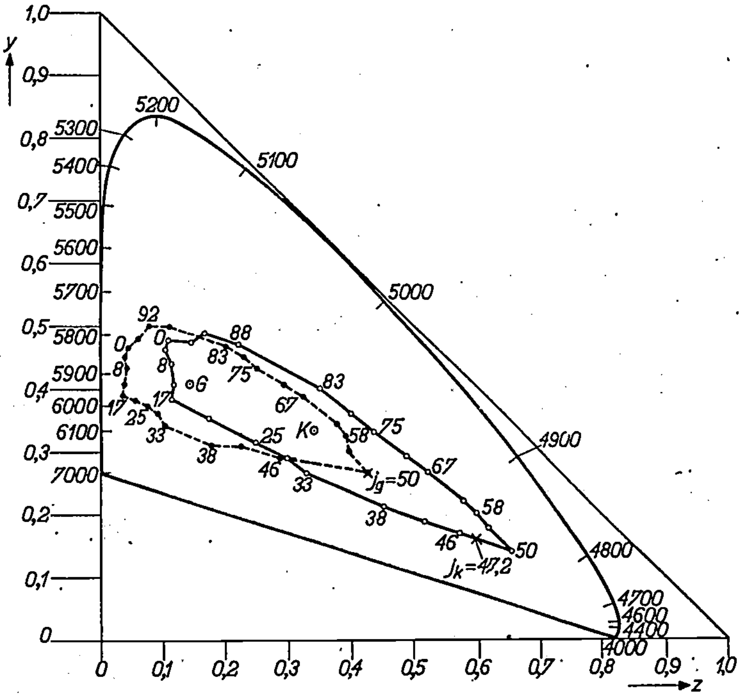


Fig. 2. Positions of the colour points of the nc-cards of Ostwald in the I.C.I. mixture diagram. The broken-line polygon connects up the colour points obtained under incandescent lamplight, whilst the fully-drawn polygon appertains to artificial daylight. $G$ and $K$ are the colour points of the two kinds of light themselves.

Provided the sources of light are not too vividly coloured, owing to chromatic adaptation a very large white object, *e.g.* a table cloth, gives the impression of its being white under any kind of light (*cf.* the experiment described in the article quoted in footnote [6]). Of course the point representing the "white" object in the mixture diagram differs according to the kind of light. Since for a white-coloured object the spectral reflection factor does not depend upon

[6] For a more exact formulation see S. M. Newhall, D. Nickerson and D. B. Judd, J. Opt. Soc. Amer. **33**, 385-418, 1943.

the wavelength, we simply find for the "white points" under different kinds of light the points representing the light emitted by the respective sources of light. Further, by approximation the rule applies [6]) that:
*when illuminated by a certain kind of light all colours lying on one straight line with the corresponding white point have the same hue.* Thus we can imagine a pencil of rays radiating from each white point, each ray representing a line for a given hue.

Two rays, starting from different white points, which belong to the same hue and thus pass through the points representing corresponding specimens of cards are called corresponding rays.

When we come to compare an "unknown" kind of light $X$ with a "known" kind of light, *e.g.* $K$, we have to do with two white points and two pencils of rays starting from those points. The hues belonging to the rays of the pencil $K$ are known from their points of intersection with the polygon belonging to the light source $K$ (see above). Now, given the point representing the colour stimulus from an object under the light $X$ and thus also the corresponding ray of the pencil $X$, we can predict the hue under the light $X$ by finding the corresponding ray in the pencil $K$.

What we have to do, therefore, is to find a rule for determining the corresponding rays of two pencils. This can be done, for instance, with the aid of the locus of their point of intersection.

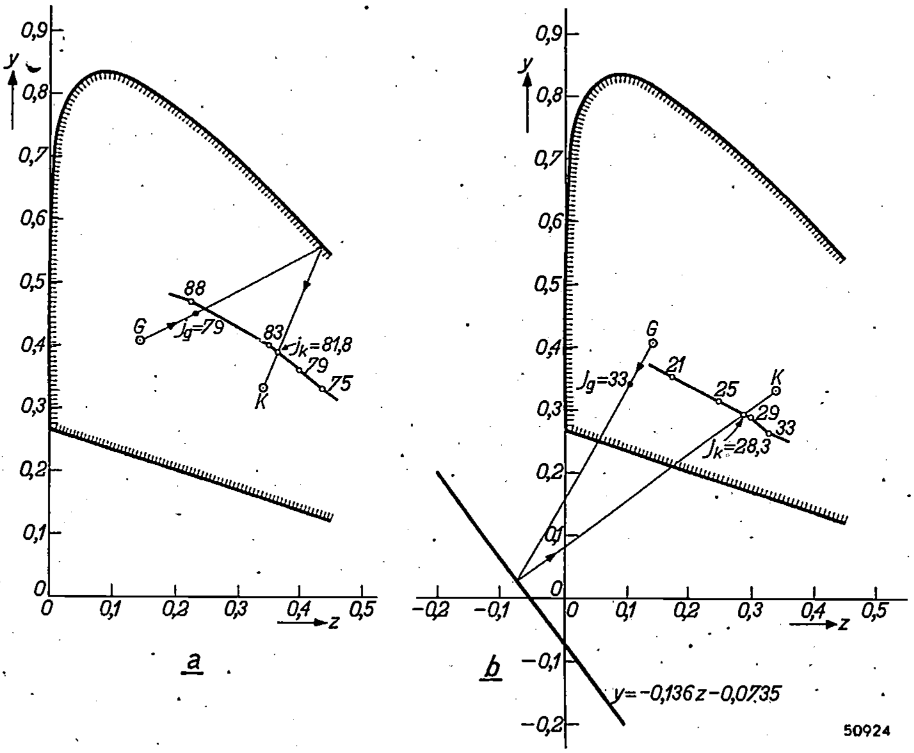


Fig. 3. Illustration of two old recipes for judging equality of hue.
a) Points of intersection of each pair of corresponding rays on the locus of spectral colours. The pairs of points given as an example ($j_g$ = 79 for incandescent lamplight and $j_k$ = 81.8 for artificial daylight) indicate two cards which according to this prescription should have the same hue (corresponding specimens).
b) Intersection of corresponding rays on a straight line (Judd's method). According to this prescription for instance $j_g$ = 33 and $j_k$ = 28.3 are corresponding specimens.

### Locus of the point of intersection of corresponding rays

Various suppositions have been made regarding this locus, each of which has led to a recipe for determining corresponding rays.

1) It has been supposed, for instance, that two corresponding rays intersect on the locus of spectral colours. When we consider the example of the comparison of incandescent lamplight with artificial daylight we find two corresponding rays by connecting any point on the locus of spectral colours with the points $K$ and $G$ representing the two sources of light (see *fig. 3a*).

The corresponding specimens of the two series of colour sensations obtained in experiments under artificial daylight and incandescent lamplight respectively are then found from the intersection of the rays found with the two polygons drawn in fig. 2. Thus it appears that card 79, for instance, under incandescent lamplight would evoke, according to this recipe, the same hue sensation as card 81.8 under artificial daylight.

Further we notice that the point of intersection of two corresponding rays for both kinds of light has the same hue. Consequently the supposition in this section could also be formulated in a seemingly very plausible way by saying that the hue of the spectral colours is not dependent upon their surroundings.

2) If we suppose, as Judd has done [7]), that corresponding rays intersect on a certain straight line (see fig. 3*b*) and not on the locus of spectral colours, we arrive at a second recipe. Judd chose the straight line $y = -0.136\ z\ -0.0735$, where $y$ and $z$ are the coordinates in the I.C.I. mixture diagram used here. According to this prescription card 33, for instance, would give the same hue under incandescent lamplight as card 28.3 under daylight.

When passing over to the U.C.S. system (uniform chromaticity scale) introduced by Judd, where equal distances in the plane of the mixture diagram represent as near as possible equal colour differences, the straight line $y = -0.136z$ $-0.0735$ passes into the infinite straight line. This means that in the U.C.S. system the pairs of corresponding rays become parallel pairs.

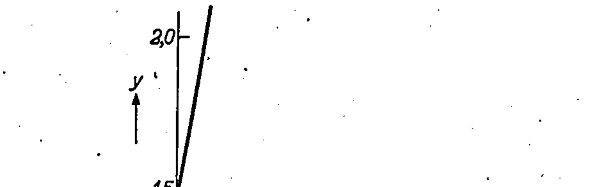


Fig. 4. Experimentally determined points of intersection (×) of corresponding rays in the two pencils of rays centered on $G$ and $K$. The locus of the points of intersection may by approximation be represented by a hyperbola. With the hyperbola most closely matching the emperical points (drawn in the figure) the standard error of the "predicted" hue number $N$ is only 0.42.

The experiments described above now enable us to determine empirically the true nature of the locus of the points of intersection. For the comparison of incandescent lamplight with artificial daylight we construct corresponding rays by drawing straight lines from the two points $G$ and $K$ through the points representing two specimens corresponding according to experiments. The point of intersection of each pair of corresponding rays is a point of the curve sought. *Fig. 4* gives a number of points of intersection determined in this manner. It appears that these points lie neither on the locus of spectral colours nor upon a straight line. They lie rather with a good approximation on a hyperbola which also passes through $G$ and $K$. We therefore use for our representation as the locus of the points of intersection the hyperbola that best matches the actual observations [8]).

The fact that the hyperbola must pass through the points $K$ and $G$ is quite evident; the line joining $K$ to $G$ is a ray of the beam from $K$ and corresponds to a certain ray coming from $G$. The point of intersection of these two corresponding rays lies in $G$ itself. The same applies for $K$. Thanks to the fact that the two "white points" ($K$, $G$) themselves lie on the hyperbola, each ray drawn from one of the white points has only one other point of intersection on the hyperbola, so that the construction of the corresponding ray is unambiguous.

Incidentally the following may be noted. Where corresponding rays intersect on a conic section this means that a so-called projective relation exists between the two pencils or "fans" of rays. This satisfies an essential condition for the application of the theory of Helmholtz and von Kries, to be mentioned below. This is one of the reasons why we have chosen a conic section as an approximation for the points empirically found for the locus of intersection.

With the help of the three prescriptions mentioned above we can calculate from the experimental $N$-values for the light $K$ (see table $I$) the $N$-values to be expected with the light source $G$ and we can then investigate in how far these calculated $N$-values

[7]) D. B. Judd, J. Opt. Soc. Amer. 30, 2-32, 1940.

[8]) This method of representation has been described by the authors in the article already quoted (Physica, 10, 36, 1943), where it is further explained how the hyperbolas were calculated.

deviate from the values found experimentally for *G* (as given in table I). The results of a similar calculation are given in *table II*.

Table II.

Standard (R.M.S.D.) and maximum (Max. D) deviations between calculated and experimental values of the hue number *N* (average of the observers *A*, *B*, *C*,) with the light source *G*.

| Locus of the points of intersection | R.M.S.D. | Max. D. |
|---|---|---|
| Curve of the spectral colours . . . | 2.3 | 5.0 |
| Judd's straight line . . . . . . . . | 2.8 | 7.2 |
| Empirical hyperbola . . . . . . . | 0.42 | 0.79 |

The deviations found in the case of the old recipes are very great indeed. As regards Judd's method the maximum deviation found, for instance, means that a colour which according to his prescription should be called yellowish green would be classified experimentally as orange with a considerable amount of yellow. For practical purposes, the first two recipes are therefore quite useless. With the new method, on the other hand, we find a standard deviation about 7 times smaller than for the old methods. The fact that the calculated hyperbola passes some of the points of intersection found experimentally at a fairly great distance does not appear to result in more than relatively small deviations in the predicted hue.

## Hues under different kinds of light

In the preceding section we have been speaking about the comparison of the two kinds of light first used in our experiments, *viz.* incandescent lamplight and artificial daylight. The next step is to sce how other kinds of light compare.

The observer *B* carried out colour-naming tests, in the same way as described above, under natural daylight and under five experimental low-pressure mercury lamps the bulbs of which were covered with different fluorescent substances. *Fig. 5* gives the points ("white points") *K*, *G*, *D* and *1-5* representing the direct light for the eight kinds of light used.

We now come to the question whether the comparison of any two of the eight kinds of light can also be carried out by means of a conic section as used above. For this purpose it will suffice to choose out of the eight kinds of light two which are greatly different and compare all the others with these. Most suitable for these two were the incandescent lamplight (*G*) and the artificial daylight (*K*) that we used in the previous comparison.

The result of these comparisons is given in *table III*, which shows the standard error and the maximum error occurring when the experimental points of intersection, determined similarly to those in fig. 4, are replaced by the points on the conic section best fitting the tests for each two kinds of light compared.

Table III.

Standard (R.M.S.D.) and maximum (Max. D.) deviations between calculated and experimental values of the hue number *N* for the observer *B*, when comparing natural daylight *D* and five coloured kinds of light *1-5* with incandescent lamplight *G* and with artificial daylight *K* respectively.

| Kind of light | Compared with *G* | | Compared with *K* | |
|---|---|---|---|---|
| | R.M.S.D. | Max. D. | R.M.S.D. | Max. D. |
| *G* | — | — | 0.18 | 0.48 |
| *K* | 0.17 | 0.37 | — | — |
| *D* | 0.25 | 0.59 | 0.13 | 0.32 |
| *1* | 0.20 | 0.45 | 0.14 | 0.39 |
| *2* | 0.25 | 0.57 | 0.20 | 0.52 |
| *3* | 0.21 | 0.49 | 0.17 | 0.36 |
| *4* | 0.52 | 1.39 | 0.46 | 1.18 |
| *5* | 0.62 | 1.35 | 0.60 | 1.23 |
| Average | 0.32 | — | 0.27 | — |

It is seen that the errors are of the same order of magnitude as found when comparing incandescent lamplight with artificial daylight, so that here, too, it is quite justified to express the results by such a conic section.

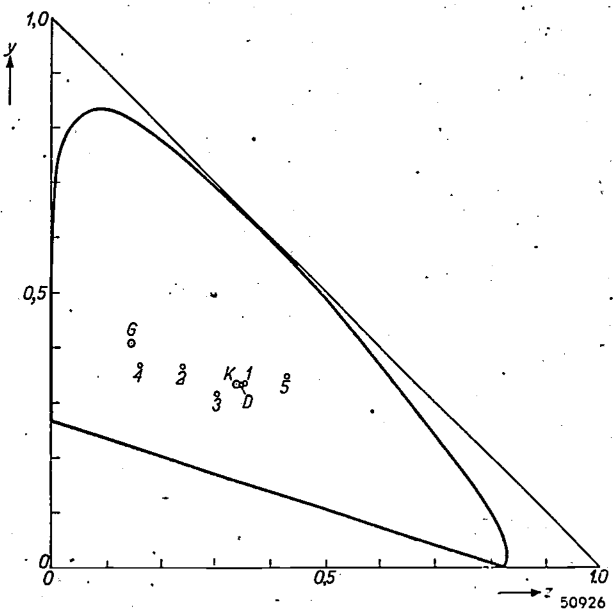


Fig. 5. The colour points of the eight sources of light used here. *G* = incandescent lamp, *K* = artificial daylight, *D* = natural daylight, *1-5* experimental low-pressure mercury lamps coated with different fluorescent substances.

The larger errors for the kinds of light *4* and *5* were only to be expected, because in relatively extensive parts of the spectrum these sources of light yield no energy at all. The result is that some of the cards gave a sensation of only very small saturation; this is a very disturbing element for an accurate judgement of the hue.

Bearing in mind the smallness of the difference in hue that corresponds to a deviation of 1 in the naming of the hue, the average value of the standard error goes to prove that table III as a whole confirms the possibility of applying the new prescription in practice.

### Coordination of all the conic sections found

We have seen that the results of a comparison of hues under a pair of kinds of light can always be coordinated with the aid of a conic section. To get a general review of the whole it will still be necessary to find a general formula for predicting the position of that conic section for any given pair of light sources. To arrive at such a formula we will follow an old theory of Helmholtz and von Kries, which says that the chromatic adaptation simply consists in a change of the mutual sensitivity ratio of the tree chromatic processes which according to Helmholtz take place in the eye. This change in the sensitivity ratio is supposed to go so far as to cause a white plane entirely filling the field of vision to give automatically after a change of light the same impression of whiteness as before. Applying this hypothesis to the conditions of our experiments it follows that there must be three points in the plane of the mixture diagram — *viz.* the physiological basic colours according to Helmholtz — where the colour sensation is not affected by the surroundings. These three points of the mixture diagram are characterised by the property that only one of the three chromatic processes is stimulated at a time, so that a change in the mutual sensitivity ratio of the three processes cannot influence the colour sensation and in particular the hue of these basic stimuli. Now we have already found such an invariance of hues when determining the locus of the points of intersection of corresponding rays: when changing over from one kind of light to another we find for a point on the conic section belonging to that transition one and the same hue before and after the change, since each point of the conic section is the point of intersection of two corresponding rays.

The conception according to Helmholtz and von Kries can therefore only be reconciled with the experiments if the physiological basic colours lie on the said conic section. Since this argument holds for all pairs of lights we come to the conclusion that if this hypothesis is to agree with our experiments it is necessary to have *three fixed points in the plane of the mixture diagram where all conic sections intersect.*

Before proceeding to investigate whether the existence of these fixed points is compatible with our experimental data, we must consider for a moment a refinement of the method hitherto followed. It was presumed that in the case of a large white field the colour sensation of that field would remain unchanged when replacing artificial daylight by another kind of light. We have already shown that this supposition is not quite correct when we have to do with kinds of light differing considerably from daylight. Mathematical reasoning (which we cannot go farther into here) shows that for a kind of light differing strongly from daylight this incomplete adaptation can be taken into account by taking as the centre of the pencil of rays that has to give the corresponding rays, not the point representing the light itself but a point shifted somewhat along the conic section into the direction of the daylight point. With this refinement in the calculation we obtained a better agreement between theory and experiment. In fact this refinement has already been applied when calculating the values given in table III.

If the conic sections that were calculated for table III are drawn it appears that three such fixed points can indeed be indicated. These points we call the red, green and blue points ($\alpha$, $\gamma$, $\beta$) according to the spectral colours nearest to which the points lie. Naturally the conic sections do not pass exactly through these points. The closest approximation is that of the blue point $\beta$; *cf. fig. 6*: all conic sections without exception pass through the circle indicated, whilst the conic section showing the greatest deviation ($p$ in fig. 6) is that belonging to the comparison of two kinds of fluorescent lamps (*K* and *1* of table III and fig. 5) which so closely resemble each other as to make the corresponding hyperbola very uncertain. For the "red point" $\varrho$ and the "green point" $\gamma$ the agreement was less satisfactory, but nevertheless

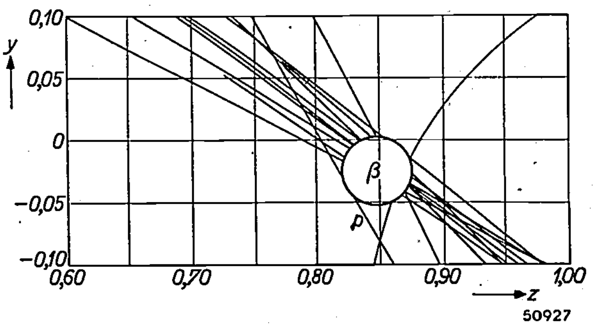


Fig. 6. Determination of the "bluepoint" $\beta$ from the series of conic sections, derived from the comparison of the light sources *D* and *1-5* with *G* and with *K* (observer *B*).

it was possible to determine also these points with a reasonable degree of accuracy (see table IV).

Now a hyperbola is fully determined as soon as five of its points are given. If the three points $\varrho$, $\gamma$, $\beta$, are known and we are further given the points $I$ and $II$ representing two kinds of light, we can therefore construct the hyperbola for the change from light $I$ to light $II$ without any further experimental data. (As the hyperbola must pass through the points of $I$ and $II$, this holds also when the refinement mentioned above is applied.)

Table IV.

Coordinates of the fixed points $\varrho$, $\gamma$, $\beta$ from which one will have to start to coordinate the experimental results according to the theory of Helmholtz and von Kries.

| | $y$ | $z$ |
|---|---|---|
| $\varrho$ | 0.223 | —0.117 |
| $\gamma$ | 1.2 | 0.408 |
| $\beta$ | 0.036 | 0.871 |

This procedure affords us a means of checking the points $a$, $\gamma$, $\beta$ found above, since the hyperbola originally found empirically, *e.g.* when comparing the kinds of light $K$ and $G$ (see fig. 4), can now be replaced by the hyperbola passing through the five points $K$, $G$, $\varrho$, $\gamma$, $\beta$, and we can then calculate to what extent the deviations given in table III have increased owing to the requirement that the hyperbola is obliged to pass through $a$, $\gamma$ and $\beta$.

When this is done for all conic sections it is seen that the average deviations of table III are certainly increased slightly but not to an appreciable extent, *viz.* from 0.32 and 0.27 to 0.40 and 0.38. We therefore conclude that the whole of our experimental material can be coordinated with a reasonably small standard error by means of conic sections passing through $\varrho$, $\gamma$ and $\beta$.

The complete prescription for finding corresponding rays can now be formaluted as follows. For the comparison of two kinds of light $I$ and $II$ each point of the conic section passing through $\varrho$, $\gamma$, $\beta$ and the points representing $I$ and $II$ yields one pair of rays. For a comparison of all sorts of light with one given kind of light, say $K$, we have to use the family of conic sections having as basic points $K$, $\varrho$, $\gamma$ and $\beta$. This family of conic sections (hyperbolas) takes the place of the single curve spoken of in the old recipes, *i.e.* the locus of spectral colours or the straight line according to Judd.

According to the line of thought developed here the three fixed points $\varrho$, $\gamma$, $\beta$ should be the three physiological basic points of Helmholtz. These basic points have also been determined by others, *i.a.* by the testing of colour-blind people [9]). The basic points found in this way differ rather strongly from our fixed points $\varrho$, $\gamma$, $\beta$, for what reason we do not know.

This discrepancy, however, does not alter the fact that the experimental results obtained by the observer $B$ can indeed be coordinated with the aid of the three fixed points given in table III. The question whether the same fixed points are valid also for other normal observers — such is to be expected if the three fixed points are actually identical with Helmholtz's basic points — can only be decided by further investigations. Anyhow it is reasonable to expect that the principle of the method (if necessary with other fixed points) can be applied for any observer.

We might mention here that the whole series of tests described in this article could also be carried out in another way, using, instead of different kinds of light and one white-coloured background, one kind of light and different plain-coloured backgrounds. Provided that too great deviations from white are excluded it will be found that here again the colour sensation of the background will be "white" thanks to chromatic adaptation. The variations of the hues then occurring with the cards placed on the background are related to those found to accompany the phenomenon of the so-called simultaneous contrast [10]). The interpretation of the experiments taken in this way is even somewhat simpler than that of those actually carried out, because when only the background is changed the points representing the cards in a mixture diagram remain unaltered.

When it comes to actual practice we have, indeed, to do with different sources of light, and it is on this account that the whole problem has become of actual importance, — but we have to reckon also with changes of the background. As a rule the background will not be white and even not plain, but its effect upon chromatic adaptation will have to be expressed as the effect of an equivalent plain background with a certain "average colour". According to the saturation of this average colour (difference from white) one will have a continuous transition from the phenomena to be placed under "chromatic adaptation" up to the cases where one has to speak of simultaneous contrast.

Further, if the possibility of predicting hues is to

[9]) A. König and G. Dieterichi, Z. Psychol. Physiol. der Sinnesorgane, 4, 241-347, 1893. H. E. Ives, J. Frankl. Inst. 195, 23-44, 1923. P. J. Bouma, Physica, The Hague, 9, 773-784, 1942.

[10]) The phenomenon that the difference between the colour sensations produced by two pigments or spots of light becomes greater if the planes are observed side by side.

be applied in practice, the restriction has to be made that in the naming of hues psychic influences sometimes play an all-important part. For instance a person is not likely to ascribe to grass any other colour than green and it would take a lot to persuade him otherwise.

### The effect of chromatic adaptation upon the sentation of saturation

Chromatic adaptation influences not only hue but also the sensation of saturation, as may be illustrated by an example. Looking at fig. 2 one would expect that under incandescent lamplight the yellow card (number 4) would give a more saturated impression than under daylight. In practice, however, due to chromatic adaptation just the reverse effect is observed: under incandescent lamplight yellow gives more the impression of being "almost white" than under daylight. This explains why thin lines of yellow ink on white paper are hardly distinguishable under incandescent lamplight.

Here again we find in the literature various "recipes" for determining when the same sensation of saturation is to be attributed to two colours, but since we have no accurate experimental data available we shall not go into this question here.

# SMALL SELENIUM RECTIFIERS

by J. J. A. PLOOS van AMSTEL.

621.314.634

A selenium rectifier consists essentially of a layer of selenium, a blocking layer and a layer of metal acting as cathode. The properties of such a rectifier depend to a large extent upon the reaction of the latter metal when the selenium is applied. Some metals used as cathode give rectifiers with an exceptionally low resistance in the transmitting direction whilst other metals make the rectifier suitable for rectifying relatively high voltages. This rectifying is further promoted by applying an additional or artificial blocking layer. With these methods Philips have developed three kinds of selenium rectifiers, two of which are made exclusively in small sizes (a few millimeters) and the third also in larger dimensions. Here the main properties and applications of the small rectifiers will be discussed.

In the course of time blocking-layer rectifiers have been developed into electrical switching elements with a great variety of properties, dimensions and uses. An article dealing with the general composition of a blocking-layer rectifier and the theory of its working was published in this journal in 1939 [1]), followed a year later by another article on their uses and in particular the application of selenium cells in rectifiers [2]). Here we shall go somewhat farther into some details of the construction of these selenium rectifiers. It will be found that the properties of the final product can be influenced in a certain direction by the choice of certain materials and their method of treatment. In this way it is possible to attain the best solution for a given object.

As a rule a rectifier is required to have the lowest possible resistance in the transmitting direction, the highest possible resistance in the blocking direction and the least possible capacity. These requirements are partly contrary to each other, for a reduction of the resistance in the transmitting direction, for instance, is accompanied by reduction of the resistance also in the blocking direction. The choice of preference depends entirely upon which property is to be stressed for the application in view. In the following it will be shown that means are available to attain this end. We shall consider here in particular small selenium rectifiers (the largest size is only a few mm) because they have the most varied applications.

Let us first recall the general composition of a selenium rectifier (*fig. 1a*): between a layer of semi-conducting selenium and a layer of good conducting metal is a very thin insulating blocking layer. Electrons appear to pass more easily into the blocking layer from the metal, rich in free electrons than from the selenium, which is poor in electrons. Thus a positive current is more readily brought about in the direction of the arrow (transmitting direction) than the other way (blocking direction).

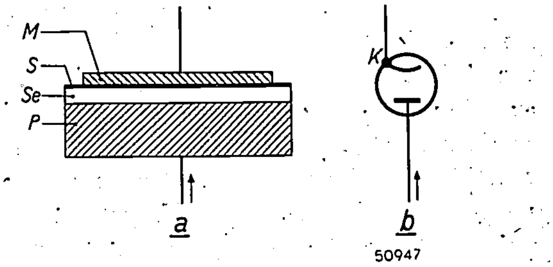


Fig. 1. *a*) Diagrammatic representation of a selenium rectifier. $P$ = metal carrier plate, $Se$ = layer of semi-conducting selenium, $S$ = blocking layer, $M$ = layer of good conducting metal or alloy.
*b*) For comparison a valve with hot cathode. The emitting cathode $K$ corresponds to the metal $M$ likewise emitting electrons.

If a selenium rectifier through which current flows in the transmitting direction is compared with a hot-cathode valve (fig. 1*b*) one may say — without going into the difference in mechanism — that the metal corresponds to the hot cathode (hence the term cathode metal or cathode layer). In many cases it is well worth bearing this in mind.

## Influence of the compositions of a selenium rectifier upon its properties

A selenium rectifier may be made in the following way: molten selenium is poured onto a metal carrier plate, serving to lend strength to the whole, then pressed flat and subjected to a heat treatment, thereby forming on the surface of the selenium an insulating layer called the genetic or natural blocking layer (in contrast to the artificial blocking layer which will be mentioned farther on). Finally a layer of good conducting metal (maybe an alloy) is added in some way or other, for instance

[1]) W. Ch. van Geel, Blocking-layer rectifiers, Philips Techn. Rev. 4, 104-110, 1939.
[2]) D. M. Duinker,The use of selenium cells in rectifiers, Philips Techn. Rev. 5, 199-207, 1940.

by extrusion, by vaporization in vacuo or by atomization in a gas discharge.

*Reactions between cathode metal and selenium*

The rectifying properties of a rectifier made in this way appear in the current-voltage characteristics, which can be recorded in both directions (how this can best be done will be shown later). Another property of great importance in some applications is the capacity that can be measured between the carrier plate and the cathode layer, these forming the external electrodes. It appears that both the magnitude of this capacity and the trend of the current-voltage characteristics depend to a high degree upon the nature of the metal or the alloy from which the cathode layer is made and also upon the treatment of the rectifier. In the main the result is determined by the reactions taking place between the cathode metal and the selenium. Two cases are to be distinguished; that where the reaction between the selenium and the metal forms non-conducting substances and that where the reaction products are indeed conductive. Generally such reactions can be promoted by heating the rectifier to a high temperature. Furthermore, the reactions forming non-conducting compounds can be promoted by heating to beyond the melting point of the cathode metal and at the same time applying a voltage making the metal positive with respect to the selenium. This causes positive metal ions to pass over to the selenium.

Where we have to do with insulating (or at least very poorly conducting) Se-metal compounds these may be regarded as an accretion of the blocking layer. When a rectifier in which such compounds may arise is subjected to heat treatment one may observe a reduction of the conductivity in both directions and also of the capacity, corresponding to the picture of a thickened blocking layer. Examples of metals forming with selenium poorly conducting compounds are cadmium, magnesium and aluminium. We will come back to the action of cadmium presently.

If, on the other hand, we use in the cathode layer metals which with selenium form highly conductive compounds — examples of such metals are gold, silver and antimony — then the layer forming these compounds may be regarded as a continuation of either the cathode layer or the selenium. The blocking layer itself may in the first instance remain unchanged. Nevertheless, also in this case it will as a rule be observed that the characteristics have undergone changes during the reaction, often in the sense of a lower resistance in both directions. This is utilised in one of the kinds of rectifiers manufactured by Philips.

*Artificial blocking layer*

Instead of leaving it at a genetic blocking layer extended or not with insulating Se-metal compounds, before precipitating the cathode layer on the genetic blocking layer an artificial blocking layer can be applied to the latter. Various substances can be chosen for this (especially certain organic compounds). Such an artificial blocking layer, which is always relatively thick, checks considerably the reactions mentioned above, so much that in as far as they take place at all they have little effect. The capacity between the electrodes of such a rectifier, per surface unit, is right at the outset considerably lower than that of a rectifier without artificial blocking layer, but at the same time the internal resistances in both directions are much higher.

## Methods of measuring

Before proceding to show what types of selenium rectifiers Philips are making according to these aspects, we will briefly deal with the methods of measuring.

*Recording the dynamic current-voltage characteristics*

The characteristic (current as function of voltage, either in the transmitting direction or in the blocking direction) can be recorded by varying an applied direct voltage step for step and measuring the current at each step. In the article quoted in footnote [2]), however, it was stated that a static characteristic derived in this manner will as a rule differ from the relation existing between simultaneous momentary values of current and voltage when the rectifier is subjected to a rapidly alternating voltage (dynamic characteristic). Since this is the case in by far the majority of the applications of these rectifiers, we are in the first place interested in the dynamic characteristics. These characteristics can be made visible by means of a cathode-ray oscillograph. *Fig. 2* shows the circuiting arrangement for recording these dynamic characteristics.

*Measuring the capacity*

Capacity can be determined with the aid of the bridge circuit illustrated in *fig. 3*. A variable direct voltage $E$ is laid on the rectifier in the blocking direction and on that a small alternating voltage of say 10 mV is superposed. Since the measuring result depends somewhat upon the frequency it is prefer-

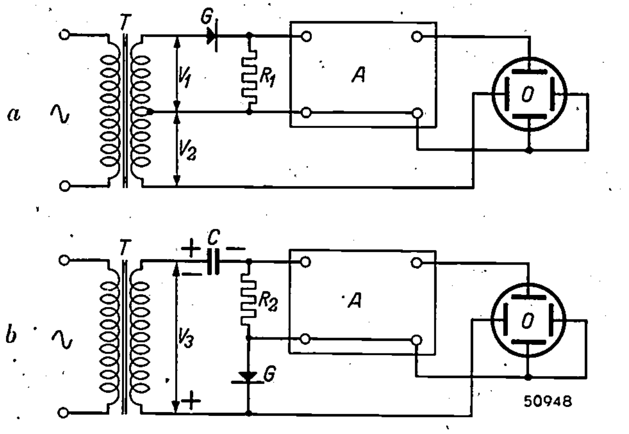


Fig. 2. *a*) Circuiting schemes for recording dynamic characteristics of blocking-layer rectifiers: (*a*) in the transmitting direction, (*b*) in the blocking direction. Since different scales are desired for the transmitting and the blocking characteristics, the recordings are taken from different circuits. In (*a*) the resistance $R_1$ is small compared with the resistance in the transmitting direction of the valve *G*; the voltage $V_1$ (about 1 V) is therefore practically equal to the voltage on the rectifier. The higher voltage $V_2$, supplied by the same transformer *T* as gives $V_1$, is thus at the same time a measure for the voltage on the rectifiers; it supplies the horizontal deflection on the oscillograph tube *O*. The voltage at $R_1$, which is a measure for the current passing through the rectifier, is conducted *via* the amplifier *A* to the plates for the vertical deflection.

*b*) The valve *G* carries the sum of the alternating voltage $V_3$ and a direct voltage from the capacitor *C*. The leakage current passing through the rectifier when $V_3$ has the given polarity causes in the resistor $R_2$ a voltage loss which supplies *via* the amplifier *A* the vertical deflection on the oscillograph tube *O*, whilst the voltage on *G* provides for the horizontal deflection direct.

able to carry out this measurement with a frequency within the range in which the rectifier is to be used (in the case of certain modulator cells for instance 60 kc/s). The bridge is balanced as far as possible by adjusting the resistor *R* and the capacitor *C*. The value read for *C* is then the capacity required. This depends for a great deal upon the direct voltage *E*, which fact must be taken into account when comparing one rectifier with another.

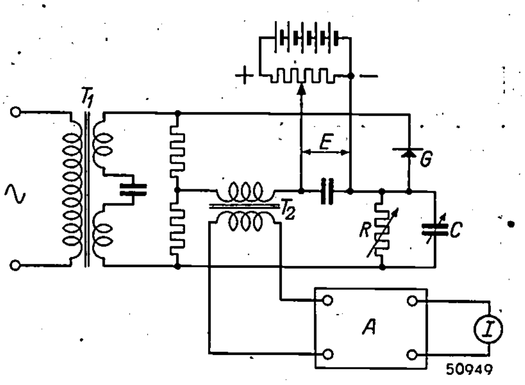


Fig. 3. Bridge circuit for measuring the capacity of blocking-layer rectifiers. $T_1$ = transformer supplying on the secondary side a voltage of 2 × abt. 10 mV with a frequency of say 60 kc/s; *G* = valve to be tested; *R* and *C* = variable resistor and capacitor for bringing the bridge into equilibrium; $T_2$ = output transformer connected to an amplifier *A* and a recording instrument *I*; *E* = variable direct voltage, upon which the capacity found greatly depends. The capacitor in series with the secondary winding of $T_1$ prevents direct current flowing through this coil.

For the explanation of the fact that the capacity depends upon the voltage, we refer to what is stated in the article quoted in footnote [1]) about insulators and semi-conductors. In an insulator each electron is bound to its place (*i.e.* a certain atom) whereas in a semi-conductor the electrons have freedom of movement, due to the fact that there is either an excess of electrons or else an electron is lacking in places. In the latter case conduction is brought about by the movement of electrons from one open place to the next. If all the open places were occupied by electrons there would be no conductivity.

One can imagine something similar happening in the selenium when a direct voltage is connected to a selenium rectifier (Se negative, metal positive, *fig. 4*): electrons in the selenium are then attracted in the direction of the blocking layer; the thin layer of the semi-conductor immediately adjacent to the blocking layer thus becomes saturated with electrons. The higher the voltage applied, the farther this charge extends into the semiconductor, causing a corresponding drop in the capacity between the two electrodes.

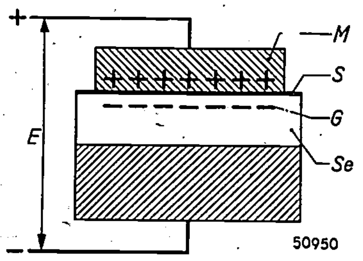


Fig. 4. The voltage *E* in the circuit of fig. 3 causes charges on either side of the blocking layer *S*, positive in the metal at *M* and negative in the semi-conducting selenium *Se*. This gives rise to an area *G* in the latter layer where all open places are occupied by electrons and which consequently no longer has any conductivity, This area *G* becomes thicker the higher the voltage *E*, which accounts for the drop then found in the capacity.

## A discussion of three kinds of selenium rectifiers

The demand in various practical fields for blocking layer rectifiers with different properties has led to the manufacture of three kinds, which we will refer to in this article as I, II and III, each of which is made in the dimensions most suitable for a certain purpose. Per unit of effective surface the capacity decreases and the resistance in the transmitting direction increases in the order of I-II-III. These rectifiers will be discussed in the order of II-III-I.

### *Type II*

The oldest is type II, in which an alloy of tin, cadmium and bismuth is applied direct to the genetic blocking layer. As already remarked, together with selenium the cadmium in this alloy forms an insulating compound which can be considered to

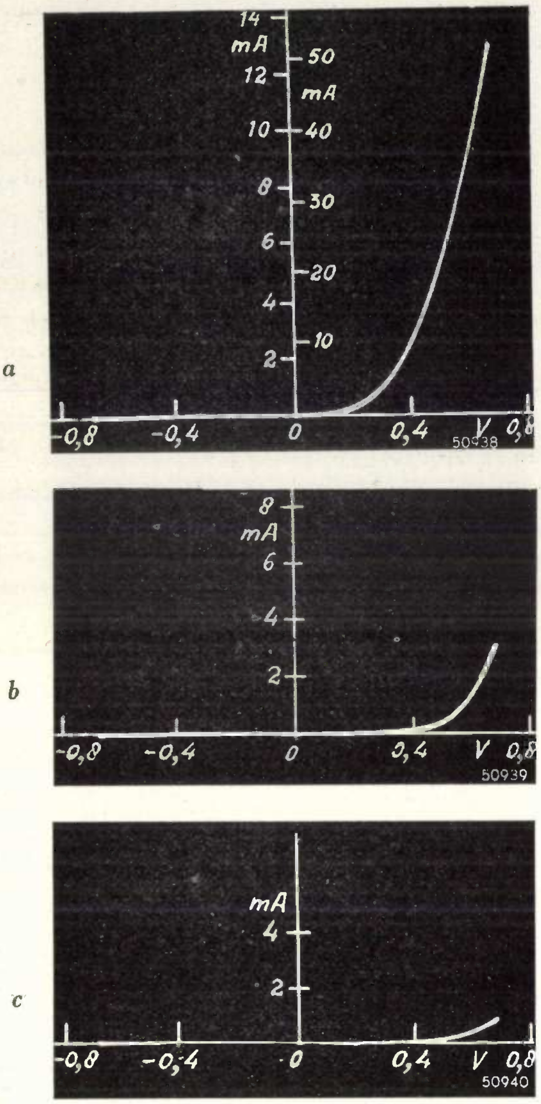


Fig. 5. Dynamic characteristics for the transmitting direction derived with the circuiting scheme of fig. 2*a* and, as in the case of the characteristics shown in figs. 6 and 7, recorded with an oscillograph of the type GM 3156. Amplitude of $V_1$: 0.68 V.

a) Cell of type I, diameter of cathode layer 1.5 mm.

*b*) Cell of type II, diameter of cathode layer 3 mm.

*c*) Cell of type III, diameter of cathode layer 3 mm.

The left-hand current scale in (*a*) indicates the currents measured, whilst the right-hand scale, which would apply for a diameter of 3 mm, is added to facilitate comparison with the 4 times as large cells of (*b*) and (*c*). To the left of the *i*-axis is to be seen a part of the blocking characteristic, which practically coincides with the negative axis of abscissa.

belong to the blocking layer. Such a compound has a favourable action owing to the raising of the resistance in the blocking direction and reduction of the capacity; against this, however, is a higher transmitting resistance.

The reason why cadmium is used in combination with tin and bismuth is that with these three metals an alloy can be made which has a low melting point, so that it is easy to spray, and which can be kept in liquid form without oxidising too much. Furthermore, this alloy can be used for soldering one of the connecting wires.

Dynamic transmitting characteristics of a rectifier of the type II are given in *figs 5b* and *6b* respectively for 0.68 V and 1.34 V amplitude of the voltage $V_1$ (*cf.* fig. 2*a*). The dynamic blocking characteristic is represented in fig. 7*b*. In *fig. 8* curve *II* shows the trend of the capacity as function of the direct voltage applied *E*. We shall revert to these figures when discussing types III and I.

### *Type III*

In the rectifiers of type III the genetic blocking layer is reinforced with an artificial blocking layer

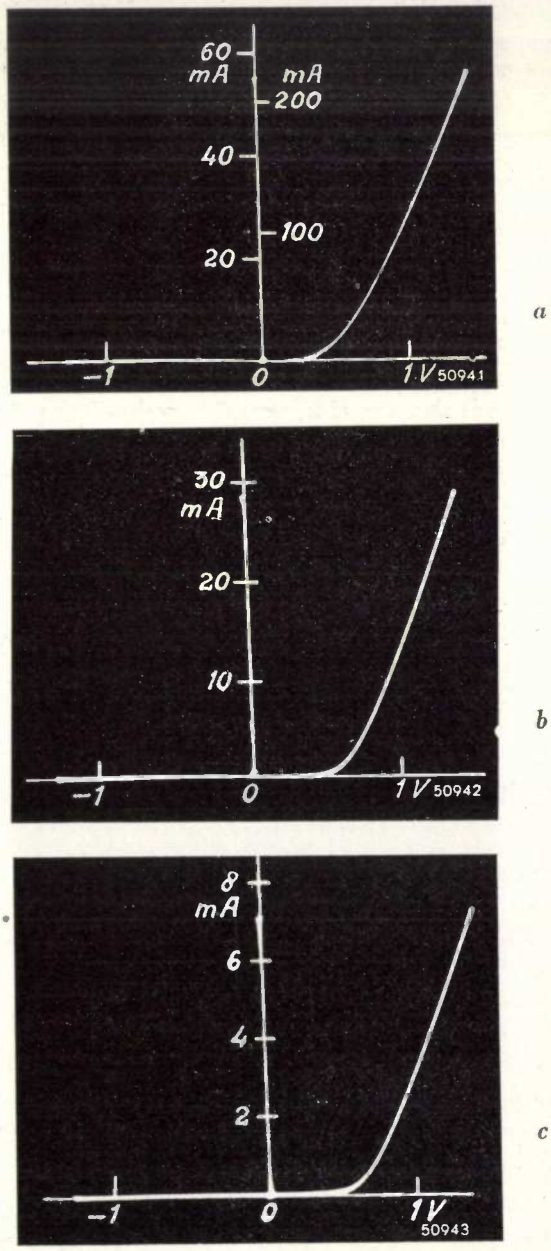


Fig. 6. The same as fig. 5 but with a voltage amplitude of 1.34 V.

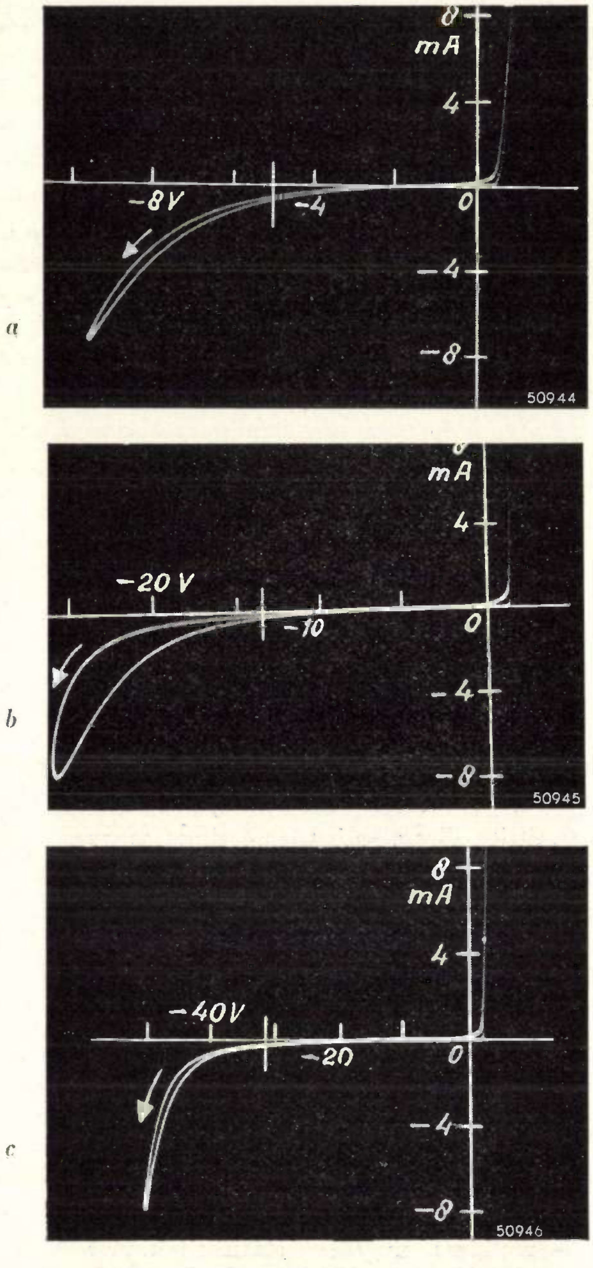


Fig. 7. Dynamic characteristics for the blocking direction measured on the three cells I, II and III in the circuiting arrangement of fig. 2*b*. In this case the cells were heavily overloaded as regards voltage. It appears that the blocking characteristic often shows a peculiar loop shape, as is particularly the case with cell II; the direction of flow through the loop is indicated by the arrow. The maximum blocking voltage permissible under normal working is indicated by a vertical stroke. To the right of the axis of ordinates a part of the transmitting characteristic is to be seen.

before applying the Se-Cd-Bi alloy. This improves, in the first place, the blocking properties, as may appear when comparing fig. 7*c* with fig. 7*b*, both of which are the characteristics of rectifiers having 7 mm² effective surface. It is seen that with type III the voltage can be raised much higher without causing a rapid increase in the leakance (which may be regarded as an indication of breakdown). This is particularly of importance for the rectification of high voltages, for when using the type III rectifiers a smaller number in series suffices.

As was to be expected, with the improved blocking properties one has to take into the bargain a less satisfactory transmitting characteristic: the "threshold voltage" (at which the current becomes noticeable) is, it is true, only little higher in type III than in type II (compare fig. 5*c* with fig. 5*b* measured with rectifiers of the same size and represented on the same scale), but the slope of the rest of the characteristic differs considerably (compare fig. 6*c* with fig. 6*b*, where the scales of current are not equal).

In the case of the type III rectifiers one will also expect a smaller capacity than that of type II. This is confirmed by the measurements (compare curve *III* with *II* in fig. 8).

Rectifiers of the type III have been standardised in a series of sizes from 7 mm² up to 14 000 mm² effective surface [3]).

### *Type I*

Last of all we have type I, which in its construction and in its properties differs appreciably from

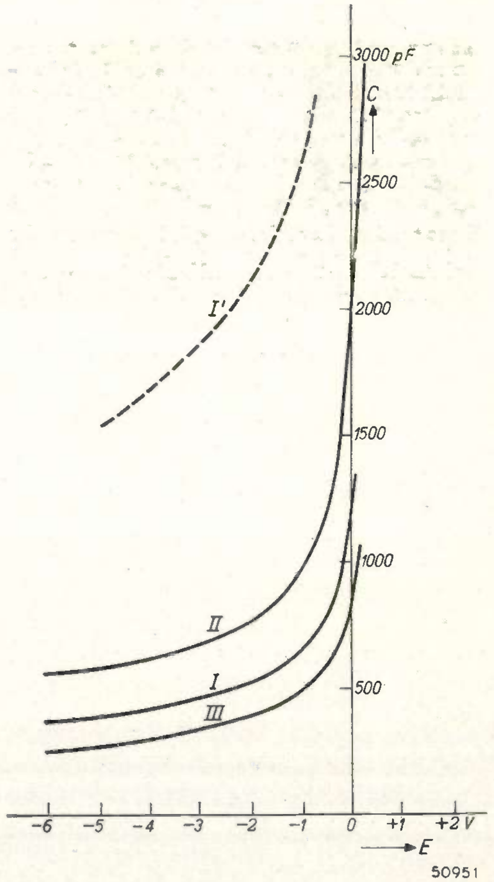


Fig. 8. The capacity *C* measured with the circuiting scheme of fig. 3 plotted as a function of the applied direct voltage *E*, for the rectifiers *I*, *II* and *III*. The curve *I* has been derived by multiplying the ordinates of *I* by 4.

[3]) The selenium cells dealt with in the article quoted in footnote 2 are of type III.

types II and III. As a matter of fact it was developed for quite a different purpose, *viz.* to get the most favourable characteristic possible in the transmitting direction. The use of the cathode metal that forms a conductive compound with selenium has been most succesful. Since, however, the melting point of the metal in question is much higher than that of selenium, the cathode layer cannot be applied in liquid form, as is the case with the alloy previously mentioned. It was found possible to precipitate a thin layer of the metal by atomization, but then the problem arose how to make a good electric contact with it. A contact spring is not reliable in the long run. Soldering on such a very thin layer is out of the question, whilst furthermore soldering over the metal would lead to reactions which are undesirable in this case, because they would result in the favourable properties in the transmitting direction being lost.

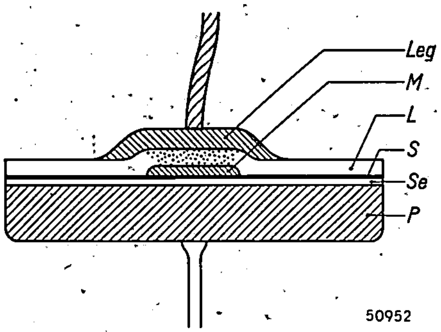


Fig. 9. Cross section of a cell of type I. $P$ = carrier plate, $Se$ = selenium layer, $S$ = genetic blocking layer, $M$ = cathode obtained by vaporization, $L$ = lacquer, $Leg$ = alloy for soldering the inlet lead. The lacquer allows enough of the alloy to pass through to ensure a good electric contact; the obnoxious Cd-Se compound is formed on only a very small part of the effective surface.

This difficulty was solved in the manner sketched in *fig. 9*. After a layer of metal with a diameter of about 1.5 mm is applied by vaporization (the reason why this diameter has been chosen so small will be clear farther on) the whole plate (diameter 6 mm) is coated with a thin layer of a suitable lacquer. Then a layer of the aforementioned Se-Cd-Bi alloy is applied in the middle, with which the lead wire can be soldered. Enough of the alloy penetrates through the pores of the lacquer to make a reliable electric contact in a number of places with the first layer of metal. At these places some reaction will take place between the selenium and the solder through the very thin cathode layer (forming the undesirable Cd-Se compound), but this takes place on such a small fraction of the effective surface as to have scarcely any effect upon the characteristic of the rectifier.

The characteristics measured for such a rectifier for the transmitting direction are to be seen in figs. 5*a* and 6*a* (the small current scale applies for a cell with 1.5 mm diameter of the cathode layer with which the measurements were carried out, whilst the large current scale applies for a surface 4 times as large and has been added to facilitate a comparison with the figures 5*b* and *c* and 6*b* and *c*, relating to rectifiers with a diameter of 3 mm). It is seen that the threshold voltage is less than half that of the rectifiers of types II and III and that the rest of the characteristic has a much steeper slope.

Fig. 7*a* shows that the rectifiers of type I can bear less voltage in the blocking direction than the other types, but this is of no importance in applications where the voltage on the rectifier does not go beyond a few volts. What is more serious is that the capacity (taken for the same surface) is much higher than in the case of types II and III (see curve $I'$ in fig. 8). As a matter of fact this is one of the reasons why the diameter has been reduced from 3 mm to 1.5-1.2 mm; the capacity is then at least 4 times as small (curve $I$, fig. 8) [4]). Since with a given current intensity the density of current in the transmitting direction is then 4 times as great, the rectifier works in a less curved part of the characteristic, which for some applications is a great advantage: the power involved is so small that the heat dissipation — in large rectifiers one of the main factors restricting the admissible current density — becomes quite negligible.

*Fig. 10* shows some selenium rectifiers in different stages of production and mounted in various ways.

## Applications of small selenium rectifiers

The most important applications of small selenium rectifiers may be classified under three headings:

1) in electrical measuring instruments,
2) in modulators in carrier-wave apparatus,
3) in "small rectifiers".

### 1) *Selenium rectifiers in electrical measuring instruments*

The uses to which these rectifiers are put in measuring instruments lie in quite different fields. In the first place blocking layer rectifiers are often used for converting an alternating current that is to be measured into a pulsating direct current, the

[4]) The thickness of the layer of lacquer compared with the blocking layer is such that the capacity between the protruding edge of the alloy and the selenium can be ignored compared with the capacity between the cathode layer and the selenium.

average value of which is measured with the aid of a moving-coil instrument. The main advantages of this method are that the meter uses very little current (as compared with most other measuring methods) whilst the scale is more or less linear. Another field of application comprises the use of a blocking-layer rectifier as shunt for a measuring instrument, for instance to guard it against overload. These two fields will be discussed here briefly, confining our remarks, as far as the first field is concerned, to voltmeters.

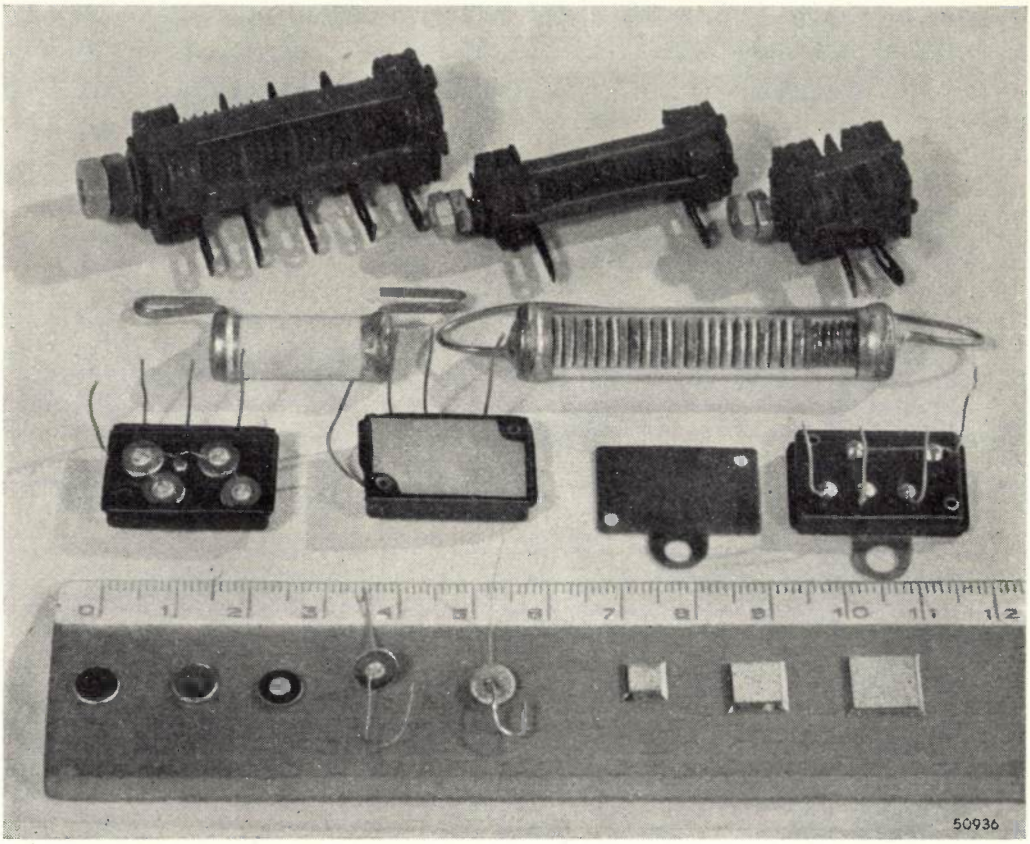


Fig. 10. Kinds of selenium rectifiers, some mounted and some not. From left to right: *(front row)* five cells of type I in different stages of construction (without cathode layer, with cathode layer, with lacquer and alloy, with inlet leads) and three square cells of type III;
*(second row)* four rectifying cells destined for a voltmeter in Graetz circuit housed in a box of "Philite"; the box filled with insulating material; the lid of the box; the complete rectifier;
*(third row)* hermetically sealed ceramic tube containing a modulator cell; a glass tube containing 25 cells connected in series (laboratory design often used, *inter alia*, during the war in clandestine radio receivers);
*(fourth row)* three units with square cells connected in different ways.

The circuit of an alternating voltage meter with rectification is represented in *fig. 11a*. The moving-coil instrument is connected to the D.C. terminals of 4 rectifiers in Graetz circuit. The alternating voltage to be measured is connected to the series connection of a resistor and the two other terminals of the Graetz circuit.

If the rectifiers were free of capacity than the meter reading would be independent of the frequency. The measuring instrument would then indicate the average value of the alternating current $I_0$, given by $I = V/(R + r)$, where $V$ represents the alternating voltage to be measured, $R$ the series resistance and $r$ the sum of the resistance $r_m$ of the instrument and twice the transmitting resistance $r_d$ of a rectifier (twice because the current always flows through two rectifiers in series). Actually, however, the rectifiers do possess capacity. This fact can be reckoned with by taking into account the capacity $C$ drawn in fig. 11$a$ in a broken line. The alternating current $I$ now introduced in the Graetz circuit differs from the current $I_0$ that it is desired to measure. From the replacement scheme of fig. 11$b$ it follows, after a small calculation, that:

$$I = \frac{I_0}{\sqrt{1+\left(\dfrac{\omega\, Cr}{1+\dfrac{r}{R}}\right)^2}} \quad . \; . \; . \; . \quad (1)$$

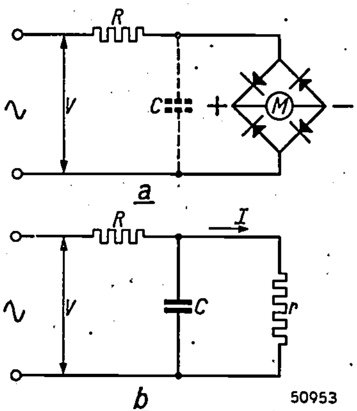


Fig. 11. *a*) Circuit for measuring alternating voltages with the aid of a moving-coil meter *M*, a Graetz circuit of four cells and a series resistor *R*. By means of the broken line capacity *C* allowance can be made for the capacity of the cell.

*b*) Replacement scheme where *r* represents the resistance of the meter and of two cells connected in series. *V* = the alternating voltage to be measured, *I* = the current, the average value of which is measured.

Let us first consider the case where the series resistance *R* is large compared with *r* (this will indeed be so in the case of meters for not too low voltages; the meter for 50 V, for instance, giving full deflection at a current of 1 mA, already has *R* = approx. 50 000 Ohms, whilst *r* = approx. 1000 Ohms). If, moreover, $\omega Cr \ll 1$, then for equation (1) one may by approximation write:

$$I \approx I_0 \left\{1 - \tfrac{1}{2}(\omega Cr)^2\right\} \quad . \; . \; . \; . \; . \; . \quad (2)$$

With *r* again 1000 Ohms and *C* = 1000 pF we find for a frequency of 50 c/s $\omega Cr \approx 3\cdot 10^{-4}$, so that the frequency error $^1/_2\, (\omega Cr)^2$ is only about $5\cdot 10^{-8}$, thus amounting to $5\cdot 10^{-6}$%. For a frequency of 50 000 c/s however, $\omega Cr \approx 0.3$, corresponding to an error in indication of about 5%. Therefore, to keep the error below a certain value within the largest possible frequency range one must make *Cr* as small as possible. Now *C* is proportional to the effective surface of the rectifier, so that it is obvious to select a small surface. This is, it is true, accompanied by a higher value of *r* but the increase of *r* is much less than proportionate to the reduction of the surface, in the first place because $r = r_m + 2r_d$ partly consists of the unchanged resistance $r_m$ of the moving-coil meter and further because the rectifier resistance $r_d$ increases less rapidly than one would expect. This last feature is related to the curvature of the characteristic, which decreases according as the current density is raised. *Fig. 12* illustrates that when halving the effective surface of a rectifier the voltage loss, for the same current, increases by less than a factor 2.

The relation between the momentary values of voltage loss and current is not constant; $r_d$ is meant as an average value of this relation. It is therefore clear that for given valves $r_d$ will depend upon the current amplitude.

This brings us to the deviations from the linearity of the scale, which are particularly evident in the case of voltmeters for low voltages where $R \ll r_m + 2r_d$ no longer applies, so that the non-linearity of $r_d$ has some effect. Here, too, it is therefore of importance to select a high current density and thus to use small rectifiers [5]).

This explains why the diameter of the discs of type I — which have been particularly developed for use in combination with moving-coil meters — has been fixed at only 1.2 mm.

Before we leave the application of selenium rectifiers in meters let us consider for a moment the safeguarding of meters against overload. Let us suppose that we have a rectifier connected

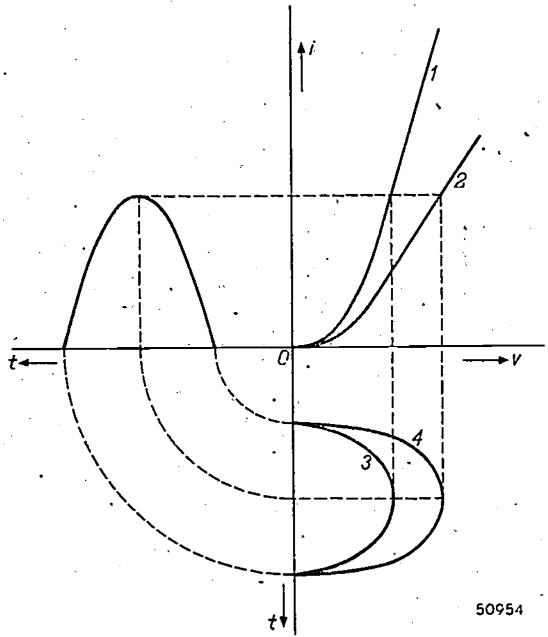


Fig. 12. (*1*) and (*2*) are characteristics of the current *i* in the transmitting direction as functions of the voltage *v* of two similar cells whose effective surfaces are as 2 : 1. The voltage loss in the cells when the same current passes through (current *i* represented on the left as function of the time *t*) is given by the curves (*3*) and (*4*). The amplitude of (*4*) is less than twice that of (*3*).

[5]) From this it follows that a meter will show a smaller frequency error at the full deflection than at a smaller one. As regards the influence of frequency in the case of low voltage meters this is less according as *R* is smaller compared with *r*, as is easily seen both from fig. 11 and from equation (1).

in parallel to a moving-coil instrument in the manner shown in *fig. 13a*. The dimensions of the meter and the rectifier are such that the voltage loss occurring at full deflection lies below the threshold voltage of the rectifier. The resistance of the rectifier is then much higher than that of the meter, so that the connection of the rectifier does not influence the deflection of the meter. Now if the total current $I$ (fig. 13$a$) increases to such an extent that in the absence of the rectifier the meter would be damaged, then by connecting the rectifier in parallel this danger is considerably reduced. The fact is that with sufficiently high voltage the resistance of the rectifier drops to a fraction of the meter resistance, so that only a small portion of the total current passes through the meter. A normal type of meter for 0.1 mA for instance has a resistance of 15 000 Ohms, so that for the full deflection 0.15 V is required. At that voltage the resistance of the rectifier is approx. 0.5 Mohm, thus more than 300 times as high as the meter resistance. With a current 6 times as strong flowing through the meter — which it can withstand for some time — the voltage becomes 0.9 V, at which level, as follows from the characteristic of the rectifier, more than 9 mA flows through the rectifier and the total current thus amounts to approx. 10 mA, *i.e.* 100 times the nominal value.

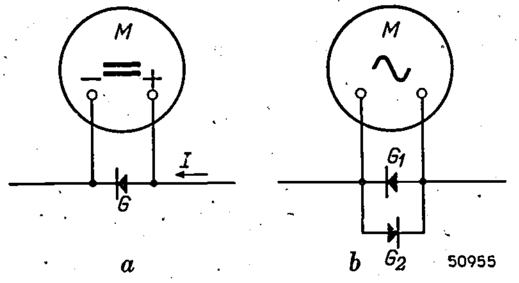


Fig. 13. *a*) Direct current meter $M$ safeguarded against overload by the rectifier $G$ connected in parallel.

*b*) For the protection of alternating current meters two rectifying cells ($G_1$, $G_2$) are used, connected anti-parallel.

For the safeguarding of A.C. meters two rectifiers are used in anti-parallel connection (fig. 13$b$).

A somewhat analogous application of selenium rectifiers has been previously described in this journal [6]), where use has been made of the shape of the blocking characteristic to get a linear decibel scale on a measuring instrument.

## 2) *Selenium rectifiers in modulators*

In carrier-telephony circuit elements are needed that have a non-linear characteristic, both on the transmitting side in order to modulate the low-frequency audio vibrations on one of the carrier waves, and at the receiving end for the reverse process. As already described in this journal [7]), selenium rectifiers lend themselves for this purpose, for instance in the so-called double push-pull circuit (*fig. 14*).

It is to be recalled that on the output side of this circuit certain undesired components (*inter alia* those with the carrier-wave frequency itself), which otherwise occur in other circuits are absent here. A condition is, however, that the four rectifiers used in such a modulator must have the same characteristic and equal capacity. This condition is better satisfied with rectifiers of type I than those of type II originally used for this purpose. Moreover, type I has the advantage of a lower threshold voltage and a smaller differential resistance in the transmitting direction, so that a smaller carrier-wave power suffices. In the third place type I constitutes an improvement because at voltages above the threshold value the characteristic is less curved; as a result the output voltage of the modulator is less dependent upon fluctuations in the carrier-wave amplitude.

Cells of the type III also find application as modulator cells, where higher voltages are required than the other types are able to withstand. Such is the case with the modulator in the so-called signal receiver; this is a component part of the signalling mechanism in carrier-telephony [8]).

## 3) *Selenium cells in small rectifiers*

We will not conclude this summary without at

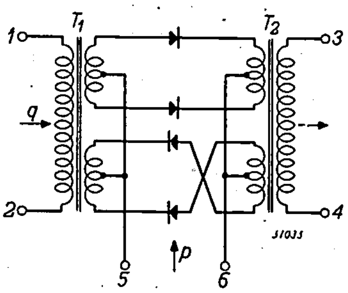


Fig. 14. Modulator in double push-pull circuit. Voltage with the audio frequencies $q$ is applied to the terminals *1* and *2*, whilst voltage with the carrier frequency $p$ is applied to the terminals *5* and *6*. The modulated voltage is taken off at terminals *3* and *4*.

[6]) F. de Fremery and J. W. G. Wenke, The measurement of peak voltages in a studio installation, Philips Techn. Rev. 7, 20-23, 1942.

[7]) F. A. de Groot and P. J. den Haan, Modulators for carrier telephony, Philips Techn. Rev. 7, 83-91, 1942.

[8]) In the case of a telephone link signalling is understood to mean the apparatus required for exciting and transmitting signals for calling, dialing, etc; see F. A. de Groot, Signalling in carrier telephony, Philips Techn. Rev. 8, 168-176, 1946.

least mentioning an important but only vaguely definable field of application of small selenium cells. We refer to those cases where the voltage obtained by rectification serves, for instance, for the activation of a relay or as grid or anode voltage of an amplifying cell. With the supply and charging rectifiers discussed in one of the articles already quoted (see footnote [2]) it is of course impossible to indicate a sharply defined limit. The applications in question are more or less incidental. Several of them have already been mentioned at some place or other in this journal; for instance there is the cascade circuit illustrated in *fig. 15*, which transforms 220 V alternating voltage into 1200 V direct voltage without a transformer [9]), and further the rectifier for supplying the anode voltage in radio receivers with extremely small dimensions [10]).

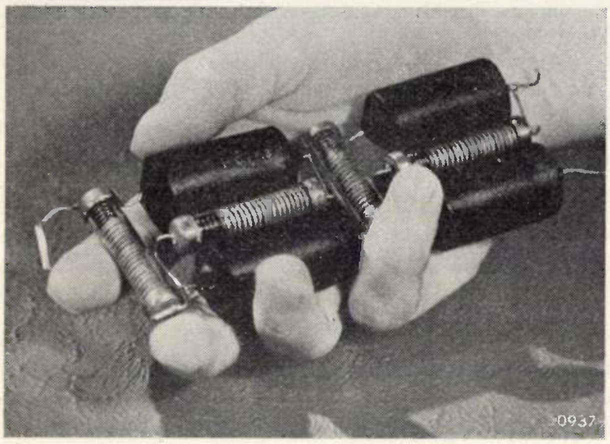


Fig. 15. Cascade connection consisting of four tubes, each containing about 25 selenium cells in series, and four capacitors. With this device the direct voltage of 1200 volt is obtained from 220 V alternating voltage.

Where it is a matter of voltages of some tens of volts or more, rectifiers with an artificial blocking layer, type III, will as a rule be indicated, since they can bear higher voltages in the blocking direction than the other types, so that a smaller number of them are needed.

[9]) Philips Techn. Rev. 6, 78, 1941.

[10]) Philips Techn. Rev. 8, 338, (fig. 3), 1946 (No. 11).

# STRESSES IN GLASS AND THEIR MEASUREMENT

by A. A. PADMOS and J. de VRIES.

666.115:539.319

In the manufacture of glassware it is usually necessary to liquefy the glass by heating and then to let the shaped objects cool down. In this process stresses will often arise which may cause the glass to crack. The same is the case when fusing together glass parts or glass to metal when the coefficients of expansion of the component parts differ appreciably. The determination of stresses in glass objects can best be done by investigating the resulting double refraction. Usually this is performed with the aid of crossed nicols as used when studying the optical properties of plates of crystal. It should first be investigated, however, whether the materials possess those properties which make them suitable for fusing together. Philips are employing a method where small plates of the glass or metal to be used are fused to a standard glass and the stress is deduced from the double refraction, from which it can then be concluded whether the material in question is suitable for the purpose.

The durability of a glass object is to a high degree determined by the stresses in the glass. These stresses are forces acting permanently in the glass body tending to compress it in one place and expand it in another. When an object containing such stresses is subjected to external forces (*e.g.* mechanical compressive or tensile force, or a sudden change in temperature, which is always attended by expansion or shrinkage) these forces will much more readily cause cracking or fracture then when the object is not already affected by internal stresses.

In the manufacture of articles consisting wholly or partly of glass it is therefore of great importance to know something about these stresses and their magnitude.

In this article we shall first give an idea of some of the causes of stresses in glass and then discuss the method which best lends itself to an investigation of the stresses, whilst finally it will be indicated how harmful stresses in the finished products can be avoided by systematic testing of materials.

## How stresses arise

At room temperature and under normal load glass behaves as an elastic solid, that is to say it recovers its original form as soon as a force ceases to act on it. Above a certain temperature glass no longer has this elasticity, and this temperature level differs considerably for various kinds of glass; generally speaking it lies between 300 and 1300 °C, but for the most common kinds lies between 350 and 500 °C. This means that at such a temperature the mass is more or less kneadable, whilst at higher temperatures it is still more easily shaped, and upon a further rise in temperature becomes semiliquid and finally liquid. Thus there is a continuous transition from the solid to the liquid state, contrary to what is found to be the case with melting ice or melting metals and in general with crystalline substances, which show a discontinuity in the transition.

In many cases the possibility of being able to mould glass depends upon this property. In glassblowing a hollow object is blown out of a mass of viscous glass at the end of a hollow pipe and this object hardens in the shape in which it is blown. In glass-drawing a tube often more than 10 meters long is drawn from a previously blown and more or less cylindrical hollow mass of glass, and during this drawing process the tube gradually cools down and becomes more solid until at last the mass has solidified to a glass tube. In glass-moulding a viscous mass of glass is placed in a mould which may said to be the negative of the shape that the mass is required to assume; after the mass has cooled down to a temperature at which no plastic distortion can occur, the object (*e.g.* a tumbler or a lemon squeezer) is taken out of the mould and given the final finishing touches.

The methods described above, which are typical for glass and are made possible by the continuous transition from the solid to the liquid state and *vice versa*, are highly conducive to the formation of stresses. When a more or less viscous mass of glass is cooled down this is naturally accompanied by a drop in temperature from the inside to the outside. This is promoted, moreover, by the lack of heat conductivity, which roughly speaking is only 0.01 to 0.002% of the heat conductivity of many metals. This temperature drop is all the greater according as the cooling takes place quicker. The condition where the glass becomes only elastically mouldable is reached earlier in the surface of the mass than deeper inside it. When upon further cooling also the inside has become solid a state of stress begins to arise, the glass at the surface having meanwhile reached a still lower temperature. The shrinkage taking place in the outer layer as it cools down to room temperature does not correspond to that taking place inside the mass, where the temperature has not yet dropped so low. This explains why in many cases the cooling of glass objects after moulding or shaping has to be done with the utmost care; this often takes place in so-called cooling ovens specially constructed for the purpose. The thicker the glass the more slowly cooling has to be done. It is not until afterwards that it can be determined by testing for stresses whether this cooling has been done properly. If the stress is still found to be too high then the object has to be heated up again and more carefully cooled until the desired stress-free condition is reached.

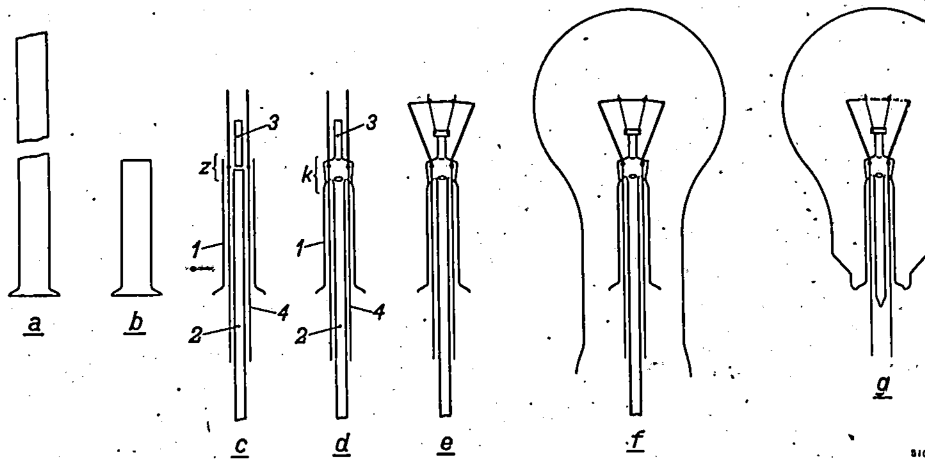


Fig. 1. Some stages in the assembling of the "mount" of the electric incandescent lamp. *a*) The bottom end of a glass tube is conically flanged by heating. *b*) The flanged tube ready cut. *c*) The unit that is to form the bridge; *1* the flange, *2* the pump stem, *3* the glass rod, *4* the conducting wires. Heating the zone *z* and pinching the soft mass of glass falling inwards between two parallel metal jaws forms the "pinch" *k* of the bridge as shown in *d*. *e*) The mount complete. *f*) The mount is placed inside a bulb. *g*) The lamp is completed by fusing the mount into the bulb.

Another cause of the formation of stresses arises in the fusing together of glass parts or of glass to metal. Such joints have frequently to be made for instance in the blowing of laboratory glassware and in the manufacture of incandescent lamps, gas-discharge lamps, radio valves for receiving and for transmission, etc. To give an example we have represented in *fig. 1* the manner in which an incandescent lamp is made. The flanging of a tube, the forming of the pinch, the fusing of the stem onto the pinch, etc. are all parts in the manufacturing process causing stresses in the glass. *Fig. 2* shows a transmitting valve, indicating how metal caps are often used which have to be fused air-tight onto the glass.

In all these cases it is essential that the thermal expansion, or rather the shrinkage, of the materials to be fused together does not exceed a certain maximum difference. If this requirement is not satisfied then, upon cooling, the difference in shrinkage must necessarily give rise to great stresses which result in breakage or at least make the article extremely sensitive to temperature changes. By means of a quantitative investigation of the stresses arising we can form an idea of the differences in shrinkage. But, as will appear farther on in this article, it is also possible from the examination of stresses in materials to form an idea as to their serviceability for the purpose in view.

When dealing in the foregoing with some of the causes of stresses it was assumed that the object was to eliminate these stresses as far as possible, but such need not always be the case. As a matter of fact in the manufacture of lamps and valves and also in the manufacture of some other products it is purposely aimed at setting up certain stresses in order to lend more strength to the article. Such is the case, for example, with toughened glass as used for the window panes of motorcars. With this kind of glass the two surface layers are given a high compressive stress by means of a special cooling method, whilst the inside of the glass is in a state of tensile stress. The object of this is to strengthen the glass against external blows. When a pane of this glass is struck, part of the surface (*e.g.* that opposite where the blow is struck) is always brought into a state of tensile stress. Small irregularities in the surface (scratches and the like), which are always present, then act as centres where this tensile stress accumulates and are therefore apt to cause a break. If, however, the surface layers are previously brought into a state of compressive stress then the stress at those places must first pass the zero value before they become dangerous and cause those

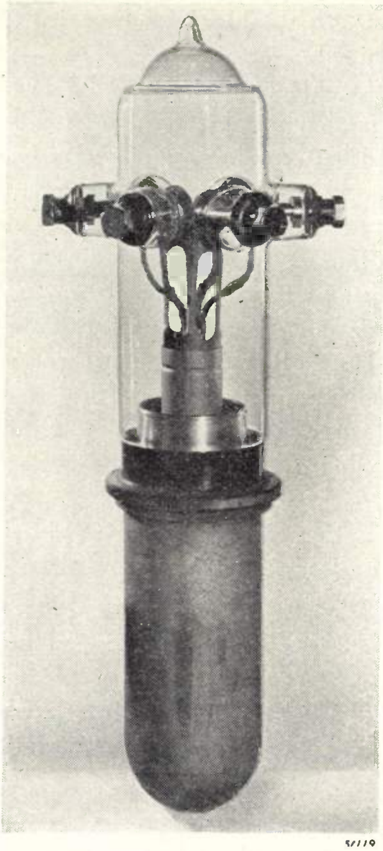


Fig. 2. A transmitting valve. A number of metal caps are fused air-tight onto the glass. At the bottom a so-called "can" is affixed to the glass by means of a chromium iron intermediate piece.

irregularities to act as centres of fracture. The zone of tensile stress inside, having no surface with scratches, does not in the first instance take any part in a possible fracture. It will be understood that by means of this process the glass pane is considerably strengthened. Furthermore, this strongly stressed state has the advantage that in the event of a glass window breaking it will crack up into hundreds of particles of a few millimeters in size which generally still hang together, thus precluding the danger of flying pieces of glass.

The measuring of stresses is indispensable for investigating the process of manufacture of toughened glass, and also in other examples quoted here.

We shall now first consider how stresses can be determined and measured with a more or less high degree of accuracy.

### Determining stresses in glass objects

In glass stresses cause double refraction. By this is understood the phenomenon whereby an incident ray of light is split into two rays of linear polarized light. These rays, which are distinguished as the "ordinary ray" and the "extraordinary ray", have wave planes at right-angles to each other.

This double refraction was first observed in crystals and has been studied by various methods.

#### *Double refraction in crystals*

To examine the double refraction of a crystal the plate is usually placed between two crossed nicols, one acting as polarizer and the other as analyzer, whilst the waves which they allow to pass through are at right-angles to each other.

If the crystal shows no double refraction then no light will be passed through in any position between the crossed nicols. If there is double refraction then the light that the polarizer allows to pass through will always undergo a change in the crystal unless the two wave directions $RR'$ and $SS'$ (*fig. 3*) coincide with the wave directions $PP'$ and $QQ'$ of the two nicols.

If the wave directions of a crystal make an angle with those of the nicols then the beam of light coming from the polarizer is resolved into two beams whose wave directions $Or$ and $Os$ (*fig. 4*) correspond to those of the crystal. In this case the analyzer will let through the components $Or'$ and $Os'$ of a certain wave $Op$ that has passed through the polarizer. Since the beams $Or'$ and $Os'$ are coherent they will interfere. Owing to the double refraction, however, the waves $Or$ and $Os$ advance along the two rays in the crystal at different velocities. As a consequence the one ray falls behind the other, having a retardation (difference of phase) $\delta$, which is expressed as

$$\delta = d\,(n_2 - n_1),$$

where $d$ is the thickness of the crystal plate and $n_2$ and $n_1$ represent the indices of refraction of the crystal for the two rays.

If $\delta = n\,\lambda$ ($n$ = the whole number and $\lambda$ is the wavelength of the incident light) then the waves $Or'$ and $Os'$ will attenuate each other, whereas if $\delta = (2n\text{-}1)\,\lambda/2$ they will amplify each other. The brightness of the emerging light is the maximum when the wave directions of the crystal make an angle of 45° with those of the nicols. When a double-refractive crystal is turned 360° between crossed nicols the field of vision changes four times from light to dark.

A crystal specimen whose wave directions are at an angle to those of the nicols will produce under white light a bright

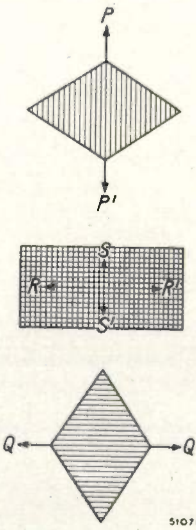


Fig. 3. A crystal plate between crossed nicols. $PP'$ and $QQ'$ are the wave directions that the polarizer and the analyzer respectively allow to pass through. $RR'$ and $SS'$ are the wave directions possible in the crystal.

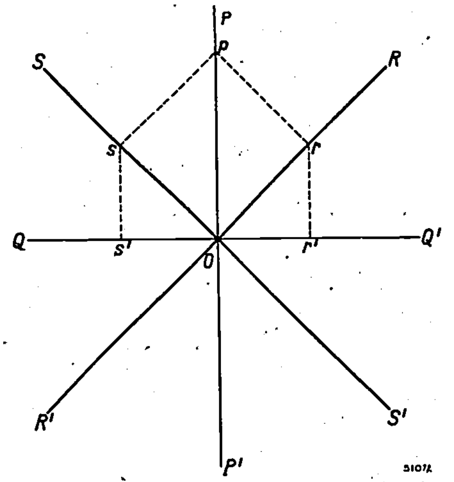


Fig. 4. The wave directions of the crystal ($RR'$ and $SS'$) make an angle with those of the niclos ($PP'$ and $QQ'$). Of a wave $Op$ that has passed through the polarizer, the analyzer allows the components $Or'$ and $Os'$ to pass through. These components will interfere.

field of vision and show a certain colour (interference colour) due to the extinction of the kinds of light having a retardation $\delta = n\lambda$.

Its colour depends upon its thickness. As the thickness gradually increases from zero a whole series of interference colours is observed [1]), which colours are classified under a number of "orders".

Substantially the same methods as employed for studying crystal specimens are applied to determine the existence of double refraction in glass with a view to finding out whether stresses are present. Should a specimen have a weak double refraction then it should show one of the colours of the first order, usually denoted successively by the names iron-grey, lavender-gray and greyish-blue. It is then difficult to see the difference from the black field of vision corresponding to the absence of double refraction. In such a case a plate having a high double refraction and giving a bright interference colour is first placed between the nicols, so that when the specimen in question is then brought into the path of the rays the presence of the weak double refraction is easily seen from the changing colour. For this purpose a so-called "red plate" is often employed, this being a plate of quartz or gypsum of such a size as to produce a retardation of 530 m$\mu$ between the "ordinary" and the "extraordinary" rays. When it is placed between two crossed nicols under white light one observes the interference colour "red of the first order". This is a saturated purple colour making it easy to observe changes due to double refraction in the object being examined.

Such glass objects as have to be examined for stresses in factory laboratories are usually too large to be placed between crossed nicols under a polarizing microscope. Use is then often made of what is called a strain-viewer (*fig. 5*), in which an incandescent lamp is employed as the source of light, part of the light from which falls on a glass plate and is polarized at a suitable angle of incidence. The beam of polarized light passes first through the glass object to be examined (*e.g.* a bottle), then through a red plate and a nicol (or a "Polaroid" filter), finally reaching the eye of the observer *via* an eye-piece.

So long as there is no object lying in the path of the rays one sees through the eye-piece the purple interference colour. Now suppose that the object to be examined has double refraction causing a retardation of 100 m$\mu$. When the ray with the highest velocity in the glass has the same wave direction as that in the red plate the resultant retardation is 630 m$\mu$, as a consequence of which the light having this wavelength (the yellowish-red) is extinguished, so that what one sees is the complementary colour blue. If, on the other hand, the ray with the highest velocity of the glass has the same wave direction as that of the slower ray in the plate, then in this example the light having a wavelength of 430 m$\mu$ (violet) is extinguished and a yellow interference colour is observed. A stronger or weaker double refraction gives rise to other colour effects and the trained observer judges the stresses accordingly, taking into account the thickness of the glass.

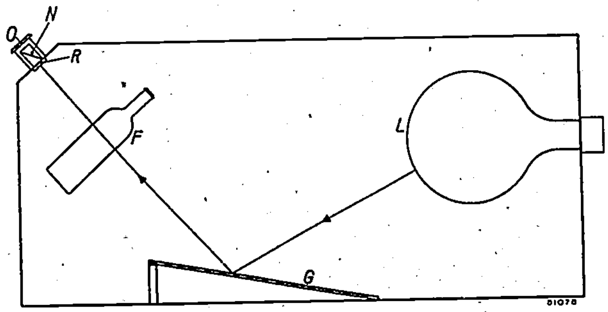


Fig. 5. A strain-viewer for determining and measuring stresses in glass articles. With the aid of a glass plate $G$ the light from the lamp $L$ is polarized and passes through the glass bottle $F$ being tested, a red plate $R$ and a nicol $N$, reaching the observer's eye through an eye-piexe $O$. The stresses present in the bottle can be determined by comparing the colour of the field before and after introducing the bottle.

In the foregoing we have assumed that the polarization planes of the two light rays in the glass were parallel to those of the plate. If that is not so then

[1]) An interesting picture of this series of interference colours is obtained by representing the calculated or experimentally observed colours in a colour triangle as done by J. A. Prins and J. J. M. Reesink, see Physica, The Hague, 12, 396-401, 1946 (No. 6).

we see less intensive colour effects. In practice the object to be examined will be held in the light beam in all sorts of positions until that position is found at which the colour effects show to their maximum.

If it is desired to ascertain whether a stress is compressive or tensile this can be done by introducing a specimen possessing tensile stress (*e.g.* a bent glass plate) in the path of the rays and seeing whether this amplifies or attenuates the effect of the double refraction.

### Measuring stresses in glass

Sometimes something more is wanted than an answer to the question whether the stresses lie below a permissible maximum, and one requires a quantitative determination of the stresses.

This can be done by measuring the magnitude of the double refraction. The velocity difference expressed in $m\mu/cm$ provides a measure for this. This magnitude of the double refraction is proportional to the mechanical stress in the glass, a constant of proportionality varying for different kinds of glass. By way of illustration it is to be noted that in the case of lime glass a velocity difference of 265 $m\mu/cm$ corresponds to a stress of 1 kg/mm$^2$.

In such a quantitative test a so-called compensator is employed in the place of the red plate. As a matter of fact the whole arrangement calls for more care. As polarizer of the light, instead of a glass plate a nicol or sometimes a "Polaroid" filter is used. As a rule a polarizing microscope is employed with small magnification. Usually plan-parallel faces are ground on the test object at the place to be examined. This avoids troublesome reflections at the surface due to oblique incident rays, and also total reflection, which may seriously impede the passage of light. It is then at the same time possible to determine exactly the thickness of the glass. One should further take into account the orientation of the polarization planes in glass with respect to those in the compensator.

As an example of such a compensator we would mention the Berek compensator (*fig.6*). This contains a small plate of Iceland spar which by means of a micrometer screw can be placed at an angle in the pencil of rays. The magnitude of the double refraction thereby caused is determined by the slope of the plate with respect to the optical axis of the microscope. With the aid of a table this double refraction can be calculated from the reading of the scale on the micrometer screw. This table can be compiled by taking measurements, for instance, on a specimen with known double refraction. With this compensator the double refraction of the test piece is compensated when the eye-piece is directed upon the spot to be examined and shows a dark field.

The methods described here can be applied for determining or measuring stresses in finished glass articles. These stresses, as we have seen, mainly arise when different kinds of glass are fused together or when glass is fused onto a metal. It is more recommendable to try to reduce stresses by previously examining the materials to see whether they are suitable for fusing together. This is what we shall now discuss.

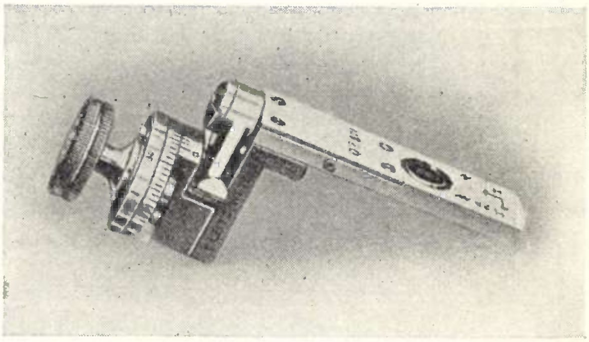


Fig. 6. Berek's compensator. A small plate of calcareous spar is placed obliquely in the pencil of rays by means of a micrometer screw. The magnitude of the double refraction thus set up is determined from the reading of the scale on the screw.

### The prevention of stresses by pre-testing the materials

The occurrence of stresses between two kinds of glass or between a metal and a glass results from a difference in the thermal expansion coefficients of the materials.

When studying the thermal expansion of different substances one must bear in mind that the expansion coefficient depends upon the temperature. This is further illustrated in *fig.* 7. The lines *a* and *b*, representing the expansion curves of two kinds of glass, show a bend at their point of transformation, *i.e.*, roughly speaking the lowest temperature at which stresses can be neutralized by lengthy heating. Below this temperature these graphs sometimes assume a straight line, whilst in other cases they are curved. The average expansion coefficient of the glass *a* between room temperature $t_{,}$ and the temperature $t_1$ is found by dividing the length $Pt_1$ by the temperature difference $t_1$-$t_0$.

When two materials having exactly the same expansion curve are fused together and cooling is done properly there is no reason to fear that differences in shrinkage and thus stresses will arise. In practice, however, this will very seldom be the case. With glass situations may arise as indicated in fig. 7;

in the case of a metal we often have a perfectly straight graph. The question is then at what temperature the expansion coefficients should so closely correspond as to ensure that the final product will have the least possible stress or no stress at all. When fusing glass to metal this temperature is that at which the glass is only just elastically mouldable, or in other words that at which it solidifies. This is not, it is true, a sharp definition for that temperature but it can be used by approximation. In cases where glass is fused to glass it is the temperature at which the glass having the lowest softening point solidifies, for stress only begins to arise when both the kinds of glass have solidified and start shrinking in different ways.

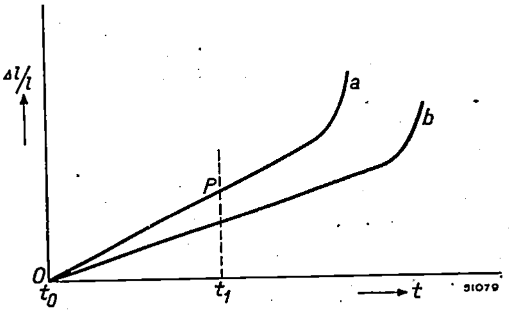


Fig. 7. Graphic representation of the expansion of two kinds of glass, *a* and *b*. The horizontal axis gives the temperature in °C ($t_0$ is room temperature) and the vertical axis the expansion per unit of length. The expansion coefficient between $t_1$ and $t_0$ for glass *a* is found by dividing $Pt_1$ by $t_1 - t_0$.

Therefore, in order to get an insight into the stresses to be expected as a result of differences in expansion, it is necessary that two factors should be known:

a) the temperature at which a glass solidifies, and
b) the expansion coefficients of the two materials to be fused together in the range from room temperature up to the aforementioned solidifying temperature.

This is no simple matter. Only a critical observer with sufficient knowledge and experience can determine the expansion coefficient accurately. In measurements up to 400-500 °C an error of 1% can only be avoided by taking very careful precautions. When such an error is made in determining an expansion coefficient of say 96 × $10^{-7}$, a value that frequently occurs, this means an error of 1 × $10^{-7}$, which is prohibitive.

Now this measuring of expansion coefficients can be replaced by a stress test. This simplifies matters a great deal and in a mutual comparison of the expansion of materials allows of an accuracy of ± 0.15 × $10^{-7}$. In the Philips works a method has been developed on this basis for the routine testing of glasses and metals for fusing [2]). It has been applied successfully for 9 years already. It can be carried out by a person without any special training, provided he works accurately. One man can deal with 15 to 20 samples per day. Compared with the direct determination of expansion coefficients this method means considerable saving in time.

In practice the test is carried out in the following way. The specimen of glass to be tested is fused onto a standard glass, both plates being cut to the size of about 1 × 1 × 0.3 cm. This and also the following processes are carried out in the flame of an ordinary glass-blower's lamp, using lighting gas and air or if necessary oxygen. These plates are ground flat on one side, that of 1 × 0.3 cm. They are then fused together under the flame so as to form one single plate of 2 × 1 × 0.3 cm. The boundary plane between the two plates has been in the liquid state for such a short time as to make the transition between the two glasses clearly visible.

While it is still hot the plate is placed in an oven where the temperature is 100-150 °C higher than the transformation temperature of the glass to be tested. It is then cooled down at a rate of about 2°C per minute [3]) to a temperature of 200-250 °C and then taken out of the oven, leaving it then to cool down rapidly to room temperature. After this treatment there is only a permanent stress, if any, in the plate caused by the difference in expansion coefficient of the two kinds of glass. This permanent stress is the greatest at the boundary plane of the two glasses and is measured in that plane in the middle of the plate.

*Fig. 8* indicates how the plate is placed in a polarization microscope with crossed nicols. We have to determine the stress parallel to the boundary surface, taking care that this boundary surface makes an angle of 45° with the directions of the polarizer and the analyzer. The field will then lighten up when this lighting effect is caused to disappear by means of a compensator. It goes without saying that the double refraction in the plate and that in the compensator must be opposed to each other; should that not be the case than the plate has to be turned 90°.

This is the manner in which the stress in the test glass is measured with respect to the standard glass. The standard glass is so chosen that its expansion

[2]) The fundamental elements of this method were laid down by J. Smelt, following upon a publication by F. Späte Glastechn. Berichte 2, 1-19, 1924.

[3]) When varying the rate of cooling from 1°C to 4°C per minute no differences were observed in the measuring results.

coefficient does not differ too much from that of the glass to be tested. For kinds of glass having a greatly deviating expansion we choose another more appropriate standard glass.

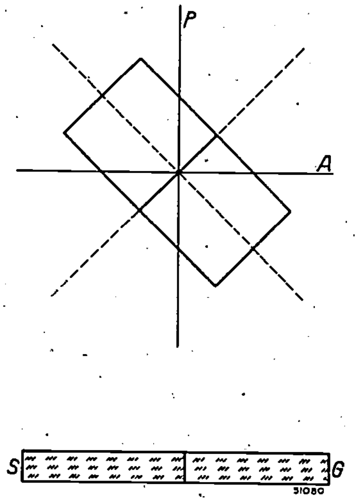


Fig. 8. The manner in which a test plate of glass $G$ fused onto a standard glass $S$ is placed in the polarizing microscope. $P$ and $A$ are the respective directions of the waves allowed to pass through the polarizer and the analyzer.

When we find compressive stress in the standard glass this means that the expansion coefficient of the glass to be tested is greater than that of the standard glass, and if we find a tensile stress it is just the other way round.

These measurements make it possible also to compare tested glasses one with the other. A glass $A$ which causes a stress $a$ in the standard glass; sets up in glass $B$, which produces a stress $b$ in the same standard glass, a stress $a$-$b$ ($a$ and $b$ are positive in the case of tensile stress and negative with compressive stress).

These measurements lead to an effective means of testing materials. There are a certain number of types of fusings and from years of experience we know that stresses found by this method to exist between the materials are admissible and what stresses cause cracking. These practical experiences are thus decisive for the question whether a certain material is useful and whether certain fusings are practicable or not. It is impossible to lay down general rules for this, because the admissible stresses are too closely related to the fusing technique followed and to various circumstances such as the shape and nature of the articles and the purpose for which these are eventually to be used.

At first sight this empirical method would seem to be unsatisfactory. But it is impracticable to apply a more exact method where the transformation temperatures and elasticity constants of respective materials are first determined so as to calculate from these data what stresses may be expected and in how far these approximate the tensile strength of the glass. Even if the expansion coefficients are known with sufficient accuracy it is not possible to detemine by any theorectical method whether the materials are useful when taken together. Against such a method there are the following objections:

a) The nature of the fusing is usually such as to make any calculation of the expected stress out of the question.
b) The temperature changes occurring in the further processing or subsequently in practical use are so inadequately definable that they cannot at all be approximated theoretically.
c) The tensile strength of the glass is itself a quantity difficult of determination, for it depends to a large degree upon the nature as well as the shape of the surface and upon the duration of activity of the tensile forces.

If instead of determining stresses one were to adopt the method of measuring the expansion coefficients one would still have to come to an empirical criterion for what is permissible and what is not. In that case stress measurements are more attractive, as explained above, for the results can be expressed in terms of m$\mu$/cm retardation.

The tests described here are being applied both for glasses and for metals, though in the latter case in a somewhat different way, but there again a standard glass is fused on and used as transparent measuring object.

Experience teaches that as a rule two materials are not suitable for fusing together when they give rise to stresses in equivalent standard kinds of glass for which the respective retardations differ by more than 300 m$\mu$/cm. In simple fusings (*e.g.* for ordinary incandescent lamps) materials are, however, sometimes used which show a stress of 400-500 m$\mu$/cm retardation.

*Fig. 9* shows to what extent the properties of glass from one and the same tank are subject to fluctuations. On the vertical axis we have the observed stress expressed in m$\mu$/cm. Three graphs are drawn, the upper and lower ones indicating the extreme values occurring each week, whilst the middle line indicates the arithmetical average of the values observed. In addition the number of random tests taken is shown for each week. The horizontal broken lines represent the monthly averages and the horizontal fully-drawn line gives the quarterly average.

In the manufacture of X-ray tubes, transmitting valves and rectifying valves metal caps are often used which have to be fused onto the glass bulb vacuum-tight. Examples have already been shown in fig. 2. The quality of the vacuum-tight seal depends for a great part upon the question whether the shrinkage of the metal and that of the glass match each other sufficiently, and this has to be determined in advance by stress testing. Materials like chronium iron (an alloy with expansion coefficient 95-100 × $10^{-7}$, which is suitable for fusing onto soft kinds of glass) and fernico (a ferro-nickel-cobalt alloy with expansion coefficient of about 50 × $10^{-7}$, which is used for harder kinds of glass) when heated to too high a temperature or cooled down to a very low temperature (*e.g.* - 40 °C) undergo transformations of the iron modifications which considerably affect their expansion. The stress test very quickly shows the consequences of such changes. It offers a means of continually checking the sensitivity of the materials for such temperature changes.

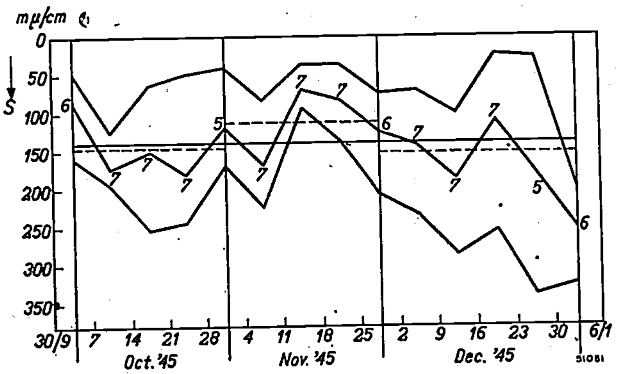


Fig. 9. Example of a graph showing how the stress varies in samples of glass taken at regular intervals from the same tank. The stress is indicated on the vertical axis. Of the three graph lines the upper and lower ones represent extreme values occurring each week, while the middle one gives the arithmetical average of the values found. It is also indicated in numbers how many random tests have been taken each week. The horizontal broken lines represent the monthly averages and the horizontal fully-drawn line the three-monthly average.

In an article recently published in this journal [4]) it was explained how glazing can be applied for the fusing together of glass parts in the manufacture of small radio valves. In such a case it is essential that the glazing should accurately match the glass. This is likewise investigated by taking measurements under the polarizing microscope.

## Stress purposely applied

In the beginning of this article we mentioned that in the case of window panes made of so-called unbreakable glass a compressive stress is purposely set up in the surface to lend strength to the glass. A similar technique is known in the manufacture of lamps and valves, though it is not applied on an extensive scale. We will conclude this article by describing an instance where this artifice is applied, though it must be noted that here the desired state of stress is brought about in a manner somewhat different from that applied for panes of glass.

In mercury lamps with high pressure a discharge takes place in a sealed glass tube mounted inside the lamp. The electric conductor serving to conduct the current from the outer bulb to the inner tube is made of the same metal wire as the electrode (*fig. 10*). The discharge taking place on the extreme point of the electrode (at *o*) heats this point intensely and the direct metallic connection from that point to the so-called bead making the gas-tight seal between the metal and the glass is a good conductor for the heat. Since, therefore, when the lamp is in use that metal wire has a much higher temperature than the glass, the bead is endangered. Assuming that the metal and glass used have the same coefficient expansion and, further, that there is a stress-free state at room temperature, then owing to its higher temperature the wire will tend to draw the glass apart in the axial direction.

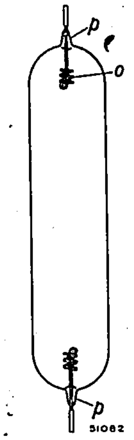


Fig. 10. The inner tube of a discharge lamp containing mercury vapour under high pressure. *p* is the bead for the electrode inlet; *o* is the point of discharge.

With the aid of a tension test we select a glass having a somewhat smaller expansion coefficient than the metal. When the bead joint is made then in the cooling process the metal will shrink more than the glass and tend to compress the glass in a longitudinal direction. When later on under the working conditions the metal expands more than the glass then the existing axial compressive stress will first have to pass through zero before changing into a tensile stress, which as it grows would constitute a danger. Here again bursting of the bulb is avoided and a reliable fusing is obtained by a suitable choice of materials.

[4]) G. Alma and F. Prakke, A new series of small radio valves, Philips Techn. Review, 8, 289-295, 1946 (No. 10).

# EPICYCLIC GEARING FOR LOW POWERS

by A. VERHOEFF.

621.831.061.1

When the motion of a shaft has to be transmitted to another shaft required to rotate at a slower speed, usually worm gcar or toothed gear is employed. In this article a transmission mechanism is described which offers certain advantages when the transmission ratio is high (*e.g.* 500 : 1) and the power to be transmitted is low.

In all sorts of machinery and mechanical plant it is necessary to transmit the rotation of one shaft to another. For this purpose it is a common practice to use worm gear and toothed gear, the ratio of the numbers of teeth on the wheels engaging each other being so chosen as to accelerate or reduce the movement in the desired proportion. In this article we shall confine our considerations to speed-reducing gearing. When the transmission ratio is high (*e.g.* 500:1) such a reduced transmission is sometimes difficult to attain. It necessitates the application of either a multiple gear wheel combination or a system of different toothed gear wheels. As a result, especially in the former case, not only is the efficiency reduced but often also constructional difficulties arise. In such cases, in so far as the power to be transmitted is low, it is advantageous to employ the speed-reducing gearing described below.

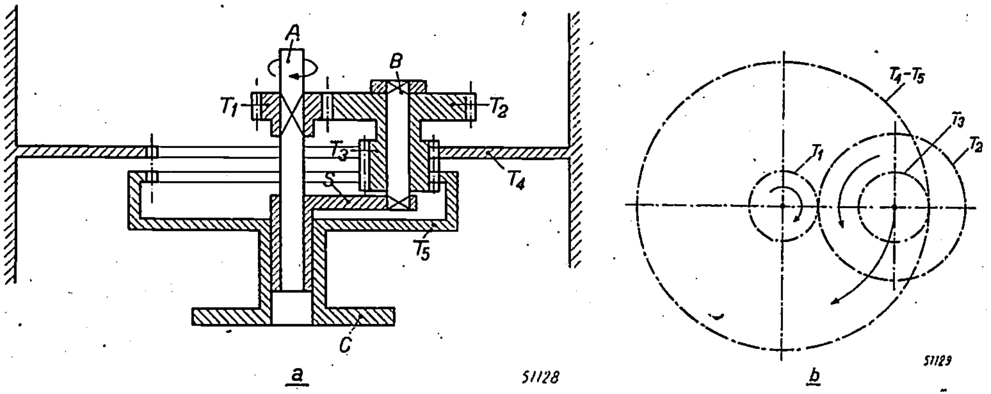


Fig. 1. Diagrammatic representation of the reducing gearing: *a*) cross section of the retarding mechanism; *b*) pitch circles and direction of rotation of the various toothed wheels. The wheel $T_1$ is fixed on the driving shaft $A$. $T_2$ and $T_3$ rotating about the secondary shaft $B$ have commen bearings in the disc or crank $S$ rotating about the shaft $A$. $T_4$ is a stationary toothed rim with inner gearing, whilst $T_5$ has a rotating toothed wheel with inner gearing coupled to the flange $C$ that has to be driven.

The construction is diagrammatically represented in *figs. 1a* and *b*. When studying this diagram it should be borne in mind that $T_4$ is a fixed toothed gear (toothed rim with internal gear) mounted with sufficient rigidity.

By means of the small toothed wheel $T_1$ mounted on the driving shaft $A$ the toothed wheels $T_2$ and $T_3$ on the secondary shaft $B$ are set in motion in the direction of the arrow. These mutually coupled toothed wheels, whose shaft $B$ runs in bearings in a balanced fork or disc $S$ (able to rotate loosely around the shaft $A$), rotate in the direction of the arrow freely about the driving shaft, the teeth of the wide, small, toothed wheel $T_3$ engaging both in the internal gear of the wheel $T_5$. The latter wheel is mounted on the flange $C$, which has to be driven. $T_5$ has the same pitch circle as $T_4$ but, for instance, one tooth more than the latter. As a consequence after every revolution of $S$, owing to the ratio of the gear wheel $T_3$, the toothed rim $T_5$ is shifted to the extent of one tooth with respect to $T_4$. For every revolution of $S$ the gear wheel $T_5$ together with the flange $C$ therefore makes $1/t_5$ revolution, $t_5$ indicating the number of teeth on $T_5$. The speed-reducing gearing thereby reached can now easily be calculated. One must bear in mind that the disc $S$ with the gear wheels $T_2$ and $T_3$ has the same direction of rotation as $T_1$ (see fig. 1*b*). During one complete revolution of $S$, $T_3$ will make one extra revolution around its own axis, so that the number of revolutions found from the ratio of the numbers of teeth $t_4$ and $t_3$ has to be increased by one.

If we term the number of teeth of the five wheels $t_1$, $t_2$, $t_3$, $t_4$ and $t_5$ and

$n$ = the number of revolutions per minute of the driving shaft $A$ connected with $T_1$,

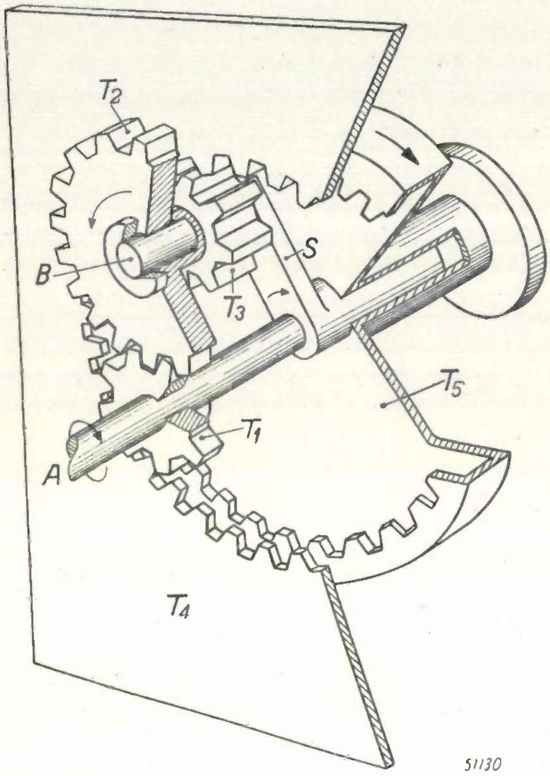


Fig. 2. Schematic drawing of the new epicyclic gearing in perspective. For the meaning of the letters see the text and explanation of fig. 1.

$N$ = r.p.m. of the flange $C$ driven *via* $T_5$, we then find

$$\frac{n}{N}=\left(\frac{t_2}{t_1}\times\frac{t_4}{t_3}+1\right)\times\frac{t_5}{t_5-t_4}.$$

If we now take another example:

$$t_2 : t_1 = 2 : 1,$$
$$t_4 : t_3 = 5 : 1,$$
$$t_5 : t_4 = 45 : 44,$$

we find

$$\frac{n}{N}=\left(\frac{2}{1}\times\frac{5}{1}+1\right)\times\frac{45}{45-44}=495.$$

In this case the reducing ratio is thus 495 : 1.

*Fig. 2* shows how the transmission described is realized. From this illustration and the description given it appears that the high-speed and low-speed shafts lie in the extension of each other's axis and that the mechanism is compact in construction, two properties which make such a construction attractive. As to the practical application it is to be noted that in order to get the two toothed rims $T_4$ and $T_5$ with the same pitch circle and a small difference in the number of teeth, these teeth can be cut with the same chisel, using different divisions on the dividing head of the slotting machine. If $t_5 = t_4 + 1$ the tooth thickness of $T_5$ must, as is readily seen, be $2/t_5$ of the tooth thickness of $T_4$ smaller than that of $T_4$.

Transmission systems employing gear wheels with the same pitch circle but with a small difference in the number of teeth have already been described in literature before and put into practical application. The essential improvement obtained with the system described here consists in the fact that $T_3$ is driven *via* an intermediate transmission $T_1$—$T_2$. This means a considerable gain in the reducing ratio.

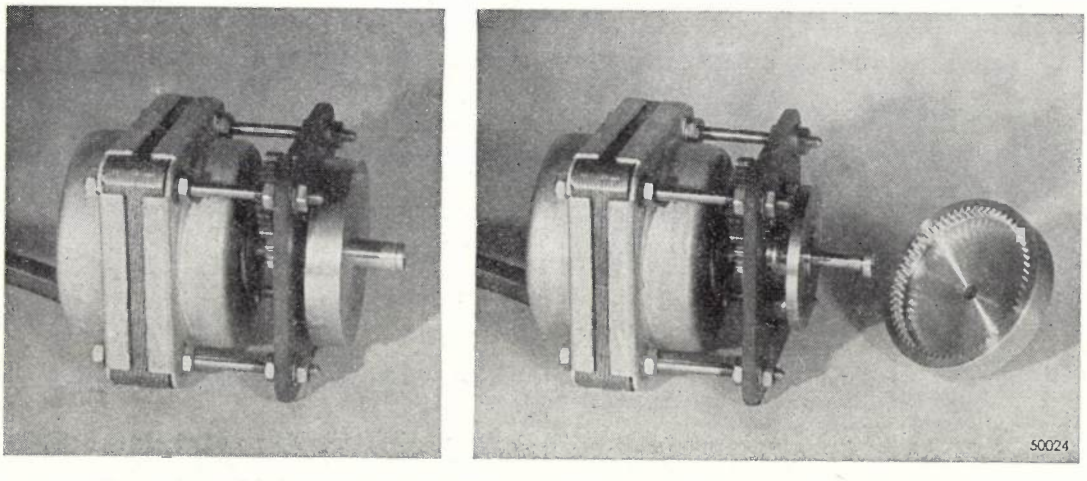


Fig. 3. A speed-reducing gearing mounted on the shaft of a small engine. *Left*: the apparatus ready for use. *Right*: the end wheel $T_5$ has been removed for the sake of clarity. Here the transmission ratio is 500 :1. In more recent constructions a modification has been made in that two or three similar toothed wheels $T_3$ connected with each other move within the toothed rim $T_4$, which makes for smoother running.

The reducing ratio can be varied between wide limits by varying the numbers of teeth $t_1$ and $t_2$ on the wheels $T_1$ and $T_2$. This reducing transmission, however, offers still other possibilities. So far we have assumed that the geared rim $T_4$ was fixed. If, however, this rim with internal gearing is made rotatable and connected with a transmission to the driving shaft one has unlimited possibilities for varying the reducing ratio. Not only can $T_5$ be kept stationary, but also this wheel can be made to rotate in a direction opposite to the original direction.

*Fig. 3* shows a reducing mechanism according to the system described here which is in use for various purposes in the Philips Physical Laboratory. This mechanism is mounted on the shaft of a small engine and is used, for instance, with high tension installations for the sliding of contacts in regulating transformers operated by remote control, and also for the driving of stirring devices used in chemical experiments.

In such applications as these the efficiency factor is of little importance, but measurements have shown the efficiency to be very satisfactory under small loads (about 80% for a power of 0.1 W). In the transmission of higher powers the efficiency drops rapidly, as is in fact also the case with worm wheel constructions and with the usual gear wheel systems.

## BOOK REVIEW

**"Zendbuizen"**[1], by J. P. Heyboer, 320 pages, 285 illustrations, Philips Technical Library, Part VII of the series on Electronic Valves.

The author of this book had for a number of years taken an important part in the development of Philips transmitting valves. Before sacrificing his life in the underground movement in 1945 he had put his knowledge and experience down on paper and by the publication of this book his knowledge has now been made available to others. The reader will repeateedly sense that the discussions of the various problems are based on the author's own experience and acquired insight.

After two introductory chapters dealing with the components parts and construction of transmitting valves and their classification, chapter III deals with the triode as a transmitting amplifier. Calculations of the current components occurring with impulse excitation lead to extensive considerations of the power output and efficiency, a close examination being made of the various factors limiting these quantities. The results from schematic valve characteristics are compared with those arrived at from characteristics actually measured. In chapter IV the tetrode and pentode are dealt with as amplifying valves in a similar manner.

In the next three chapters the manner of working of the transmitting valve in other applications is examined. Chapter V deals with the modulation of the transmitter amplifier, whilst chapters VI and VII give a discussion of the transmitting valve as oscillator and as frequency-multiplier.

In chapter VIII some special subjects (grid emission, discharges in transmitting valves, etc.) of practical importance in transmitters are discussed. The last chapter is devoted to transmitting valves for ultra-high frequencies, first dealing with the excitation of ultra-high frequencies with feed-back circuits, then discussing the effect of the inertia of electrons on the functioning of a transmitting valve, whilst in the last part of the chapter the action of the velocity-modulation valve and the Haeff amplifying valve is explained.

Finally there is an appendix dealing with some questions concerning the power, efficiency and distortion of low-frequency amplifiers which have to supply the modulating power for telephony transmitters.

[1] The English translation of this book on Transmitting Valves is in course of preparation.

The reducing ratio can be varied between wide limits by varying the numbers of teeth $t_1$ and $t_2$ on the wheels $T_1$ and $T_2$. This reducing transmission, however, offers still other possibilities. So far we have assumed that the geared rim $T_4$ was fixed. If, however, this rim with internal gearing is made rotatable and connected with a transmission to the driving shaft one has unlimited possibilities for varying the reducing ratio. Not only can $T_5$ be kept stationary, but also this wheel can be made to rotate in a direction opposite to the original direction.

*Fig. 3* shows a reducing mechanism according to the system described here which is in use for various purposes in the Philips Physical Laboratory. This mechanism is mounted on the shaft of a small engine and is used, for instance, with high tension installations for the sliding of contacts in regulating transformers operated by remote control, and also for the driving of stirring devices used in chemical experiments.

In such applications as these the efficiency factor is of little importance, but measurements have shown the efficiency to be very satisfactory under small loads (about 80% for a power of 0.1 W). In the transmission of higher powers the efficiency drops rapidly, as is in fact also the case with worm wheel constructions and with the usual gear wheel systems.

## BOOK REVIEW

"**Zendbuizen**"[1]), by J. P. Heyboer, 320 pages, 285 illustrations, Philips Technical Library, Part VII of the series on Electronic Valves.

The author of this book had for a number of years taken an important part in the development of Philips transmitting valves. Before sacrificing his life in the underground movement in 1945 he had put his knowledge and experience down on paper and by the publication of this book his knowledge has now been made available to others. The reader will repeateedly sense that the discussions of the various problems are based on the author's own experience and acquired insight.

After two introductory chapters dealing with the components parts and construction of transmitting valves and their classification, chapter III deals with the triode as a transmitting amplifier. Calculations of the current components occurring with impulse excitation lead to extensive considerations of the power output and efficiency, a close examination being made of the various factors limiting these quantities. The results from schematic valve characteristics are compared with those arrived at from characteristics actually measured. In chapter IV the tetrode and pentode are dealt with as amplifying valves in a similar manner.

In the next three chapters the manner of working of the transmitting valve in other applications is examined. Chapter V deals with the modulation of the transmitter amplifier, whilst chapters VI and VII give a discussion of the transmitting valve as oscillator and as frequency-multiplier.

In chapter VIII some special subjects (grid emission, discharges in transmitting valves, etc.) of practical importance in transmitters are discussed. The last chapter is devoted to transmitting valves for ultra-high frequencies, first dealing with the excitation of ultra-high frequencies with feed-back circuits, then discussing the effect of the inertia of electrons on the functioning of a transmitting valve, whilst in the last part of the chapter the action of the velocity-modulation valve and the Haeff amplifying valve is explained.

Finally there is an appendix dealing with some questions concerning the power, efficiency and distortion of low-frequency amplifiers which have to supply the modulating power for telephony transmitters.

[1]) The English translation of this book on Transmitting Valves is in course of preparation.

## CONTENTS OF THE LAST TWO NUMBERS OF PHILIPS RESEARCH REPORTS

### Vol. 2, No. 4, 1947

*R 49:* On a non-linear noise problem .............. F. L. H. M. Stumpers

*R 50:* Hardening of metals by internal oxidation, part II .............. .......................... J. L. Meijering and M. J. Druyvesteyn

*R 51 :* Minimum-cost chokes ............ T. H. Oddie and J. L. Salpeter

*R 52:* On the amorphous and crystalline oxide layer of aluminium ......... ............................ A. J. Dekker and W. Ch. van Geel

### Vol. 2, No. 5, 1947

*R 53:* Method of measurement of noise ratios and noise factors ........... ........................................... A. van der Ziel

*R 54:* Reflections in electron tubes ........................ J. L. H. Jonker

*R 55:* The temperature-dependence of the fluorescence of tungstates and molybdates in relation to the perfection of the lattice .. F. A. Kröger

*R 56:* Some characteristics of trigonal selenium crystals obtained from the vapour phase .................................... F. de Boer

*R 57:* On the electrical conductivity of selenium crystals ...... F. de Boer

*R 58:* Elastic relaxation and some other properties of the solid solution of carbon and nitrogen in iron .......................... L. J. Dijkstra

*R 59:* The dissociation of nitrogen in the welding arc .......... J. D. Fast

Readers interested in any of the above mentioned articles may apply to the Administration of the Philips Physical Laboratory, Kastanjelaan, Eindhoven, Holland, where a limited number of copies are available for distribution.

## CONTENTS OF COMMUNICATION NEWS

### Vol. IX, No. 1, 1947

Rotating field beacons .................................. F. de Fremery

Frequency control in transmitters .. H. B. R. Boosman and E. H. Hugenholtz

VOL. 9 No. 10, pp. 289-320 1947/1948

# Philips Technical Review

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

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

## THE METAL-DIAZONIUM SYSTEM FOR PHOTOGRAPHIC REPRODUCTION

by R. J. H. ALINK, C. J. DIPPEL and K. J. KEUNING.

773.79

The metal-diazonium system for photographic reproduction, which has been developed in Philips' laboratory in Eindhoven in the course of the last few years, is based on the discovery that when a solution of a diazonium compound and a metal salt, say mercurous nitrate, is exposed to light, atomic metal — *in casu* mercury — is separated. The "latent" mercury image thus obtained can be transformed by physical development into a silver image and intensified. The light-sensitive system is obtained in the form of a film or sheet by impregnating a suitable carrier, say a strip of cellophane 40 μ thick, in a homogeneous solution of the said materials. The metal-diazonium system possesses an extremely high resolving power (> 1000 lines/mm) and allows of working with a very high gamma (6—8) whilst on the other hand also low gammas (1—2) can easily be obtained by varying external factors, *viz.* the moisture content or the intensity of exposure. The light-sensitivity of the system is in cellophane several times (in paper some tens of times) greater than that of the usual diazotype printing papers. This system, which was originally intended only for producing distortion-free copies of Philips-Miller sound film, lends itself excellently, *inter alia*, for the copying of picture-sound films, thanks to the external variability of the gamma. The impregnating of the cellophane base takes place on printing machines designed for the purpose, whilst at the same times these machines are fitted with a device for regulating the moisture content of the base. The good photographic properties of the system and the very low cost of materials open great prospects for its application on a large scale in all sorts of fields. In addition to the sound film and the picture-sound film also the field of micro and macro documentation is regarded as an important domain for the application of the system.

The metal-diazonium system is a new light-sensitive system that has been worked out by a group of scientists in the Philips Laboratory at Eindhoven. It possesses a number of remarkable properties making it exceptionally suitable for the photographic reproduction of pictures as well as of sound. Something has already been said about this system in a previous article [1]), where it was shown by a comparison with the usual methods of reproduction what place the new system could occupy.

Here we shall give a more detailed description of the fundamental principles of the metal-diazonium system and after going more deeply into its properties show how it is realized and p.actised, then concluding by dealing with a number of perspectives for the application of this new reproduction material.

### Principle of the system

*The light-sensitive material*

Diazonium salts have the general chemical formula

$$[R—N \equiv N]^+X^-$$

in which R is an aromatic radical and X some anion. These compounds have been used for quite a time already for photographic reproduction methods. They are most familiar and most widely used in diazotype, a light-printing process for the multiplication of technical drawings and suchlike (tracings) made on transparent material.

Just as with all reproduction methods based on diazonium compounds, the fundamentals of this light-printing process lie in the following properties:

[1]) C. J. Dippel and K. J. Keuning, Problems in Photographic Reproduction, in particular of Sound-films, Philips Techn. Rev. 9, 65-72, 1947 (No. 3).

a) Coupled with certain phenols or amines, the diazonium salts form what are known as azo-dyestuffs.

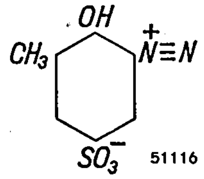
b) When a diazonium salt is exposed to light in the presence of water (water vapour) dissociation takes place and nitrogen is released. Schematically the reaction takes place as follows:

$$RN_2X + H_2O + h\nu \rightarrow \text{l.d.p.} + N_2 + HX \quad (1)$$

The light-decomposition product (l.d.p.) thus formed is no longer capable of forming a dyestuff. If, therefore, one exposes a "diazotype paper" on which a tracing is laid and then causes the reaction sub a) to take place, a dyestuff is formed only on those parts of the papcr that have not been exposed to the light. In this way a positive copy is obtained direct from the tracing.

The difficulties and limitations referred to in the previous article [1]) as being inherent in photographic reproduction with the usual silver-bromide and silver-chloride systems exist in a still higher degree with the diazonium processes hitherto known. They have not met with any appreciable success, for instance, in the reproduction of picture-sound films.

The new photo-chemical system described here is based on the discovery that the light-decomposition products obtained from certain diazonium compounds according to equation (1) are capable of releasing the metal from suitable metal salts, for instance mercury from mercurous nitrate, gold from aurochloride, etc. As diazonium compound one may use for instance o-cresol-diazonium-sulphonic acid:



When an aqueous solution of this compound and, for example, mercurous nitrate, $Hg_2(NO_3)_2$, applied in a thin flat layer is exposed with light of a short wavelength (*e.g.* 3650 Å) one may observe under the microscope small drops of metallic mercury at the places struck by the light. Thus a faint "mercury picture" is formed which might be compared to the "latent" picture in the usual method of photography with silver halogenides. And here, too, the "latent" picture can be intensified and made durable by developing. Contrary to the custom with silver halogenide photography, however, a so-called physical developing process is applied with our system. In view of the important part this developing method plays in the metal-diazonium system a separate chapter will be devoted to it below.

In order to show properly the difference from the old diazonium processes the point is stressed that only the second of the two characteristic properties of diazonium salts mentioned above, *viz.* the light decomposition, is utilized in the metal-diazonium system, the other property — formation of dyestuffs — playing no useful part in our system. The diazonium compound left on the unexposed parts is removed, while the exposed parts turn black (separation of metal). Thus we get instead of a positive copy a negative copy, as is the case with silver halogenides.

As to the practical realization of the system we shall revert to this later, but it is necessary to say something here about the manner in which the thin layer of the light-sensitive material is obtained. In this respect the new system differs fundamentally from the silver halogenide systems. With the lattei systems an "emulsion" (more correctly: a suspension) of the light-sensitive substance, *e.g.* crystalline silver bromide, is made in gelatine and after a complicated and most precise ripening process this emulsion is cast on a celluloid film or a glass plate. In our case, on the other hand, a homogeneous solution is made of the diazonium compound and the metal salt and a suitable carrier is saturated with it. As carrier one may use for instance paper or, as we have done in the most important applications, a transparent base of regenerated cellulose. This latter material (which is more commonly known under the name of cellophane and is widely used as packing material for shop goods) is used by us in the form of a reel of film 0.04 mm (0.0016 inch) thick. But in principle one may also use with our system a gclatinous layer on celluloid or glass. One may also use the cellulose acetate film (so-called safety film) commonly applied for sub-standard films and by saponification make the surface suitable to absorb the aqueous solution of our light-sensitive system.

## *Physical developing process*

What takes place in the so-called physical developing process in photography may be resolved into a phenomenon often met with in nature and in technology and which might be described as follows. When particles from an over-saturated solution or vapour begin to precipitate they show a preference for places where certain "nuclei" are present. If these nuclei are distributed locally in such a way as to form a "picture", maybe so faint that the eye cannot see it, then the picture is intensified by this selective precipitation of particles and may thereby be made visible.

A familiar phenomenon known to everyone is the picture formed by condensation on a smudged glass window when the window is breathed upon or cold air blows upon it from one side, water vapour condensing on the traces of dirt acting as nuclei and making visible figures [2]). Another example that may be given is that of the Wilson camera, where droplets of water from an oversaturated vapour condense on electrically charged particles and thus make visible the path followed by an ionizing particle.

Turning particularly to photography we see that both chemical and physical developing methods aie in use, the latter more particularly in cases where negatives have to be intensified. Let us first consider what ordinary chemical developing comprises. A silver bromide film or plate is placed in a developing bath containing a reducing agent, for instance metol or hydrochinon, in a generally alkaline solution. The grains of silver bromide, in which silver nuclei have been formed by the exposure (groups of say 100 atoms of metallic silver), are reduced by the developer entirely to silver, whilst the unexposed or too weakly exposed grains remain untouched. After this developing, as is known, the remaining silver bromide is removed with the aid of sodium thiosulphate (fixing).

If the picture thus obtained is not dense enough it can be intensified by physical development. The film is placed in a weak acid solution containing silver nitrate in addition to a reducing agent, for which metol or hydrochinon can again be used. In the solution the silver nitrate is gradually reduced to silver, the solution becoming oversaturated, as it were, with atomic silver, which precipitates preferably on those places where "silver nuclei" are already present, *i.e.* on the exposed parts of the film [3]).

It is perhaps of interest to note that this physical developing process can be applied also to the "latent", not yet chemically developed image on the exposed silver bromide film. One can start by fixing the exposed film, thus removing all the silver bromide and leaving only the silver nuclei formed by the exposure, the latter then acting as nuclei for the subsequent physical development as described above.

The oldest photographic process, daguerreotype, was likewise based upon a physical method of developing; an image was formed on an iodized silver plate and the exposed plate was treated with oversaturated mercury vapour. Mercury was thereby condensed on the exposed parts of the plate and the picture became visible.

After these examples the method of physical developing applied with the metal-diazonium system does not need much more explanation. The carrier, containing for instance o-cresol-diazonium-sulphonic acid and mercurous nitrate, is placed after exposure in a suitable solution of silver nitrate and a reducing agent. Reaction then takes place between the locally formed metallic mercury and the silver nitrate, the mercury dissolving and silver precipitating at those places, the latent mercury picture thus being transformed into a latent silver picture. Moreover, by reduction metallic silver is gradually formed in the solution. The local metal deposits already present act as nuclei upon which more and more silver is deposited as new silver is formed in the solution, thereby developing the picture.

One of the essential factors in this process is the rapidity with which the silver is formed in the solution by reduction. If this takes place very rapidly one gets instead of a selective deposit a more or less evenly spread deposit of silver, a phenomenon that can even be turned to advantage for making homogeneous silver mirrors. By giving the solution, for instance, a suitable degree of acidity one can regulate the speed of the spontaneous reduction of the silver nitrate and cause practically all the silver formed during the developing process to precipitate on the nuclei of the picture, to the exclusion of almost all undesired precipitation of silver on the unexposed parts. The negative may not be kept in the developing bath longer than the time taken for developing, because the developer is an unstable system and liable in time to cause a spontaneous flocculation of all the silver, which then precipitates anywhere.

We will not conclude this explanation of the

[2]) In the familiar game of writing or drawing on a glass window with a wet finger, when some days later the writing is made visible by breathing upon it, we have examples of negative pictures; the window is more or less uniformly covered with "dirt", which is removed by the finger (or at least partly so) or else a thin film of grease is left behind, so that less vapour is condensed on the writing than on the rest of the window.

[3]) In this case the physical developing process comprises also a chemical process. From this it appears that the name "physical development", which we use because it has already been introduced, does not express the essential difference from the actual chemical proccss of developing. The difference as we see it lies in the fact that in chemical development the metal from which the picture is built up is already present in the appointed place prior to develloping in any form whatever, whereas in the physical development it is only brought into its appointed place by the developing process.

principles of the metal-diazonium system without remarking that here only a very rough and greatly simplified representation of the processes has been given. To understand the connection between the photographic properties obtained and the numerous variable factors of the system it has been necessary to study deeply the mechanism of the light decomposing reaction [equation (1)] and of the developing process. It may be possible at a later date to go more deeply into the problems arising, some of which are still unsolved.

## Properties of the metal-diazonium system

### *Resolving power*

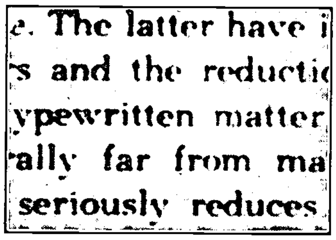
In the article previously quoted [1]) it was explained that for photographic reproduction, especially of picture-sound films, a high resolving power is favourable. However, there is a limitation in this respect, in that it tends to spoil the quality of reproduction (unsharp pictures, distortion of high frequencies), or, when trying to avoid this drawback, it necessitates a greater length of film (the picture may not be too small), the same applying for the speed of the film on account of the sound.

The positive films commonly used for copying picture-sound films have a resolving power of between 50 and 75 lines/mm, so that when projecting gratings finer than what corresponds to this number of lines per mm the lines run into each other and therefore can no longer be made visible separately. Some new Kodak films for special purposes have a resolving power as high as 160 lines/mm. *The metal-diazonium system is quite capable of resolving 1000 lines/mm.* It is even possible that its resolving power is still greater, for the optical means used in determining this power have themselves a limited resolving power, which in our case did not reach further than the said limit of 1000 lines/mm.

In other laboratories too, *i.a.* Kodak, it has for some time been possible to make reproductions with an extremely high resolving power, of the order of 1000 lines/mm [4]). Collodion plates or so-called Lippmann emulsions are used. Except for the already known application of wet collodion plates for making autotypes, as far as we know none of these processes has been suitably developed for general use.

The high resolving power of the metal-diazonium system is due for a large part to the fact already mentioned that the light-sensitive material is not used in the form of an emulsion but in that of a homogeneous solution; the unexposed system is quite free of grains, so that there is very little diffusion of light in the sensitized material. If a large quantity of light reaches one point then in the circle round about that point where undesired light reaches, owing to diffusion, the quantity of that light is extremely small. Moreover — and this is the second cause of the high resolving power — that circle (the diffusion halo) remains extremely small owing to the fact that with the chosen concentration of the chemical components in the light-sensitive solution the active light is very strongly absorbed. Consequently the diffused light does not reach far.

Owing to the strong absorption the direct light thrown upon the carrier upon exposure stays in the top layer. Consequently the metal forming the picture is limited to a thin layer. This is likewise of importance for the high resolving power, for it is not sufficient that details are recorded well-separated in the carrier — they must also be reproduced separately either for copying or for projection. Owing to the fact that one never uses perfectly parallel light beams, the thicker the picture layer the more details are lost [5]).



e. The latter have i
s and the reductic
ypewritten matter
ally far from ma
seriously reduces

50928

Fig. 1. A reproduction, linearly enlarged about 250 ×, of a piece of a micro-document recorded with the metal-diazonium system in cellophane. The size of this piece in the micro-document was 0.012 × 0.017 cm and the height of the letters 12 microns. The good definition of the reproduced letters gives an idea of the exceptionally high resolving power of the system (> 1000 lines/mm).

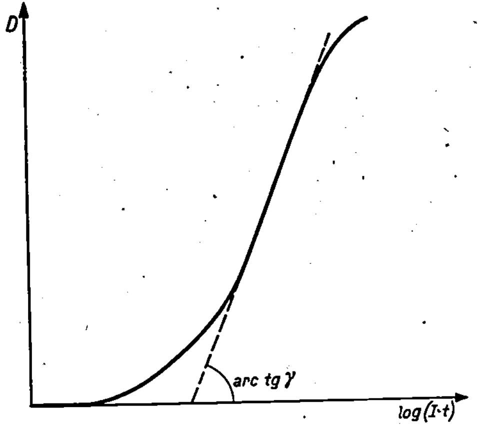
*Fig. 1*, the magnified reproduction of a micro-document made with the new system, demonstrates the high resolving power.

[4]) See for instance J. Sci. Instr. 18, 66-67, 1941, where an account is given of the "Kodak maximum resolution plate", which can resolve 600 or even 1200 lines/mm. Similar results have been reported by H. Frieser, Z. wiss. Phot. 40, 132, 1941. For older methods see E. v. Angerer, Wissenschaftliche Photographie, Akad. Verl. Leipzig, 1931, p. 136 *et seq.*

[5]) Also in the old diazotype processes based on diazonium compounds the material is free of grains, but the resolving power is not very high, because, *i.a.*, the picture is rather thick: owing to the relatively small absorption a thick layer of dyestuff is necessary for adequate "density". Furthermore the formation of the dyestuff is a relatively slow reaction, so that after the exposure a noticeable diffusion takes place before the (fixed) molecules of the dyestuff are formed.

*The gamma value*

Every photographic picture, be it positive or negative, has a certain characteristic density curve which indicates the density obtained for any exposure $E = I \cdot t$ ($I$ luminous intensity, $t$ exposure time). The density $D$[6] is usually plotted as a function of log $E$; see for instance *fig. 2*. For many light-sensitive materials the density curve shows a more or less extended, practically rectilinear part, where at the same time the slope of the curve is greatest. The whole curve is in fact characterised by this slope, the gamma, because this is decisive for the gradation in the reproduced picture (reproduction of the shades of brightness). It is well-known that with the usual silver halogenide systems the value of the gamma depends largely upon the emulsion and is further particularly determined by the conditions of developing (temperature and composition of the developing bath, time taken in developing). Common values in practice vary from 0.5—2.5.



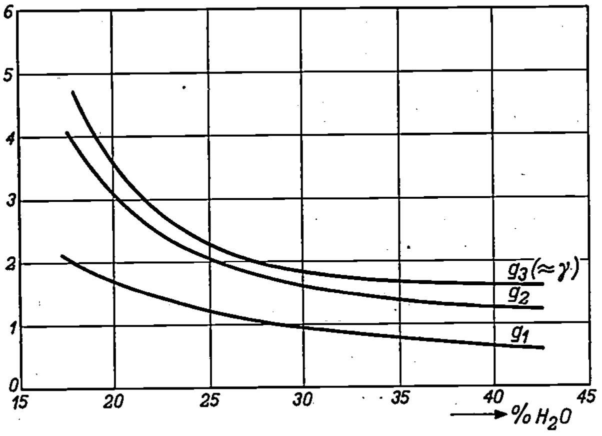
50929

Fig. 2. Example of a density curve $D = f(\log I \cdot t)$ of a photographic picture. The maximum slope of the curve lying in the practically rectilinear part in the middle is the gamma.

When applied in a suitable manner, as will be defined below, the metal-diazonium system has very *much higher gammas, e.g. 6—8*. In the previous article[1]) it has been explained what important advantages this offers for instance for sound reproduction. Since with a high gamma a relatively small reduction in the exposure intensity is sufficient to bring about a transition from the "greatest density" to the "smallest density", only the innermost part of the circle diffusion halo already referred to is noticeably blackened. Thus the high gamma promotes a high resolving power and therefore in the case of sound reproduction promotes good reproduction of the high frequencies. The high gamma yields particular advantages in the copying of Philips-Miller film, as we have seen in the previous article; the "lens effect" arising in this process is rendered harmless without any other measures being necessary[7]).

*External variability of the gamma*

In sound reproduction by the amplitude system advantage can safely be taken of a high gamma, because in principle we have only to do with two densities, a veiy low one in the transparent sound track and a very high one for the rest of the film. There are all sorts of other applications where the same holds, *e.g.* in micro and macro documentation, to be discussed below. In picture reproduction, on the other hand, a whole series of shades of brightness (half tones) has to be reproduced by corresponding, continuously varying densities. To attain this it is necessary to satisfy the Goldberg condition[8]), as a result of which in picture reproduction one has to work with relatively low gammas, *e.g.* 1.5—2.5.



50930

Fig. 3. Effect of the moisture content of the cellophane film upon the gamma of the mercury-diazonium system in the film for a certain concentration of the ingredients and a certain manner of exposure and developing. The lines plotted represent the average slopes $g_1$, $g_2$, $g_3$ of the density curve in three different density areas; $g_3$ is practically equal to the maximum slope $\gamma$.

[6]) Defined as $D = \log i_0/i$, where $i$ is the portion of an incident quantity of light $i_0$ that the blackened plate or film allows to pass through.

[7]) Also the so-called wedge effect with thePhilips-Miller film is rendered harmless by the high gamma. When the transparent sound track in a "Philimil" film is cut with the wedge-shaped chisel, which is a characteristic feature of the Philips-Miller system, an oblique edge is left on the top layer, and when copying this results in a gradual transition from the maximum to the minimum density. In itself this docs no harm, but aberrations arise if there is any variation in the thickness or the density of the covering layer. The higher the gamma, the narrower this edge transition shows on the copy and the smaller the aberration.

[8]) See the explanation in the article quoted in footnote 1.

Now such gammas and still lower values can be obtained quite easily with the metal-diazonium system. This is possible not only by a suitable choice of the composition of the system but also by making use of the following important feature: *the gamma of a metal-diazonium-cellophane system can be greatly influenced by the moisture content of the film during exposure and likewise by the duration of the exposure.* The very high gammas referred to under the previous heading occur when the cellophane film is dry, that is to say when its moisture content is not more than say 15% by weight, and when the film is exposed with a great intensity (and corresponding short exposure time). If the moisture content is increased to 25—30% by weight then one gets the low gammas required for good picture reproduction.

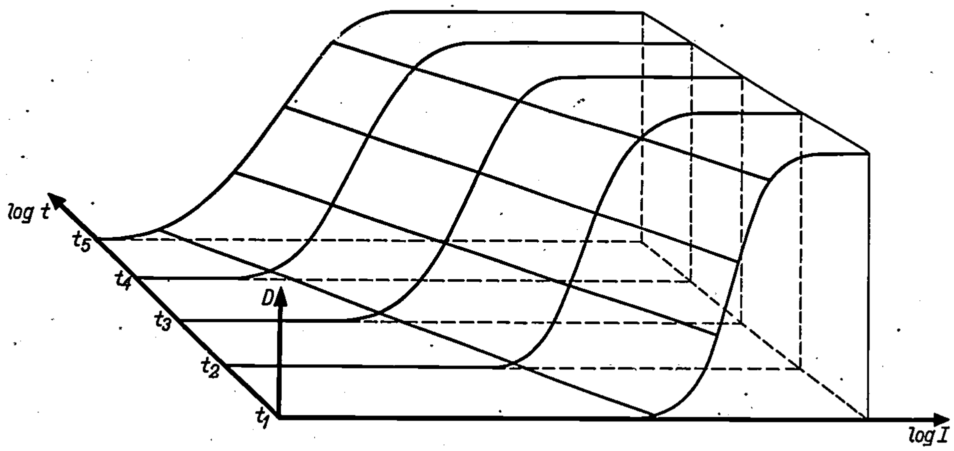


Fig. 4. The gamma of the metal-diazonium system also depends, *i.a.*, upon the intensity of exposure *I*. Therefore the density *D* is not fully determined by the product $I \cdot t$; the photographic behaviour of the system has to be described by a solid density plane $D = f\ (I, t)$. This plane is drawn here for given concentrations, moisture content, developing, etc. in perspective. Log *I* and log *t* are plotted as independent variables.

*Fig. 3* indicates the relation between the gamma value and the moisture content of a cellophane film for certain concentrations of the ingredients and under certain conditions of exposure, developing, etc. There the following refinement has been applied. Instead of showing the gamma indicating the maximum slope of the density curve, the values $g_1$, $g_2$, $g_3$ have been plotted of the average slope that the density curve assumes in three consecutive density intervals of the most importance in practice, *viz.* between $D = 0.05$ and 0.5 (high lights); $D = 0.5$ and 1.0 (intermediate tones); $D = 1.0$ and 1.5 (shadows). The value $g_3$ is generally practically equal to the maximum slope defined as $\gamma$.

This defined description of the gradation, often applied with silver halogenide systems and sometimes even extended to five intervals, is desired when one has a group of density curves which do not all have a rectilinear part with maximum slope in the density area required. This is also frequently the case with the metal-diazonium-cellophane system, with which it is possible to get a large variety of density curves under the influence of the numerous variables.

The fact that the gamma, or in other words the density curve, depends upon the duration of exposure *t* means that the density *D* is no longer a function of the product $I \cdot t$ but of *I* and *t* separately. Therefore instead of a two-dimensional density curve, to describe fully the photographic behaviour of the system we need a solid figure in which the curved plane $D = f\ (I,\ t)$ is represented. Such a plane is drawn in perspective in *fig. 4*. It must be borne in mind that by varying the moisture content, the conditions of developing, etc. a different plane is obtained every time. The "density curve" of a picture taken with an exposure time $t_1$ is the cross section of the density plane parallel to the *D* and *I* axes which is intersected by the *t* axis at $t_1$. From fig. 4 it is clearly seen that for different values of *t* one obtains density curves with a different slope (gamma). A number of these curves are drawn in *fig. 5*, whilst in *fig. 6* the slopes $g_1$, $g_2$, $g_3$ ($\approx\gamma$) of these curves are plotted as a function of *t*.

From this it also follows that the relation between log *I* and log *t* for a constant density *D* (horizontal cross sections of the curved plane at different heights; see fig. 4) is given by lines the slope of which must gradually change with varying *D*. Whereas in the simplest case, which was assumed in fig. 2, *D* was only a function of $I \cdot t$, so that for the given *D* we had $\log I + p \log t =$ constant with $p = 1$, in our case as a rule $p \neq 1$. This is the familiar Schwarzschild effect. Moreover, according to the foregoing, in our case the Schwarzschild exponent *p* (slope of the log *I*-log *t* lines in the area where these may be regarded as straight) depends also upon the density.

Figs. 3 and 6 give a clear picture of the great variability of the gamma in the metal-diazonium system. Now it is important to note that this is an *external variability*: on one and the same material and with one developing process we can make copies with a high gamma and with a low gamma. This creates not only the possibility already mentioned of copying both sound and pictures each with the most favourable gamma, but also an entirely new possibility of copying sound and picture side by side on one film and developing them together without necessitating a compromise in the gamma such as is characteristic for the present picture-sound film technique (see the previous article [1])). All that is necessary is either to select the exposure intensities separately for the copying of the pictures and for the copying of the sound track, or else to vary the moisture content of the cellophane band between the two places where in the printing machine first the picture is copied and farther on the sound track is copied (it may be that both measures have to be applied together).

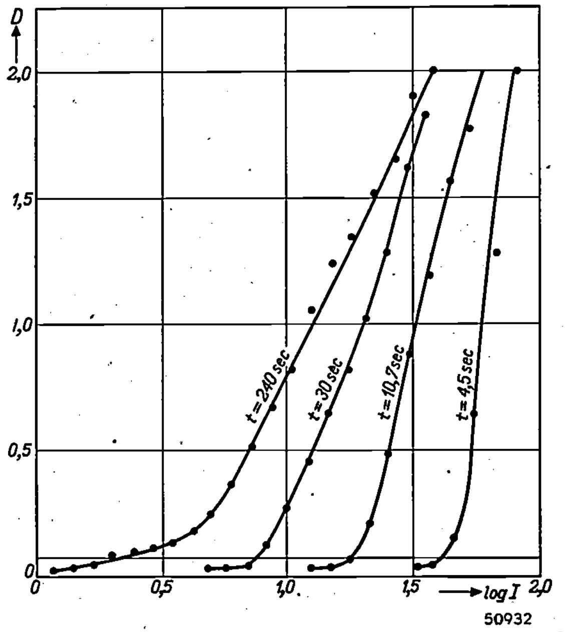


Fig. 5. From the density plane drawn in fig. 4 it is possible to find for any exposure time $t$ the density curve $D = f\ (I)$ of the picture obtained with that exposure time, by taking a cross section of the plane perpendicular to the $t$-axis at the level of that time. Such cross sections are represented here for various exposure times. It is seen that curves are obtained with different slopes (different gammas).

In a similar manner we can also print on one film the pictures of different scenes with a different gamma while requiring only one developing process! This offers the possibility of correcting any exposure or developing variations in the negatives.

*Light sensitivity*

The light sensitivity of the metal-diazonium system depends not only upon the choice of the ingredients, etc. but also varies with diffeient carriers. When cellophane is used the sensitivity may be several times greater than that of the known diazotype papers. Yet with paper as a carrier the sensitivity of the system is greater by a factor of 10. For instance with mercury as metal the metal-diazonium paper can easily be made 20—25 times as sensitive as the positive diazotype papers, a gain which just makes it possible to produce enlargements by the light-printing process (see the last chapter). Still this is a factor $10^4$ below the sensitivity of silver bromide enlargement papers.

There is, therefore, no question that the metal-diazonium system as at present developed could compete with the materials commonly used for photographic recording. As a matter of fact the spectral zone in which the metal-diazonium system is sensitive is too limited for this purpose.. The sensitivity of the mercury-diazonium system lies mainly in the near ultra-violet with a maximum in the vicinity of 3900 Å and not extending beyond the bluish green (about 5000 Å).

For application as reproduction material, however, in most cases neither the low sensitivity nor the limitation of the spectral area constitutes any objection. One only needs to use for the copying process a light source possessing great luminous intensity in the said range of the ultra-violet. Super-high-pressure mercury lamps with water-cooling are excellently suited for this purpose. With these lamps the intensity of light that can be reached when exposing the film is so great that a film can be copied at fairly great speeds, for instance 20 meters per minute.

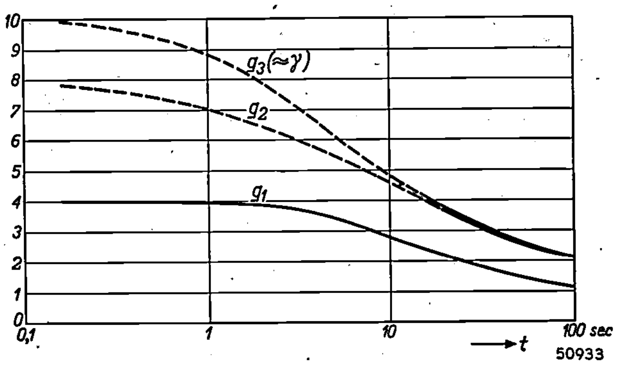


Fig. 6. The slopes $g_1$, $g_2$, $g_3$ of the curves of fig. 5 plotted as functions of the exposure time $t$.

An important advantage of the limited spectral sensitivity lies in the fact that the whole working of the system can take place under a bright sodium light, since in the wavelength range of the sodium lines (5890 Å) the sensitivity is practically nil.

A great deal of work has been done in this laboratory in respect to the question as to what determines the sensitivity of the metal-diazonium system. In the particular case where mercury is employed as a metal it was demonstrated that the process of the formation of the latent mercury picture has a quanta yield of about 50%; for an average of two exposure quanta one atom of metallic mercury is formed. This means that the "primary" sensitivity is of the same order as that of the silver halogenide systems. According to the conclusions provisionally reached from an investigation into the highly complicated mechanism, the cause of the so much smaller resulting sensitiveness of our system is to be sought rather in the further history of the nuclei to be developed; the metallic atoms combine to form larger particles. As already stated, under a very strong magnification (about 800 × ) these particles may be seen in the latent picture. Thus in this stage the metal is rather coarsely dispersed and in the physical developing process the resultant density is all the less according as the (given) quantity of metal of the latent picture is more coarsely distributed.

The influence of the moisture content and the intensity of the exposure upon the gamma is also closely related to the history of the primary metallic atoms and of the metallic nuclei prior to developing.

### *Durability*

When talking of durability in the case of the metal-diazonium system we have to differentiate between the exposed and the unexposed state. The unexposed system appears to have as yet too little durability for a sensitized cellophane film, for instance, to be kept in stock until one has need of it. For the most important applications that we have in mind, however, this forms no serious objection, as will be made clear below.

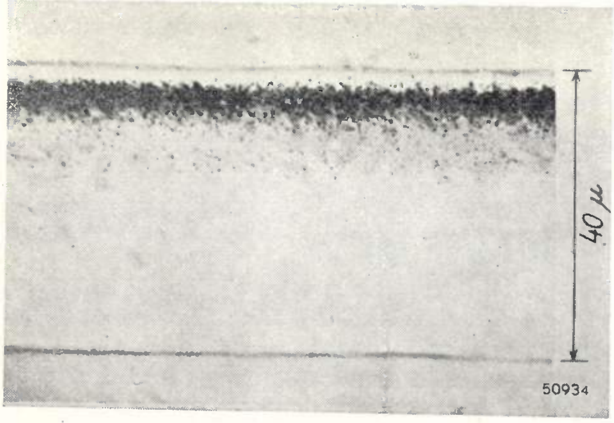


Fig. 7. Microtome cross section of a sensitized cellophane film 40 microns thick after exposure and developing. The particles of silver from which the picture is built up lie in a thin layer a few microns below the surface.

After exposure and developing one has a picture that will keep practically indefinitely: as already mentioned above, the definitive picture consists of metallic silver, just as is the case with silver halogenide systems, and thus is perfectly proof against atmospheric influences and light (such contrary to the dyestuff pictures of the old diazotype papers). Furthermore, the picture on a cellophane film is protected in a peculiar manner against mechanical damage: when a microtome section of such a film is cut and examined under the microscope it will be seen that the extremely thin layer containing the metallic silver of the image (see above) does not lie on the surface of the film but a few microns below it; see *fig.* 7. We cannot go into the explanation of this here, but it has the welcome practical advantage that the actual picture is protected against scratching etc. by the thin layer of clear cellophane covering it.

### *Economy*

It is to be expected that the cost of the metal-diazonium system will be relatively low, a factor that will prove to be of great weight for various applications. How is it that the system turns out to be so economical? In the first place there is the choice of the carrier, cellophane being the cheapest material imaginable for this purpose. Another important factor is the limited consumption of silver, it being a characteristic of the physical developing method that the sensitive material itself need not contain any silver; the silver comes from the developer and is added to the negative, and the developing can be so controlled that not much more silver need be used than is necessary for building up the ultimate silver picture. With the usual silver halogenide systems, on the other hand, a very high percentage of silver is wasted; it comes out of the exposed film in the fixing bath and cannot be recovered except at considerable expense.

Of further importance is the very simple method of manufacture, as will be explained under the next heading.

Yet another advantage to be mentioned in connection with the economy of the metal-diazonium cellophane system is the saving in volume and the attendant ease of storage and transport. A cellophane film 40 microns thick and say 300 meters long winds up into a reel about 13 cm in diameter, whereas a normal celluloid film of 300 meters length forms a reel 26 cm in diameter.

### Realization of the metal-diazonium system

The metal-diazonium system can be realized in quite different ways according to the use intended and the carrier employed. As a typical example we will consider here the application of the system for the copying of picture-sound films with cellophane as the carrier.

Just as is the case with the common silver bromide films, so with the cellophane film the actual copying process takes place on a machine where both the negative and the positive films are caused to pass along under a lamp simultaneously. Now we have already said that the cellophane film is sensitized by impregnating it in its entirety with the solution containing the light-sensitive system. Further it has been stated that the cellophane film sensitized in this manner has only a limited durability. The difficulties that this might involve have now been overcome in a very simple way by *combining the sensitizing process with the printing process*, the impregnating of the cellophane film and the exposure taking place on the same machine in succession.

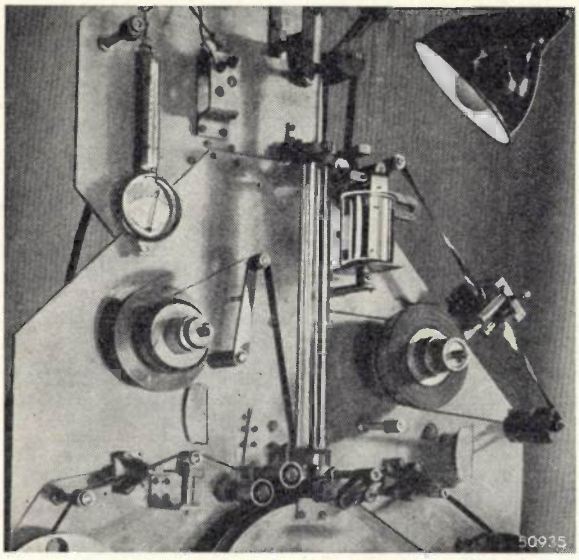


Fig. 8. Printing machine on which a cellophane film is first impregnated in the light-sensitive solution, then dried to the desired moisture content and after that exposed. The film is fed in from the right, impregnated in the bath at the top on the right, dried in the vertical tube and printed on the drum in the middle at the bottom of the photo, where it is brought into contact with the original film to be copied, the two films passing underneath the lamp simultaneously. (For drying the film when running at a high speed several tubes were used; with the latest machine drying is done by high-frequency heating.)

The fact that we have here an extremely simple and economical *modus operandi* is quite evident when comparing it with the manufacture of silver bromide film, where the preparation of the carrier, the preparation and ripening of the emulsion, the casting of the emulsion on the celluloid film and later the copying are all done in separate departments.

The first mentioned method has also been found to improve considerably the reproducibility of the properties of the film.

Between the impregnating and the exposing of the cellophane film this has to be dried to a moisture content corresponding to the desired gamma. Further, it must be possible to reduce still further the moisture content between the exposure of the picture and that of the sound track on the film if such should be desired. This can be done by passing the film through a tube with conditioned air, as seen in *fig. 8*, or by means of high-frequency heating.

### Possibilities of application

The system is still too young to allow of any data being given as to its applications, but the experience so far gained with it opens such interesting perspectives that a brief outline of some of its possible uses may well be given here.

#### *Sound film*

A stereophonic sound track has been made on 7 mm cellophane films by copying a stereophonic Philips-Miller film [9]. Thanks to the sharp definition of the mechanically recorded original and the high resolving power of the copying material an exceptionally good quality of sound is obtained, in respect to the reproduction of the high frequencies and the absence of non-linear distortion (see the article quoted in footnote [1]). This good quality of reproduction together with the enhanced "naturalness" obtained by stereophony seem to us to constitute the requisites for imparting to "mechanical" music the original musical character.

Partly by reason of the low cost of the reproduction method, it is in principle possible that this ideal method of reproducing music will come within the reach of everyone for use in the home. An attraction of the cellophane films used for this method is that a music film with a playing time of one hour forms a reel no more than 18 cm in diameter (playing speed 32 cm per sec); see *fig. 9*. This is due on the one hand to the extreme thinness of the cellophane film (40 microns) and on the other hand to the fact that it is possible to print on a 7 mm film two stereophonic sound tracks (thus in all 4 tracks).

[9] K. de Boer, Stereophonic Recording on Philips-Miller Film, Philips Techn. Rev. **6**, 80-84, 1941.

A piece of music of one hour can be reproduced without any of the interruptions that are unavoidable with the gramophone even when using an automatic record-changer.

*Picture-sound film*

Apart from the cinema there is a wide field of possibilities awaiting the "talkies". It could be used on a large scale for entertainment in the home, for educational purposes in schools, for advertising, etc. provided it is cheap and of good quality. The 8 mm film, which would seem to lend itself best to this purpose, has not yet been widely used because it is still too expensive and owing to the limited resolving power of the common emulsions the pictures are not sharp enough; furthermore it does not leave any room for the sound track. Though the sond track is applied on a 16 mm film the quality of the sound reproduction is not all one would desire.

Now the metal-diazonium system as a copying material presents a situation that is more favourable in many respects: apart from the possibility of making a sound track even on an 8 mm film, on a 16 mm film a better sound quality can be obtained than has hitherto been possible, partly due to the high resolving power and partly by reason of the variability of the gamma. Both of these factors also help in improving the picture quality — as is clearly noticeable on a contact print of a very fine-grained film, *e.g.* Isopan FF — although a limit is set to the sharpness of the picture by the limited resolving power of the original film (about 55—75 lines/mm) and possibly of the optical system of the camera. Finally the film could be cheaper.

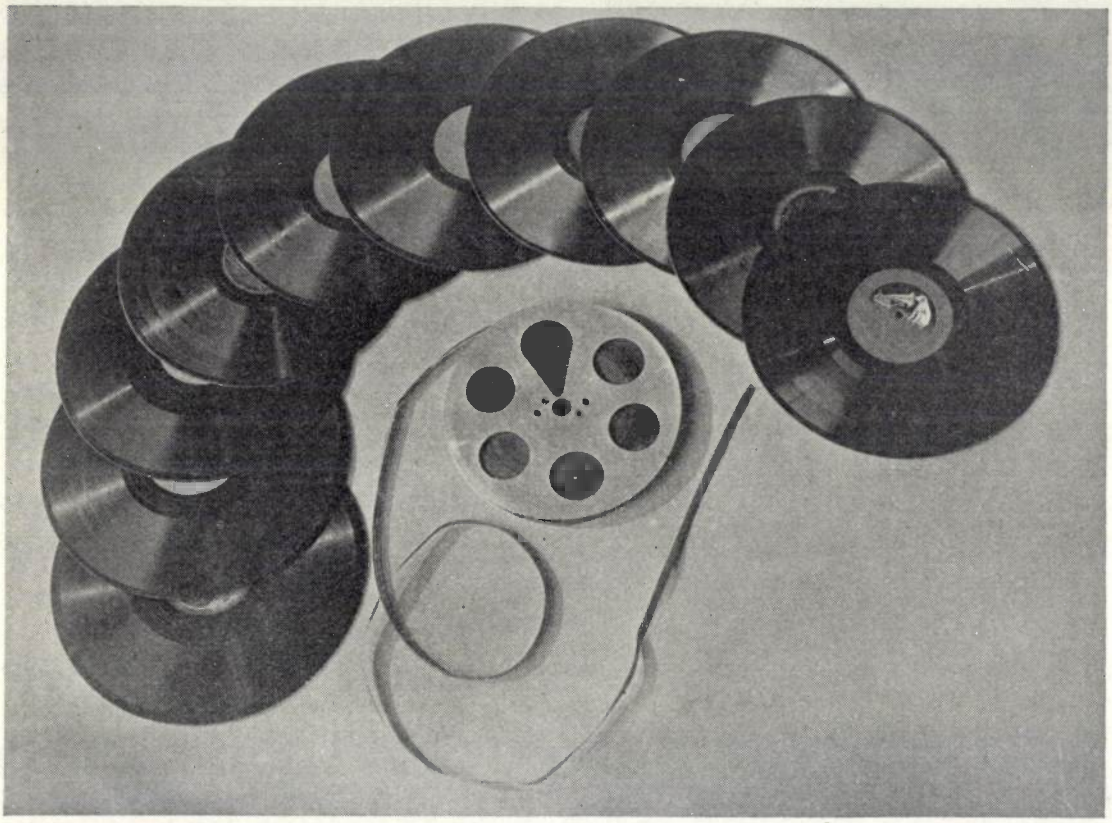


Fig. 9. An illustration of the compactness of a recording of music on cellophane film. Music that takes one hour to play can be recorded stereophonically on a film reel of the size of that shown in the illustration. For the same playing time (without stereophony!) 10 gramophone records of 25 cm diameter are required.

The same considerations apply for the copying of standard 35 mm films on the new system.

For playing the very thin cellophane films special projectors are required. *Fig. 10* shows a model of a home cinema equipped with such a projector. If normal projectors are still to be used then the metal-diazonium system will have to be applied in or on a thicker carrier (say 130 microns thick), but then of course one loses the advantage of the great compactness of film reels with a long playing time.

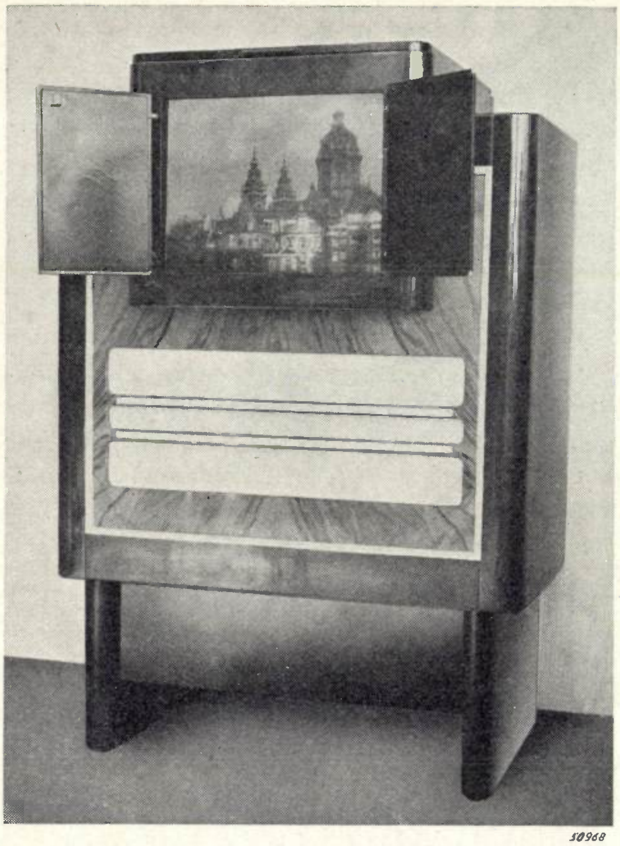


Fig. 10. Model of a home cinema fitted with a special projector (mounted in the cabinet) for cellophane film. The projected picture satisfies high requirements regarding sharpness as well as gradation.

### *Micro-documentation*

Micro-documentation is a comparatively young branch of the technique of reproduction, but it seems that a surprising development may be expected in this very direction.

The name itself already expresses its meaning: the recording of documents on a very small scale. The need for this may be due to various reasons. In some cases it is resorted to because the documents are so bulky or would become so bulky as to constitute a problem for their filing, storage, handling or transportation; examples are the catalogues of large libraries, card index systems of large telephone exchanges or of the registers of births, deaths and marriages of large cities, etc. There are other cases where micro-documentation is applied because reproduction on the normal scale is too expensive, as for instance the copying of publications by libraries, or for the air-mailing of documents where weight is a big consideration, etc.

Obviously the higher the resolving power of the material used for the photographic reproduction, the smaller the size of the reduced document. With the extremely high resolving power of our metal-diazonium system it is possible to make a perfect record of a whole page of printing of the size of this journal on an area of 0.6 by 0.9 mm, the height of the letters being about 12 microns.

It cannot be predicted whether one will ever go as far as such an extremely small size with the present stage of development of micro-documentation. For the present the sizes of for instance 5 × 7 or 2 × 3 mm seem desirable.

If one keeps to the larger dimensions of say 5 × 7 mm per page then the high resolving power of the metal-diazonium system is not utilized to its fullest extent, but even so this material will prove to be of great advantage owing to its low cost. Large tabular works, encyclopedia, etc., which can now practically only be consulted in libraries, could be reproduced on such a small scale on this inexpensive material as to be brought within the reach of anyone having occasion to read such works from time to time. Micro-reproductions can be read with a simple reading apparatus. The saving in volume is astonishing, even with the relatively

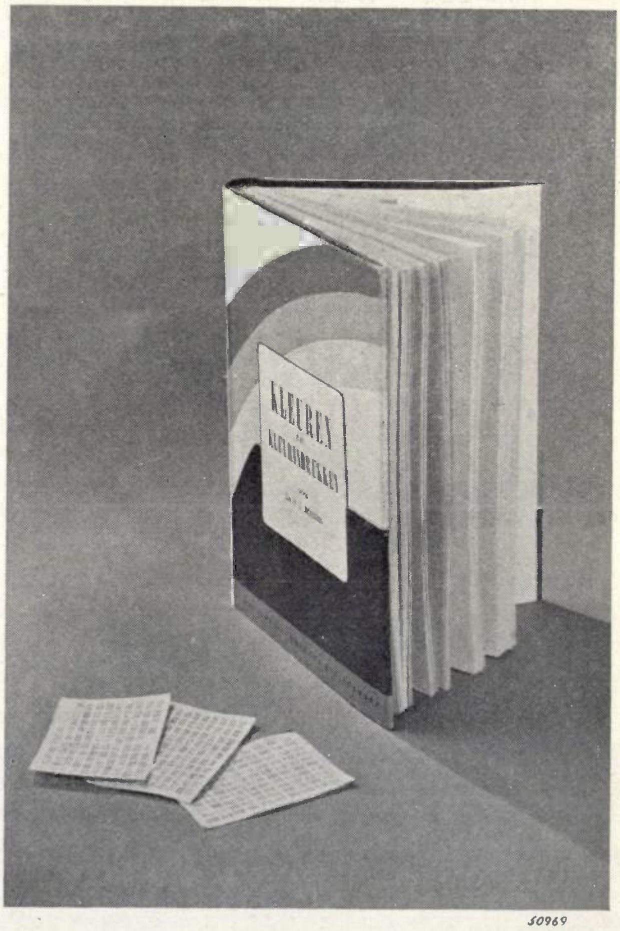


Fig. 11. Three small sheets of metal-diazonium paper on which the complete contents of a book of 330 pages have been reduced. Each page of the book is reduced on the paper to a size of 5 × 7 mm. The reproduced book is quite legible with a simple reading apparatus.

large size of 5 × 7 mm, for a series of large books totalling 10 000 pages can be reduced to a pocket-size booklet of 100 pages. Three of such pages, compared with the original normal book, are shown in *fig. 11.*

*Macro-documentation*

As the last field of application for the metal-diazonium system we would mention that of macro-documentation, which has already become of common usage in the form of diazotype and blue-printing and photocopying directly on a legible scale. Owing to the nature of the documents to which this process is applied (drawings, specifications, etc.), for this purpose the metal-diazonium system would be employed in paper. The paper is sensitized on both sides, so that it can be used on both sides for different copies; it does not curl up, and on account of its two-sided use it sometimes means a considerable saving in volume. Furthermore the image (a silver picture) gives a very rich contrast, with a pleasant tone (neutral grey), and is quite stable, properties which are often so lacking in diazotype and blue-printing processes. Of particular importance, however, is the possibility of making enlargements with metal-diazonium paper with exposure times that are practicable (for instance 1-5 seconds, depending upon the size). Thus we get a very efficient working method: wherever such is desired on account of frequent use, documents recorded on a micro-film can be enlarged again on metal-diazonium paper.

# OPTICAL ABERRATIONS IN LENS AND MIRROR SYSTEMS

by W. de GROOT.

535.317.6

When dealing with technical-optical problems, such as X-ray screen photography and television projection, the need is felt of an optical system with large aperture. Apart from the special lens systems developed for this purpose, Schmidt's mirror system, consisting of an aspherical correction plate, demands attention.

In this article the principal aberrations (so-called third order aberrations) of optical systems in general and of the spherical mirror in particular are discussed, and it is shown how with one exception (the curvature of field) all the third order aberrations of a spherical mirror can be eliminated with the aid of a correction plate.

## Introduction

In the field of illumination, photography, picture projection, etc. innumerable technical problems arise where the need of an optical system with large aperture is felt. Certain special applications, as for instance X-ray screen photography and television projection, have emphasized the need of such a system in the Philips laboratory.

Complicated lens systems with large relative aperture have already been developed for this purpose, but now a second solution, the application of mirror systems, is demanding more and more attention.

Following the invention by Schmidt, who for astronomical purposes fitted a spherical mirror with an aspherical correction plate eliminating the most important aberrations, Philips have developed a mirror system with a correction plate which can be applied for the purpose mentioned above and which is simple to construct.

The object of this article is to examine more closely the aberrations of optical systems in general and of spherical mirrors in particular. The properties and construction of the correction plate will be discussed in a subsequent article.

## Causes of the aberrations

It is already known that, with the aid of a lens or system of lenses whose limiting surfaces are centred spherical surfaces, and also with aspherical concave mirrors, waves of light emitted from a point source can be approximately concentrated at another point (the image point). The fact that this imaging is imperfect is due to two causes.

In the first place after refraction or reflection from a convex surface the spherical light waves emitted from the point source will generally be no longer truly spherical and consequently will not converge into one point.

The wave surfaces are always limited by the edges of the lenses or mirrors and often also by a diaphragm expressly fitted for the purpose. As a rule the aberrations due to the aspherical shape of the wave surfaces become greater the larger the diameter of the limiting apertures.

On the other hand, as a result of this limitation so-called diffraction phenomena arise which also cause lack of definition in the image. The effect of these phenomena is the greater, the smaller the diameter of the limiting aperture relative to the wavelength of the light used. It therefore depends upon the circumstances whether the first or the second cause predominates in the resulting lack of definition. With photographic lenses and with the mirror systems that will be discussed in this article one may, as a rule, ignore the diffraction effects. It is then permissible to substitute for the conception of light waves that of light rays, that is to say very narrow pencils of light which pass through the system independently of each other and are everywhere at right-angles to the wave surfaces. In other words one is satisfied with the approximation offered by geometrical optics.

## First-order aberrations

*Fig. 1* is a representation of the well-known elementary construction of the passage of rays through an optical system with rotational symmetry. The rays emitted from point $P$ at a distance $x_1$ in front of the first focus $F_1$ and a distance $y_1$ below the axis converge at a point $Q$ at a distance $x_2$ behind the second focus $F_2$ and a distance $y_2$ above the axis, where the equations

$$y_2 : y_1 = f : x_1 = x_2 : f = v.$$

apply. Thus the plane through $P$ at right-angles to the axis is uniformly projected with a constant linear magnification $v$ upon a plane through $Q$ at right-angles to the axis. This is only correct to a sufficient approximation when all the rays pass so close to the axis ("paraxial") that the angles formed by all parts of the rays with the axis are so small that their sines and tangents may be interchanged. The quantity $v$ is called the paraxial enlargement.

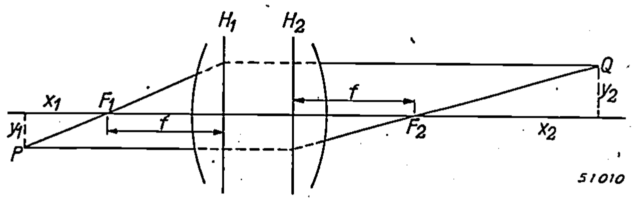


Fig. 1. Elementary construction of the image $Q$ of a point source $P$ in the case of paraxial passage of the rays. $H_1$ and $H_2$ represent the principal planes of the system; the focal distance is $f$.

Even in the simple case of the paraxial passage of rays represented above one may speak of aberrations. Let us suppose (*fig. 2*) that a narrow pencil of rays converges at a point $Q$ in the plane $V$ and a distance $y_0$ from the axis. If the rays are collected in another plane $W$ not coinciding with $V$ then instead of a point of light a small patch is observed in this plane, displaced with respect to the point $Q$. In particular, a ray intersecting a plane $U$ at a point $D$ having as coordinates $y = H \cos \varphi$, $z = H \sin \varphi$, will strike the plane $W$ at a point having the coordinates

$$y = y_0 + a_1 H \cos \varphi + a_2 y_0, \quad . \quad . \quad (1a)$$

$$z = a_1 H \sin \varphi, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1b)$$

in which $a_1$ and $a_2$ are proportional to the distance $d$ between $V$ and $W$.

Let $\varphi$ assume all values between 0 and 2 $\pi$, then the point $y_u$, $z_u$ describes a circle with radius $H$ in the plane $U$. The ray $DQ$ thus describes a conic surface. Next let $H$ assume all the values between zero and a maximum value $H_1$. Then we get a cone entirely filled with rays. One may also imagine this having been brought about by the introduction in the plane $U$ of a circular diaphragm with radius $H_1$. We shall therefore call the plane $U$ the diaphragm plane.

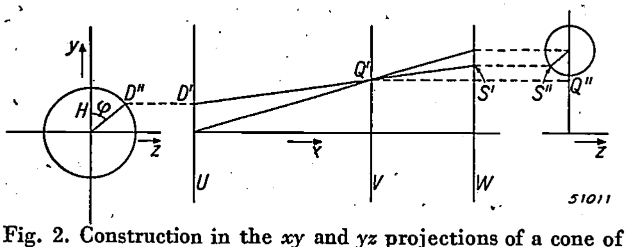


Fig. 2. Construction in the $xy$ and $yz$ projections of a cone of rays with apex $Q$ (projections $Q'$. $Q''$) intersecting the plane $U$ in a circle with radius $H$. This cone intersects the plane $W$ in a circle with respect to which the $yz$ projection is displaced from $Q$ to $Q''$ (adjustment aberration).

We may now call $y—y_0$ and $z—z_0$ ($= z$, since $z_0 = 0$) the aberrations for the ray determined by $y_u$, $z_u$ and $Q$. These aberrations result from the choice of the plane $W$. In practice they may occur, for instance, with a photographic camera when the distance between the frosted glass plate and the lens is incorrectly adjusted. As a matter of fact the aberrations described by equations (1a, 1b) are sometimes called first-order aberrations, because they are described by an expression of the first degree in $H$ and $y_0$. There are two kinds of first-order aberrations. The term with $y_j$ still remains even when the magnitude of $H$ is zero. It then results in an increased distance between the image point and the axis, the increase being in the proportion of $(1 + a_2) : 1$. This may be called the first-order enlargement aberration. The terms governed by $H$ increase with $H$, thus with the opening of the cone of rays. These may be called first-order aperture aberrations. Obviously both enlargement aberrations and the aperture aberrations are proportional to the distance $d$ between $V$ and $W$ and they disappear when $W$ is coincident with $V$ ($d = 0$).

### Higher-order aberrations

If (*fig. 3*) the point source $P$ is chosen on the axis of the optical system, but the condition that the ray must make a very small angle with the axis is dropped, then owing to the rotation-symmetry the ray will lie entirely in a plane through the axis. After refraction, however, it will no longer intersect the axis in the paraxial image point $Q$ but in a point $Q_1$ and upon extension meet the plane $V$ at a point $S$. If, to determine the ray, we again choose a plane $U$ and a point of intersection therein having the coordinates $y_u = H \cos \varphi$, $z_u = H \sin \varphi$, then the coordinates of $S$ will be

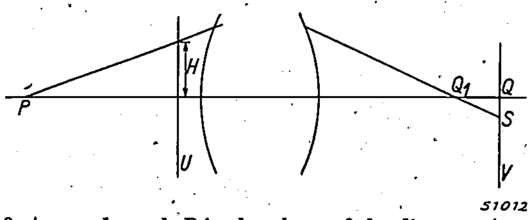


Fig. 3. A ray through $P$ in the plane of the diagram ($\varphi = 0$), making a large angle with the axis, intersects the axis not at the paraxial image point $Q$ but at $Q_1$ and the plane $V$ at $S$, where to a first approximation $QS$ is proportional to $H^3$. $H$ is the distance from the point of intersection of the ray with the plane $U$ to the axis.

$$y = A \cos \varphi, \quad z = A \sin \varphi, \quad . \quad . \quad . \quad (2)$$

where $A$ is a function of $H$, which may be developed in a series of the form

$$A = c_1 H^3 + e_1 H^5 + \ldots \qquad (3)$$

The series necessarily has odd terms only, because when $H$ is replaced by $-H$ and $\varphi$ by $\varphi + \pi$, which does not cause $y_u$ and $z_u$ to change in value, $y$ and $z$ must also remain unchanged.

The aberrations (2) are known as spherical aberration. The terms with the coefficient $c_1$ are called third-order spherical aberration, those with the coefficient $e_1$ fifth-order spherical aberration, and so on.

Let us now consider a point source $P$ off the axis but in the same $xy$ plane (*fig. 4*). If the laws of paraxial imaging were still to hold, the image point would be in $Q$ at a distance $y_0$ from the axis. An arbitrary ray determined by its point of intersection $y_u = H \cos \varphi$, $z_u = H \sin \varphi$ on a plane $U$ will now strike the paraxial image plane $V$ at a point $S$, the $xy$ projection ($S'$) and the $yz$ projection ($S''$) of which are given in fig. 4.

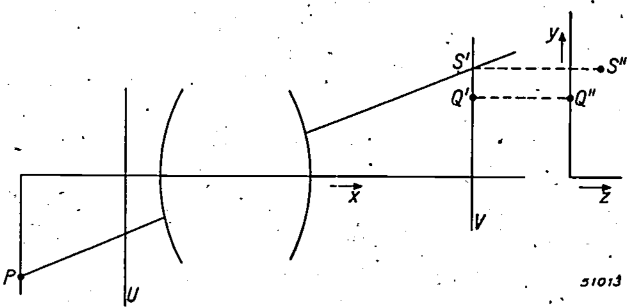


Fig 4. Construction in the $xy$ projection of a ray through $P$ intersecting the plane $U$ at the point $y_u = H \cos \varphi$, $z_u = H \sin \varphi$. $P$ lies in the plane of the figure which at the same time is the plane of symmetry, but the ray in question does not lie in that plane. This ray intersects the plane $V$ at a point $S$ the $xy$ projection $S'$ and the $yz$ projection $S''$ of which are indicated and which deviates from the paraxial image $Q$ (projections $Q'$ and $Q''$) of $P$ (see formulae 4a and 4b). The distance from $Q'$ and $Q''$ to the axis is $y_0$.

Calculation shows that the coordinates $y$ and $z$ of $S$ are given by

$$y = y_0 + c_1 H^3 \cos \varphi + c_2 H^2 y_0 (2 + \cos 2\varphi) + + c_3 H y_0^2 \cos \varphi + c_4 y_0^3 + \dots \quad (4a)$$

$$z = c_1 H^3 \sin \varphi + c_2 H^2 y_0 \sin 2\varphi + + c_3' H y_0^2 \sin \varphi + \dots \quad (4b)$$

in which the constants $c_1 \dots c_4$ depend not only upon the nature of the optical system but also upon the position of $P$ and the choice of the plane $U$.

Since we shall here confine our considerations to third-order aberrations the series expansion has been broken off at the terms of the third degree.

If one wishes to find the point of intersection of the same ray on a plane $W$ parallel to $V$ at a distance $d$, then the right-hand terms of equations (4a) and (4b) are increased by terms of the form (1a) and (1b), whilst if $d$ is small enough the coefficients $c$ may be regarded as remaining unchanged.

The fact that the equations for third-order aberrations are indeed such as given in (4a) and (4b) can most easily be deduced by a reasoning that originated with Conrady [1]).

Let us consider for this purpose (*fig. 5*) refraction at a single convex boundary plane (centre $M$, radius $r$, index of refraction left: $n_1$, right: $n_2$). We are not only interested, as we were above, in all points of light situated in a plane at right-angles to the axis of the system, but rather in those lying on a sphere with radius $R_1$ and centre $M$. Let one of these points be $P$, so that $MP = R_1$. It is clear that in this special case, owing to the spherical symmetry, the line $PM$ may equally well be considered as an axis as any other. If we confine ourselves to rays which are paraxial with respect to the new axis $PM$ then the image point $Q$ appertaining to it will be at a distance $-R_2$ from $M$, in which case it follows from the elementary theory that

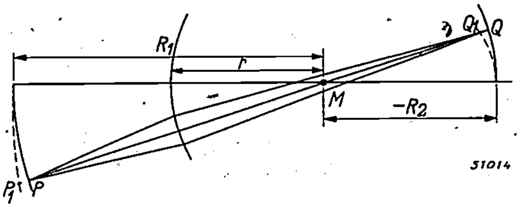


Fig 5. Refraction at a spherical boundary plane (radius $r$). The point $P$ is imaged by rays paraxial with respect to $PM$ at the point $Q$. Displacement of $P$ to $P_1$ causes the image point $Q$ to shift to $Q_1$.

$$\frac{1}{n_1 R_1} - \frac{1}{n_2 R_2} = \frac{1}{r}\left(\frac{1}{n_1} - \frac{1}{n_2}\right) = -\frac{1}{r} \Delta \left(\frac{1}{n}\right).$$

The quantities $R_1$ and $R_2$ are to be taken as positive or negative according to whether the concave side of the lens faces to the right or to the left. If $P$ is situated anywhere on the sphere $R_1$ then the locus of the point $Q$ lies on the sphere $R_2$.

If we then consider the ray $PA$ (*fig. 6*), which is not paraxial with respect to the axis $PM$, we find that under the influence of spherical aberration this ray does not cut the sphere at $Q$ but at a point $S$ having as coordinates

$$y' = ka^3 \cos \psi \quad \text{and} \quad z' = ka^3 \sin \psi,$$

in which $y'$ is the distance from the $y$ projection to the axis $PM$ and $z' = z$ is the distance to the plane of delineation, whilst $a$ represents the distance from $A$ to the axis and $\psi$ is the angle between the plane $PAM$ and the plane of delineation.

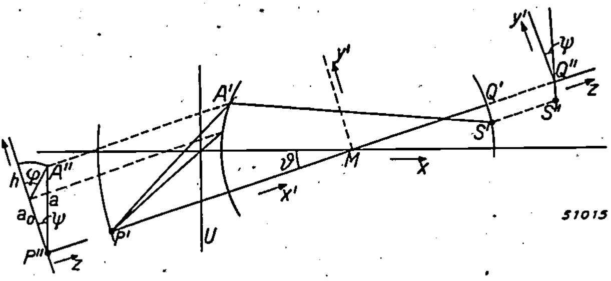


Fig. 6. Explanation of the third-order aberrations arising from spherical aberration according to Conrady. The ray from $P$ to $A$, of which points the $x'y'$ projections $P'A'$ and the $y'z'$ projections $P''A''$ are indicated, strikes the sphere through $Q$ (projections $Q'$, $Q''$) at $S$ (projections $S'$, $S''$).

[1]) A. E. Conrady, The five aberrations of lens-systems, Monthly Not. Roy. Astron. Soc. 79, 60-66, 1918.

With the aid of the $\mathfrak{y}'z'$ projection drawn on the left-hand side of fig. 6 we easily find that

$$a \sin \psi = h \sin \varphi \quad \text{and} \quad a \cos \psi = a_0 + h \cos \varphi .$$

The meanings of $a_0$, $h$ and $\varphi$ are given in the diagram. Since, further

$$a^2 = a_0^2 + h^2 + 2a_0\, h \cos \varphi ,$$

we find for $S$:

$$\mathfrak{y}' = k[a_0^3 + 3a_0^2\, h \cos \varphi + a_0 h^2\, (1 + 2 \cos^2 \varphi) + h^3 \cos \varphi] \quad (5a)$$

$$z' = k[a_0^2\, h \sin \varphi + 2a_0 h^2 \sin \varphi \cos \varphi + h^3 \sin \varphi] \; . \; . \; . \; . \; . \quad (5b)$$

Since it is only a question of small quantities and we break off the expansion at terms of the third degree, the right-hand expressions represent at the same time the aberrations $\mathfrak{y}—\mathfrak{y}_0$ and $z—z_0$ of the point $S$ in the system of coordinates $x\mathfrak{y}z$. (In the transformation from $y'$ to $y$, one may take $\cos \vartheta = 1$ because higher terms only affect aberrations of the fifth and higher orders.) Bearing in mind, further, that to a sufficient approximation (and again the factors ignored affect only the higher order aberrations)

$$a_0 = \alpha\, y_0 \quad \text{and} \quad h = \beta H,$$

where $\mathfrak{y}_0$, $H$ and $\varphi$ have the same meaning as before. Then we find

$$\mathfrak{y}—\mathfrak{y}_0 = \gamma_1 H^3 \cos \varphi + \gamma_2 H^2 y_0\, (2 + \cos 2\varphi) + \\ + 3\gamma_3 H y_0^2 \cos \varphi + \gamma_4\, \mathfrak{y}_0^3 , \; . \; . \; . \; . \quad (6a)$$

$$z—z_0 = \gamma_1 H^3 \sin \varphi + \gamma_2 H^2 \mathfrak{y}_0 \sin 2\, \varphi + \gamma_3 H \mathfrak{y}_0^2 \sin \varphi , \; . \; . \quad (6b)$$

by which equations (4) are obtained, but with this difference that the coefficients of $H\mathfrak{y}_0^2$ in (4) are not, as here, in the ratio 3 : 1. This is because in this case the coordinates are determined from the point of intersection of the refracted ray on the sphere $R_2$, whilst the image point is chosen on the sphere $R_1$. If $P$ is moved along $MP$ to $P_1$ on a sphere with radius $R'_1$ whichi s tangent to the sphere with radius $R_1$, then the point $Q$ is displaced to $Q_1$. Equations (6) now represent to no less good an approximation (for the term with $y_0^3$ only is this not absolutely correct) the coordinates on a sphere with radius $R_2'$ tangent to the sphere with radius $R_2$, for which we again have:

$$\frac{1}{n_1 R_1'} - \frac{1}{n_2 R_2'} = \frac{1}{r}\Big(\frac{1}{n_1} - \frac{1}{n_2}\Big) = -\frac{1}{r}\, \Delta\Big(\frac{1}{n}\Big) .$$

A flat object plane at right-angles to the principal axis of the system is obtained by taking $R_1' = \infty$. $R_2'$ thereby assumes a certain value $R_p$, for which the expression

$$-\frac{1}{n_2\, R_F} = -\frac{1}{r}\, \Delta\Big(\frac{1}{n}\Big)$$

applies.

This is the so-called Petzval surface. The aberration in the paraxial image plane, *i.e.* the plane surface at right-angles to the main axis tangent to the sphere $R_2$, differs from the expression (6) by an amount proportional to the distance from $Q_1$ to that plane ($= \mathfrak{y}_0^2/2R_F$) and also proportional to the aperture parameter $H$. When this correction is made an equal amount is added to the two coefficients $3\gamma_3$ and $\gamma_3$ and the simple ratio of 3 : 1 is lost.

When dealing with a series of refracting surfaces (a system of lenses) instead of one such surface, one may proceed further and consider the spherical surface with radius $R_P$ as the object plane for the next refraction. One then still arrives at an equation of the form of (6), which then applies to a sphere with radius $R_P$, for which one finds:

$$\frac{1}{n_{l+1} R_P} = \sum_{i=1}^{i=l} \frac{1}{r_i}\, \Delta_i \Big(\frac{1}{n}\Big) .$$

Here $l$ relates to the last refracting surface; $R_P$ is the radius of the Petzval surface of the system. Here again the ratio 3 :1 exists between the coefficients of $H\mathfrak{y}_0^2$ but it is lost again when the corrections are made which are necessary for the transition to a plane image field.

## Further considerations of third-order aberrations

A better insight into the significance of the terms of the equations (4a) and (4b) is obtained by studying the intersection figures formed when, with $H$ constant, $\varphi$ is made to assume all possible values.

It must then be borne in mind that although the analytical treatment proposed gives a complete insight into the geometry of the refracted pencils of rays this is no longer the case if, as we shall do presently, each of the terms is viewed separately. The fact that this has often led to incorrect conclusions being drawn in the literature on the subject has been pointed out already by Gullstrand [2]).

Even though one may have an absolutely true picture of the course of the pencils of rays after refraction, one may not immediately draw conclusions therefrom about the distribution of intensity in the beams. Here the approximation of geometrical optics fails more or less, since it does not allow for the coherence between the rays. This has been pointed out by Picht [3]) and Zernike [4]) among others.

For the sake of simplicity we shall now examine equations (4a) and (4b) one term at a time.

### *Spherical aberration*

In the terms with $H^3$ one recognises the spherical aberration previously discussed. This therefore remains unaltered for points outside the axis. When this defect occurs alone then as $\varphi$ varies the point $S$ describes a circle around the point $Q$, the radius of which is proportional to $H^3$ (*fig. 7*).

### *Coma*

The terms with $H^2 y_0$ all come under the name of coma. If these terms alone were present, then with a variation of $\varphi$ and constant $H$, owing to the term $2\ \varphi$, the point of intersection $S$ would describe a circle in such a way that if the zone in the diaphragm plane is passed through once then that circle would be described twice. The centre of this circle lies at a distance $2c_2 H^2 y_0$ from the paraxial image point. Owing to the proportionality with $y_0$ one may here speak of an enlargement aberration, which,

[2]) See for instance A. Gullstrand, Naturwiss. 14, 653-664, 1926.

[3]) J. Picht, Optische Abbildung, Braunschweig, 1931.

[4]) See B. R. A. Nijboer, thesis, Groningen, 1942.

however, is proportional to $H^2$. If $H$ be given different values in succession then one obtains (*fig. 8*) a system of circles the common tangents of which intersect each other at an angle of 60°.

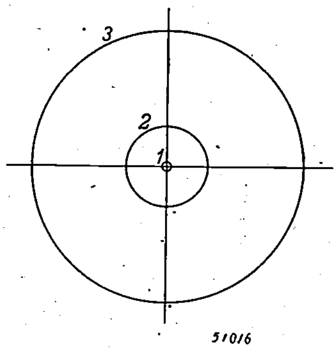


Fig. 7. Intersection figures on the plane $V$ corresponding to the terms with $H^3$ (spherical aberration) for values of $H$ which are as 1 : 2 : 3.

### *Astigmatism and field curvature*

Owing to the proportionality with $H$, the terms containing $Hy_0^2$ are to be regarded as aperture aberrations, which, however, depend upon $y_0$. The intersection of the rays of a zone ($H$ constant, $\varphi$ variable) with the plane $V$ is an ellipse with axes $c_3Hy_0^2$ and $c_3'Hy_0^2$. When we consider the intersection upon a plane $W$ at a distance $d$ from $V$ then the terms $a_1H \cos \varphi + a_2y_0$ and $a_1H \sin \varphi$ respectively are added to the aberrations, thereby altering the relation between the axes of the ellipse. By a suitable choice of $d$ one can reduce either the coefficient of $\cos \varphi$ or the coefficient of $\sin \varphi$ to zero, in consesequence of which the figure of intersection converges into a small line at right-angles to the $y$ axis and the $z$ axis respectively (*fig. 9*). The centres $M_s$ and $M_t$ of these degenerate ellipses are called respectively the sagittal and the tangential or meridional image points. The locus of the sagittal image points when $y_0$ varies is a circle in the $xy$ plane and a convex surface in space, the so-called sagittal image surface; likewise the tangential image points lie on another convex surface, the tangential image surface.

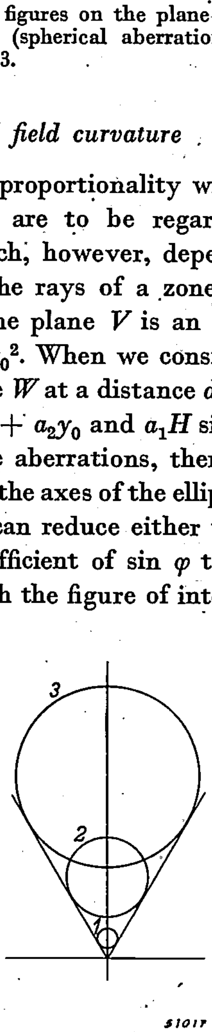


Fig. 8. Intersection figure on the plane $V$ corresponding to the terms with $H^2y_0$ (coma) for values of $H$ which are as 1 : 2 : 3.

The curvatures of the image surface are given by

$$\frac{1}{R_t}=\frac{2c_3}{a_1/d} \text{ resp. } \frac{1}{R_s}=\frac{2c_3'}{a_1/d} \quad \ldots \quad (5)$$

The equation

$$\frac{1}{R_m}={}^1\!/_2\left(\frac{1}{R_s}+\frac{1}{R_t}\right) \quad \ldots \quad (6a)$$

represents what is called the average field curvature. The figure of intersection on the sphere corresponding to this curvature is a circle. The quantity

$$A={}^1\!/_2\left(\frac{1}{R_t}-\frac{1}{R_s}\right), \quad \ldots \quad (6b)$$

determining the distance of the points $M_s$ and $M_t$ is called the astigmatism. An interesting relation arises between $R_s$ and $R_t$ when these quantities are related to the quantity $R_P$, the so-called Petzval radius of curvature of the system.

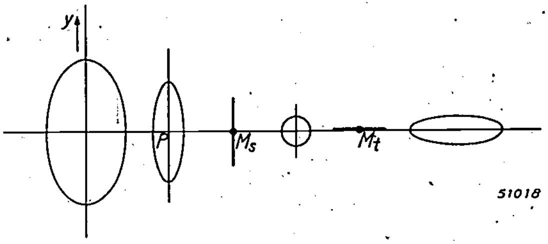


Fig. 9. Series of cross sections ($H$ = constant) with a number of planes $W$ ($d$ negative, increasing in absolute valuc) corresponding to the terms $Hy_0^2$ (field curvature and astigmatism). $M_s$ is the sagittal, $M_t$ the tangential or meridional image point. Halfway between $M_s$ and $M_t$ the cross section is a circle. At $P$ (the intersection with the Petzval-surface, $PM_s = 1/3\ PM_t$) the ratio of the axes is 1 : 3.

As we have seen above, this is simply related to the data of the system. Instead of the expression given above one may also write

$$\frac{1}{n_{l+1}R_P}=-\sum_{k=1}^{k=l+1}\frac{\Phi_k}{n_k},$$

in which $\Phi_k$ represents the strength of the lens (placed in a vacuum) formed by the two successive boundary surfaces, and $n_k$ the index of refraction of the medium between those boundary surfaces. Here $\Phi_k$ has to be taken for zero thickness so that

$$\Phi_k=(n_k-1)\left(\frac{1}{r_{k-1}}-\frac{1}{r_k}\right),$$

whilst $1/r_0$ and $1/r_{l+1}$ have to be taken equal to nil.

The Petzval curvature $1/R_P$ and the two field curvatures are related thus:

$$\frac{1}{R_P} = \frac{1}{R_m} - 2A = \frac{1}{R_s} - A = \frac{1}{R_t} - 3A. \quad (7)$$

*Distortion*

Finally the term $y_0^3$ represents an enlargement aberration dependent upon $y_0$ but independent of $H$ and thus existing for an infinitely narrow beam. It results in a non-linear radial distortion of the image. A square lying in the object plane symmetrically with respect to the axis is distorted into the form of a barrel (barrel distortion) or into the form of a cushion (pincushion distortion) according to whether $c_4$ is negative or positive.

One must always bear in mind that the aberrations dealt with here and given separate names are equivalent terms in a series expansion and that the true deviation of a point of intersection with respect to the paraxial image point is obtained by summing all the terms. Consequently the true intersection figure of the rays of a zone upon the plane $V$ or upon a plane $W$ may be a complicated curve (in general of the fourth degree). Further it must be remembered that the coefficients $c$ depend not only upon the properties of the system in general but also upon the position of the object and the choice of the diaphragm plane.

Finally it is obvious that the third-order aberrations discussed here represent only an approximation and that to obtain a complete insight it would be necessary to consider also the terms of the fifth and higher orders.

## Chromatic aberrations

One would be inclined to deduce from the foregoing tbat in order to obtain a good image formation in a flat image plane with the desired enlargement it is only necessary to ensure that the system has the right focal distance and that the first and third-order aberrations, *i.e.* the coefficients $a_1 \ldots c_4$, are zero. This would involve satisfying a number of conditions by giving the right values to the curvatures and the distances of the lens surfaces and to the indices of refraction of the kinds of glass, since the coefficients $a_1 \ldots c_4$ are functions of the latter quantities. No allowance would then have been made, however, for the fact that the indices of refraction and hence the coefficients $a_1 \ldots c_4$ depend on the wavelength $\lambda$. If the focal distance for a given $\lambda$ has one value, for some other $\lambda$ it will be different. Consequently for the same adjustment, thus for the same position of the image plane, even though the adjustment errors for a given $\lambda$ are nil, for any other wavelength aberrations of the first order will arise. These are called first-order chromatic aberrations. The coefficients $c_1 \ldots c_4$ will also depend upon the wavelength $\lambda$. These conditions can be met by requiring that the coefficients $a_1 \ldots c_4$ for a number (two or three) of values of $\lambda$ must be zero, in the hope that they will not differ much from $\lambda$ zero for intermediate wavelengths. This, however, considerably increases the number of conditions that have to be satisfied. Moreover, with an objective having a large aperture (large $H/f$) and an extensive field (large $y_0/f$) not only the first and third-order aberrations but also those of higher (fifth and seventh) order must be considered.

## The designing of a system of lenses

It is evident that from a mathematical analysis of the aberrations alone it is almost impossible to design a system of lenses that has to satisfy high demands. In practice, therefore, one does not proceed in this way at all. The designing of optical systems is more often than not a matter of practical experience rather than one of theory and to some extent more of an art than a science. Nevertheless, the result will always be put to the test by applying an analysis of the aberrations to a solution arrived at empirically. It will then usually be found that the third-order aberrations are not zero but that there is a combination of these together with aberrations of a higher order, such as to make the result on the average as satisfactory as possible.

The difficulties caused by chromatic aberrations do not arise when mirrors are used for the projection. This is one of the reasons why in some cases a mirror system is preferred to one of lenses. Another reason is that the spherical aberration of a concave mirror is less than that of a refracting surface of the same strength.

## The spherical mirror

Let us imagine that we have a spherical mirror with circular rim. The axis of symmetry is chosen as the main axis of the optical system. The paraxial rays radiating from a point $P_0$ on this axis converge upon a point $Q_0$ on the same axis. A non-paraxial ray from $P_0$ after reflection meets the axis at a different point $Q_0'$ and reaches the paraxial image plane at a point $S$, so that $Q_0S$ $(= c_1H^3)$ again represents the spherical aberration.

We shall now consider a point source $P$ outside the axis but at the same distance from the centre $M$ as $P_0$. The line $PM$ for the convex surface of which the mirror forms a part will then likewise be an axis of symmetry. Rays which are paraxial with respect to $PM$ will converge upon $PM$ at a point $Q$, where $QM = Q_0M$. The rays which are not paraxial with respect to $QM$ strike this line at another point $Q'$. Thus the aberrations in the imaging of the point $P$ may be described as spherical aberrations with respect to the secondary axis $PM$.

The situation is now absolutely the same as that in the case described above (see figs, 5 and 6) of refraction by a single convex refracting surface. To this, too, one can apply the theory of Conrady and in this way deduce the values of the coefficients of the third-order aberrations.

### Diaphragm at the centre of curvature

We shall not go into the details of the general case but discuss the special case in which the diaphragm plane passes through $M$. Further we shall choose the distance $P_0M = \infty$ and thus consider the imaging of an infinitely distant plane. The paraxial image point of $P_0$ then coincides with the focus of the spherical mirror. It is easily seen (*fig. 10*) that owing to the particular position of the diaphragm all beams passing through the diaphragm are subject to the same conditions (apart from the fact that from the point of view of a beam falling obliquely on the circular aperture of the diaphragm the aperture assumes the form of an ellipse having slightly different axes). Rays which are paraxial with respect to $PM$ converge at a distance $|f| = |R/2|$ from $M$. The other rays have the same spherical aberration with respect to the sphere with radius $|f|$ as that shown by the rays from $P_0$ with respect to that sphere. We may therefore say that for rays which are paraxial with respect to a secondary axis the image formation is sharp in a curved image plane (radius $|f|$) and that the other rays only display a spherical aberration.

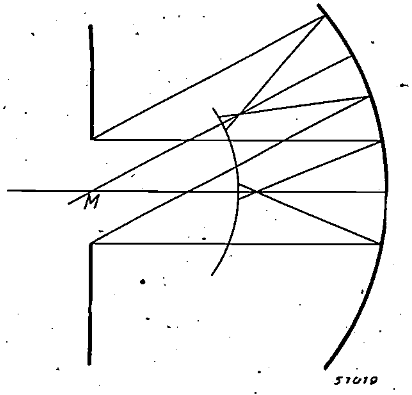


Fig. 10. A diaphragm the centre of which coincides with the centre of the mirror $M$ brings about a practically sharp image on a sphere with radius $|f| = R/2$. Depending on the diameter of the diaphragm, however, spherical aberration occurs. The point sources $P_0$ and $P$ are to be imagined as the infinitely distant points of the main and secondary axes respectively.

We may, however, also — as was done above for a system of lenses — consider a flat image field through the focus (paraxial image plane with respect to the main axis $P_0M$) and express the aberration with respect to this plane by the coefficients $c_1 \ldots c_4$. It then appears that when $y_0 = f \tan \vartheta$ ($\vartheta = \angle P_0MP$) and $z_0 = 0$ the point of intersection of the reflected ray from $P$ on the image plane is given by

$$y—y_0 = \frac{1}{2R^2} H^3 \cos \varphi — \frac{2}{R^2} H y_0^2 \cos \varphi,$$

$$z—z_0 = \frac{1}{2R^2} H^3 \sin \varphi — \frac{2}{R^2} H y_0^2 \sin \varphi,$$

so that $c_1 = {}^1/_2R^2$, $c_3 = c_3' = —2/R^2$ and $c_2 = c_4 = 0$. From this we find again that although there is field curvature there is no astigmatism nor any coma or distortion. Further the only defect is the spherical aberration.

As may easily be proved, $a_1 = d/f = 2d/R$ and the field curvature is thus

$$\frac{2\,c_3}{a_1/d} = \frac{2}{R}.$$

The radius of the curved image plane is thus $|R/2| = |f|$. Since astigmatism is zero, this surface is at the same time the Petzval surface. This may be deduced, according to a common method from the general formulae given above for the Petzval curvature, which apply for the case of refraction, by putting $n$ prior to reflection = 1 and after reflection = —1, so that $\Delta\ (1/n) = 2$.

### Spherical mirror with correction plate

If when using a spherical mirror one had a means of causing the incident rays to change direction when passing through the diaphragm plane in such a way as to eliminate the spherical aberration, then a perfect projection would be obtained on a sphere with radius $|R/2|$.

Such a possibility is of great practical importance. The means of attaining this to a high degree of approximation is to be found in Schmidt's correction plate, which is optically flat on one side and has an aspherical surface on the other side. It is clear that with this a system can be devised which combines the advantages of a large aperature with those of a wide image field. This provides a simple way of obtaining a result which would otherwise entail a highly complicated system of lenses. It is even possible to exceed considerably the light flux of the present lens systems. One must, however, remember that the image field is not flat but curved.

The construction and applications of the correction plate will be dealt with in subsequent articles.

**Appendix: Note on the parabolic mirror**

Finally some remarks are to be added regarding the parabolic mirror.

*Fig. 11* represents the cross section of the plane of the diagram for a parabolic mirror whose radius of curvature at the top is $R$. If the normal is drawn from a point $P$ on the mirror surface it will intersect the axis at a point $N$, such that the distance from $N$ to the foot $Q$ of the perpendicular from $P$ on the axis (the so-called sub-normal) is $R$, contrary to the case with the sphere with centre $M$, where the length of the normal itself is $R$ while all normals pass through $M$. This difference in behaviour of the two curves has the effect that rays striking the axis obliquely show no spherical aberration but all pass exactly through the focus $F$, whereby $MF = FO = {}^1/_2\,|R|$. Without further thought one might suppose that a parabolic mirror with a diaphragm at the centre of curvature $M$ would automatically be free from aberrations, but this is not so, because with the parabola there is not such a high degree of symmetry with respect to the point $M$ as there is with the sphere. In fact one finds by calculation that

$$c_1 = 0,\ |c_2| = 1/R^3,\ |c_3| = 8/R^2,\ |c_3'| = 4/R^3,\ |c_4| = 4/R^2.$$

Except for spherical aberration — which is naturally absent — all other aberrations do indeed occur.

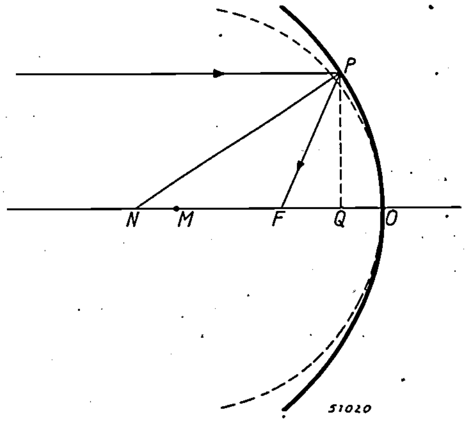


Fig. 11. The passage of rays with a parabolic mirror. A ray through $P$ parallel to the axis, after reflection, strikes the axis at $F$, $MF = FO = {}^1/_2\,|R|$. Here $R$ is the radius of curvature at $O$ (broken line). $NP$ is the normal drawn from $P$ on the parabola; $NQ = MO = |R|$.

# CONSTANT AMPLIFICATION IN SPITE OF CHANGEABILITY OF THE CIRCUIT ELEMENTS

by J. J. ZAALBERG van ZELST.

621.394.645.3

In a previous article [1]) amplifier circuits were discussed which make the amplification very little dependent upon the slope of the valves, assuming the other circuit elements to be constant. The present article deals with the problem of designing such a circuit as to minimise also the effects of variations in those other circuit elements upon the amplification, or, to be more precise, to keep the amplification constant within limits narrower than the tolerances of the variable elements. The solution of this problem is built up upon the method described in the previous article, where the amplifier is caused to oscillate with a non-disturbing auxiliary frequency. As feedback network a quadripolar system is used which in the event of variations taking place in the component parts possesses a practically constant attenuation, provided that at the same time the (auxiliary) frequency is modified in such a way that the phase difference between the input and output voltages of the quadripolar system remains constant. This condition is automatically satisfied in the oscillator, since the latter can only oscillate when the said phase difference amounts to 180°. In this manner circuits can be made with an amplification per valve of, for instance, approximately 8 ± 1% for a tolerance of + or — 5% in the values of the component parts. A number of such stages can quite well be placed in cascade.

## Introduction

An amplifier consists of one or more amplifying valves and a number of other circuit elements such as resistors, capacitors, etc. The magnitude of the amplification depends, as far as the valves are concerned, mainly upon the slope (mutual conductance) and, as regards the other elements in a given circuit, upon their impedance. In a previous article in this journal [1]), which we shall further refer to as article I, we discussed a number of circuits where variations in slope (within certain limits) result in a very much smaller relative change in amplification. For all practical purposes we were therefore able to speak of "constant" amplification. It was expressly presupposed, however, that the values of the other circuit elements were constant.

We will now go a step further and aim at altering at least one of the circuits discussed in such a manner as to make the amplification very little subject to the influence of the values of the resistors, capacitors, etc., while retaining of course insensitivity to changes of the slope.

Let us first for a moment consider another problem, the construction of a direct voltage souce of which the voltage remains constant within very narrow limits. To arrive at a solution at all practicable we should have to have recourse to some material constant, for instance an electro-chemical EMF (in the case of a standard cell) or a normal cathode fall (in the case of a stabilizing tube). Now in practice it is impossible to get conditions under which material constants are actually constant, for as a rule they are dependent upon other physical or chemical quantities over which one has not complete control, e.g. the temperature, or the purity of the materials used, which are apt to change in the course of time and have their effect upon the result. But even if the principle of the voltage source is based upon suitably chosen material constants, the inconstancy of the voltage cannot be made smaller than that of the practically least variable material constant — both inconstancies being taken, of course, relatively.

What has been said here not only applies to voltages but likewise holds for many other physical quantities. One might, therefore, easily be led to ascribe general validity to it. In the case of the amplifier already mentioned above one would also be inclined to assume *a priori* that at a given frequency the amplification cannot be more accurately determined than the slope, resistance, capacitance, etc. deviating the least from the prescribed value (with the possibility of all these quantities varying). Amplification, however, is a non-dimensional quantity; it may well be found less difficult to imagine a relation of two similar quantities as having a constant value than to suppose the same of a physical quantity itself, but it is not so simple to be able to say why the former is more readily acceptable. However this may be, it does indeed appear possible to construct alternating vol-

[1]) J. J. Zaalberg van Zelst, Stabilised Amplifiers, Philips Techn. Rev. 9, 25-32, 1947 (No. 1).

tage amplifiers whose amplification satisfies the requirement of stability, that is to say whose amplification differs from a given value relatively less than the relative deviation that we wish to allow for each of the component parts from their prescribed values. In other words, if in such an amplifier a tolerance of + or — $p\%$ is allowed in the values of the component parts, then the amplification obtained should deviate from the nominal value by less than $p\%$. How this is to be attained is the subject of this article.

Here we will proceed to build up upon the method discussed sub IIb in article I, where the amplifier oscillates with a non-disturbing auxiliary frequency.

### The amplifier with negative feedback

In addition to this method there is another means by which it seems at first sight possible to arrive at the object in view. We have in mind an amplifier possessing a feedback resistance $R_k$ in the cathode line and a resistance $R_a$ in the anode circuit, where the voltage on these two resistances in series serves as the output voltage $V_0$. By choosing $SR_k$ large in comparison with unity ("infinitely strong feedback"; $S$ = slope of the valve) the amplification becomes approximately equal to $1 + R_a/R_k$ and thus independent of $S$. Moreover, if $R_a$ is chosen only small with respect to $R_k$, it can be made as little dependent upon the resistance values $R_a$ and $R_k$ as one wishes; the amplification then drops, however, to scarcely more than unity. If we consider how large $R_a$ may be taken while still satisfying the requirement of constancy, we find $R_a = R_k$; the nominal amplification is then 2, thus still only very modest. Although amplifiers with such a low amplifying factor may be useful in certain cases, we shall leave them out of consideration here partly for the following reason.

If one tries to reach higher amplifications by employing stages in cascade one finds that the amplification still does not work out higher than a nominal value of 2 so long as the constancy requirement is complied with — apart from the practical difficulties connected with a cascade circuit of stages not having a common input and output terminal.

## The oscillating amplifiers

We would recall — see for instance article I — that in order to make a valve oscillate a feedback is necessary, bringing a voltage derived from the anode current to the input of the circuit. In *fig. 1* it is indicated diagrammatically how a voltage $V_{t1}$ when brought to the input of an amplifier $A$ gives rise to an anode current $I_a$ from which the feedback network $T$ derives a voltage $V_{t2}$. When the output terminals of $T$ are properly connected with the input terminals of the amplifier the condition will be maintained (*i.e.* the amplifier begins to oscillate) if $-V_{t2} = V_{t1}$. If we define the (total) slope of the amplifier as $S = I_a/V_{t1}$ and the transfer impedance $Z$ of the feedback network as $Z = V_{t2}/I_a$, then $V_{t2} = ZI_a = ZSV_{t1}$. The condition that must be satisfied for $-V_{t2} = V_{t1}$ is thus:

$$ZS = -1 \quad \ldots\ldots\ldots \quad (1)$$

The minus sign expresses that a phase difference of 180° must exist between $V_{t1}$ and $V_{t2}$ (where positive directions are chosen as indicated in fig. 1).

In article I we discussed circuits where the relation (1) is used to determine the slope $S$ of the amplifier as the reciprocal of a given transfer impedance $Z$ in spite of possible changes in the valve slope. This was found to be attainable in the best way by causing the valve to oscillate with a non-disturbing auxiliary frequency. From the oscillating voltage a regulating voltage was derived which compensates the changes in slope.

As a rule the transfer impedance $Z$ is a function of the frequency. For the frequency at which relation (1) is satisfied, *i.e.* for the oscillation frequency, the rule therefore applies that the total amplification in the circuit formed by the amplifier and the feedback network in cascade is exactly equal to unity; in other words the amplification in the amplifier itself is just compensated by the attenuation in the feedback network.

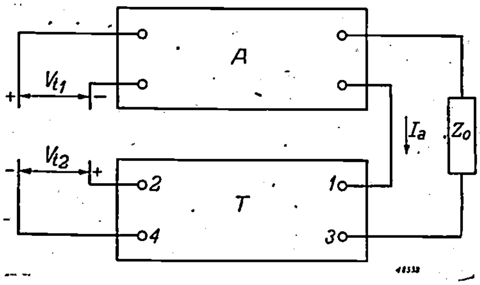


Fig. 1. If for a certain frequency the transfer impedance $Z = V_{t2}/I_a$ of the feedback network $T$ assumes the value $-1/S$ ($S$ = slope of the amplifier $A$, $= I_a/V_{t1}$) then the voltage supplied by $T$ is just identical with the voltage required at the input of the amplifier to produce $V_{t2}$. When the output of $T$ is connected to the input of $A$ the amplifier starts oscillating in that frequency.

If one succeeds in stabilizing this attenuation in spite of changes in the circuit elements — and it will be seen that this is possible owing to a property of certain networks — then one will at the same time have obtained a constant amplification.

### *Correlation between the amplitude and phase characteristics of some quadripolar systems*

By the amplitude characteristic of a quadripolar system we understand the absolute value of the ratio of the output voltage to the input voltage as a function of the frequency, and by the phase characteristic is understood the phase difference $\varphi$ between these two voltages likewise as a function of the frequency. If one or more elements in a quadripolar system are altered then at a given frequency $f_0$ the phase-angle $\varphi$ will as a rule change. By giving the frequency a certain value $f'$ the phase angle $\varphi$ can be returned to its original value. The correlation just referred to which occurs with certain quadripolar systems consists in the fact that *at the frequency $f'$ the attenuation of the quadripolar system with the altered components is practically equal to that of the quadripolar system with the original values at the frequency $f_0$*. We shall presently give an example of a quadripolar system possessing this property.

If we now use such a quadripolar system as feedback network so arranged that the phase difference $\varphi$ becomes 180° for a certain frequency ($f_0$) then the circuit will oscillate with that frequency (assuming that the amplification is adequate). If we then change one or more elements of the quadıipolar system the oscillation frequency will automatically assume the value — just called $f'$ — at which $\varphi$ remains constant (in this case 180°, since this is one of the conditions for oscillation). Thus we have obtained a circuit which in the event of a change in the elements varies its oscillating frequency in such a way as to keep the attenuation of the feedback network practically constant. *A fortiori* this will therefore also be the case with the amplification, since this, according to another condition for oscillation, is just as great as the attenuation.

If one should now succeed in conducting the signal to be amplified — whose frequency strongly deviates from the oscillation frequency — to the amplifier and carrying off the amplified signal in such a way as to retain the above-mentioned property of the feedback network, then one has indeed reached an amplification substantially constant for the signal frequency.

Before going further into these signal circuits we shall give an example of a quadripolar system possessing the property described above; see *fig. 2a*. The fact that it does indeed possess this property can be proved by a not difficult but extensive calculation. Rather than follow this calculation here we shall demonstrate how the said correlation between the amplitude and phase characteristics comes about with the circuit according to fig. 2*a*. Here we assume that $R_1C_1$, $R_2C_2$ and $R_3C_3$ have been so chosen that for a certain frequency $f_0 = \omega_0/2\pi$ there arises in each of the three sections a phase shift of 60° between the output and input voltages of a respective section. Further it is assumed that $R_3 \gg R_2 \gg R_1$, so that the load of one section upon the previous one may be ignored. Under these simplified conditions the vector diagram of *fig. 3* applies. Since the voltage vectors $OB$ and $BA$ are at right-angles to each other, $B$ lies on a circle with $OA$ as diameter. Similarly one finds that $C$ lies on a circle with $OB$ as diameter and $D$ on a circle with $OC$ as diameter. Owing to the phase shift of 3 × 60° $OD$ comes to lie in the extension of $AO$; in other words the output voltage $V_3$ is in counterphase to the input voltage $V_0$, so that the circuit in which this quadripolar system acts as feedback network will oscillate with the frequency in question (the frequency $f_0$ for which the phase shift per section of the network is 60°, thus for which $2\pi f_0C_1R_1 = \tan 60°$ applies).

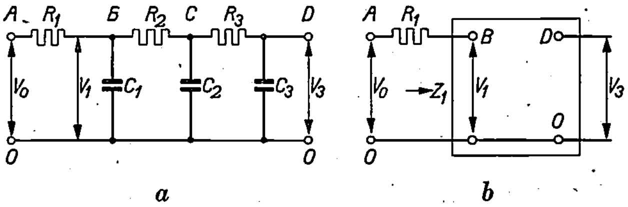


Fig. 2. *a*) Example of a network possessing the following property: when one or more elements change then the ratio $V_0/V_3$ of the input voltage to the output voltage does not, in the first instance, change provided the frequency is at the same time changed in such a way that the phase difference between $V_0$ and $V_3$ returns to its original value.

*b*) The same network as in (*a*) but with the part $OBOD$ drawn as a separate quadripolar system, the impedance of which, viewed from $OB$, is called $Z_1$.

If, now, one or more of the circuit elements is changed then the oscillating condition $\angle AOD = 180°$ remains valid. The oscillation frequency will therefore adjust itself in such a way that in spite of the alteration of the elements the sum of the angles $AOB$, $BOC$ and $COD$ is still 180°. The angles

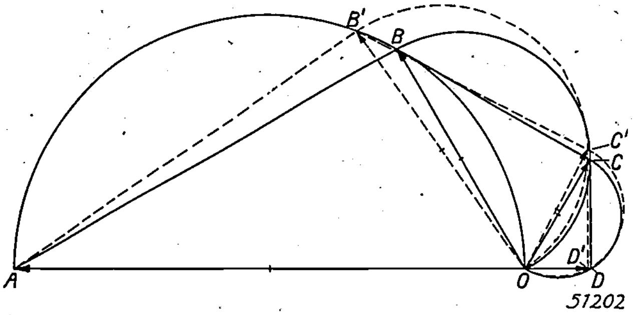


Fig. 3. Vector diagram of the voltages in the circuit of fig. 2*a* when each of the sections $R_1$-$C_1$, $R_2$-$C_2$ and $R_3$-$C_3$ produces a phase shift of 60° between the input voltage and the output voltage of the respective section (corresponding to the angles $AOB$, $BOC$ and $COD$) and $R_3 \gg R_2 \gg R_1$. When the elements are changed and at the same time also the frequency is changed in such a way that $OD$ continues to lie in the extension of $AO$, a figure is produced like that shown in broken lines, where $B$ moves along the circle to $B'$, $C$ to $C'$ and $D$ to $D'$. This broken-line figure has been constructed for the case where $R_1$ becomes 25% smaller while the other elements have their nominal values. $OD'$ appears to be only 2% smaller than $OD$.

may, however, deviate slightly from 60°, $B$ moving along the circle described on $AO$, for instance to $B'$. Although $OB'$ may differ noticeably from $OB$ in length the projection from $OB'$ towards a direction $OC'$ approximately coinciding with that from $OC$ will be practically equal to $OC$, because $BC$ is a tangent to the circle. This applies to a still greater degree for the final result, the projection $OD'$ from $OC'$ on the extension of $AO$, which then also coincides by good approximation with the original vector $OD$.

As may easily be seen from fig. 3, the nominal attenuation of the network, *i.e.* the absolute value of the ratio of the input voltage to the output voltage, is equal to $1/\cos^3 60° = 8$.

*Example*

Applying the index 0 to indicate the values appertaining to the condition in which all elements of the circuit have their nominal value, then:

$$C_{10}R_{10} = C_{20}R_{20} = C_{30}R_{30}, \text{ put} = T.$$

When the elements are varied $C_{10}R_{10}$ changes into $C_{10}'R_{10}'$, say $= p_1T$, $C_{20}R_{20}$ into $p_2T$, and so on. As a consequence the angular frequency $\omega_0$ of the oscillation will assume a value $\omega'$, say $q\omega_0$, and the angles $AOB = \alpha_0$, $BOC = \omega_0$ and $COD = \gamma_0$, originally 60° each, become $\alpha'$, $\beta'$ and $\gamma'$. Also these last-mentioned angles sum up to 180°:

$$\alpha' + \beta' + \gamma' = \pi. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

Now $\tan \alpha' = \omega C'_1R_1' = q\omega_0 \cdot p_1T = p_1q\omega_0T$, $\tan \beta' = p_2q\omega_0T$, $\tan \gamma' = p_3q\omega T$, whilst $\omega_0T = \omega_0C_{10}R_{10} = \tan \alpha_0 = \sqrt{3}$. Equation (2) thus becomes:

$$\tan^{-1}(p_1q\sqrt{3}) + \tan^{-1}(p_2q\sqrt{3}) + \tan^{-1}(p_3q\sqrt{3}) = \pi. \quad (3)$$

For given variations of the elements, thus for given values of $p_1$, $p_2$ and $p_3$, we get from (3) the value of the factor $q$ by which the frequency changes. With this value given one can calculate the angles $\alpha'$, $\beta'$ and $\gamma'$, the cosines of which determine the magnitude of the attenuation; for the ratio of the input voltage to the output voltage of the network (fig. 3) is:

$$OA/OD' = 1/\cos \alpha' \cdot \cos \beta' \cdot \cos \gamma' \quad . \quad . \quad . \quad . \quad . \quad (4)$$

By way of example let us suppose that only $R_1$ changes and that this resistance becomes 25% smaller. Then $p_1 = 0.75$, $p_2 = p_3 = 1$. From (3) we then find $q = 1.106$, from which it follows that $\alpha' = \tan^{-1}(0.75 \cdot 1.106 \cdot \sqrt{3}) = 55°9'$, $\beta' = \gamma' = \tan^{-1}(1.106 \cdot \sqrt{3}) = 62°\ 25.5'$. According to (4) the attenuation is now 8.167, *i.e.* only 2% greater than the original value of 8.000, in spite of the great change (25%) in $R_1$.

*Choice of the angles $\alpha_0$, $\beta_0$ and $\gamma_0$*

Let it be asked that the nominal values of the angles $\alpha$, $\beta$ and $\gamma$ shall be so chosen that when these angles — which together always make 180° — vary the product $\cos \alpha \cdot \cos \beta \cdot \cos \gamma$ shall change as little as possible. To that end the nominal values ($\alpha_0$, $\beta_0$, $\gamma_0$) will have to be each 60°, as is to be understood from the following.

Let us first suppose that $\gamma$ = constant, $= \gamma_0$. Then

$$\cos \alpha \cdot \cos \beta \cdot \cos \gamma = \cos \alpha \cdot \cos(\pi - \alpha - \gamma_0) \cdot \cos \gamma_0 =$$
$$= \frac{-\cos(2\alpha + \gamma_0) + \cos \gamma_0}{2} \cdot \cos \gamma_0.$$

In this equation the only term containing the variable angle $\alpha$ is $\cos(2\alpha + \gamma_0)$. This is the least affected by changes in $\alpha$ when $2\alpha + \gamma_0 = 0$ or $\pi$, from which it follows that $\alpha = {}^1/_2\,(\pi - \gamma_0)$ so that $\beta = \pi - \alpha - \gamma = {}^1/_2\,(\pi - \gamma_0)$, thus $= \alpha$.

Following an analogous reasoning it can be proved that if $\gamma$ is also taken to be variable then the nominal value $\gamma_0 = 0$, thus that each of the angles must be chosen at 60°.

In *fig. 4* we have as feedback network a quadripolar system very much resembling that of fig. 2*a*. There is, however, a small difference: as will be seen, in fig. 4 the resistor $R_1$ is connected parallel to $C_1$, which is not the case in fig. 2*a*. If we introduce an impedance $Z_1$ as indicated in fig. 2*b* then:

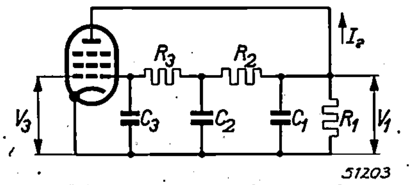


Fig. 4. The quadripolar system of fig. 2*a* as feedback network in a slightly modified form ($R_1$ parallel to $C_1$).

$$V_1 = \frac{Z_1}{R_1 + Z_1} V_0.$$

In the scheme of fig. 4 the anode current $I_a$ flows through the parallel circuiting of the resistor $R_1$ and the impedance $Z_1$, so that here the voltage $V_1$ amounts to

$$V_1 = -\frac{R_1Z_1}{R_1 + Z_1} I_a.$$

Thus the two circuits are equivalent provided $V_0 = -R_1I_a$. From fig. 4 it is also seen that $I_a = SV_3$, and from fig. 3 that $V_0 = -V_3/\cos^3 60° = -8\ V_3$. The condition $V_0 = -R_1I_a$ therefore leads to

$$-8V_3 = -R_1SV_3,$$

or

$$R_1S = 8. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

From the oscillation condition $ZS = -1$ it follows that in combination with (5) the transfer impedance $Z = -R_1/8$, as may easily be deduced direct from fig. 4. In the first instance $R_1$ is therefore the only element in the network that governs $Z$. This is another way of formulating the property of the network referred to above.

In order to satisfy this condition also when the valve slope or the resistance $R_1$ changes, a regulating voltage derived by rectification from the oscillating voltage is caused to act upon one of the grids of the valve. The circuit then oscillates with an amplitude such that the regulating voltage corresponds to the required slope. By taking suitable measures, which we shall not enter upon here [2]), it can be ensured that this will be the case already at a very small amplitude of oscillation.

As a consequence the amplified signal — about which more will be said under the next heading — will not be modulated to any extent by the oscillation voltage.

*Applying an input circuit and an output circuit*

We shall now use the relation $R_1S = 8$ to amplify the signal eight times with the oscillating valve. The frequency with which the amplifier oscillates must be so chosen — as already remarked above — as to prevent any troublesome interference arising, thus for instance much lower than the signal frequency [3]).

Let us first consider the circuit represented in *fig. 5*, which has been obtained from that of fig. 4 by adding the circuits $L_1$—$K_1$ and $L_2$—$K_2$ tuned to the signal frequency. The signal voltage $V_i$ laid on to the input gives rise to an anode current $I_a' = SV_o$ ($I_a$ is used to distinguish this from the anode current $I_a$ having the oscillation frequency). The output voltage $V_o$ then becomes $V_o = Z_2SV_i$, where $Z_2$ represents the impedance of the circuit $L_2K_2$ for the signal frequency. Then, however, the signal amplification $V_o/V_i = Z_2S$ is proportional to $Z_2$ and thus, however constant $S$ may be kept, is dependent upon the elements of the circuit $L_2$—$K_2$.

Moreover, with the circuit according to fig. 5 neither of the signal circuits is earthed, and this constitutes an objection if it is desired to connect several stages in cascade. (One could, it is true, couple the output of one stage, for instance, inductively with the input of the next stage, but then we should have the coupling coefficient acting as a factor the variations of which are not compensated.)

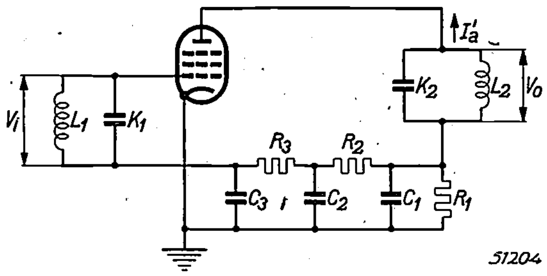


Fig. 5. The same circuit of fig. 4 with the addition of two circuits $L_1$-$K_1$ and $L_2$-$K_3$ tuned to the signal frequency.

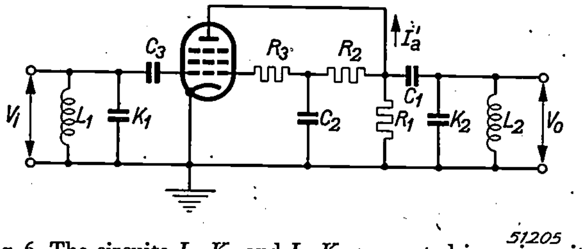


Fig. 6. The circuits $L_1$-$K_1$ and $L_2$-$K_2$ connected in series with $C_3$ and $C_1$ rcspectively in such a way that both are earthed and the resistor $R_1$ is connected parallel to the circuit $L_2$-$K_2$.

These objections can be overcome by proceeding according to the diagram of *fig. 6*, where the circuit $L_1$—$K_1$ is connected in series with $C_3$ and $L_2$—$K_2$ in series with $C_1$ in such a way that both circuits are earthed. If, now, the oscillation frequency is much lower than the signal frequency then the self-inductances $L_1$ and $L_2$ will practically act as shortcircuits for the currents with the oscillation frequency, so that as regards the latter the feedback network may be considered as unchanged. The signal current $I_a'$ now traverses an impedance formed by the parallel circuiting of $R_1$, $R_2$, $L_2$ and $K_2$ ($C_2$ and $C_1$ may be imagined as being short-circuited for the signal frequency). Since the circuit $L_2$—$K_2$ behaves, in respect to the signal frequency to which it is tuned, as a resistance which is much greater than $R_1$, and since, according to a supposition already previously made, $R_2$ is also large in comparison with $R_1$, the anode impedance for the signal frequency equals approximately $R_1$. Hence the output voltage is by approximation $R_1I_a' = R_1SV_i = 8V_i$, so that the amplification does indeed amount to approximately 8 and is practically independent of the values of all the circuit elements employed.

It is to be remarked here that the amplification is not limited just to one value, *i.e.* 8. This is the value found from $1/\cos^3 60°$. In the first place it is not necessary to obtain the phase shift of 180° in three sections of 60° each, for this can also be done with four sections of 45°, five of 36°, etc. or in general with $n$ sections of $180°/n$. The amplification then becomes $1/\cos^n (180°/n)$, *i.e.* for $n = 3, 4, 5, \ldots$ respectively 8, 4, 2.89, ...; thus it is greatest for $n = 3$, in which case the circuit is also the simplest.

Other possibilities are presented when it is borne in mind that oscillation can be brought about not only with a total phase shift of 180°, but likewise with an odd multiple of 180° and, if an oscillator of two valves is employed, also with even multiples of 180°. For instance in the case of two stages a phase

[2]) What is mcant is that the oscillating voltage can be amplified with an auxiliary valve before a regulating voltage is derived from it by means of rectification; see fig. 7 of article I.

[3]) One can also work with an oscillating frequcncy much higher than the signal frequency provided the necessary alterations are made in the circuit.

shift of 360° can be obtained with 5 sections of 72°, which would lead to an amplification of $1/\cos^5 72° \approx 355$.

Finally it is to be remembered that in the foregoing circuit diagrams we have ignored the possibility of a mutual load between the *RC* sections, it being supposed that $R_2 \gg R_1$, $R_3 \gg R_2$ and so on. If we drop this restriction, which has only been introduced for the sake of simplicity, then the amplification may deviate rather considerably from the above mentioned value $1/\cos^n (m \cdot 180°/n)$, in which $m$ represents the multiple of 180° used. In our example with 3 × 60° one is not restricted to an 8-fold amplification but can give it any value between about 4 and 10 by choosing suitable values.

### *A practical example*

Thus we have indicated in broad lines the fundamental idea of the process to be followed, which may be summarized as follows: just as in article I, an amplifier is employed which oscillates with a non-disturbing auxiliary frequency, while the oscillation voltage supplies a regulating voltage that compensates variations in slope. As feedback network we have a quadripolar system possessing the above-mentioned correlation between the amplitude and the phase characteristics, so that when changes take place in the values of the components the attenuation in this quadripolar system remains practically constant. Equal to this attenuation is the amplification of the signal.

In actual fact account must of course be taken of the factors which have so far been ignored, for which small corrections have to be made. For instance allowance will have to be made for the loads of the second and third *RC* sections (see fig. 2*a*) on the preceding section(s). Account must also be taken of the influence, be it ever so small, of the series resistance of the coils upon the oscillator, and also of the variation in the high-frequency anode impedance due to the deviations allowed in the values of the circuit elements. It would lead us too far to go into all these finesses here.

Suffice it to give here an example of practical application and to mention the maximum deviation to be expected from the nominal amplification. This example is by no means claimed to be the most favourable case that could be found.

In *fig. 7* we have a diagram of an amplifying stage for an angular frequency of $3 \cdot 10^6$ rad/sec. This diagram is complete except for the regulating voltage mechanism, for which an auxiliary amplifier can be used (*cf.* footnote [2])). The values of the components given in the text underneath the illustration are to be regarded as prescribed values. We shall assume that in the actual construction components may be used the value of which is not greater than 1.05 times and not smaller than 1/1.05 times the prescribed value; thus, for instance, where a capacity of 1000 pF is indicated a capacitor of between 1050 and 952.4 pF may be used. Under these conditions the amplification of this stage will deviate less than 1% plus or minus from the nominal amplification, which amounts to 8.1336 [4]).

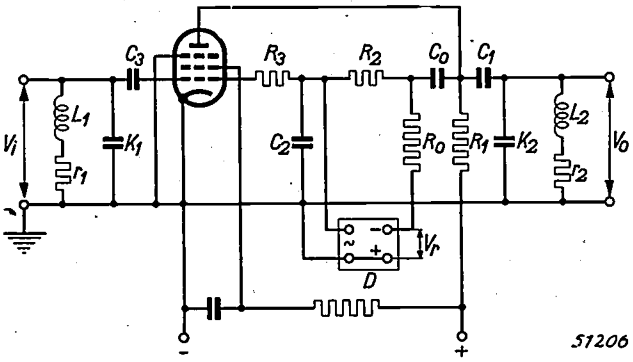


Fig. 7. The circuit of fig. 6 in a further developed stage. From the oscillation voltage on $C_2$ the part *D* (not shown in detail) derives a regulating voltage $V_r$ which is brought, *via* the resistors $R_0$, $R_2$ and $R_3$, to the control grid so as to limit the amplitude of oscillation and compensate the slope variations. $C_0$ is a blocking condenser. The resistance of the coils $L_1$ and $L_2$ is indicated by $r_1$ and $r_2$ respectively; the other symbols are the same as in fig. 6.

As an example the following case is taken. If the values of the components lie between 1.05 times and 1/1.05 times the prescribed values, then with a signal frequency of $3 \cdot 10^6/2\pi$ the amplification is 8.1336 ± 1%, for which the following components are required:

| | | | |
|---|---|---|---|
| | $R_0$ | = | 1 Mohm |
| | $R_1$ | = | 2 004 ohm |
| | $R_2$ | = | 30 000 ohm |
| | $R_3$ | = | 750 000 ohm |
| | $C_0$ | = | 10 μF |
| | $C_1$ | = | 394 000 pF |
| | $C_2$ | = | 24 000 pF |
| | $C_3$ | = | 1 000 pF |
| | $L_1 = L_2$ | = | 2 mH |
| | $r_1 = r_2$ | = | 30 ohm |
| | $K_1 = K_2$ | = | 55,5 pF |
| Nominal valve slope without regulating voltage | $S$ | > | 6 mA/V |
| Internal resistance of the valve | $R_i$ | > | 1 Mohm |

The angular frequency of the oscillation = abt. 2300 rad/sec.

### *Stages in cascade*

From the fact that the amplification per stage varies so much less than the variations of the component parts it follows that a certain number of stages can be connected in cascade without fear of the deviation in the total amplification exceeding the tolerance of the elements. Consequently in the practical example just given, where the ratio of these relative deviations is more than 5, one may go as far as five stages in cascade. One then reaches an amplification of well over 35000 times, exact to within a tolerance of less than 5%

[4]) Ar first sight there does not seem to be any sense in giving the amplification figure to four decimals when deviations up to 1% are possible. It would, however, be incorrect to round off say 8.1336 to 8.13, because the limit values of 8.13 ± 1% would not correspond to the guaranteed limit values of 8.1336 ± 1% (or rather: 8.1336 ± 1.00%).

in the least favourable case. If the values chosen for the elements are well within the tolerance then the deviation from the nominal amplification may be appreciably less.

With these stages in cascade the output circuit of one stage can serve as input circuit of the next stage, thanks to the fact that all these circuits are earthed.

The oscillation frequencies should be chosen far enough apart as to ensure that the stages do not interfere with each other.

### *What is the least favourable case?*

One might well expect the least favourable case to be that where all the circuit elements show the maximum permissible deviation, perhaps with some in the positive direction and others in the negative sense, but this need not always be the case.

In order to demonstrate this in a simple way we will suppose for a moment that only two of all the circuit elements, for instance one resistor $R$ and one capacitor $C$, have a certain tolerance and that all the other parts have the prescribed value. With $R$ and $C$ as coordinates we can draw lines of constant amplification, which, as the calculation proves, are curves as shown in *fig. 8*. From this diagram it is to be seen that the maximum amplification, in this case 8.3, is reached when $R$ has the minimum value $R_{\min}$ and $C$ the maximum value $C_{\max}$, whereas the minimum amplification, 8.0, is obtained with $R = R_{\max}$ and $C =$ a value between $C_{\min}$ and $C_{\max}$; with $R = R_{\max}$ and choosing $C = C_{\min}$ or $C_{\max}$ the amplification was increased again (up to about 8.15 and 8.05 respectively).

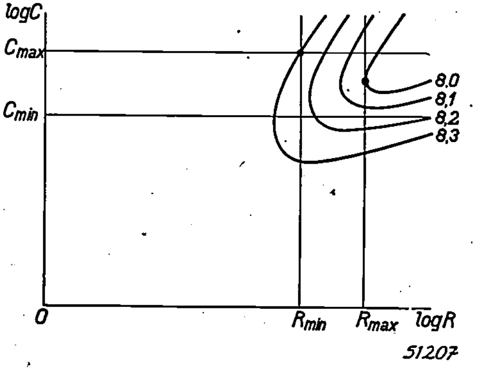
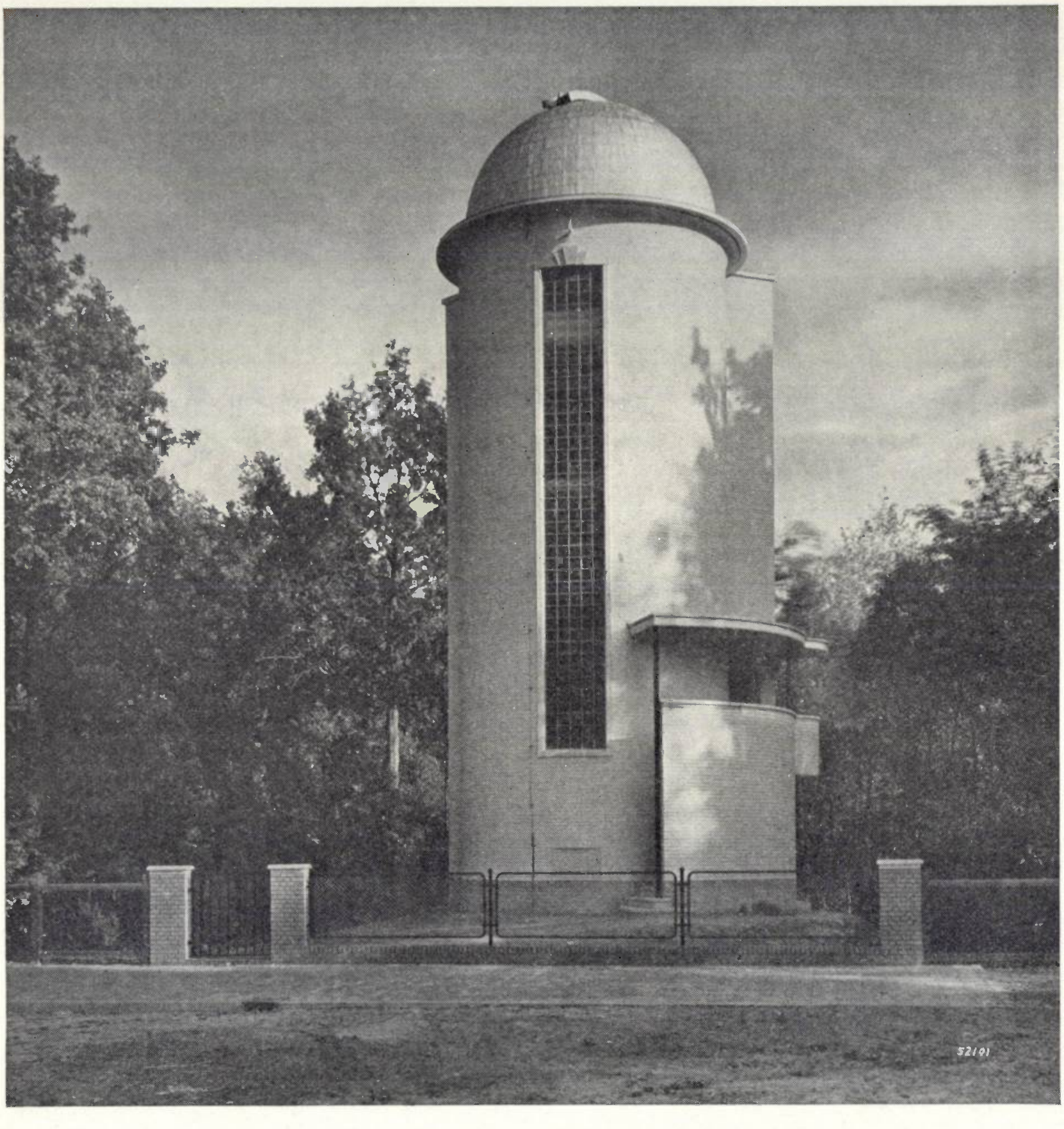


Fig. 8. $R_{\min}$, $R_{\max}$, $C_{\min}$, $C_{\max}$ = limits between which the value of one resistor $R$ or of one capacitor $C$ respectively may lie in the diagram of fig. 7; the other components have the prescribed values. With log $R$ and log $C$ as coordinates the lines of constant amplification are the curves indicated. It is seen that the smallest amplification (8.0) is obtained when $C$ has a value between $C_{\min}$ and $C_{\max}$.

In principle the same applies when there are more than two variable elements, only then we have to deal with just as many coordinates as there are variable elements.



The Eindhoven Observatory, which for a number of years already has been cooperating with the scientists attached to the Philips Physical Laboratory.

# MEASURING THE RATE OF WATCHES WITH A CATHODE-RAY OSCILLOGRAPH

by H. van SUCHTELEN.

681.112.47.001.4:621.317.361

A description is given of an experimental apparatus with which it is possible to compare the rate of a timepiece with a standard frequency with the aid of a normal cathode-ray oscillograph. By means of a microphone the ticking of the timepiece is converted into voltage peaks which produce the vertical deflection on the oscillograph. The time-base voltage of the oscillograph is synchronized with a fixed standard frequency (60 c/s) derived by frequency division from the highly constant frequency of a quartz oscillator. From the speed at which the peak moves across the oscillograph screen it can be determined how much the timepiece is gaining or losing. A small rate error, of for instance a few seconds per 24 hours, can be determined within a very short space of time (a few minutes).

The usual practice is to test the rate of a watch by comparing it with that of a standard chronometer. To reach an accuracy of say 1 second per 24 hours (for which the watch must of course have a seconds hand) it is necessary to keep the timepiece under observation for 24 hours. After adjustments have been made it is as a rule necessary to check the rate again. Furthermore it is often desired to ascertain to what extent the rate is affected by the position of the timepiece and by the fact whether the spring is wound up or run down. A thorough test may therefore take several days and in many cases this obviously has its objections.

A much quicker method is possible by electrical means, by comparing the frequency of the escapement with a considerably higher standard frequency. This method is applicable not only to complete watches but also to the escapements alone, which are generally made in special factories and have to be tested before delivery to the watchmakers. Apparatus with which these tests can be made have been known for a number of years already. In the Philips laboratory an experimental apparatus has been devised for this purpose which makes use of a cathode-ıay oscillograph and which is briefly descıibed in this article.

The watch or escapement[1]) to be tested is placed in a holder and clamped onto a microphone which picks up the sound of the ticking. This signal, consisting of a strongly damped vibration and some background noise, is amplified about 10 000 times and applied to a relay valve (a gas-filled triode; see *fig. 1*), in the anode circuit of which sharply defined current impulses are released corresponding to the ticking, which normally should have one of the standardized frequencies of 4, 5 or 6 c/s.

A suitable frequency for comparison could be taken in a similar manner from the ticking of a chronometer, but a still greater degree of accuracy is to be obtained with the aid of a quartz oscillator. To this end use is made of one of the characteristic vibrations of a piezo-electric quartz plate which in the electrical sense behaves as an *LC* circuit of very high quality. In combination with a triode with feedback continuous oscillations with a highly stabilized frequency can be obtained. The extent to which this frequency is stabilized depends, for a given manner of oscillations, upon the dimensions of the quartz plate, the angle of its planes with respect to the axis of the crystal, and upon the temperature. The relative frequency variation due to fluctuations in temperature can be kept at about $10^{-6}$ per °C[2]) by a suitable choice of the

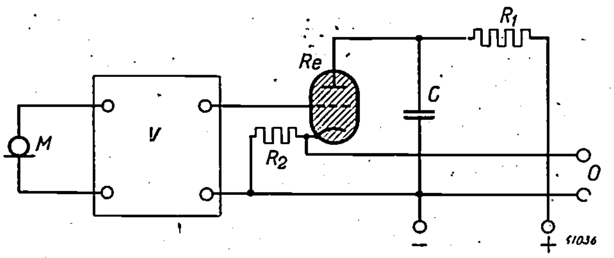


Fig. 1. A circuit supplying sharply defined voltage impulses with the frequency of the ticking of the watch being tested. The circuit formed by the resistor $R_1$, the capacitor $C$ and the relay valve *Re* connected in parallel to the latter can generate a relaxation oscillation with a frequency approximately equal to the ticking frequency of the watch (4-6 per second). *Via* the microphone *M* and the amplifier *V* the ticking produces an alternating voltage on the grid of the relay valve, which synchronises the relaxation oscillation with the ticking. At every tick contact is made in the relay valve and the capacitor $C$ is discharged across the valve and the resistor $R_2$. The capacitor is then recharged *via* the resistor $R_1$, discharged again upon the following tick and so on. In this way $R_2$ receives voltage impulses which cause the vertical deflection on a cathode-ray oscillograph connected at *O*.

[1]) The damping of an escapement alone is so extremely small that after it is put in motion by a slight touch it continues to run long enough for the test to be made, which, as will appear farther on, takes only a few minutes.

[2]) The frequency of a quartz oscillator can be accurately determined by comparing it with one of the standardized frequencies transmitted by various radio stations, for instance 200 000 c/s used by Droitwich of the B.B.C., which shows a relative deviation less than $10^{-8}$.

cut and the manner of oscillation. This variation corresponds to a gain or loss of no more than one second in 12 days.

The characteristic frequency of a quartz plate of reasonable dimensions (thickness of the order of a few millimeters, diameter of the order of a few centimeters) lies at about $10^4$—$10^6$ c/s, which is much too high for direct comparison with a frequency of 4—6 c/s. By applying the principle of frequency division, however, it is possible to derive from the characteristic frequency a frequency that is a whole number of times lower. Such a division can be brought about in various ways. In the case in question here use has been made of synchronised oscillators with sinusoidal oscillations [3]), the action of which will be briefly considered here.

In order to dcrive from the frequency of the quartz oscillator, which in our case amounts to 72 900 c/s, for instance a frequency three times as low (24 300 c/s) a simple valve oscillator is adjusted as closely as possible to a characteristic frequency of 24 300 c/s and then coupled in a suitable way to the quartz oscillator. Provided a certain condition that will presently be discussed is satisfied, the synchronism will be maintained — *i.e.* the frequency of the second oscillator will remain exactly 1/3 of the frequency of the quartz oscillator — even if the characteristic frequency of the second oscillator should vary within certain limits (due to temperature changes or fluctuations in the supply voltage).

The effect of the injection of a high-frequency voltage into the oscillator with a lower frequency averages out to nothing if that voltage should be active during the whole cycle of the low frequency. Consequently this has to be avoided, and it is possible to do so in a simple way by applying the high-frequency to the grid of the valve in the low-frequency oscillator and at the same time applying to that grid such a high negative bias as to cause anode current to flow only during a small part of each cycle (*fig. 2*). As a result it is only during these brief intervals that the synchronising oscillation with the higher frequency can make its effect felt (provided suitable amplitudes are chosen). The synchronising voltage is then active only once in each $p$ cycles (by division $p$ : 1) of that fiequency, thus only then contributing to the supply of energy to the oscillator circuit which is tuned to the lower frequency.

If it is desired to give a certain ratio of division $p$ strong preference over neighbouring ratios ($p$—1 and $p + 1$) then a limit has to be set for the order of $p$. If the frequency $f_0$ of the quartz oscillator weie divided by a factor of for instance $p = 10$ then the risk of the characteristic frequency of the second oscillator varying from the nominal value $f_0/10$ to a neighbouring frequency fraction $f_0/9$ or $f_0/11$ and the oscillation frequency jumping to one of these values is much greater than when one chooses for instance $p = 3$, where the adjacent frequency fractions are relatively farther removed from the nominal value. A high value of the ratio $p$ therefore makes high demands as regards stability of the component parts of the circuit and of the supply voltage.

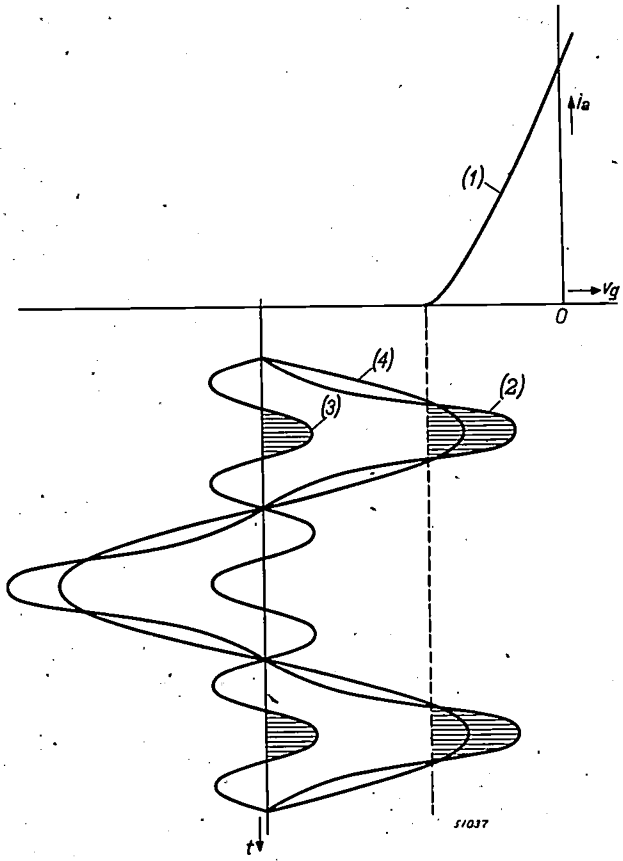


Fig. 2. Curve (*1*) is the anode current $i_a$ as a function of the grid voltage $v_g$ of the oscillator valve in an oscillating circuit the frequency of which must be $p$ times smaller than a control frequency $f_0$ (here $p = 3$). The grid alternating voltage (*2*), plotted as a function of the time $t$, consists of two components (*3*) and (*4*) with frequencies of $f_0$ and $f_0/p$ respectively. The adjustment of the valve and the amplitudes of (*3*) and (*4*) are so chosen that an anode current flows only during a small part of the cycle of (*4*). Only one of each $p$ positive half-waves of (*3*) contributes to the anode current, so that the oscillator is excited with a frequency $f_0/p$.

Since in our case any jump of the frequency would lead to entirely erroneous conclusions any possibility of this occurring must be precluded. For this reason $p$ is chosen not higher than 5. By connecting

[3]) I. Koga, A new frequency transformer or frequency changer, Proc. I.R.E. 15, 669-678, 1927.

a number of dividing stages in cascade an arbitrarily low frequency can also be obtained with small values of $p$.

What then is the most suitable value for the standard frequency? As already stated, for the escapements of watches three frequencies have been standardized: 4,5 and 6 c/s. Obviously it is preferable that only one frequency should be needed as comparison for these standardized frequencies. Therefore it should be a multiple of 4,5 and 6, thus say 60 c/s. With the experimental apparatus described here the standard frequency for comparison is in fact 60 c/s. It is derived by frequency division in six stages from the frequency of 72 900 c/s supplied by the quartz oscillator. The principle of the set-up is indicated in broad lines in *fig. 3*, while a photograph of the arrangement is given in *fig. 4*.

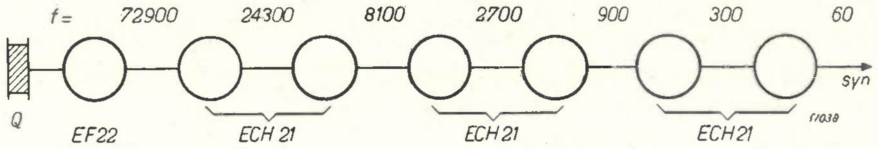


Fig. 3. By means of frequency division one may derive from the frequency 72900 c/s of a quartz plate $Q$, oscillating by means of a pentode EF 22 with feedback, the frequency of 60 c/s with which the time-base voltage of a normal cathode-ray oscillograph is synchronized. In this instance the frequency division takes place in five stages with triple division and one stage with quintuplicate division. The numbers in the diagram indicate the frequencies $f$ in the various stages in cycles per second. By using valves of the ECH 21 type, containing two electrode systems (a triode and a heptode), the total number of amplifying valves required to effect the frequency division 72900 : 60 can be reduced to three.

The time-base voltage of a normal cathode-ray oscillograph is synchronized with this standard frequency, whilst the impulse of the ticking of the timepiece under test provides the vertical deflection, as is to be seen in *fig. 5*. When, therefore, a watch with the nominal frequency of 5 ticks per second is tested each tick is recorded on every twelfth time-base line described on the screen. If the rate of the timepiece is correct then the image of the tick impulse will appear exactly at the same place every time, whereas if the watch regularly gains or loses then the image shifts regularly to the left or to the right, whilst if the rate is irregular this will be reflected in an irregular displacement along the time-base line.

Let us now considei the case where a timepiece (with a nominal frequency of $n$ ticks per second) has a small erro. of say 5 seconds per 24 hours. How much time will it then take to produce with this apparatus an easily discernible displacement of the impulse on the oscillograph screen, for instance over the whole length of the time-base?

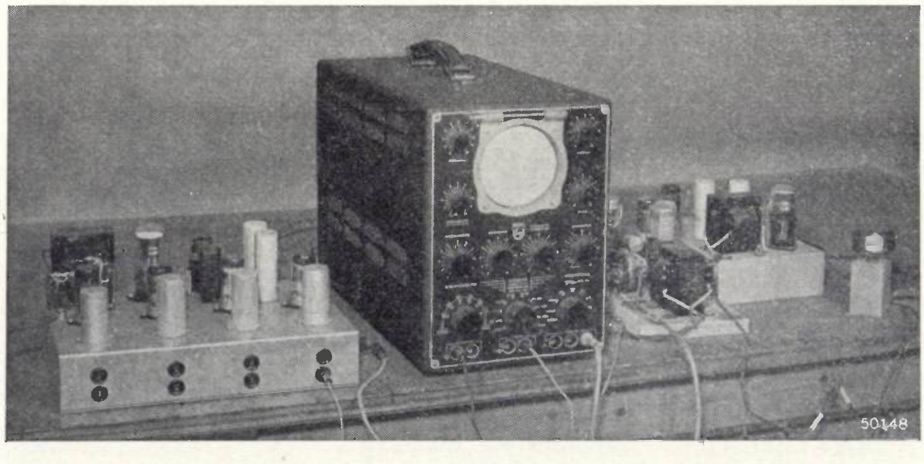


Fig. 4. Experimental arrangement of the apparatus for measuring the rate of watches. In the centre the cathode-ray oscillograph, to the left the apparatus supplying the synchronizing voltage with a frequency of 60 c/s. (see fig. 3). On the extreme right the microphone with a watch placcd on top. Between the microphone and the oscillograph are shown the microphone amplifier and the apparatus for converting the ticks of the watch into voltage impulses (see fig. 1).

In 24 hours there are 86400 seconds. The presumed error in the rate of the timepiece is therefore 5 : 86400. Between two successive ticks the length of the time-base should be described exactly $60/n$ times. The displacement of one impulse with respect to the previous one is therefore $(5/86\,400) \times (60/n)$ times the length of the time-base. This displacement occurs in $1/n$ second (the interval between two ticks). The displacement of the impulse grows to the length of the time-base in

$$\frac{86\,400}{5}\cdot\frac{n}{60}\cdot\frac{1}{n}\text{ sec} = 288\text{ sec} \approx 5\text{ min.}$$

Since it is not necessary to wait until the impulse has been displaced over the entire length of the time-base, the error mentioned can easily be observed within a few minutes. The fact that only such a short time is needed is due to the standard frequency being 10—15 times as high as the ticking frequency, for this means that the time scale is extended as it were by a factor of 10—15, thus showing the displacement all the sooner. Even a complete series of tests with the timepiece in different positions or fully wound up or almost run down takes only a relatively short time. It may be remarked here that a watch with the presumed error of only 5 seconds in 24 hours is regarded as a very good time-keeper. A timepiece running less accurately shows a certain displacement of the impulse in a correspondingly shorter time.

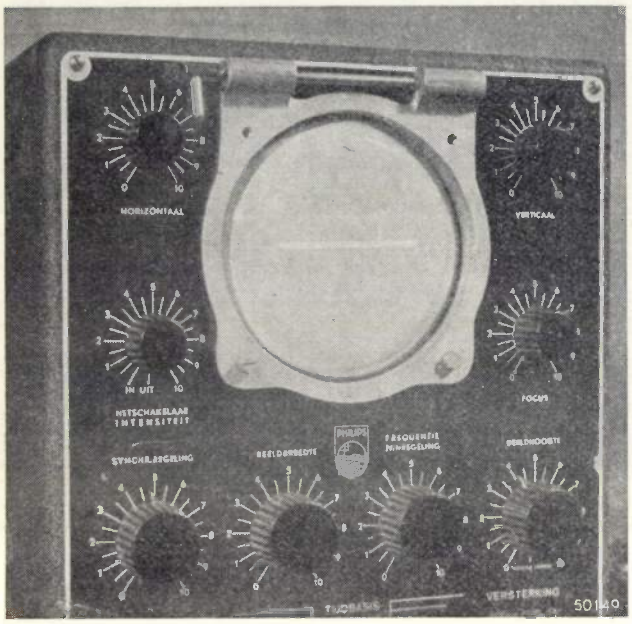


Fig. 5. The voltage impulses generated by the ticking of the watch are visible on the screen of the oscillograph in the shape of peaks, which if the watch is keeping correct time appear at the same place every time but shift when the watch is gaining or losing. The speed at which the peaks shift is a measure for the deviation from the correct time-keeping. Even a small deviation can already be measured in a few minutes.

VOL. 9 No. 11, pp. 321-348 1947/1948

Philips Technical Review

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

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

# THE LIGHT-EMISSION OF X-RAY SCREENS

by H. A. KLASENS *) and W. de GROOT.

537.531:535.371

This article reviews the various ways in which fluorescent screens are employed in radiography and discusses the factors determining the brightness of a viewing screen or the darkening of a photographic image obtained with the aid of an intensifying screen. The factors referred to are:

1) the absorption of X-rays in the screen;
2) the conversion of absorbed X-ray energy into energy of fast electrons formed in the mass of the screen;
3) the conversion of the electron energy into light energy or photographically active radiation;
4) the absorption and the scattering of light in the screen;
5) the selectivity of the eye or of the photographic emulsion relative to the spectral distribution of the radiated fluorescent light.

The optimum thickness for viewing and intensifying screens and the most suitable luminophores are investigated. It is shown that for intensifying purposes, when using low tube voltages (less than 70 kV), ZnS screens are to be preferred to $CaWO_4$ screens.

## Introduction

Fluorescent screens play an important part in radiography. In the first place they are used as viewing screens for diagnostical purposes and for the examination of materials. In the second place a fluorescent screen is needed in screen photography, when the image formed on the screen is photographed by a camera. In the third place fluorescent screens are used as intensifying screens for producing a direct image on a photographic film, where the fluorescing layer is in immediate contact with the sensitized material.

In all cases the fluorescent screen consists of a layer of very small crystals of substances such as zinc sulphide (ZnS) or calcium tungstate ($CaWO_4$), which have the property of converting the absorbed X-ray energy into visible or photographically active rays. These crystals are held together by a binder, usually an organic substance, which cements the crystals together in a smooth layer. This layer of crystals with binder is applied on a thin base of white cardboard to give it the necessary firmness. Cardboard is chosen because in many cases it is necessary that the base should not obstruct the passage of the X-rays; it will be seen farther on why the cardboard must be white.

*Fig. 1* illustrates diagrammatically the use of a fluorescent screen as a viewing screen or in screen photography. The X-rays coming from the left, possibly reduced in strength by the object *A*, strike the cardboard base (*1*) of the screen *S* and pass through it into the fluorescent layer (*2*). Owing to the differences in transmission between the parts of the object there are differences of brightness in the fluorescent layer, so that a "shadow image" of the

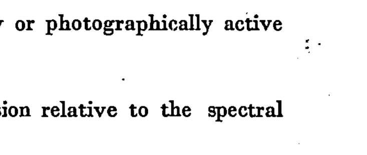


Fig. 1. Diagrammatic representation of the arrangement of a viewing screen or an arrangement for screen-image photography. *A* = object, *C* = the eye or camera, *1* = cardboard base, *2* = fluorescent layer.

*) Material Research Laboratory, Philips Electrical Ltd., London, England.

object can be observed in all its details. $C$ is the eye or the camera.

*Fig. 2* illustrates the various ways in which an intensifying screen can be used. In fig. 2*a* we see a front screen with the cardboard side facing the X-ray tube as in fig. 1. The photographically active layer (*3*) applied on a celluloid film (*4*) is in immediate contact with the fluorescent layer (*2*).

Fig. 2*b* shows a fluorescent screen used as a back screen, where the X-rays first pass through the celluoid (*4*) and the sensitized layer (*3*) of the film and then through the immediately adjacent fluorescent layer (*2*). Fig. 2*c* shows a frequently applied combination of a front screen and a back screen: the film is usually covered with a sensitized layer on both sides (*3*) and (*5*), against which fluorescent layers (*2*) and (*6*) of the screens are laid.

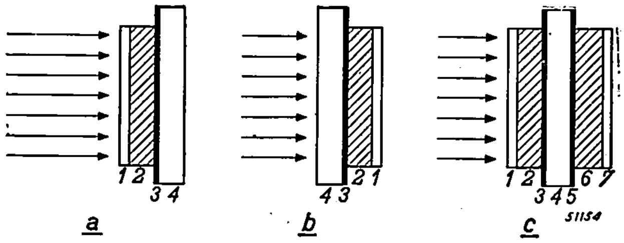


Fig. 2. Diagram showing how intensifying screens are used. *a*) front screen, *b*) back screen, *c*) front and back screens. *1* and *7* cardboard base, *2* and *6* fluorescent layer, *3* and *5* sensitized layer, *4* celluloid.

The X-rays absorbed in the fluorescent layer release electrons which cause that layer to glow. Primarily the luminous intensity of the image is therefore proportional to the amount of X-ray energy absorbed. But not all the light coming from the crystals of the fluorescent substance emerges from the fluorescent layer itself, because a part of it is scattered and absorbed. Owing to this scattering and absorption the luminous intensity is reduced and moreover there is a lack of sharpness in the image observed or photographed. The thicker the fluorescent layer, the more pronounced these two effects become. When the screen is used as an intensifying screen then the sensitized layer of the film is also struck directly by the X-rays, but these only contribute towards the darkening to a small extent (3 to 10%), mainly because the X-rays used in radiography are only to a small extent absorbed in the photographic emulsion layers. The visible or ultra-violet light generated in the fluorescent layer thereby causes an intensification (shortening of the exposure time) by a factor of 10 to 30.

This factor would be considerably smaller if a larger fraction of the X-ray energy could be absorbed in the photographic emulsion itself, for instance by making it much thicker. This, however, is not very feasible, for practical reasons. In crystal structure analysis X-rays of much longer wavelength than those used in Röntgenology are used. These rays are much more strongly absorbed in the normal photographic emulsion and consequently the use of intensifying screens is of little advantage in this case.

With a given intensity of X-rays it is obviously of importance to obtain the highest possible luminous intensity or actinic action of the screen. The object of this article is to investigate the factors influencing the brightness of the screen.

## The absorbed X-ray energy

We may imagine X-rays as being a rain of quanta, each having an energy $h\nu$ ($\nu$ = frequency, $h$ = Planck's constant = $6.62 \times 10^{-34}$ watt sec$^2$).

Upon striking a substance these X-rays are attenuated in two ways, firstly by absorption, the energy $h\nu$ of the quantum (less the ionization energy) being transmitted to an electron which is thereby released from the atom struck, and secondly by scattering, the rays being only changed in direction without loss of energy. Since, however, with the substances from which fluorescent screens are made the tendency of X-rays to be scattered is small compared with the absorption, this scattering will be left out of consideration.

The consequence of scattering is that the average path travelled by a quantum in the layer is somewhat greater than the thickness $D$ of the screen, and as a result there is a slight increase in the total absorption.

For a screen consisting of a certain kind of atoms the number of quanta absorbed per second per square centimeter in a layer of a thickness d$x$ will be:

$$N_0\, n\, a\, \mathrm{d}x,$$

in which $N_0$ represents the number of incident quanta per cm$^2$, $n$ the number of atoms per cm$^3$ and $a$ the cross section for absorption per atom.

$$\text{Since } n = \frac{\text{mass per cm}^3}{\text{mass per atom}} = \frac{\varrho}{A/N_A}\ .$$

($\varrho$ = density in g/cm$^3$, $A$ = atomic weight, $N_A$ = $6.023 \times 10^{23}$ = the Avogadro number),

$$N_0\, n\, a\, \mathrm{d}x = N_0 \frac{\varrho a N_A}{A}\, \mathrm{d}x = N_0 \mu \mathrm{d}x, \quad . \ . \quad (1)$$

where $\mu = \varrho a N_A / A$ is the so-called absorption coefficient.

In the case of a compound consisting of more than one kind of atom

$$a_{mol} = n_1 a_1 + n_2 a_2 + \ldots ,$$

where $n_1$, $n_2$, .... represent the number of atoms per molecule, whilst

$$\mu = \frac{\varrho a_{mol} N_A}{M},$$

where $\varrho$ represents the density of the compound and

$$M = n_1 A_1 + n_2 A_2 + \ldots .$$

is the molecular weight of the compound.

The quantity $a$, as we have seen, is related to the release of an electron from the atom. It depends upon the frequency $\nu$, and thus also upon the wavelength $\lambda = c/\nu$ ($c$ = velocity of light), such that

$$a = C \cdot Z^4 \cdot \lambda^3 . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

in which $Z$ is the atomic number of the element absorbing the X-rays. As the wavelength is shortened (frequency raised) $a$ decreases until a frequency is reached at which a more tightly bound electron can be released. The constant $C$ then suddenly jumps to a higher value, and the frequency or the wavelength at which this jump takes place is called the absorption limit or absorption edge (*fig. 3*).

On the so-called $K$-edge, corresponding to the release of a $K$-electron (an electron bound in the innermost "shell" of an atom), with which we are most concerned in our case, $C$ ($\lambda$ in Å) jumps from the value

$$0.3 \cdot 10^{-26} \text{ for } \lambda_K < \lambda < \lambda_L$$ [1]

to

$$2.3 \cdot 10^{-26} \text{ for } \lambda < \lambda_K .$$

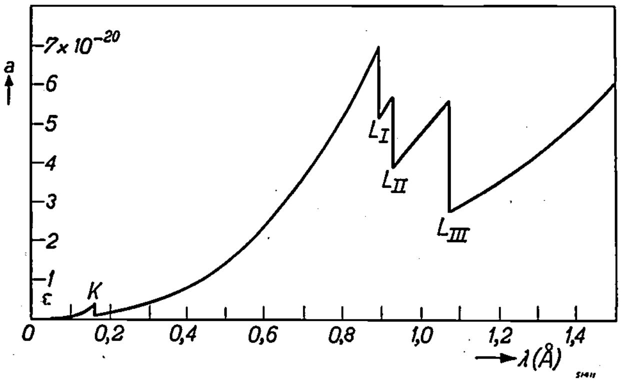


Fig. 3. Graph representing the absorption $a$ as function of the wavelength $\lambda$ for the element Pb. The single $K$-absorption edge and the three-fold $L$-absorption edge occur when the quantum of X-rays is large enough to release an electron from the $K$ and the $L$ shells respectively.

*Fig. 4* gives the absorption coefficients of the compounds $CaWO_4$ and ZnS. The absorption edge ($K$-edge) in the curve for $CaWO_4$ is due to the $W$-atoms. The absorption edges of Ca, O, Zn and S and the $L$-edge of W fall outside the graph.

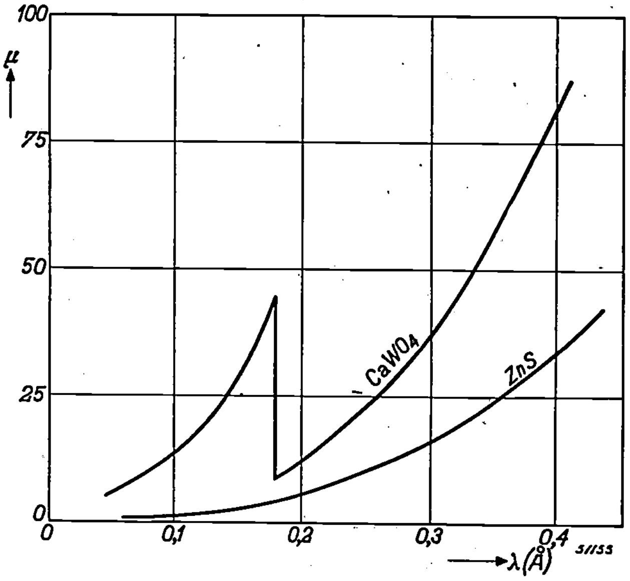


Fig. 4. Graphical representation of the absorption coefficient $\mu$ (in cm$^{-1}$) as function of the wavelength $\lambda$ (in Ångström) for $CaWO_4$ and ZnS.

Taking "monochromatic" X-rays, the number of quanta passing through a finite layer of the thickness $D$ is $N_0 e^{-\mu D}$ and the energy absorbed is:

$$N_0 h\nu \, (1 - e^{-\mu D}) \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

In practice we never have monochromatic X-rays, and this makes the calculation more complicated. The distribution of the continuous X-ray energy (the characteristic radiation does not play an important part here) among the various wavelengths is represented by the curves in *fig. 5*. The shortest wavelength in the continuous X-ray spectrum is given by the equation

$$\lambda_{\min} = \frac{12.4}{V}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (4a)$$

the wavelength $\lambda$ being expressed in Ångström and the tube voltage $V$ in kV. When, as is usual, the rays are filtered through a thin metal plate, the long-wave rays are suppressed, the maximum intensity then being about $\lambda = 20/V$. The calculation can be simplified by assuming the radiation to be monochromatic with a wavelength

$$\lambda_{\text{eff}} = \frac{20}{V} \cdot \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4b)$$

With a tube voltage of 50 kV, thus with $\lambda_{\text{eff}} = 0.4$ Å, the value of $\mu$ in a ZnS screen is 30 cm$^{-1}$ and

[1] For the sake of simplicity the three-fold absorption edge is regarded as a simple edge.

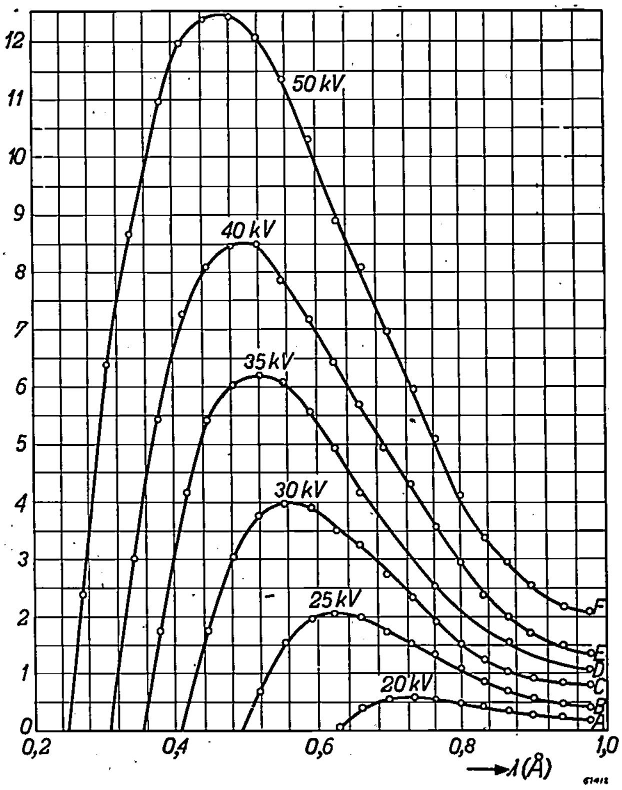


Fig. 5. Energy distribution in the continuous X-ray spectrum as a function of the wavelength $\lambda$ for various voltages between 20 and 50 kV. With increasing tube voltage the peak of the graph is shifted to the left and the value of $\lambda_{min}$, the shortest wavelength in the radiation, becomes smaller. The intensity is given on the vertical axis on an arbitrary scale.

in a $CaWO_4$ screen 80 cm$^{-1}$. In a layer of zinc sulphide 0.03 cm thick, which is reasonable for a fluorescent screen, about 60% of the X-ray energy is then absorbed.

### The conversion of X-ray energy into light energy

As we have seen, the X-ray energy is converted into light energy in two stages:

1) conversion of the absorbed quantum of X-rays into a rapid electron;
2) conversion of the energy of the electron into the energy of the fluorescent radiation.

From one energy unit of absorbed X-rays $k$ energy units of light radiation are obtained ($k < 1$), for which we write:

$$k = k_{re} \cdot k_{el} \quad . \ . \ . \ . \ . \ . \ . \ . \quad (5)$$

in which $k_{re}$ relates to the process 1) and $k_{el}$ to 2). We shall now consider these two coefficients $k_{re}$ and $k_{el}$ separately.

#### *The conversion of X-ray energy into electron energy*

The quantity $k_{re}$ is easily calculated. Assuming that a $K$-electron is released from the atom, the energy expended is $h\nu_K$, in which $\nu_K$ is the frequency of the $K$-edge. The energy imparted to the electron is then $h\nu—h\nu_K$, so that in this case the yield is

$$\frac{\text{electron energy}}{\text{X-ray energy}} = \frac{\nu - \nu_K}{\nu} = k_K .$$

Similarly, for the release of an $L$-electron we have:

$$\frac{\nu - \nu_L}{\nu} = k_L .$$

Since $\nu_L$ is usually very much smaller than $\nu$, this last fraction may be taken as being equal to 1, whilst if $\nu > \nu_K$ the chance of $L$-absorption may be ignored. Hence per atom one may write:

$$k_{re} = \frac{\nu - \nu_K}{\nu} \quad (\nu > \nu_K, \lambda < \lambda_K) \quad . \quad (6a)$$

$$k_{re} = 1 \quad (\nu < \nu_K, \lambda > \lambda_K) \quad . \ . \ . \ . \quad (6b)$$

By determining the average of these coefficients for all kinds of atoms we find the value of $k_{re}$ for the whole screen.

In *fig. 6* $k_{re}$ is plotted as a function of the wavelength for two screen substances ($CaWO_4$ and ZnS).

When a $K$-electron is ejected from the atom then an $L$-electron may fall into the unoccupied $K$-orbit, with the emission of a quantum whose frequency is less than $\nu < \nu_K$ (X-ray-fluorescent radiation or characteristic radiation). If this quantum is absorbed in the layer it cannot release another $K$-electron but can release an $L$-electron. We shall disregard this process, although actually $k_{re}$ is thereby increased a little and, if the frequency is greater than $\nu_K$, approaches more closely to the value 1.

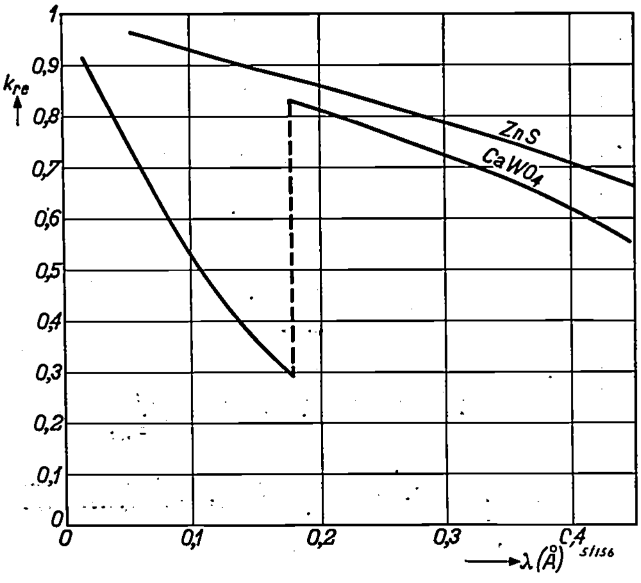


Fig. 6. The yield $k_{re}$ from the conversion of X-rays into electron energy as function of the wavelength $\lambda$ for $CaWO_4$ and ZnS.

### *Conversion of electron energy into light*

As to the conversion characterized by $k_{el}$, much is already known from measurements of the light yield of cathode-ray tubes. In these tubes a fluorescent screen is struck directly by fast electrons. It has been found that in the case of electrons having an energy of 50 kV and more the light energy generated may amount to 10% of the electron energy.

Further, $k_{el}$ can of course be deduced from measurements of the luminous intensity of a fluorescent screen irradiated with X-rays or with gamma rays [2]). From these measurements, too, it is to be deduced that

$$k_{el} = 0.05 — 0.1.$$

### Scattering and absorption of light in the screen

Let us assume that we have a screen of the thickness $D$ which is homogeneously irradiated with X-rays, and confine our attention to one particular crystal at a distance $x$ below the surface which is brought to luminosity by the radiation. On its way to the free side of the screen the light from this crystal will be scattered (but without loss of energy) and subsequently absorbed (energy being thereby lost). Consequently the available light energy will be less than

$$k \cdot N_0 h\nu \,(1—e^{-\mu D}).$$

Extensive calculations concerning the influence of absorption and scattering have been made by Hamaker [3]), who also took into consideration the reflection of the light from the cardboard layer and, where present, the photographic film. It if were not reflected by the cardboard then a considerable portion of the light energy would be lost on the "wrong" side of the fluorescent layer, and that is why white cardboard is used as base.

Hamaker's formulae for the light energy emitted per cm$^2$ are as follows:

$$J_B = A \cdot kN_0 h\nu \,\frac{\mu}{\mu + \sigma}\,(1—e^{-(\mu+\sigma)D}) \;.\;.\;.\;.\;.\;.\; (7)$$

$$J_F = A \cdot kN_0 h\nu \,\frac{\mu}{\mu - \sigma}\,(e^{-\sigma D}—e^{-\mu D}) \;.\;.\;.\;.\;.\; (8)$$

$$J_S = B \cdot kN_0 h\nu \,\frac{\mu}{\mu - \sigma}\,(e^{-\sigma D}—e^{-\mu D}) \;.\;.\;.\;.\;.\; (9)$$

Equation (9) applies to the case of fig. 1 (viewing screens), equation (8) to that of fig. 2*a*, and (7) to that of fig. 2*b* (intensifying screens). The coefficients $A$ and $B$ are generally 0.5 to 1, whilst $\sigma$ is a coefficient of attenuation, related both to the absorption and to the scattering of the light.

The equation for the attenuation coefficient $\sigma$ is:

$$\sigma = \sqrt{\mu_l\,(\mu_l + 2\sigma_l)} \;.\;.\;.\;.\;.\;.\;.\; (10)$$

in which $\mu_l$ represents the absorption coefficient (in diffuse light) and $\sigma_l$ the scattering coefficient of the layer, both for light.

The coefficients $A$ and $B$ are equal to

$$A = \frac{1 + \varrho_1}{2\,(\varrho_1 + \beta)}, \qquad B = \frac{1}{1 + \beta},$$

where $\varrho_1 = (1—r_1)/(1+r_1)$, in which $r_1$ represents the reflectivity of the film, whilst $\beta = (1—R_\infty)/(1+R_\infty)$, $R_\infty$ indicating the reflection coefficient of an infinitely thick fluorescent layer. In the derivation of equations (7) to (9) it has been assumed that $\beta = \varrho_2 = (1—r_2)/(1+r_2)$, where $r_2$ represents the reflection coefficient of the cardboard. If this condition does not exist then the equations become much more complicated [3]). This assumption, however, is by no means imaginary, for it appears that both $\beta$ and $\varrho_2$ are very much less than 1, or $r_2$ and $R_\infty$ are approximately equal to 1. Further, absorption of the X-rays in the cardboard is ignored.

Hamaker's formulae show a great similarity to the formulae previously deduced by Glocker and his fellow-workers [4]), viz:

$$J_B = kN_0 h\nu \,\frac{\mu}{\mu + \tau}\,(1—e^{-(\mu+\tau)D}) \;.\;.\;.\; (11)$$

$$J_F = kN_0 h\nu \,\frac{\mu}{\mu — \tau}\,(e^{-\tau D}—e^{-\mu D}) \;.\;.\;.\;.\; (12)$$

These formulae have the advantage that they can easily be deduced by integration, by assuming that a fraction $e^{-\tau x}$ or $e^{-\tau(D-x)}$ of a quantity of light generated at a depth $x$ in the fluorescent layer emerges at the surface (here the layer is assumed to be without a cardboard base), so that

$$J_B = kN_0 h\nu \int_0^D \mu e^{-\mu x} \cdot e^{-\tau x}\,\mathrm{d}x,$$

$$J_F = kN_0 h\nu \int_0^D \mu e^{-\mu x} \cdot e^{-\tau(D+x)}\,\mathrm{d}x.$$

Here $\tau$ has the character of a pure absorption in contrast to $\sigma$, which quantity depends upon absorption and scattering.

The quantity $\sigma$ can be determined cxperimentally by measuring the quantity $J_B/J_F$, which depends only upon $\mu$ and $\sigma$; if the chemical composition of the screen is known, $\mu$ can be calculated in the manner indicated above.

[2]) In a recently published article by J. W. Coltman, E. G. Ebbinghausen and N. Altar (J. Appl. Physics **18**, 530-544, 1947, No. 6) it has been deduced from the luminous intensity of a CaWO$_4$ screen irradiated with X-rays that the value of $k_{el}$ for CaWO$_4$ is 0.05. J. H. Born, N. Riehl and K. G. Zimmer. (Reichsber. Phys. **1**, 154-158, 1945) give for ZnS-Cu a much higher value, viz: $k_{el}$ = 0.71, but the value given by Riehl himself for the conversion of alpha rays into light, viz: $k_{al}$ = 0.80, is much higher than the values quoted by other authors (see H. A. Klasens, Trans. Faraday Soc. **42**, 666-668, No. 11). When the corrections indicated by Klasens are applied to the calculation one finds from the same experimental data for ZnS $k_{el}$ = 0.05 to 0.1.

[3]) H. C. Hamaker, Philips Res. Rep. **2**, 55-67, 1947 (No.1).

[4]) R. Glocker, E. Kaupp, H. Widmann, Ann. phys. Leipzig, **85**, 313-332, 1928.

It appears that $\sigma$ decreases with increasing crystal size. With different X-ray screens its value varies rather considerably, being of the order of 10 to 100 cm$^{-1}$.

### *Light intensity as a function of D*

From equations (7) to (9) three conclusions can be drawn:

1) From (7) it appears that when a fluorescent screen is used as a back screen the amount of light energy radiated per cm$^2$ approaches a limiting value with increasing thickness of the screen, this being equal to 0.99 of the maximum light energy available when the thickness $D = 4.6/(\mu + \sigma)$; for $\mu = \sigma = 50$ this thickness is 0.046 cm.

2) When the fluorescent screen is used as a front screen or as a viewing screen it is possible to find the value of $D$ at which the light energy is a maximum by differentiating formula (8) or (9) with respect to $D$:

$$\frac{\mathrm{d}J}{\mathrm{d}D} \sim \frac{\mu e^{-\mu D} - \sigma e^{-\sigma D}}{\mu - \sigma} \quad . \; . \; . \; . \; . \quad (13)$$

Taking this differential quotient as being equal to zero, it follows that:

$$D_{\mathrm{opt}} = \frac{\ln (\sigma/\mu)}{\sigma - \mu} \quad . \; . \; . \; . \; . \quad (14)$$

In *fig. 7* $D_{\mathrm{opt}}$ is plotted as a function of $\mu$ for some values of $\sigma$.

By substituting the value of $D_{\mathrm{opt}}$ in (9) for instance we get

$$(J_S)_{\max} = B \cdot kN_0 h\nu \cdot F\left(\frac{\sigma}{\mu}\right) \quad . \; . \quad (15)$$

in which

$$F\left(\frac{\sigma}{\mu}\right) = \left(\frac{\sigma}{\mu}\right)^{\frac{\sigma}{\mu-\sigma}} \quad . \; . \; . \; . \; . \quad (16)$$

This quantity is represented in *fig. 8* as a function of $\sigma/\mu$.

3) The case of a combined front and back screen is somewhat more complicated. First of all it must be said that the use of a film sensitized on both sides and with two adjacent screens offers great advantages over a film sensitized on one side used with one screen. In the first place, given the same quantity of light energy, the photographic density is more intense.

This is due to the fact that with increasing light energy $L$ the density $z$ increases less than proportionately with $L$, as is represented, for instance, by the empirical formula

$$z = E \log (1 + pL), \quad . \; . \; . \; . \; . \; . \quad (17)$$

in which $E$ and $p$ are constants. From *fig. 9*, in which $z$ is plotted as a function of $L$, it appears that

$$z\,(L_1) + z\,(L_2) > z\,(L),$$

when

$$L_1 + L_2 = L.$$

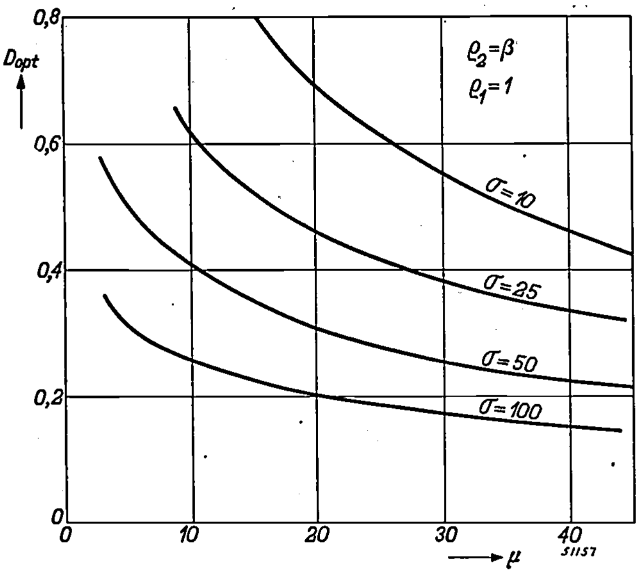


Fig. 7. The optimum thickness (in mm) of a viewing screen as function of the absorption coefficient $\mu$ for several values of the attenuation coefficient $\sigma$. Here it is assumed that $\varrho_2 = (1-r_2)/(1+r_2)$ (in which $r_2$ = the reflectivity coefficient of the cardboard) is equal to $\beta = (1-R_\infty)/(1+R_\infty)$ (in which $R_\infty$ = the reflectivity coefficient of an infinitely thick fluorescent layer) and that $r_1$, the reflectivity of the film for light, is so small that $\varrho_1 = (1-r_1)/(1+r_1)$ can be taken as equal to 1.

This can also be easily demonstrated analytically. At the same time it appears that

$z(L_1) + z(L_2)$ has a maximum when $L_1 = L_2 = L/2$.

The second advantage in using more than one screen lies in the fact that with two thin screens placed one behind the other more light is obtained than from one screen of the same thickness as that of the two thin ones together. In both cases, it is true, the same amount of X-ray energy is absorbed, but when two thin screens are used more

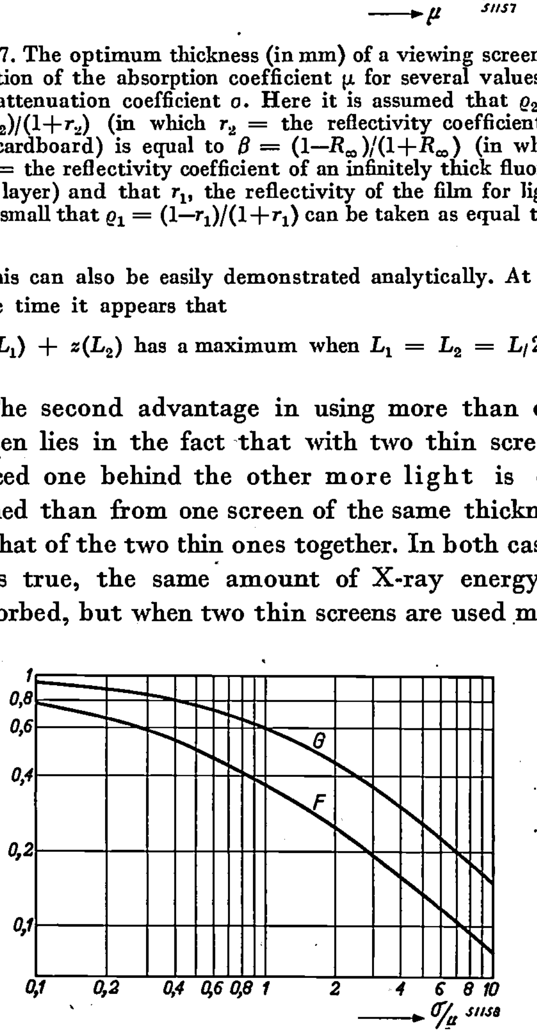


Fig. 8. The quantities $F(\sigma/\mu) = (\sigma/\mu)^{\sigma/(\mu-\sigma)}$ and $G(\sigma/\mu) = (^1/_2 + \sigma/2\mu)^{\sigma/(\mu-\sigma)}$ as functions of $\sigma/\mu$, where $\sigma$ represents the attenuation, are used for determining the optimum thickness of viewing screens and of front screens respectively.

light energy is released, because there is less absorption of light.

Theoretically one might therefore expect a maximum photographic effect when a large number of films are used, each being separated from the next by a thin screen. The light rays would then have to pass through no more than the thickness of a thin intermediate screen before acting upon a film. If, on the other hand, one thick screen, is used the light has to pass through deep layers where it is strongly absorbed before emerging. An improvement is obtained when we use a thick fluorescent layer as back screen and a thin foil as front screen as is done in practice. The truth of this argument has been experimentrlly confirmed by Barth and Eggert[5]) from an investigation with a combination of two films and four screens.

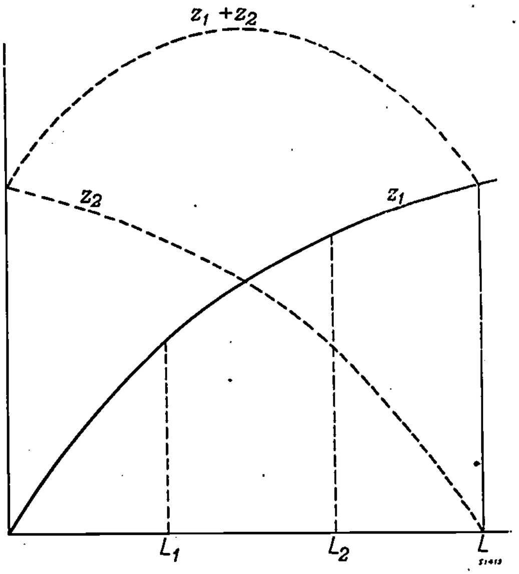


Fig. 9. Graphical representation of the blackening $z$ as function of the light energy $L$ according to formula (17). It is seen that $z(L_1) + z(L_2) > z(L)$ for $L_1 + L_2 = L$. The left-hand member of this inequality reaches its maximum for $L_1 = L_2 = L/2$.

In the same way as has been done for viewing screens and for simple front screens, we can now investigate what the optimum thickness of a front screen is when this is used in combination with the back screen.

If the back screen is taken to be of "infinite thickness" then the available energy is

$$J_B + J_F = A \cdot kN_0h\nu\mu \left(\frac{e^{-\mu D}}{\mu+\sigma} + \frac{e^{-\sigma D} - e^{-\mu D}}{\mu-\sigma}\right). \quad (18)$$

Differentiation with respect to $D$ yields:

$$D_{\text{opt}} = \frac{\ln\,[2\mu/(\mu+\sigma)]}{\mu-\sigma} \quad . \; . \; . \; . \quad (19)$$

In *fig. 10* $D_{\text{opt}}$ is plotted as a function of $\mu$ for several values of $\sigma$. When we compare fig. 10 with fig. 7 it is found that with equal values of $\sigma$ and $\mu$ and an infinitely thick back screen the optimum thickness of a front screen is smaller than that of a scanning screen.

By substitution in (18) we obtain:

$$(J_B + J_F)_{\max} = A \cdot kN_0h\nu \cdot G\left(\frac{\sigma}{\mu}\right), \; . \; . \quad (20)$$

in which

$$G\left(\frac{\sigma}{\mu}\right) = \left(\frac{\mu+\sigma}{2\mu}\right)^{\frac{\sigma}{\mu-\sigma}} \quad . \; . \; . \; . \; . \quad (21)$$

This quantity is also shown in fig. 8 as a function of $\sigma/\mu$.

Further, it appears that when the maximum effect is reached[6])

$$(J_B)_{\text{opt}} = (J_F)_{\text{opt}}.$$

The thinner the back screen is made, the larger is the optimum thickness of the front screen, until finally, when there is no back screen at all, it becomes equal to the optimum thickness of a viewing screen.

## The utilisation of light energy by the eye or by the photographic emulsion

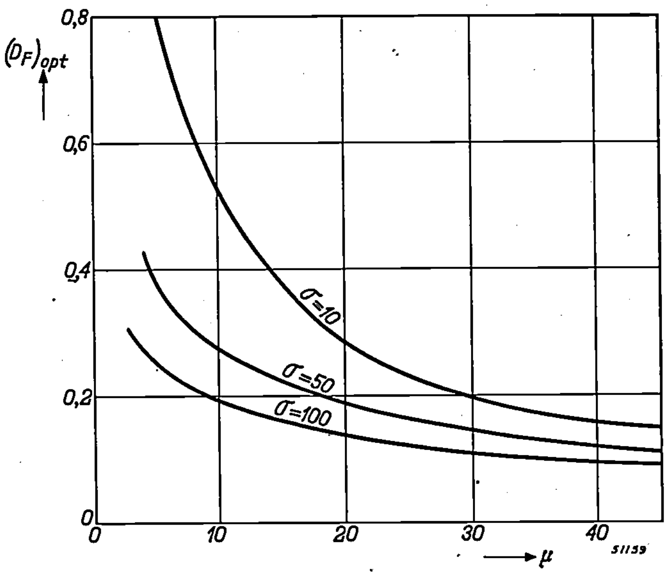


Fig. 10. The optimum thickness of the front screen (in mm) as function of the absorption coefficient $\mu$ for three values of the attenuation coefficient $\sigma$. Here it is assumed that the back screen is of infinite thickness.

From the foregoing it can be deduced what factors have to be taken into account in order to

[5]) W. Barth and J. Eggert, Fortschr. Rö. 39, 88-100, 1929.

[6]) It is interesting to note that this result agrees with the condition for a maximum photographic effect (the splitting up of a given quantity of light into two equal parts).

get a maximum amount of energy in the form of visible or photographically active light. The light stimulus brought about by this energy in our eye, or the darkening of the sentitized layer, depends, however, to a certain extent also upon the selective properties of the eye, or of the sensitized layer respectively, and the spectral composition of the fluorescent light.

With a high brightness (> 1 milli-Lambert) the maximum sensitivity of the eye lies at 5500 Å and with lower brightnesses ($10^{-2}$—$10^{-5}$ milli-Lambert) such as occur with viewing screens it drops gradually to 5200 Å (Purkinje effect). For viewing screens one must therefore use substances whose maximum light emission lies around this wavelength.

Formely these screens used to be covered with barium platinocyanide (it was with the help of this substance that Röntgen discovered his X-rays) or with willemite ($ZnSiO_4$-$Mn_2SiO_4$), but these substances have a relatively low light yield. More suitable are substances such as ZnS-Cu or (60% Zn 40% Cd) S-Ag, which possess a high light yield. Moreover, this latter substance has a high $\mu$ value. From fig. 8 (curve $F$) it may be seen that given an equal value of $\sigma$ (e.g. $\sigma = 50$) about twice as much light is obtained with a ZnCdS-screen ($\mu = 22$ for $\lambda = 0.25$ Å, $V = 80$ kV) as with a ZnS-screen ($\mu = 10$).

Sulphide screens are subject to slight decomposition when exposed to daylight, the resultant blackening being due to the ultra-violet and the blue rays of the sunlight. This effect can be counteracted by adding to the binder of the fluorescent substance a dyestuff (e.g. auramine) which absorbs the short-wave part of the sunlight but allows the light generated in the screen to pass through.

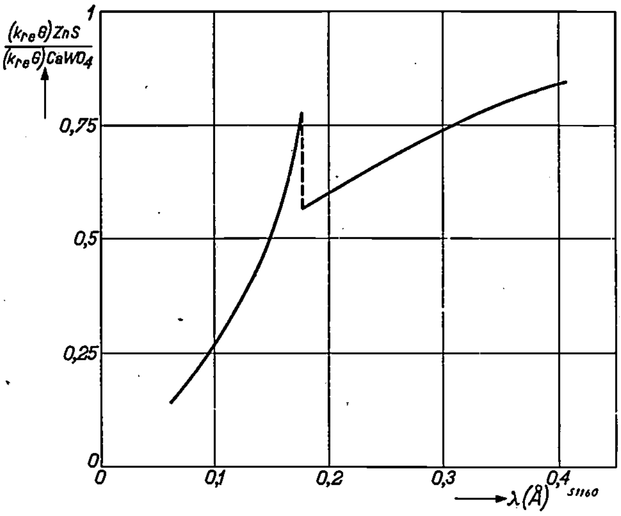


Fig. 11. The ratio of the quantities $(k_{r e}G)$ ZnS and $(k_{re}G)$ $CaWO_4$ (for $\sigma = 50$) as function of the wavelength $\lambda$. Here $k_{re}$ represents the efficiency of the transition from X-rays into electron energy, whilst $G$ is the function illustrated in fig. 8.

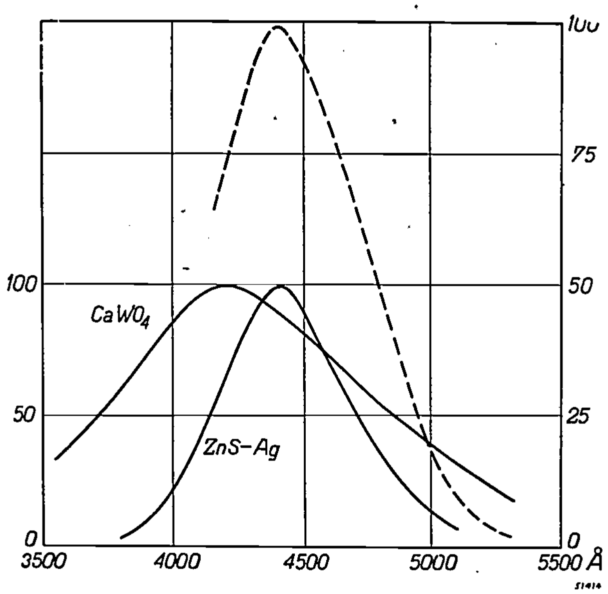


Fig. 12. The full lines represent the spectral distributions of the light emitted by a $CaWO_4$ screen and a ZnS-Ag screen. The intensity is given along the vertical axis on the left on an arbitrary scale. The dotted line represents the relative spectral sensitivity of a non-sensitized photographic film, derived from measurements with Agfa Spectral and Ilford Spec. Rapid films; the corresponding scale is given on the right..

Sulphides are apt to give a troublesome phosphorescence, but this can be suppressed by adding a trace of nickel (a so-called "killer") to the luminescent substances.

If it is desired to photograph a viewing screen then the sensitivity of the photographic material must correspond to that of the eye, which means that orthochromatic or panchromatic material must be used. If, however, it is not necessary for the screen to be viewed directly and only screen-image photography is to be applied, or if direct photographs are to be taken with an intensifying screen then a non-sensitized film suffices, since there are enough substances to be found which have a powerful emission in the violet or in the near ultra-violet, where this film is most sensitive. For such purposes ZnS-Ag and $CaWO_4$ are suitable.

With the aid of the data given in this article we shall try to decide which of these two substances is to be preferred for making intensifying screens. If it is assumed that the values of the constants $A$ and $B$ in equations (7) to (9) differ very little for these two substances, then these factors can be disregarded, and we have to compare the values of $kG$ for these two luminophores.

Fig. 4 gives $\mu$ as a function of $\lambda$, while fig. 8 gives $G$ as function of $\sigma/\mu$ and fig. 6 $k_{rc}$ as function of $\lambda$. By a combination of these data *fig. 11* has been constructed to show the ratio of $(k_{re}G)_{ZnS}$: $(k_{.e}G)_{CaWO_4}$ $CaWO_4$ as function of the wavelength. From this figure it is to be seen that with equal values of $\sigma$ (e.g. 50) and of $k_{el}$ a ZnS-screen yields less energy than one of $CaWO_4$. On the other hand, however, owing to the larger crystals in ZnS-screens the value of $\sigma$ is usually smaller; moreover, the light emission of ZnS-Ag matches the sensitivity of the photographic plate better (see *fig. 12*). In practice one finds the following conditions [7]):

| Tube voltage in kV | Ratio of the amplifying factors ZnS : $CaWO_4$ |
|---|---|
| 40 | 3 |
| 55 | 3 |
| 70 | 2.5 |
| 85 | 1.5 |
| 100 | 1.25 |

From these figures one will therefore be inclined to prefer zinc sulphide screens to calcium tungstate screens where these are to be used as intensifying screens with low tube voltages, but the fact must be borne in mind that with zinc sulphide screens the X-ray image is less sharp.

This screen-unsharpness, which is inherent in the use of intensifying screens and plays an important part also in viewing screens, will be discussed in a subsequent paper.

[7]) Taken from a publication of Ilford Ltd. The figures relate respectively to Ilford Standard Tungstate and Ilford Fluorazur Screens.

# A CONDENSER-MICROPHONE FOR STEREOPHONY

by A. RADEMAKERS.

621.395.616 : 534.76

During recent years a condenser-microphone has been employed in Philips laboratory for stereophonic sound reproduction. For this purpose the condenser-microphone has to satisfy certain requirements, viz: 1) up to 14 000 c/s the frequency characteristic must follow a horizontal straight line; 2) the dimensions of the microphone may not exceed the wavelength at the highest frequency to be reproduced; 3) the threshold value must lie below 30 phons (i.e. sounds stronger than 30 phons must come through undisturbed). The first requirement is met by ensuring that the resonance frequency is 12 000 c/s, by utilizing the "stiffness" of the air cushion behind the diaphragm. The resonance peak is eliminated by introducing an air-damping. In order to obtain, with the small dimensions, a sufficiently low threshold value, the distance between the diaphragm and the electrode must be small (13 μ). The sensitivity appears to be 3 mV/μbar for a direct voltage of 100 V. Finally the construction is described.

## The various types of microphones

In the course of years many different principles have been applied for the construction of microphones, each of which in turn was popular for a time. First of all the common telephone working on the electro-magnetic principle was used as microphone, but its low sensitivity was a drawback, because there were then no amplifiers available. Then the carbon microphone offered a solution, until the advent of radio broadcasting made heavier demands upon the quality of the microphone, whilst with the employment of amplifiers sensitivity became of less importance. One of the oldest quality microphones consisted of a coil in the shape of a flat helix suspended in the wide air-gap of an electromagnet; this may be regarded as the forerunner both of the coil microphone and of the ribbon microphone, both of which types are still at the present day considered to be first-class microphones and are frequently used.

For quite a time a prominent place was occupied by an improved design of the carbon microphone of much better quality due to its filling of fine-grained carbon powder and large diaphragm. With large volumes, however, the sound was too badly distorted.

A remedy of this evil was found — though at a sacrifice of sensitivity — in the condenser-microphone, which then ousted the carbon microphone. The objection, however, of having to mount the first amplifying valve in the immediate vicinity of the microphone led in turn to the advent of the electro-dynamic microphone, whilst also the piezo-electric principle opened the possibility of constructing good microphones. But all these types of microphones have their typical drawbacks, so that not a single one of them has definitely done away with the others, and thus various types are still in use for different purposes [1]).

The innovation of frequency modulation, accompanied by the aim at extending the frequency characteristic up to the highest tone limit, and, further, the necessity of placing two microphones in a dummy head for stereophonic sound reproduction [2]), led to the desire to restrict the dimensions of the microphone.

As to expansion of the frequency characteristic, when the dimensions of the microphone are larger than the wavelength of the sound and this microphone is introduced into a sound field then the original sound field is disturbed by the occurrence of diffraction phenomena. As a consequence, the actual frequency characteristic of the microphone (so-called field calibration) is different from that obtained when the microphone is exposed to a frequency-independent sound pressure (pressure calibration). Since this is undesirable, it is necessary that the microphone should be smaller, at least in two dimensions, than the wavelength at the highest frequency to be reproduced.

For stereophony a dummy head is used, so that the two microphones pick up the sound with the same differences as are received in the ears of the human head when one is listening normally. It would, therefore, be obvious to choose dimensions for the microphone of the same order as those of the human ear.

Electro-dynamic microphones with their rather

[1]) An exhaustive discussion of various types of microphones is to be found in an article by J. de Boer, Philips Techn. Rev. 5, 140-148, 1940.

[2]) K. de Boer, Stereophonic sound reproduction, Philips Techn: Rev. 5, 107-114, 1940, and Stereophonic images, ditto, 8, 51-56, 1946 (No. 2).

voluminous permanent magnet cannot be made sufficiently small. The Rochelle salt crystals used in piezoelectric microphones have certain peculiarities making these microphones less suitable for quality reproduction. We have therefore tried to make the small condensor-microphone that has been in use a long time already for measuring purposes suitable for the reproduction of music. Owing to the limit set in the reduction of the dimensions by the accompanying decrease of sensitivity of the microphone, the dimensions could not actually be kept smaller than the wavelength at the highest frequencies. We have chosen a diameter equal to the wavelength at a frequency of about 14 000 c/s, i.e. 25 mm, so that the active part of the diaphragm may be about 15 mm in diameter. With these dimensions differences may arise between the field and the pressure calibration at frequencies above 2000 c/s, at most 5 db. when sound waves of 14 000 c/s strike the diaphragm perpendicularly; corrections must therefore be made when these microphones are used for measuring purposes.

When the microphone is used in a dummy head it is not the dimensions of the microphone that govern the diffraction phenomena but rather those of the dummy head itself. Nevertheless, the diaphragm may not be too large because otherwise different parts of it would be activated in different phases. With the chosen diameter of 15 mm the deviations due to this effect are kept below 2 db.

As already remarked, a reduction in dimensions is accompanied by reduced sensitivity, the latter being understood as the voltage supplied per microbar pressure [3]). In itself this is of little importance because with the present-day amplifying technique it is always possible to amplify a signal to the desired strength. Sensitivity only plays a part in as far as it influences the so-called threshold value of the microphone.

By threshold value is understood the loudness level [4]) of the sound that produces on the output side of the amplifier the same voltage as function of time as the noise of the circuit. It is essential that the weakest sounds to be transmitted should be stronger than the noise of the amplifiers. For that reason this threshold value is a measure for the usefulness of the microphone. It determines the weakest sounds that the microphone can reproduce without interference and should, therefore, really lie below the threshold value of hearing. Since, however, there is inevitably a certain noise level in the studio and in the concert hall, one can safely take the threshold value of the same order as that noise level. For the present we have to be satisfied with a threshold value of 30 phons, though this is higher than what can be reached with some other microphones [5]).

## The working of the condenser microphones

The first condenser microphone of the form in which it is commonly used at present was made by Wente [6]) in 1917. This type has a vibrator, facing a fixed elctrode; together they form a condenser. The vibrator must therefore be electrically conductive. It consists of a very thin foil of metal (e.g. aluminium 15 μ thick) stretched taut. *Fig. 1* is a schematic diagram of the arrangement of a condenser-microphone. Owing to the movement of the vibrator the capacitance of the condenser undergoes changes and as the flow of the charge is obstructed by the resistance potential differences arise between the points $A$ and $B$.

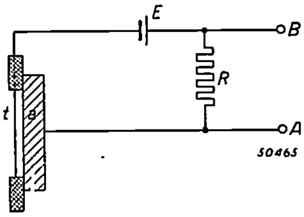


Fig. 1. Diagrammatic representation of the arrangement of a condenser microphone. The flexible electrode or vibrator $t$ and the fixed or rear electrode $a$ are connected via a resistor $R$ to the poles of a battery $E$. The movement of the vibrator causes capacity changes of the condenser which result in voltage variations between the points $A$ and $B$ of the circuit.

Let us imagine, for simplification, that the resistance is so great as to keep the charge $Q$ of the microphone constant. In that case the voltage $V$ on the microphone will always be inversely proportional to the capacitance $C$ ($V = Q/C$). In order to obtain a horizontal electrical frequency characteristic it is desired that equal sound pressures with different frequencies should give the same variations in capacitance, thus the same movements of the diaphragm. This is the reason why the foil is stretched taut. The resonance frequency is then high, and sufficiently far below this frequency the devia-

[3]) 1 μbar sound pressure means an R.M.S. alternating pressure of 1 dyne per cm$^2$.

[4]) The loudness of the sound is expressed in phons, the given value indicating by how many decibels an equally loud sound of 1000 c/s lies above the intensity of $10^{-16}$ W/cm$^2$. A tone of 1000 c/s and a pressure of 1 μbar has a loudness of 74 phons. The noise in a "quiet" concert room amounts to about 34 phons, and a normal conversation corresponds to 60 phons.

[5]) An article will shortly be published in this journal on a circuiting scheme worked out in Philips laboratory which when applied to the condenser-microphone gives a very low noise level.

[6]) E. C. Wente, Physical Review, 10, 39-63, 1917.

tion of a resonator with the same force is independent of the frequency.

In the case of a simple resonator having the mass $m$ and stiffness $s$, upon which a sinusoidal force acts which varies with the time $t$ and has the amplitude $K$, the equation of motion applying is:

$$m\ddot{x} + sx = K \sin \omega t,$$

where $\omega$ represents the angular frequency of the force. Resonance occurs when $\omega_0 = \sqrt{s/m}$, whilst when $\omega < {}^1/_2\, \omega_0$ with a deviation less than 2 db

$$x = \frac{K}{s} \sin \omega t.$$

From this it follows that with frequencies sufficiently far below the resonance frequency the unloaded microphone will give a horizontal characteristic. Confining ourselves to this area, we shall give farther on in this article a calculation of the sensitivity of the microphone, when it will be shown at the same time how the non-linear distortion is restricted to imperceptible values. This will be followed by a discussion as to how the horizontal part of the frequency characteristic can be extended to higher frequencies, whilst finally we shall deal with the influence of the electric circuit. We then have to do with the case where the microphone is working under load, a state which influences the frequency characteristic. Moreover, the circuiting involves noise voltage, which affects the threshold value of the combination of microphone and amplifier.

### Calculation of sensitivity

When we regard the diaphragm as a simple resonator we must assume that its deflection can be described with one coordinate. This is indeed possible as long as the frequency does not go far beyond that of the first resonance and all points of the diaphragm therefore move in phase with a practically constant characteristic deflection.

If this characteristic deflection is known then an equivalent amplitude can be ascribed to it, for instance by determining the average of the amplitude over the whole area ($O$). One then finds the equivalent mass $m_{\text{eq}}$ and the equivalent stiffness $s_{\text{eq}}$ of the system with one degree of movement by writing the kinetic and potential energies of the diaphragm proper respectively as

$$E_K = {}^1/_2\, m_{\text{eq}} \cdot \dot{x}^2_{\text{eq}} \qquad \text{and} \qquad E_P = {}^1/_2\, s_{\text{eq}} \cdot x^2_{\text{eq}}.$$

The force exercised upon the diaphragm by the sound pressure $p$ has an equivalent value equal to the actual force $p \cdot O$, since the work performed for a certain deflection can be written directly as the product of the actual force and the average deflection.

If we regard the metal foil as being entirely without any bending stiffness, thus as a stretched ideal diaphragm, then the deflection curve is fairly easily calculated. When a constant pressure is applied on one side (thus for the static case) the deflection is parabolic, whilst for the first characteristic vibration it is in the form of a Bessel function $J_0$.

The table below gives for these two cases:

1) the amplitude $x$ as function of the distance $\varrho$ to the centre ($R$ = radius of the diaphragm),
2) the maximum amplitude $x_0$ (for $\varrho = 0$) divided by the average amplitude $\bar{x}$,
3) the resultant $m_{eq}$ divided by the actual mass $m$.

| $x(\varrho)$ | $x_0/\bar{x}$ | $m_{\text{aeq}}/m$ |
|---|---|---|
| $x_0\left(1 - \dfrac{\varrho^2}{R^2}\right)$ | 2 | 1.33 |
| $x_0\, J_0\left(2.4 \cdot \dfrac{\varrho}{R}\right)$ | 2.32 | 1.45 |
| average: | 2.16 | 1.4 |

The actual shape of the curve is not easy to describe. In the first place the bending stiffness of the material undoubtedly plays a part. This rigidity tends to concentrate the deflection towards the centre, so that $x_0/\bar{x}$ rises (we have in mind here a flat clamped edge). On the other hand, however, there is the influence of the air between the diaphragm and the electrode, which tends to restrict most the greatest amplitudes and thus to reduce $x_0/\bar{x}$; this factor will be discussed later on.

We take it, therefore, that the figures given in the table above do not differ much from the reality, and we can now make an approximate calçulation of the sensitivity of the microphone in order to form an idea of the order of size of the actual figures.

An aluminium foil with a diameter of 15 mm and thickness of 15 $\mu$ has a mass of about 7 mg. This gives an equivalent mass of 7 × 1.4 = 10 mg. If this is to have a resonance frequency $\nu$ of 12 000 c/s then an equivalent stiffness is needed of $5.6 \cdot 10^7$ dyne/cm = 5.6 kg/mm. With a sound pressure of 1 $\mu$bar (which at 1000 c/s corresponds to a loudness of 74 phons) the equivalent force acting upon the diaphragm is 1.8 dynes. The r.m.s. value of the average amplitude is then $1.8/5.6 \cdot 10^7$ cm = 3.2 Å.

This is of the order of the spacing of the atoms in a crystal. It is astonishing that the microphone should still be able to give undisturbed reproduction with a 100 times smaller amplitude (34 phons) in spite of the fact that the movements of the diaphragm are then smaller than the dimensions of the atoms.

With the strongest sound occurring the amplitudes are 200 times greater than the first mentioned value.

The peak value of the deflection in the centre of the diaphragm is then 0.13 $\mu$, which is still to be regarded as small compared with the distance $a$ between the diaphragm and the fixed electrode, which usually exceeds 10 $\mu$ and in our case amounts to about 13 $\mu$.

We shall now calculate the capacitance between the diaphragm and the electrode for the case of a parabolic deflection with an amplitude $x_0$ in the centre and thus $^1/_2x_0$ as average value:

$$C = \int\limits_0^R \frac{2\pi\varrho \cdot \mathrm{d}\varrho}{4\pi\left[a + x_0\left(1 - \dfrac{\varrho^2}{R^2}\right)\right]} =$$

$$= \frac{O}{4\pi a}\left(1 - \frac{1}{2}\frac{x_0}{a} + \frac{1}{3}\frac{x_0^2}{a^2} - \ldots\right).$$

If the diaphragm is considered as one flat plane displaced with the average amplitude $^1/_2x_0$ then the capacitance would be:

$$\frac{O}{4\pi a\left(1 + \dfrac{1}{2}\dfrac{x_0}{a}\right)} = \frac{O}{4\pi a}\left(1 - \frac{1}{2}\frac{x_0}{a} + \frac{1}{4}\frac{x_0^2}{a^2} - \ldots\right).$$

Since $x_0/a$ is always less than 1% the variation in capacitance (which in both series is represented by the terms containing $x_0$) can be determined within an error of less than $2^0/_{00}$ by assuming that the diaphragm vibrates as one flat plane with the given amplitude.

In our microphone the fixed electrode has a diameter of 12 mm, thus smaller than the diaphragm, for reasons of insulation. For the variation in capacitance we have only to reckon with this part of the diaphragm. If we take for this $x_0$ divided by the average amplitude we find $x_0/\bar{x} = 1.4$. The average amplitude of 3.2 Å for the whole diaphragm thus corresponds to an average of 4.5 Å taken over the active part.

If we assume that the charge of the microphone remains constant and that the voltage between the diaphragm and the fixed electrode is 100 V (any higher voltage involves the risk of shorting) then for the amplitude mentioned we arrive at a calculated alternating voltage of 3.5 mV, since with constant charge and thus constant field strength the voltage varies proportionately with the distance. This voltage value agrees well with the measured value of the sensitivity.

### Damping and air-rigidity

In the foregoing we have already mentioned that the resonance frequency of the diaphragm is 12000 c/s. This has been arrived at as a compromise between two requirements, viz:

a) that frequencies up to 14 000 c/s must be reproduced with a reasonably horizontal characteristic (deviations less than 2 db);
b) that the threshold value should lie below 30 phons and sensitivity should be sufficiently high to attain this.

The first requirement could easily be satisfied by choosing a resonance frequency higher than 14 000 c/s, but the higher stiffness that this would require would reduce the sensitivity too much to be able to maintain the second requirement.

We therefore followed another method, fixing the resonance frequency in the highest range to be reproduced and damping off the resonance peak in such a way that the characteristic extends horizontally to 14 000 c/s with only small deviations.

When a diaphragm of duralumin of the dimensions already quoted is stretched as tightly as is possible without fracture, the resonance frequency lies approximately at 8000 c/s. This can be raised to 12 000 c/s in one of two ways, viz:

a) The bending stiffness could be increased by using a thicker material. This, however, means increasing the mass and then the stiffness would have to be further increased; this has a very adverse effect upon the sensitivity.
b) A better method is to arrange for the space behind the diaphragm to be so small that the movement of the diaphragm causes a perceptible compression of the enclosed air, thereby providing additional rigidity.

One might go even as far as to consider using a very thin and thus light diaphragm, and deriving the whole of the rigidity from the air pocket. This would, it is true, improve the sensitivity, but on other grounds it is less desirable: the high voltage between the diaphragm and the electrode constitutes an attractive force between the two, with the result that in the event of a small leak in the air chamber there would be a danger of the diaphragm being drawn back onto the electrode. Now it is essential to have such a small leak, in order to avoid a variation in the distance between the diaphragm and the electrode due to gradual barometric changes. We have therefore kept to a diaphragm 15 $\mu$ thick.

If the volume of air behind the diaphragm were to consist only of the approximately 13 $\mu$ thick layer between the diaphragm and the electrode, the air rigidity would be fantastically high, as is evident from what follows.

A movement of the diaphragm over a distance $\bar{x}$ then gives, with adiabatic compression, a rise $\Delta p$ of the pressure $p$ according to the equation:

$$\frac{\Delta p}{p} = -\frac{\Delta V}{V} = -\varkappa\frac{\bar{x}}{a}.$$

where $V$ is the volume of air and $\varkappa$ is the ratio of the specific heat of air under constant pressure to that for constant volume ($\varkappa = 1.4$). The force required (apart from the sign) is therefore:

$$\Delta p \cdot O = \bar{x}\frac{p \cdot \varkappa \cdot O}{a},$$

and the stiffness is

$$s_{\mathrm{aeq}} = \frac{p \cdot \varkappa \cdot O}{a}.$$

With $p = 1$ atm $= 10^6$ dyne/cm$^2$ and $a = 13 \cdot 10^{-4}$ cm we therefore get:

$$s_{\mathrm{aeq}} = 2 \cdot 10^9 \text{ dyne/cm},$$

corresponding to a resonance at 72 000 c/s.

Obviously, in order to reduce the air rigidity the volume of air must be increased, not by increasing the distance $a$ (this would greatly reduce again the generated voltage) but by providing a separate compression space in the form of an annular channel around the electrode or as grooves or cavities in the surface.

The choice of form of this separate compression space is closely related to the damping required for flattening out the resonance peak, because for this damping we make use of the viscous flow of the air in the narrow gap between the diaphragm and the electrode from and to the compression space. Assuming that the air during this flow behaves as if it were incompressible, it is not difficult to calculate the damping action.

In the differential equation now applying for the diaphragm movement,

$$m\ddot{x} + r\dot{x} + sx = K \sin \omega t,$$

we find for the coefficient $r$ in the case of a round flat electrode with radius $R$ and surface $O$:

$$r = \frac{1.5 \cdot 10^{-3}}{2\pi} \cdot \frac{OR^2}{a^3}.$$

The degree of damping can be judged by calculating $\omega_0 \cdot r$ ($\omega_0$ is the frequency at which resonance occurs). Resolution of the differential equation shows that $\varkappa = K/s$ for $\omega \ll \omega_0$ and $\varkappa = K/\omega_0 r$ for $\omega = \omega_0$. We therefore get the right damping approximately when

$$\omega_0 \cdot r = s.$$

In our case ($R = 0.6$ cm), when $\omega_0 = 2\pi \cdot 12\,000$ $^{c}/_{s}$ is chosen

$$\omega_0 \cdot r = 3.6 \cdot 10^9 \text{ dyne/cm}.$$

This is much too high a value, since the value found above as being required for the equivalent stiffness is $5.6 \cdot 10^7$ dyne/cm.

It is therefore necessary to reduce the resistance by a factor of 150. The formula given for the resistance shows that we can easily do this by dividing the electrode into parts with smaller dimensions. The resistance is then reduced (for the same total area) by the square of those dimensions, When it has been reduced to the desired value ($r = s/\omega_0$) the air is practically incompressible in its flow from and to the grooves in the divided electrode. The rises in pressure in those grooves then react upon the diaphragm and thus produce the additional stiffness required to raise the resonance.

We have divided up the electrode by milling straight grooves in its surface. This results in a somewhat different value for the damping constant, viz:

$$r = \frac{4.3 \cdot 10^{-3}}{2\pi} \cdot \frac{O\,(^1/_2\, b)^2}{a^3},$$

where $b$ is the width of the dam between two grooves.

Although this formula does not yield exact results it is very useful in practice as giving the right relation between the damping and the various dimensions. With the aid of this formula and one experimental model it is generally possible to determine the exact shape the rear electrode should have.

## The electric circuit

When we come to consider more closely the manner in which the alternating voltage is generated by the movement of the diaphragm we start from the simplified model: a diaphragm moving to and fro as a flat plane. The capacitance in the position of rest is $C_m = O/4\pi a$, whilst the variations are described by replacing $a$ by $a\,(1 + \alpha)$, where $\alpha = \bar{x}/a$ may, for instance, vary sinusoidally with time. Parallel to the capacitance of the diaphragm is an inevitable scattering capacitance and possibly also a parallel capacitance purposely introduced; these two together are termed $C_p$.

If we again assume a constant charge then

$$V = \frac{Q}{C_p + \dfrac{C_m}{1+a}} = \frac{Q}{C_m + C_p} \cdot \frac{1+a}{1 + a\,\dfrac{C_p}{C_m + C_p}} =$$

$$= V_0 \, \frac{1+a}{1 + a\,\dfrac{C_p}{C_m + C_p}}$$

invariably applies, in which $V_0$ represents the voltage in the position of rest.

The voltage variation is

$$\varDelta V = V - V_0 = V_0\, a \cdot \frac{C_m}{C_m + C_p} : \frac{1}{1 + a\,\dfrac{C_p}{C_m + C_p}} =$$

$$= V_0 \cdot a \cdot \frac{C_m}{C_m + C_p}\left(1 - a\,\frac{C_p}{C_m + C_p} + \ldots\right).$$

From the formula we see that:

1) the voltage generated is proportional to $V_0$ and in the first instance proportional to $a = \bar{x}/a$;
2) when a parallel capacitance is introduced the voltage generated is reduced in the ratio $C_m/(C_m + C_p)$;
3) when $C_p$ is introduced there also arises a distortion factor of the order of $a\ C_p/(C_m + C_p)$.

With very strong sound this distortion factor reaches at most a value of 1%, even with a very large $C_p/C_m$ (because we always have $a \ll 1\%$), so that it is negligible in comparsion with the distortion generally occurring in the remaining part of an apparatus for sound transmission.

So far we have been assuming that the microphone has a constant charge, presuning the charge resistance to be high enough. This point, too, has to be looked into more closely.

Without parallel capacitance and with constant charge the microphone produces an "open voltage" $= V_0 a$. When the microphone is connected to the direct voltage source without any resistor in between we get a "shorting current" $= V_0 \cdot \mathrm{d}C/\mathrm{d}t = V_0 \cdot a{\cdot}j\varphi C_m$ (when $a$ is taken to be sinusoidal and terms of a higher order are ignored). According to Thévenin's theorem we may consider the microphone as a voltage source of a strength $V_0 a$ with an internal impedance equal to the open voltage divided by the shorting current, thus $= 1/j\varphi C_m$, which is the specific impedance of the microphone. This is indicated diagrammatically in *fig. 2*. With the aid of this result let us consider further how the combination of the microphone with the parallel capacitances and the resistor behaves. In the diagram given in fig. 2 the alternating voltage across $R$ is

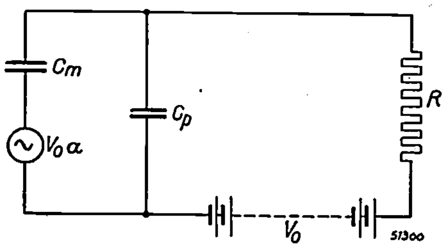


Fig. 2. The electrical circuiting of a condenser microphone in the form in which it might be regarded according to Thévenin's theorem. $V_0$ is the voltage of the current source, $R$ the resistor, $C_m$ the capacity of the condenser-microphone in the state of rest, $C_p$ the parallel capacity, and $a$ the factor determining the change in the distance between the diaphragm and the rear electrode.

$$V_R = a V_0 \, \frac{C_m}{C_m + C_p} \cdot \frac{1}{1 + \dfrac{1}{j\omega R\,(C_m + C_p)}} \,.$$

In *fig. 3* this voltage is plotted in its absolute value as a function of $RC_0\omega$, in which $C_0 = C_m + C_p$. We see that for $RC_0\omega \gg 1$ practically the full calculated voltage occurs at the ends of the resistor. At $RC_0\omega = 1$ the signal is reduced by 3 db.

If the lowest frequency to be reproduced is put at $v = 32$ c/s and we allow here a drop of 3 db then we must have

$$R\,(C_m + C_p) = \frac{1}{2\pi \cdot 32} \approx \frac{1}{200}\ \text{sec.}$$

This can be attained in various ways, the simplest of which is to choose, for the capacity of the microphone, such a resistance as will satisfy the condition set. This, however, leads to very large resistances (80 megohms, because $C_m = 55$ pF and $C_p = 10$

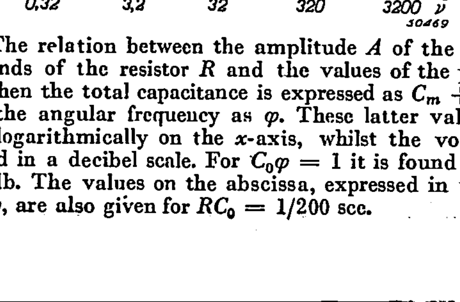


Fig. 3. The relation between the amplitude $A$ of the voltage at the ends of the resistor $R$ and the values of the product $RC_3\varphi$, when the total capacitance is expressed as $C_m + C_t = C_0$ and the angular frequency as $\varphi$. These latter values are plotted logarithmically on the $x$-axis, whilst the voltage is expressed in a decibel scale. For $C_0\varphi = 1$ it is found that $A = -3$ db. The values on the abscissa, expressed in the frequency $v$, are also given for $RC_0 = 1/200$ sec.

pF at least), and this makes heavy demands upon the insulation both in the microphone and on the grid of the first amplifying valve.

Another solution is to use a resistance not quite so high and to connect a capacitor in parallel to the microphone, but this causes the voltage to drop in the ratio $C_m/(C_m + C_p)$. It is true that a reduction in the resistance causes also a reduction in the noise voltage, but in this case the threshold value is adversely affected, whilst it has to be borne in mind that the noise voltage of the resistor is reduced only in proportion to the square root of the resistance ratio.

A. third solution could be found by applying a feed-back voltage in the manner of the much used "cathode-follower", which makes the circuit appear to have a high input resistance. Unfortunately it is found that the ratio of noise to signal is not proportionately reduced, so that this sysem does not offer any real advantage and has the drawback that the first valve does not contribute anything to the amplification.

We have calculated the ratio of the noise voltage to the signal voltage for two cases.

If the resistance $R$ is high (80 megohm) and the parallel capacitance is a minimum (about 10 pF) we find for the signal voltage 3 mV/μbar, A calculation of the r.m.s. value of the noise voltage gives as result 6.0 μV, so that this lies 54 db below the signal voltage at 1 μbar. If there were an interference voltage of this value with a frequency of 1000 c/s this would lead to a threshold value of 20 phons. Taking into account the sensitivity curve of the ear, it is evident that the noise voltage mentioned results in a threshold value lower than 20 phons.

In the second case let $R = 5$ megohm and $C_p = 1000$ pF. The signal voltage is then 0.15 mV per μbar. Under these conditions the r.m.s. value of the noise voltage lies 40 db lower than the signal voltage per μbar, namely at 1.4 μV. With a frequency of 1000 c/s of the noise tone the threshold value would then lie at 34 phons; actually it is lower than this.

From this it follows that a satisfactory threshold value can be reached if a sufficiently large resistor is used.

## The practical construction

It has already been shown that a combination of an extended horizontal frequency characteristic and a low threshold value can only be obtained by using a diaphragm as light as can be allowed on the aforementioned grounds and with the smallest possible distance between the electrode and the diaphragm as is practicable for manufacturing with a sufficient degree of uniformity. This distance must be kept as uniform as possibl- because the damping varies according to the third power of that distance; with distances of the order of 10 μ an increase of 1 μ causes a variation in the damping of more than 2 db, which manifests itself in an equally large elevation of the characteristic at the point of resonance. We have not, therefore, attempted to make the distance smaller than that mentioned, the more so because that would make it still more difficult to keep the microphone free of short-circuiting in consequence of extremely small foreign particles.

*Fig. 4* gives the frequency characteristics of three microphones with the same electrode but different distances between the diaphragm and the electrode. It appears that this distance is of great importance for the damping and thus for the shape of the characteristic. Graph *a* shows that the damping is too strong, and graph *c* that there is too little damping.

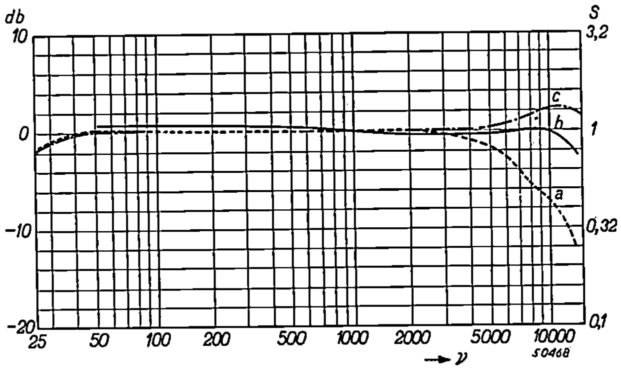


Fig. 4. Three characteristics showing that the damping can be regulated by adjusting the distance between the diaphragm and the rear electrode:
*a*) too strong damping, *b*) the right damping, *c*) too little damping. On the horizontal axis is the frequency to a logarithmic scale, whilst vertically on the right we have

$$S = \frac{\text{voltage at the frequrncy } v}{\text{voltage at 1000 c/s}}$$

and on the left the same voltage ratio in deçibels.

*Fig. 5* is a cross-sectional drawing of the condenser microphone as constructed in the Philips laboratory after consideration of the factors dealt with in the foregoing.

The diaphragm is stretched tight in the microphone casing between two rings pressed tightly up against each other by means of a nut. The electrode is pressed into another ring provided with an insulation of amber. The whole is slipped into the microphone and by means of a special nut the electrode ring is forced up until it presses out the diaphragm to get the required tension. It is previously arranged that the surface of the electrode comes to lie 10 μ lower than the rim of the ring into which it is pressed. Finally the electrode is connected to an insulated plug pin.

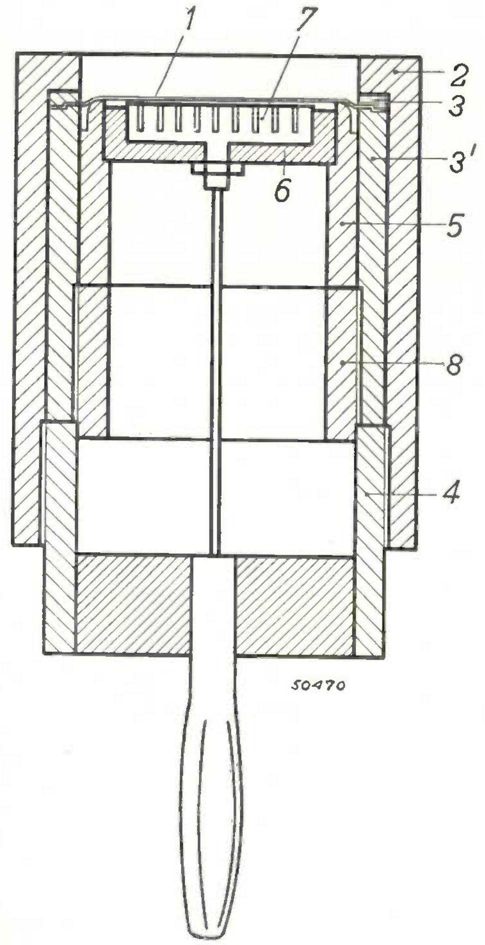


Fig. 5. Cross-sectional drawing of the new design of condenser-microphone, The diaphragm *1* is stretched in the casing *2* of the microphone between the rings *3* and *3'*, which are presesd together by the nut *4*. The rear electrode *7* is pressed into a ring *5* containing an amber insulation *6*. The whole is slipped into the microphone and pressed down until the ring *5* presses through the diaphragm far enough to give the required tension, this being adjusted by means of the nut *8*. It is previously arranged that the surface *7* lies 10 μ below the edge of *5*. The rear electrode is connected to an insulated plug pin.

The following details are given regarding the construction of the electrode system. The electrode proper, which is 12 mm in diameter, is first provided with the required number of grooves (15 grooves 1.8 mm deep and 0.15 mm wide), then placed in the amber ring, after which the whole is pressed into the holder ring; the top edge of the amber ring is then about 0.1 mm below the upper surface of the electrode. The surface of the electrode is then turned on a precision lathe until it comes to lie on a level 10 μ lower than the edge of the holder ring.

This last process has to be done very carefully. If it were done by hand the chisel would make the surface too rough, with the risk of short-circuiting. The chisel has to be driven, with a very small pitch, by a special motor. The displacement of the chisel parallel to the bed over a distance of 10 μ is usually done by hand and is read from a dial gauge mounted on the support. When the surface has been turned any small burrs are rubbed off with a chamois-leather.

The lateral movement of the chisel must of course be absolutely perpendicular to the axis of the lathe, as otherwise the surface would be conical and give rise to short-circuiting or else make the air-gap too large. As a check for accuracy a turned surface was placed underneath an interferometer plate. The tolerance was found to be less than 1 μ over a distance of 15 mm.

*Fig. 6* illustrates some specimens of the condenser-microphone.

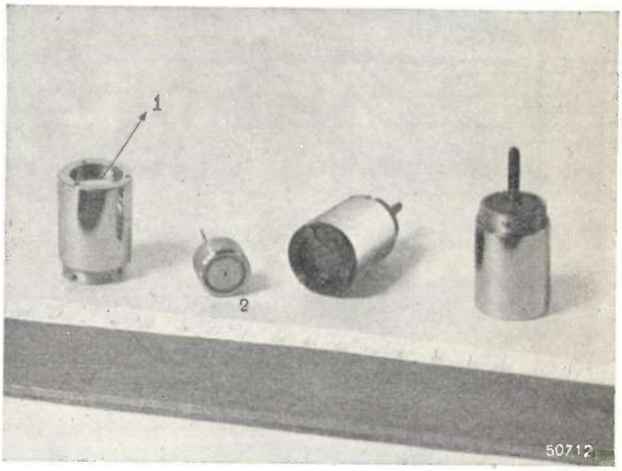


Fig. 6. Some condenser-microphones. The instrument on the left contains the diaphragm *1* stretched in the casing, and beside that is a rear electrode *2*. The microphone lying on its side has a protective gauze over the diaphragm. The ruler gives an idea of the dimensions.

A good idea of the mechanical accuracy reached can be formed when imagining the microphone to be enlarged 1000 times. The electrode then has a diameter of 12 meters, whilst parallel to it at a distance of 10 mm is a plate 15 mm thick. Capacity measurements have shown that this plate is not absolutely flat but somewhat convex. If the model just described is to answer fully the actual requirements then the distance between the electrode and the plate should be in the middle about 13 mm. This convex shape of the diaphragm is undoubtedly obtained from the pressing in of the duralumin foil when mounting the electrode in the microphone.

A good idea of the vibration of the diaphragm is obtained from a calculation of the velocity. With a sound pressure of 1 μbar and a frequency of 50 c/s the maximum velocity is:

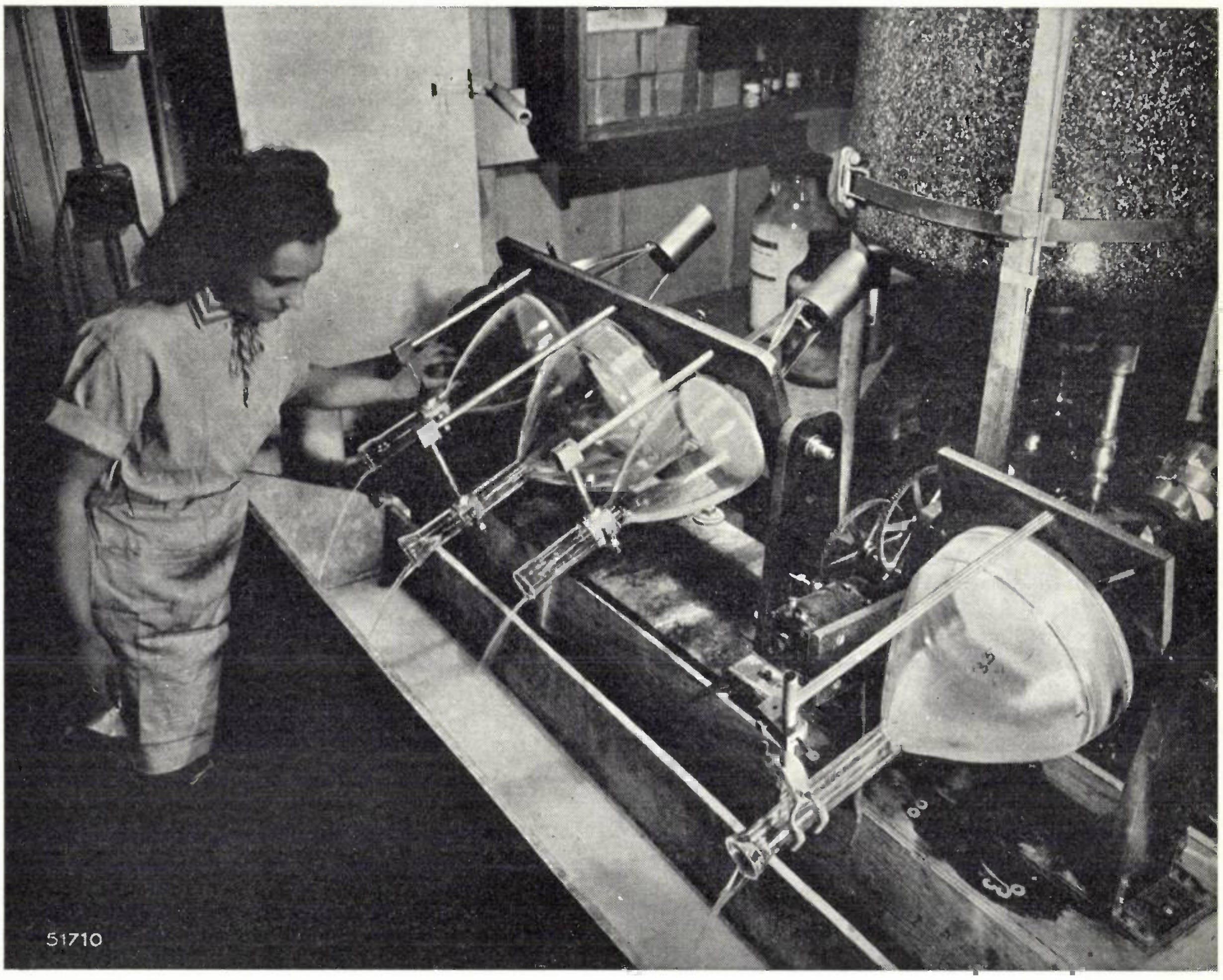
$$v_{\max} = 2\pi \cdot 50 \sqrt{2 \cdot 4.5 \cdot 10^{-4}}\ \mu/\text{sec} = 1.72\ \text{cm/day}.$$

By way of comparison it may be mentioned that a sunflower grows 2-3 cm per day.

The condenser-microphone described here is being used in the Philips laboratory in the dummy head for experiments in stereophony. It is not precluded that this instrument will find application also for the monaural reception of sound when the desirability of extending the frequency range beyond $10^4$ c/s becomes urgent. The objection of the first amplifier having to be set up in the immediate vicinity of the microphone will then have to be overcome in some way or other [7]). The graph in fig. 4 already shows that the quality of reproduction satisfies high demands. Practice has proved that this design can be made with sufficient uniformity for the microphones to be safely interchangeable.

[7]) A solution of this problem will be given in the article referred to in footnote [5]).

# TELEVISION TUBE PRODUCTION



The phosphor coating constituting the fluorescent screen in television tubes is deposited on the face of the tube by settling from a dispersing liquid. The remaining liquid must be removed in such a way as to avoid even the slightest turbulence that would tend to make the liquid wash away parts of the screen. In the Dobbs Ferry, N.Y., plant of North American Philips Co., Inc., the tubes are emptied by tilting them at a slow, uniform rate by means of a carefully balanced machine operated by a small electric motor and gear system.

# AN ELECTRONIC SWITCH WITH VARIABLE COMMUTATING FREQUENCY

by E. E. CARPENTIER.

621.317.755.06

A description is given of a new construction of an electronic switch (type number GM 4580) for the simultaneous production of two oscillograms. Various steps have been taken to avoid disturbing peaks in the output voltage. The normal oscillograph amplifier, which led to distortion of the commutating voltage, is no longer used (except as pre-amplifier for one of the voltages to be imaged). As a result it has been possible to raise the commutating frequency to 40 000 c/s, thereby improving the clearness of the details in the oscillograms. Moreover, this frequency is continuously variable, to 2 c/s downwards, which in certain cases is of advantage. The input may be asymmetrical or symmetrical as desired, the latter being useful, *inter alia*, when it is desired to produce more than two oscillograms simultaneously with the aid of electronic switches in cascade.

With the aid of a cathode-ray oscillograph [1]) the trend of a voltage can be pictured as function of another voltage. If for the latter a linear time-base voltage is used then the screen of the oscillograph shows the trend of a voltage as a function of the time. Although in many cases it is not required, it is often desirable to produce an image of two or more voltages simultaneously as functions of some other voltage, for instance when it is required to record or measure a difference in time or phase.

In a method already frequently described in literature — also in this journal [2]) — a switch is used which connects the two voltages to be oscillographed alternately to the plates for the vertical deflection in a normal oscillograph; here the switching is effected by means of electronic valves.

## Principle of the electronic switch

The principle of such an electronic switch is recalled with reference to *fig. 1*, relating to an earlier design of such an apparatus (*cf.* footnote 2). The voltages to be oscillographed $v_I$ and $v_{II}$ are conducted to the channels $I$ and $II$ of the electronic switch, the inputs of which are connected to the control grids of the pentodes $P_I$ and $P_{II}$ respectively. The two screen grids of these pentodes are connected to mutually phase-opposed alternating voltages having an approximately rectangular curve. These so-called commutating voltages are generated by a multivibrator on the Abraham and Bloch principle: their amplitudes are such that alternately one of the pentodes is in the normal working state while the other is dead. The current in the common anode resistor $R_a$ thus flows alternately through $P_I$ or $P_{II}$; consequently this current — and thus also the anode voltage — is alternately governed by $v_I$ or $v_{II}$. The anode voltage governs in turn (*via* an amplifier) the vertical deflection of the oscillograph.

The time during which the commutating voltage oscillates is chosen small in respect to the time taken by the spot of light to traverse the screen from left to right. A picture is then produced of the nature of that shown in *fig. 2* (where for the sake of clarity the commutating frequency has been chosen much too low). Each of the image curves is thus a dotted line.

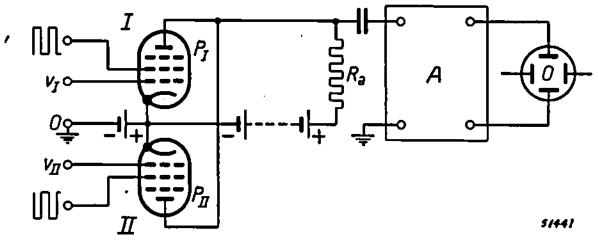


Fig. 1. Circuiting diagram of an electronic switch. $I$ and $II$ are channels functioning alternately for carrying the input voltages $v_I$ and $v_{II}$ to the amplifier $A$ and the oscillograph tube $O$. $P_I$ and $P_{II}$ are pentodes, $R_a$ is a common anode resistance.

By another and perhaps more obvious method of producing two or more oscillograms simultaneously an oscillograph tube is used in which a number of electron beams each describe an oscillogram on the screen. The tube contains the same number of cathodes — each with its appropriate focusing, accelerating and deflecting system — as the number of oscillograms it is desired to record simultaneously. The fact that in this way continuous lines are produced instead of dotted lines (fig. 2) has the advantage, *inter alia*, that no details are lost in the oscillograms and the luminous intensity obtained from one beam of electrons is not distributed over a number of curves. There are, however, also some objections, such as the unavoidable differences between the deflecting systems (small differences in sensitivity, for instance, may be very troublesome). Further, especially at high frequencies, there will

[1]) For descriptions of cathode-ray oscillographs see inter alia: Philips Techn. Rev. 1, 147-151, 1936, (type GM 3150), Philips Techn. Rev. 4, 198-204, 1939, (type GM 3152), Philips Techn. Rev. 5, 277-285, 1940, (type GM 3156), Philips Techn. Rev. 9, 202-210, 1947, No. 7 (type GM 3159).

[2]) C. Dorsman and S. L. de Bruin, An Electron Switch, Philips Techn.Rev. 4, 267-271, 1939.

be mutual reactance between the deflecting systems. But it is mainly the complexity of the construction that has prevented these multiple tubes from becoming more generally employed.

### *Upper limit of the commutating frequency*

The higher the commutating frequency is chosen, the finer is the structure of the image, so that smaller details can be reproduced. There is, however, a limit to the raising of this frequency, at least when the output voltage of the electronic switch is so small as not to be able to dispense with the amplifier of the oscillograph, for it is then necessary that this amplifier should amplify a rectangular oscillation without any distortion; this means that it must produce a large number (*e.g.* 100) of harmonics with reasonably faithful amplitude and phase. With the amplifier of a normal oscillograph, for instance of the type GM 3152, one arrives at a maximum commutating frequency of 8 000-10 000 c/s, which value is therefore used in the electronic switch (type GM 4196) described at the time in the article referred to in footnote 2.

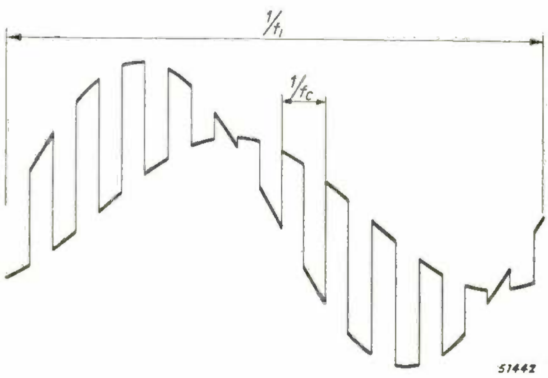


Fig. 2. Picture of two sinusoidal input voltages as it appears on the oscillograph with the circuiting according to fig. 1 when the commutating frequency $f_c$ is higher than the frequency $f_i$ of the input voltages. In reality $f_c$ is much higher than is represented here.

The fact that the shape of the curve of the commutating voltage obtained with this apparatus after amplification still leaves much to be desired is to be seen from *fig. 3a*; here the two input voltages were zero. This is manifest in the oscillograms produced with such a commutating voltage, as seen for instance in fig. 3*b*; the peaks visible in fig. 3*a* fog the image.

In the designing of the new electronic switch (type GM 4580) to be described below the object has been to improve the quality of the oscillograms as far as possible, for instance by giving the amplified commutating voltage a shape more closely approaching a rectangle. *Fig. 4a* shows the oscillogram of this voltage and fig. 4*b* an oscillogram produced with it. A comparison of fig. 4*b* and fig. 3*b* clearly shows that the troublesome fogging has been eliminated.

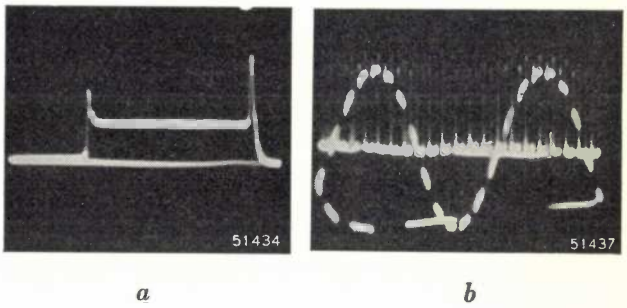


Fig. 3. *a*) Oscillogram of the output voltage of the amplifier of the oscillograph GM 3152 with the old electronic switch GM 4196 connected in front of it, the two input voltages of the latter being zero. The voltage obtained (frequency about 8000 c/s) deviates rather considerably from the rectangular shape, this being due, *inter alia*, to a peak arising at each commutation.

*b*) Oscillogram of a voltage with a frequency of about 600 c/s and a zero voltage recorded with the combination of GM 4196 and GM 3152. The peaks visible in (*a*) cause a troublesome fogging of the picture.

### Improved electronic switch

The main points of difference between the electronic switch GM 4580 illustrated in *fig. 5* and its predecessor are the following:

I) The amplified commutating voltage (*i.e.* the voltage on the oscillograph plates when both input voltages of the electronic switch are equal direct voltages) no longer has the disturbing peaks previously referred to. Distortion in the oscillograph amplifier is avoided by dispensing with the use of that amplifier (except as pre-amplifier for one of the voltages to be imaged, to which we shall revert later).

II) The commutating frequency has been made variable. On the one hand it can be raised to 40 000 c/s, so that four times as small details

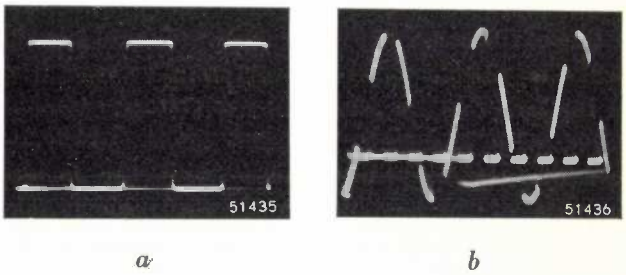


Fig. 4. *a*) Oscillogram of the output voltage of the new electronic switch GM 4580 (both input voltages zero); here the frequency was again about 8000 c/s. The rectangular shape is much more closely approximated than in fig. 3*a*.

*b*) Oscillogram of a voltage with a frequency of about 2000 c/s and a zero voltage recorded with the combination GM 4580/GM 3152.

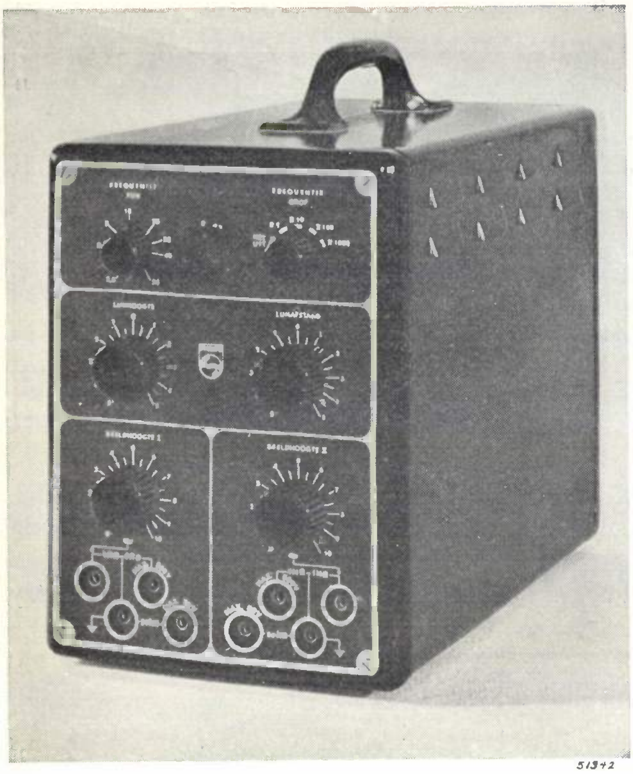


Fig. 5. The new electronic switch type GM 4580. From top to bottom: two knobs for regulating the commutating frequency, two knobs for vertical displacement of the image (left-hand one for shifting the whole image, right-hand one for shifting the two oscillograms in opposite directions), two knobs for regulating the sensitivity of the channels, terminal sockets for the symmetrical and the asymmetrical inputs.

can be made visible, whilst on the other hand it can be adjusted to extremely low values, which may be advantageous in some cases.

III) The inputs of the electronic switch can be made asymmetrical or symmetrical as desired. The latter is useful, for instance, for the simultaneous production of more than two oscillograms.

These points will be discussed further below.

### I) *Avoidance of troublesome peaks and distortion*

One of the causes of the occurrence of troublesome peaks in the amplified commutating voltage lies in the fact that the multivibrator voltage shows a curve with somewhat concave flanks, which is consequently likewise the case with the anode currents of the pentodes in the electronic switch (*figs. 6a* and *b*); this has already been mentioned in the article quoted in footnote 2. For every commutation this gives rise to a drop in the current flowing through the common anode resistor (fig. 6*c*) and a corresponding peak in the anode voltage (fig. 6*d*). We have already seen what these peaks are like in figs. 3*a* and *b*.

By applying a push-pull circuit in both channels of the electronic switch these peaks in the output voltage are avoided. The manner in which the push-pull circuit may be arranged is shown for one channel in *fig. 7* (actually the circuit is somewhat different, as will be seen farther on). In order to obtain an output voltage symmetrical with respect to earth in spite of the fact that the input voltage $v_I$ is asymmetrical, use can be made of the method previously described [3]) where a resistor $R_{kI}$ (fig. 7) is introduced in the common cathode line. As a result when $SR_{kI} \gg 1$ the output voltage is practically symmetrical ($S$ = slope of the valves).

In *fig. 8* this push-pull circuit is applied to two channels. It is also indicated how in each channel the two pentodes in push-pull can be replaced by one double pentode (type EFF 51). The circuit is shown for asymmetrical inputs, but in reality the new electronic switch has also symmetrical inputs; we shall revert to this later.

It is easily understood that with push-pull circuits the peaks previously referred to will no longer occur, for from fig. 8 it may be seen that if the two input voltages are zero then the output voltage $v_o$ amounts to:

$$v_o = (i_I' + i_{II}')\,R_a' - (i_I'' + i_{II}'')\,R_a'',$$

in which $i_I'$ . . . $i_{II}''$ represent the anode currents

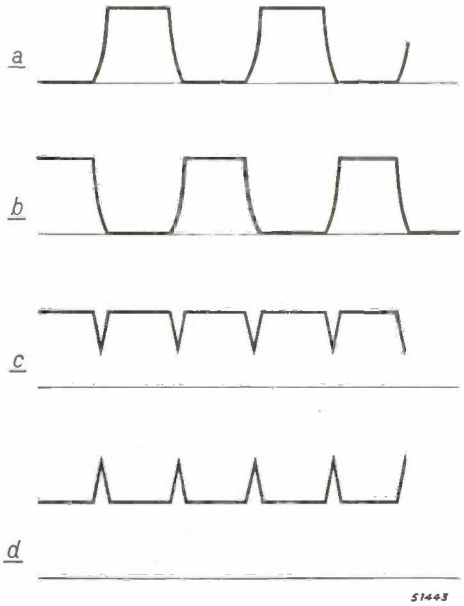


Fig. 6. Trend of the two anode currents ($a$ and $b$), the current ($c$) through the resistance $R_a$ and the anode voltage ($d$) in the circuit according to fig. 1, with equal and constant values of $v_I$ and $v_{II}$.

[3]) See the last of the articles listed in footnote 1, in particular fig. 3*b*.

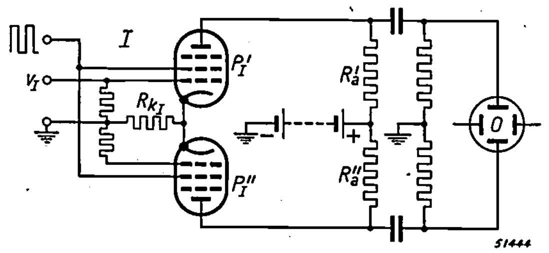


Fig. 7. Here the channel *I* of fig. 1 is in push-pull (pentodes $P_I'$ and $P_I''$). $R_{kI}$ is a common cathode resistance ensuring symmetrical action.

of the four pentodes $P_I' \ldots P_{II}''$, and $R_a'$ and $R_a''$ represent the two anode resistors. On account of the approximate symmetry of the push-pull circuits we have at any time $i_I' \approx i_I''$ and $i_{II}' \approx i_{II}''$, so that if $R_a' = R_a''$ the output voltage is indeed approximately zero.

Another cause of undesired peak voltages lies in the capacitive effect exercised upon the control grid by the screen grid to which, in the circuiting described above, a commutating voltage is applied. Every jump in the screen grid voltage results in a current impulse in the circuit formed by the capacitance $C_{sg}$ between the screen and control grids (*fig. 9*) and the impedance $Z_g$ between control grid and cathode. This current impulse gives risc in turn to a voltage impulse on the control grid. When the screen grid voltage jumps from a positive value to zero the corresponding peak on the control grid is of little consequence, since the anode current disappears at the same moment (owing to the screen grid voltage dropping to zero). A jump in the reverse direction, however, is indeed reflected in the output voltage.

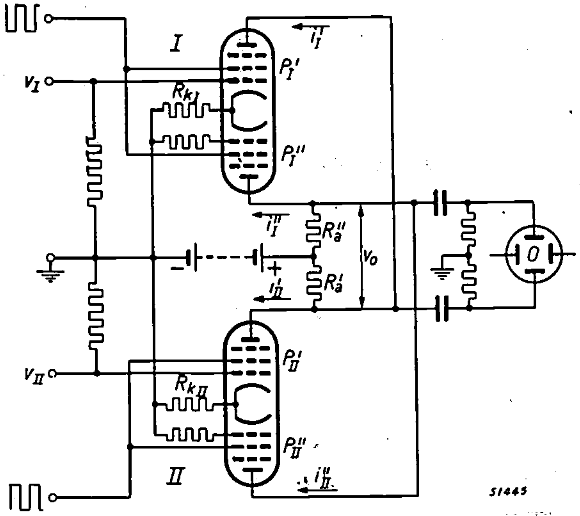


Fig 8. Both channels of fig. 1 in push-pull as indicated in fig. 7 for channel *I*. The pentodes $P_I'$ and $P_I''$, respectively $P_{II}'$ and $P_{II}''$, can be replaced by a double pentode (type EFF 52).

In this case a push-pull circuit as illustrated in fig. 7 cannot alter matters. The amplitude of the disturbing voltage on the control grid depends upon the impedance $Z_g$; in the control grid circuit of $P_I''$ this has a fixed value, but in the case of $P_I'$ it is formed partly by an external impedance (that of the source of the input voltage), which may have any value. Consequently there is not as a rule any symmetry such as is obtained in the previous case, with the result that even with push-pull circuits peaks of the kind in question here may still arise.

In the electronic switch of the type GM 4580 this phenomenon is greatly reduced owing to the fact that the commutating voltage, instead of being conducted to the screen grid circuit, is applied to the control grid circuits, parallel to $R_k$ (*fig. 10*). Owing to the presence of the capacitance $C_{gk}$ between control grid and cathode the effect does, it is true, still arise in principle; even more so, since $C_{gk}$ is about twice as large as $C_{sg}$. But this is amply compensated by the fact that a much smaller amplitude of the commutating voltage is needed on the control grid than on the screen grid, namely 10 V instead of 250 V. As a result the peaks in question become 10 times smaller and are therefore no longer troublesome.

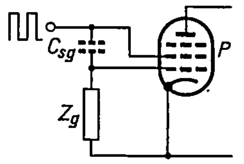


Fig. 9. The commutating voltage on the screen grid of the pentode *P* causes, *via* the capacitor $C_{sg}$ between the two grids a voltage impulse on the control grid. $Z_g$ = the total impedance between control grid and cathode.

This low value of the commutating voltage has yet another favourable effect, in that the time required for this voltage to rise from the minimum to the maximum value, or *vice versa*, can be kept short. (During these intervals of time the cathode ray does not produce any effective light on the screen, and, moreover, any details of the input voltages happening to occur just in those intervals do not become visible.) The reason why this commutating time is not zero is to be sought mainly in stray capacitances in the multivibrator exciting the commutating voltage. The adverse influence of these capacitances is all the less according as the resistances to which they are connected in parallel have a lower value. The smaller the commutating voltage required, the lower these resistances may be chosen. The fact, therefore, that in the new electronic switch only 10 V suffices (where formely 250 V was needed) leads to a much shorter commutating time. We have gone a step farther in this direction by choosing relatively high anode currents in the multivibrator, thereby lowering the resistances still further. (The multivibrator is equipped with output pentodes of the type EBL 21.)

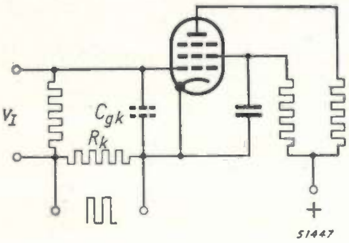


Fig. 10. The commutating voltage is conducted to the resistor $R_k$ in the control grid circuit; the screen grid is fed in the normal manner with a direct voltage. $C_{gk}$ = capacitance between control grid and cathode.

Another important difference between the old electronic switch and the new one is the fact that with the latter no use is made of the normal oscillograph amplifier between the electronic switch and the oscillograph tube, thus avoiding a serious source of distortion of the commutating voltage. In order to get an oscillogram of reasonable height also with relatively small input voltages, it has been arranged for the electronic switch itself to act as an amplifier as well. The amplification cannot, it is true, be raised very high without giving rise to the same difficulties as occur with the normal oscillograph amplifier, but with the amplifying factor of 100 applied here no trouble is experienced in this connection, whilst for many applications the sensitivity is still sufficient. Moreover, there is no objection to the oscillograph amplifier being used as pre-amplifier, placing it in front of one of the channels of the electronic switch, so that a very great sensitivity is available at least on one channel [4]).

The fact that the oscillograph plates for the vertical deflection are now connected directly to the output of the electronic switch constitutes a second argument for applying the push-pull principle in this switch, as otherwise defocusing of the electron beam and distortion of the image would occur [5]).

II) *Variable commutating frequency*

With the dispensing of the amplifier between the electronic switch and the oscillograph tube we have removed the greatest obstacle preventing the commutating frequency $f_c$ from being raised. In the new electronic switch $f_c$ has been made variable — the use of which will be explained presently — with 40 000 c/s as the upper limit, so that the structure of the image obtained is four times finer than was the case with the old type. Furthermore, $f_c$ can be regulated to very low values, it being desirable in certain cases to have a commutating frequency which is low in comparison with the frequency $f_i$ of the voltage to be screened (*fig. 11*). The screen then first shows for a time the complete oscillogram of one voltage, followed by that of the other voltage, then again the first, and so on. There is, however, a limit to which the commutating frequency $f_c$ can be reduced, at least for the visual recording of the oscillograms, because if the transition is too slow there is a troublesome flickering of the image. For this reason $f_c$ should not be chosen lower than about 50 c/s. When the voltage frequency $f_i$ is 1000 c/s or more then as a rule a slow commutation ($f_c \ll f_i$) will yield better results than rapid commutation ($f_c \gg f_i$). If, on the other hand, $f_i$ is lower than about 200 c/s, in order to avoid flickering one must have rapid commutation. In between these limits it does not matter much whether one chooses $f_c \gg$ or $\ll f_i$ [6]).

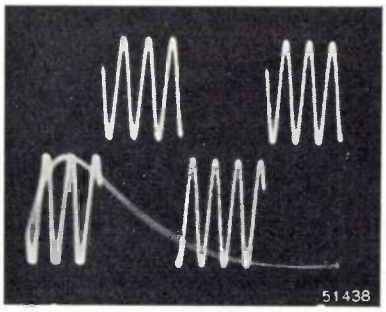


Fig. 11. Oscillograms recorded with a low commutating frequency ($f_c < f_i$). The frequency of the time-base voltage has been chosen low enough for the image to cover two cycles of the commutating frequency.

When the image is photographed this flickering is of no consequence (at least if the exposure time is greater than $1/f_c$; each of the two oscillograms requires the time $^1/_2$ $f_c$ to be described on the

[4]) If it is desired to have two channels of high sensitivity a separate pre-amplifier, *e.g.* type GM 8002 or GM 4570, can be used for the second channel.

[5]) See, for instance, the second article listed in footnote 1, in particular fig. 3.

[6]) The said limits of $f_i$ depend upon the frequency of the time-base. It is assumed that the latter is 1 to $^1/_4$ times $f_i$, thus that the oscillogram covers 1 to 4 cycles of the input voltage.

screen). In that case one can work with low values of $f_c$ even when $f_i$ is lower than 200 c/s.

So far we have been assuming that both the input voltages are periodic. Obviously, in cases where one of the two (or both) occurs only once, $f_c$ must be chosen sufficiently high for the switching to take place a number of times within the observation period (with such a frequency that the details to be investigated can be seen).

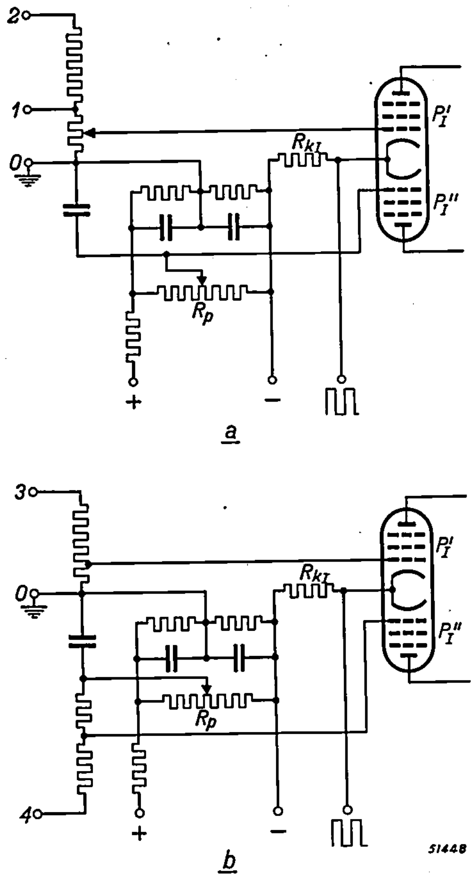


Fig. 12. The input circuit of each of the channels in the electronic switch GM 4580: *a*) for asymmetrical input voltage, *b*) for symmetrical input voltage. *0-1* input terminals for max. 50 V, *0-2* input terminals for max. 500 V, *3-4* input terminals for max. 2 × 500 V. $R_p$ is a potentiometer by means of which the oscillograms can be moved up and down.

*Table I* gives a summary of the cases which may arise. To make it possible to select the optimum commutating frequency for each of these cases, with the new electronic switch this frequency has been made variable between wide limits, *viz.* from 2 c/s to 40 000 c/s.

This range is covered by varying the grid condensers in a multivibrator in a number of stages, combined with a continuous regulation of the grid resistance. In this way it is always possible to choose a value of $f_c$ which is not exactly a multiple of the time-base frequency. The dots which go to make up the oscillograms when the commutation is rapid will then move along the curves to be described and thus traverse the whole of the oscillogram, including all details. The speed at which the dots move depends upon the value of $f_c$; by adjusting $f_c$ so as to cause these dots to move rapidly one gets more the illusion of continuous curves.

**Table I.**

This indicates how the commutating frequency $f_c$ can best be chosen in different cases. $f_i$ = frequency of the input voltage in cycles per second, $T$ = observation time in the case of non-periodic phenomena, expressed in seconds.

| Input voltage | $f_i$ | Recording | |
|---|---|---|---|
| | | visual | photographic |
| | | $f_c/f_i$ | $f_c/f_i$ |
| periodic | < abt. 200<br>abt. 200-1000<br>> abt. 1000 | $\gg 1$<br>$\gg 1$ of $\ll 1$<br>$\ll 1$ | $\gg 1$ or $\ll 1$<br>$\gg 1$ or $\ll 1$<br>$\ll 1$ |
| | | $f_cT$ | $f_cT$ |
| non-periodic | — | $\gg 1$ | $\gg 1$ |

### III) *Input circuits of the electronic switch*

The electronic switch GM 4580 is provided with symmetrical as well as asymmetrical inputs.

The asymmetrical input, illustrated in *fig. 12a* for one channel, is used when it is desired to oscillograph a voltage with respect to earth. According to the size of this voltage it is connected between the terminals *1* and *0* or between *2* and *0* (maximum values respectively 50 and 500 V). By means of a potentiometer $R_p$ a variable and polar-reversible direct voltage can be taken up in the control grid circuit of $P_I''$. Thanks to the commutation this direct voltage corresponds to a rectangular alternating voltage on the output terminals. As a consequence the oscillogram of the corresponding

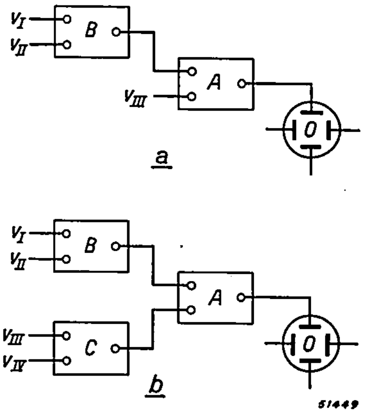


Fig. 13. Block diagram for the simultaneous recording of (*a*) three or (*b*) four oscillograms. *A*, *B*, *C* = electronic switches, *0* = oscillograph tube, $v_I \ldots v_{IV}$ = input voltages.

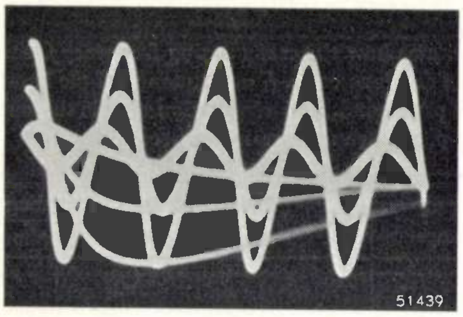


Fig. 14. Example of three oscillograms produced simultaneously.

input voltage moves upwards or downwards according to the polarity of the direct voltage. The potentiometers of the two channels are controlled with one knob and are coupled in such a way that when this knob is turned the two oscillograms move in opposite directions, so that they can easily be brought into a position that is most suitable for the observation.

The symmetrical input circuit is illustrated in fig. 12*b*. Here again we have the displacement device just described. We shall mention here two exceptional cases where the symmetrical input is useful on account of the symmetrical output of the preceding apparatus. In the first place such a case occurs when the oscillograph amplifier is used as pre-amplifier for one of the channels, as has been mentioned above. In the second place such a case is met with when it is desired to record more than two oscillograms together, by connecting electronic switches in cascade. The input of the electronic switch connected with the output of the preceding switch must then be symmetrical, because the said output is symmetrical.

This latter case is further illustrated in *figs. 13a* and *b*. According to these principles three oscillograms can be produced with the aid of two electronic switches, or four oscillograms with three such switches. An example is shown in *fig. 14.*

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

Reprints of the majority of these papers can by obtained on application to the Administration of the Reseach Laboratory, Kastanjelaan, Eindhoven, Netherlands. Those papers of which no reprints are available in sufficient number are marked with an asterik.

**1718:** J. M. Stevels: The physical properties of glasses. V. A peculiar phenomenon in the vitreous system $Na_2O$-$B_2O_3$
VI. The relationship between the volume and the number of glass-forming ions in silicate glasses.
VII. The molecular refraction of glasses.
(J. Soc. Glass Technology, Trans **30**, 303-317, 1946 Aug.-Oct.).

V. The density-concentration curve of sodium borate glasses is determined and the S-shaped curve found is discussed. At a concentration of 4 per cent $Na_2O$ a widening of the oxygen network is found which is not connected essentially with the sudden increase of the electric conductivity at this composition.

VI. It is shown that, at least in silicate glasses, there is no perceptible difference between the volumes of bridging and non-bridging oxygen ions. The volume of the interstices in the oxygen network is proportional to $R^{3/2}$ ($R$ being the ratio of the number of oxygen ions to the number of glass-forming ions) as is deduced from a simple model of the glass structure. Using these two results the volume of a block of glass containing one gram-atom of oxygen is calculated as a function of $R$. The values obtained are identical, within the limits of errors, with those derived from a formula obtained previously.

VII. It is shown that the molecular refraction of abnormal glasses can be calculated in an additive way, with the aid of atomic (ionic) refractions, provided that distinction is made between bridging and non-bridging oxygen ions. This additivity is not expected for the normal glasses. These expectations are found to be realised for a large number of pure silicate glasses containing $Na^+$, $K^+$ and $Ca^{++}$ ions only. The deviations between experimental and calculated molecular refractions for the abnormal glasses are less than 0.3 per cent. Glasses containing isolated oxygen tetrahedra are not examined.

**1748:** J. Hoekstra and C. P. Fritzius: De hechting van verf en vernis. (De Verfkroniek **20**, 195-198, 1947 Aug.) (The adhesion of paints and varnishes, in Dutch)

Paint adhesion may be defined as the energy necessary for lifting a unit surface of paint from its base. As it is practically impossible to mcasure this encrgy, a method is looked for in which a paint film is drawn from the base under specified conditions while the applied force is measured. The authors, choosing a shearing force rather than a force perpendicular to the film, describe a method in which the upper surface of two metal blocks with a metal strip between them is painted. After unscrewing, the strip is drawn off parallel with the paint layer (see also Philips Techn. Rev. **8**, 147-148, 1946).

**1749:** J. F. H. Custers: Diffusion of water into a polymer (J. Polymer Sci. **2**, 301-305. 1947)

In contradistinction to usual methods, the diffusion of water into a polymer has been investigated by starting from a disc of the material surrounded by water. In this way actual service conditions are often closely approached. Since a theoretical basis is desirable, equations are given for this type of diffusion and experiments with a wooden floor containing phenol-formaldehyde are described from which values for the permeation constant could be derived.

**1750*:** C. J. Bouwkamp and N. G. de Bruijn: The electrostatic field of a point charge inside a cylinder in connection with wave guide theory (J. Appl. Phys. **18**, 562-577, 1947 no. 6).

The field of a point charge inside a hollow, infinitely long, circular cylinder is studied. The case of an axial point charge is treated in detail. The surface density function is obtained as the solution of a Fourier type of integral equation. Then the potential caused by these charges is obtained. A second method works in the opposite direction. Here the potential is obtained as solution of a boundary value problem. A third method is based on the theory of Fourier-Bessel-Dini series. The potential is developed into discrete normal solutions of the potential equations in cylindrical coordinates. It is emphasized that the study of the potential problem can serve as a guide in questions of wave propagation in hollow circular cylinders. In this connection the third method is shown to be extremely useful. This is demonstrated in the case of acoustic waves inside a cylinder caused by a harmonically vibrating point source.

**1751:** N. Warmoltz: Variational principles in the theory of the cathode fall of a glow discharge (Physica **13**, 479, 1947 No. 8).

Attention is drawn to a recent remark of Seeliger (Phys. Z. **45**, 141, 1944) who criticizes Rogowski's theory of the cathode fall. One has to demand that a functoin, taking into account the formation (ionisation) and the transport of the ions as well, attains an extreme value for the real field ditribution. By following a procedure given by Seeliger it is possible to find an infinite number of these functions, but it seems difficult to select a function permitting a simple physical interpretation.

**1752:** C. J. Bouwkamp: A new method for computing the energy of interaction between two spheres under a general law of force (Physica **13**, 501-507, 1947 No. 8).

A new method is communicated, based on the theory of Bessel functions, for calculating the energy of interaction of two homogeneous spheres, on the assumption that $V(r)$ is the mutual potential energy of two unit point masses a distance $r$ apart. Particularly, when $V(r)$ is proportional to the $n^{\text{th}}$ power of $1/r$ an early result of Bradley is obtained. For integral values of $n$, subject to $1 < n < 8$ a limiting process leads to logarithmical terms.

**1753:** K. F. Niessen: On a cavity resonator of high quality for the fundamental frequency (Appl. Sci. Res. **B 1**, 18-34, 1947 No. 1)

The quality of a cavity resonator, the cross section of which is a parallelogram of a very special shape, is evaluated. It is shown that this cavity resonator may be used very well where a good quality is required without the possibility of lower frequencies occurring when the cavity is used as a part of a triode generator.

**1754:** M. J. Druyvesteyn: Experiments on the effect of low temperature on some plastic properties of metals (Appl. Sci. Res. **A 1**, 66-80, 1947, No. 1).

The yield value, breaking strength, elongation and hardness of a number of pure polycrystalline metals were measured at room temperature and at —183° C. The yield value was always higher at —183° C than at 20° C, the difference being relatively small (< 75 per cent) for the cubic face-centered and a number of hexagonal metals (e.g. Mg). The same difference is large for the body-centered metals and for Zn, Cd and Sn. The former metals have a larger elongation at lower temperatures; the latter, however, become more or less brittle at low temperatures. The breaking strength and the hardness generally increase with decreasing temperature.

**1755*:** J. A. Haringx: Over sterk samendrukbare schroefveren en rubberstaven en over hun toepassing bij trillingsvrije opstellingen (Dissertatie Delft 1947) (On highly compressible helical springs and rubber rods and on their application to vibration-free mountings; in Dutch).

The contents of this thesis will appear in full in Philips Res. Reports. A part of the subject has been treated in Philips Techn. Rev. **9**, 16 and 85, 1947, Nos 1 and 3.

**1756:** Balth. van der Pol: Wiskunde en radioproblemen (Simon Stevin **25**, Dec., 1947) (Mathematics and radio problems; in Dutch).

A lecture given before the Dutch "Mathematical Centre". The author treats the lack of mutual understanding sometimes found between technicians, physicists and mathematicians, each group speaking its own "language". The paper deals with the relation between Dirac's function and Stieltjes's integrals; Landau's O-symbol; Hurwitz's determinants characterising the stability of linear systems; the wave equation, diffraction round a sphere (propagation of radio-waves round the earth); continued fractions applied to filter circuits; non-linear differential equations as related to valve oscillators and, finally, modern electrical calculating machines.

**1757:** C. J. Bouwkamp: On spheroidal wave functions of order zero (J. Math. and Phys. **26**, 79-92, 1947 No. 2)

A survey is given of some of the results arrived at by the author in his dissertation (Groningen 1941; dealing with the diffraction of sound by a circular aperture and related problems): calculation of characteristic values and characteristic functions occurring in the separation of the wave equation in spheroidal coordinates; tables of numerical values.

VOL. 9 No. 12, pp. 349-380 1947/1948

Philips Technical Review

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

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

# THE MANUFACTURE OF CORRECTION PLATES FOR SCHMIDT OPTICAL SYSTEMS

by H. RINIA and P. M. van ALPHEN.

535.313

When an image is formed with optical systems employing mirrors or lenses, the imperfect imaging due to aberrations often gives rise to great difficulties. Disregarding those of a higher order, there are five defects: the spherical aberration, coma, astigmatism, curvature of the field and distortion. In 1931 B. Schmidt designed a mirror system which is corrected very well for four of these five defects, whilst the influence of the curvature of the field upon the distortion may usually be ignored. Schmidt introduced a diaphragm and a correction plate in the centre of curvature of a spherical mirror. The great advantage of this system lies in the fact that it allows of large apertures being used while giving at the same time a reasonably large useful field of vision. First the working of this Schmidt optical system and the shape of the correction plate will be discussed, followed by a description of a new method worked out in the Philips laboratory at Eindhoven for the manufacture of these correction plates. With this method a mould is used the face of which is the negative of the plate required, but with the thickness dimensions exaggerated, for instance, five times. A 20% gelatin solution is then applied on the mould and after it has dried it shows the desired shape. Schmidt's system, first applied in astronomical telescopes, is also useful for various other optical purposes.

## Introduction

Most optical instruments are designed to produce an image of an object. This image sometimes has to be enlarged, as in the case of projection lenses, or reduced, as in the case of photographic lenses but it is always required to be sharp and conformable to the object. That it is not so simple to meet this requirement is apparent from the fact that astronomical telescopes, used for producing an image of a celestial constellation, have already taken more than three centuries to develop and still cannot be said to be perfect. Aberrations still cause a certain unsharpness of the images and lack of conformity with the object.

The principal aberrations have been dealt with in an article published in a previous issue of this journal [1]), where it was shown that their magnitude depends upon the distance $y_0$ from the paraxial image point to the optical axis of the system and upon the height of incidence $H$ of the light ray.

If we confine ourselves to third order aberrations, i.e. those which are proportional to at most the third power of the parameters just mentioned, then, when using monochromatic light, there are five optical aberrations: spherical aberration, coma, astigmatism, curvature of the field and distortion. When we work with composite light then in addition to these five defects we have to contend with two chromatic aberrations.

Optical systems have been constructed in various ways in an attempt to eliminate these aberrations as far as possible. In astronomical reflectors, for instance, the hollow mirror has been made parabolic instead of spherical. When stars are observed in the axial direction no aberrations then occur even when the height of incidence $H$ is great, but when there are pencils of rays making an angle with the axis then coma and astigmatism very soon lead to large aberrations, with a parabolic mirror a large aperture can, in fact, be used but the field of vision is small.

Another possibility is that with a spherical mirror a diaphragm is placed in the centre of cur-

[1]) W. de Groot, Optical Aberrations in lens and mirror systems. Philips Techn. Rev. 9, 301-307, 1947 (No. 10).

vature. In directions making a fairly large angle with the optical axis of the system some of the aberrations are avoided, just as is the case in the direction of the axis, but the spherical aberration remains. In order to limit the effect of this spherical aberration it is necessary to give the diaphragm a small aperture; in the case of a spherical mirror with the diaphragm in the centre of curvature one can work with a large field of vision, but one must then make the aperture small, and this means low luminous intensity.

With lens systems attemps have been made to eliminate aberrations by combining lenses of different shapes and different kinds of glass. Thus composite lenses are formed which eliminate certain aberrations. But for an optical system with large aperture and large field of vision it is impossible to correct all aberrations at once in this way. Often the elimination of some aberrations is accompanied by a greater effect of the others. It is particularly the aberrations of higher orders that then play a part. Even if a lens system were designed in such a way that all third order aberrations of one part were absolutely neutralized by the corresponding aberrations of the other part, there would still be no guarantee that the image is sharp. Aberrations of the fifth and higher orders may still completely spoil the image. It is impossible to suppress these aberrations exactly together with those of lower orders. The greater the parameters $y_0$ and $H$, *i.e.* the larger the field of vision and the higher the luminous intensity, the greater is the effect of these higher order aberrations.

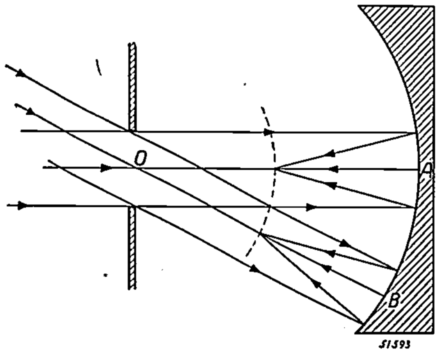


Fig. 1. A spherical hollow mirror with a diaphragm in a plane passing through the centre of curvature $O$. The images formed by the beams with parallel rays $A$ and $B$ lie on a spherical surface with $^1/_2$ $R$ as radius, $R$ representing the radius of the mirror and $O$ its centre.

An important construction designed with the object of correcting aberrations as far as possible while retaining a large field of vision and extremely high luminous intensity is Schmidt's mirror system. In the article quoted in footnote [1]) it has already been stated that in this system a spherical hollow mirror and a correction plate with an aspherical surface are used.

In the present article a new method is described for the manufacture of these plates, a method which has been worked out in the Philips laboratory at Eindhoven. First, however, we shall discuss briefly the working of this Schmidt optical system and the shape of the correction plate.

## B. Schmidt's invention

B. Schmidt, an instrument-maker of the Hamburg Observatory at Bergedorf, invented his optical system in 1931. We shall consider the points he had in mind when devising this system [2]).

Schmidt started with a spherical hollow mirror. This offers at once two great advantages. In the first place a mirror is perfectly free of all chromatic aberrations, and further, given equal focal distance and diameter, the spherical aberration of a hollow mirror is eight times smaller than that of a simple lens.

In the centre of the spherical mirror (the radius of which we shall call $R$) Schmidt placed a diaphragm. Since the direction of each incident ray can now be regarded as the optical axis, the coma and astigmatism related to the diaphragm plane are nil. The image field for beams of parallel rays is a part of a sphere with $^1/_2$ $R$ as radius, as is seen in *fig. 1.* For photographic telescopes and various other purposes (screen photography, projectors for television receivers) the fact that we have to do with a curved image field does not constitute any great objection,for the film or the screen of the cathode-ray tube can be made spherical, so that this curvature of the field is of no effect. If with a hollow mirror a diaphragm is placed in the centre of curvature and a spherical image field is used, then all aberrations except the spherical aberration are eliminated, not only those of the third order but also those of higher orders.

Strictly speaking, there is still the effect of distortion. Since this defect is independent of the diaphragm aperture it does not disappear even when the aperture is made small to counteract spherical aberration.

This occurrence of distortion is apparently contradictory to what has been stated in the article quoted in footnote [1]) about the value of the coefficients $c_1 \ldots c_4$ governing the third order aberrations. There it is argued that in our case the coefficient $c_4$ is nil, from which it is to be concluded that no distortion occurs. It must be borne in mind, however, that the coefficients $c_1, \ldots c_4$ relate to the intersection of the ray in question on a flat plane perpendicular to the axis in the focal point. The secondary axis ($H = 0$) does indeed intersect

[2]) Central-Zeitung für Optik and Mechanik, 52, No. 2, 1931.

this plane, after reflection, at the point $y_0 = {}^1/_2\, R \tan \vartheta$, where $y_0$ represents the position of the paraxial image field and $\vartheta$ the angle between the secondary axis in question and the main axis. When, however, we have a curved image field (a sphere with radius ${}^1/_2\, R$) and consider the point where the same ray intersects this sphere, then the distance from this point of intersection to the axis measured along the sphere is ${}^1/_2\, R\, \vartheta$, and since $\vartheta = t - {}^1/_3\, t^3$ when $t = \tan \vartheta$ this may be regarded as a distortion, which becomes manifest, for instance, when a strip of film originally stretched over the surface of the sphere is laid out flat and we examine the image in this spherical surface drawn out flat. As a rule, however, this distortion is so slight that its influence may be ignored.

The only image defect that has to be eliminated is the spherical aberration. Schmidt does this by introducing a correction plate in the diaphragm.

The manner in which this correction element works can be explained with reference to *fig. 2.* We assume that a source of light is placed in the focus of a hollow spherical mirror. In the top half of the drawing it is shown how the rays coming from the focus are reflected when there is no correction plate. Only the paraxial rays run practically parallel to the main axis after reflection. The greater the angle between the incident ray and the main axis, the more the reflected ray diverges from the line parallel to the main axis.

In the bottom half of the illustration it is indicated how Schmidt has eliminated this spherical aberration. We can imagine that owing to the refraction in a prismatic piece of glass each of the reflected rays is made to run parallel to the major axis of the mirror. All these glass prisms together then form the corrcction element. The division into a large number of very thin prisms is only a schematic representation. Actually the correction plate has a continuous surface. In the illustration the thickness of this plate is greatly magnified.

Schmidt realized that it is of great importance to apply the correction plate in the centre of the mirror. This gives the same great advantage as mentioned above when the diaphragm is placed in that centre. Since all the incident rays now bear the same relations in respect to the correction element, when constructing the passage of the rays through this element one may regard each direction as the optical axis. In regard to the rays not passing along the main axis of the mirror, the correction plate is not, it is true, perpendicular to the direction of incidence, but since the deflection of a ray through a prismatic piece of glass is only to a very small extent dependent upon the angle of incidence, the differences thereby arising may be regarded as an effect of a higher order.

This is more readily understood when it is borne in mind that the purpose of introducing the plate in the diaphragm is only to correct the image. The correction element is very thin, the optical strength of the system being supplied by the hollow mirror. When owing to the angle of incidence the direction of the non-paraxial rays is slightly changed in the correction plate, this only results in a change in the correction, in contrast to the case of a lens system where the angle of incidence of the boundary rays causes these to pass through an entirely different thickness of glass, resulting directly in a change in the image. That is why a Schmidt optical system can be used with an aperture much larger in relation to the focal distance than is the case with a lens system.

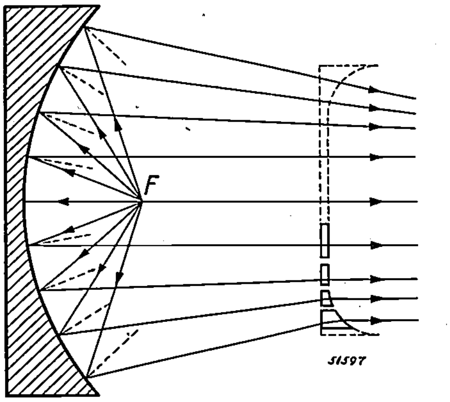


Fig. 2. The effect of a correction plate applied to a spherical hollow mirror. A light source is imagined as being situated in the focus $F$ of the mirror. The top half of the figure shows the direction given to the reflected rays as a consequence of spherical aberration. The bottom half shows how a correction plate bends the reflected rays into a parallel beam.

There is one other point that must be briefly dealt with here. It has been stated above that with a hollow mirror no chromatic aberrations occur. But as the correction plate refracts, it has dispersion and consequently gives different deviations to rays of different wavelengths. This can be taken into account when deciding upon the shape of the correction plate, taking care to keep the influence of chromatic aberration as small as possible. Consideration must also be given to the fact that the difference in cross section between the thickest and the thinnest part of the correction plate is very small, often not more than a few tenths of a millimeter, so that there need be no fear of any appreciable effect of chromatic aberration with a plate such as this.

With a parabolic mirror the spherical aberration is eliminated in only one direction, whereas with Schmidt's mirror system it is practically eliminated in all directions. Since also the other aberrations (except for the curvature of the field) are practically eliminated for all directions, this optical system can be used with a large aperture and moreover has a fairly large field of vision.

All we have to do is to construct a correction plate that meets all the requirements.

### The shape of the correction plate

Before discussing the method by which the correction plate can be made, we shall first show a simple way to calculate the required shape of the plate.

Since the spherical aberration — the angle of deviation of the rays from the desired direction — increases according to the third power of the height of incidence $H$, the correction plate compensating this aberration must have a surface whose slope likewise increases according to the third power of the distance to the axis. With a prism the deviation is proportional to the angle of refraction. The thickness of the correction plate (one side of which is imagined as being flat) must therefore increase from the middle in proportion to $H^4$. If also fifth and seventh order aberrations are taken into account, then the thickness of the plate must conform to the equation

$$d = d_0 + BH^4 + CH^6 + DH^8, \ldots \quad (1)$$

where $B$, $C$ and $D$ are still unknown coefficients.

From this formula it follows — as already appears from the construction of the passage of the rays in fig. 2 — that in the middle the correction plate has to be very flat, rapidly increasing in thickness towards the edge.

It is advantageous to reduce the difference in thickness between the middle and the edge of the plate, for it has been seen above that the correction element gives rise to two kinds of aberrations: 1) chromatic aberrations due to dispersion, and 2) aberrations due to the angle of incidence of the non-paraxial rays. These aberrations can be reduced by making the absolute value of the deviation as small as possible. One must not, therefore, use the form of plate schematically represented in fig. 2, which shows a very large slope at the edge while the middle is practically flat. Instead we have to construct a plate in such a way that the slope at the edge is reduced by giving the surface in the middle a slope in the opposite direction. This can be achieved by adding to formula (1) a term $- AH^2$ ($A$ positive), so that we get the expression:

$$d = d_0 - AH^2 + BH^4 + CH^6 + DH^8. \quad (2)$$

In this manner the correction plate is applied as it were to a flat-spherical lens. *Fig. 3* shows side by side the original correction plate, this flat-spherical lens and the plate obtained by a combination of the first two. The thickness and the slope at the edge have thereby been reduced, whilst the flat minimum in the thickness has been shifted from the middle to close to the edge. The rays passing through this zone of the minimum are not refracted. The point where after reflection from the hollow mirror they intersect the axis has now become the point where all the rays converge.

Owing to the addition of the flat-spherical lens the focal distance of the system is somewhat lengthened, as will be found when following the passage of the paraxial rays, but there is no objection to this.

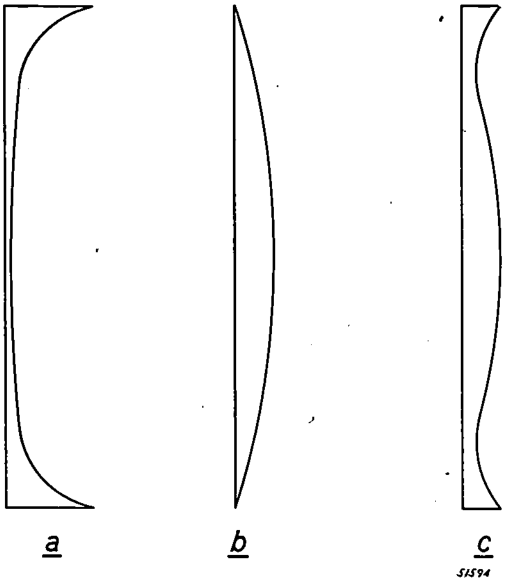


Fig. 3. *a*) A correction plate for a Schmidt optical system.
*b*) A flat-spherical lens with which the correction plate is often combined.
*c*) The shape that the correction plate assumes when combined with *b*.
The thickness of the plate is strongly exaggerated in these illustrations.

The coefficients $A$, $B$, $C$ and $D$ and $d_0$ in formula (2) are determined by calculating for a number of heights of incidence $H$ the value required of the slope and thus of the thickness of the plate.

*Fig. 4* shows diagrammatically how the correction plate in its ultimate form causes the rays of a parallel beam to converge upon one point after reflection from a hollow mirror.

The effect of the correction plate is also clearly seen from the two photographs reproduced in *fig. 5*, likewise relating to a pencil of parallel rays reflected by a hollow mirror. In fig. 5*a* there is spherical aberration, whilst in fig. 5*b* this has been eliminated by means of a correction plate.

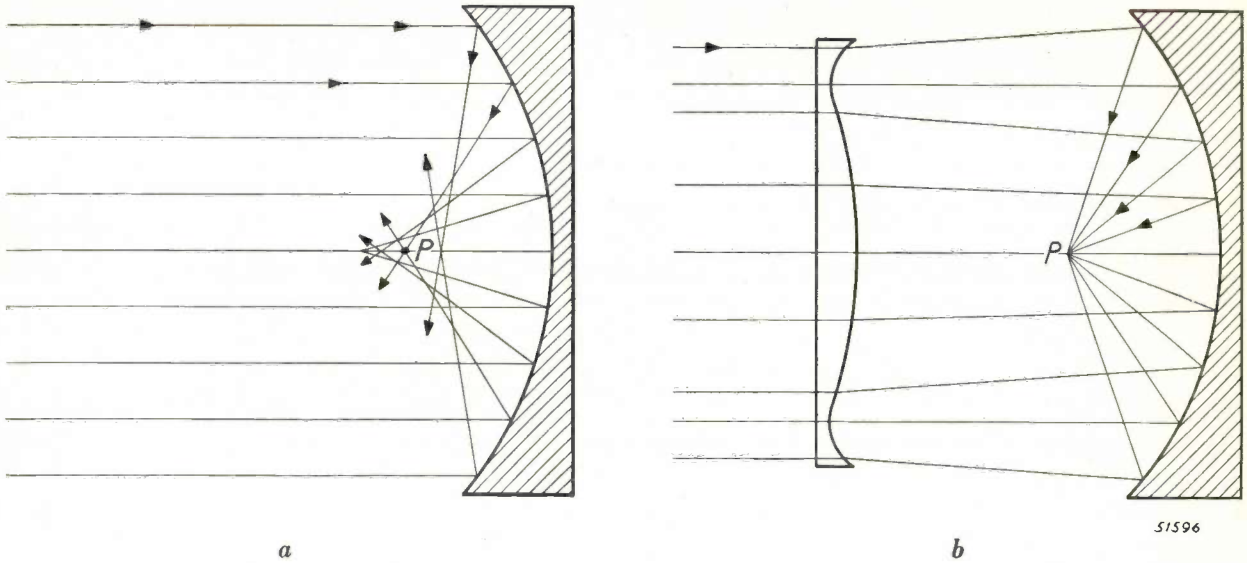


Fig. 4. *a*) With a spherical mirror the rays of the beams at different distances from the optical axis converge at different points.
*b*) The effect of the correction plate is to cause all parallel rays to converge at one and the same focal point *P*.

## The manufacture of the correction plate

Having determined the shape that the correction plate should have, we must then consider the question how to make a plate conforming exactly to that shape.

Schmidt used a plate of glass, but the correction element may also be made of any other transparent and light-refracting substance.

When Schmidt published his invention he did not say how he gave the correction plate the right shape. Many believed that the grinding of such a plate would cost hundreds of hours of work and that this would be an objection against the application of this new construction in practice.

After Schmidt's death in 1935, R. Schorr published the method employed by this Hamburg instrument maker [3]). It then appeared that Schmidt had been making his correction plates in an ingenious manner without it costing him too much time. He had realized that the curved plane of the correction plate corresponded approximately

[3]) Zeitschrift für Instrumentkunde, 56, 336-338, 1936.

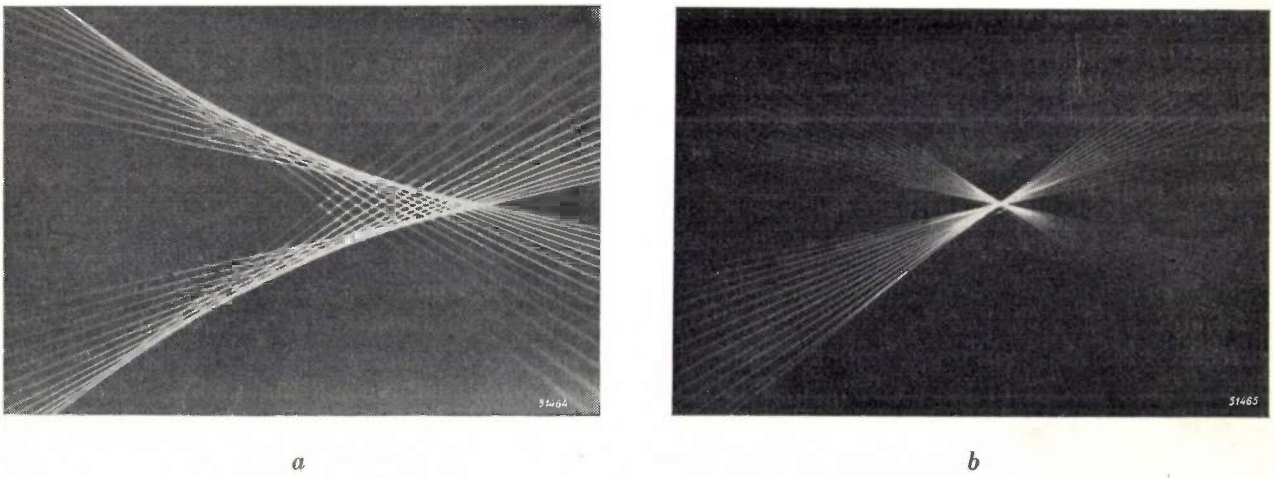


Fig. 5. The influence of a correction plate on the passage of rays reflected from a spherical hollow mirror. *a*) The course followed by the rays without the correction plate. On the left is the hollow mirror reflecting the rays, which are made visible here. *b*) The course of the rays when a correction plate is used. Before striking the mirror the rays pass through a correction plate. The rays which in *a*) form an extensive caustic curve converge in *b*) upon a sharply defined focal point. These photographs were taken with a slanting photographic plate in the focus of the mirror system. In both cases the central rays of the beam reflected by the mirror have been blocked out.

to the deflection plane that a circular plan-parallel plate supported round its edge assumes under a uniformly distributed pressure. Putting this into practical application, he laid a flat glass plate on the rim of a round vessel which he then evacuated until the necessary deflection took place. Schmidt then ground the upper surface of the glass plate in the deflected state until it was absolutely level. When he then let air flow into the vessel again, the ground surface assumed approximately the shape that it has to have for the boundary of the correction plate.

It is indeed a very ingeneous method, but A. Couder [4]) has proved that it is certainly not the right way for correction plates of small dimensions. And for large plates, such as are used for astronomical telescopes, this method is not accurate and is also rather cumbersome.

Various people have been investigating the ways and means of constructing these plates. Some have been hand-ground, in the same way as the parabolic mirror is made for a telescope, checking step by step in how far the desired result is reached and where improvements are still necessary. This method may be practicable for making one single plate for an astronomer's telescope, but it is certainly not suitable for the mass production which further technical applications require.

Another method is based on the use of transparent masses shaped in moulds under pressure, some plastic mass, for instance polystyrene, being pressed in a mould made in the desired shape. The mould has to be extremely accurate and have a surface of "optical quality". Moreover, it has to be able to withstand pressure or heating, so that the choice of material is limited. The mould is generally made and polished by hand and needs touching up in places to give it the right shape. A separate mould is thus needed for every correction plate of a different shape.

## A new method for the manufacture of correction plates

An entirely new method for the manufacture of correction plates has been worked out at Eindhoven. On a turner's precision lathe a metal mould is made the surface of which forms the negative of the shape of the plate required. In the transversal directions the shape of this mould is of the same dimensions as the correction plate is required to have, but the differences in thickness are exaggerated about five times; if the actual correction plate has a total variation in thickness of say 0.5 mm, in the mould this will be 2.5 mm.

The carefully finished and polished mould is heated to abt. 40° C with running water, after which a 20% gelatin solution [5]) is poured onto the mould and a glass plate laid over it. With the aid of a pair of set-screws this plate is kept at a small distance away from the surface of the mould, any excess gelatin thereby being pressed out. When cold water is then run through the mould it is cooled down and the gelatin very soon sets into a stiff gel. The glass plate is then raised by turning the set-screws and removcd from the mould. Owing to the strong adhesion between glass and gelatin the gel easily comes away from the mould and shows all the details of the mould.

The skin of gelatin is then hardened in formalin vapour and uniformly dried. Being held in the lateral directions by the glass plate, the gel shrinks only in thickness. The original surface is reduced in one direction, otherwise remaining unchanged. When a 20% gelatin solution is used 80% water will evaporate, leaving a glass plate covered with a thin layer of gelatin the surface of which is a five-fold reduction of that of the mould, thus exactly the required shape of the correction plate. It is astonishing how extremely accurate and uniform this shrinking is, producing a very smooth surface with every detail of the mould reduced five times

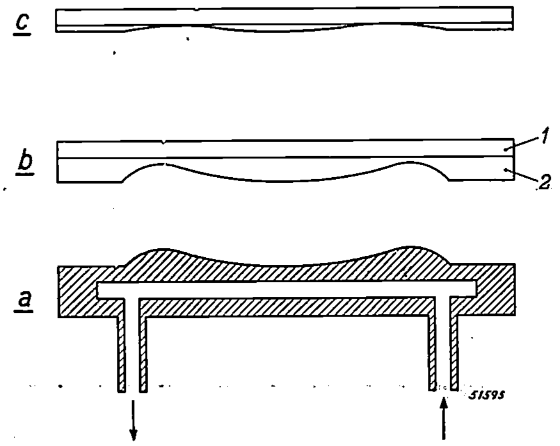


Fig. 6. *a*) A mould for a correction plate shown in cross section. This is a hollow mould with inlets and outlets for passing through cold or hot water.

*b*) A correction plate with layer of gelatin before drying. *1* represents the plan-parallel glass plate and *2* the layer of gelatin solution.

*c*) The same correction plate after drying.

In these illustrations the dimensions in the direction perpendicular to the correction plate have been greatly magniifed.

[4]) Comptes Rendus Ac. Sc. Paris **210**, 327-328, 1940.

[5]) This 20% gelatin solution is used when the differences in thickness in the mould are fivc times exaggerated. Other concentrations may be used, for instance a 10% solution, when the dimensions of thickness in the mould are exaggerated 10 times. See also footnote 6).

but otherwise true to shape. It is essential that the gelatin shows a really smooth surface after drying.

*Fig. 6* shows a mould and also a correction plate before and after the drying of the gelatin. It will be understood that the dimensions in the direction perpendicular to the plate are greatly exaggerated in these drawings.

The hardened layer of gelatin is found to be very stable. Nevertheless, it may be advisable to protect the curved surface of the correction plate with a second flat glass plate, because wet fingers or drops of water are apt to damage the gelatin.

The method described offers a numbers of advantages over those applied hitherto:

1) Since the surface ultimately obtained is a five-fold reduction of the shape of the mould, the latter may be five times less accurate. If, for instance, the final shape is required to be accurate within 0.5 $\mu$ then the shape of the mould need only be accurate to within 2.5 $\mu$. Any fine scratches in the mould are of no consequence, because these are also reduced five times when the gelatin dries; in many cases their depth becomes smaller than the wavelength of light, so that they become practically invisible.
2) The mould needs only moderate heating and cooling. Since there is no pressing there is no distortion arising therefrom.
3) Since the correction plate consists mainly of glass any distortion or mechanical deflection of the plate is precluded. The hardened layer of gelatin is highly impervious to scratching, so that the plate can safely be cleaned with a soft cloth.
4) Correction plates of different optical strengths can be made with one and the same mould, for one is not restricted to the concentration of 0.2 or 0.1. By varying the concentration of the gelatin solution one gets correction plates, after hardening, of different strength [6]).

## Applications of the Schmidt optical system

Schmidt's first spherical mirror with correction plate was made for the Hamburg observatory. It was undoubtedly astronomy that he first had in mind as the field of application for his invention, and astronomers have in fact made good use of it. On January 1st 1941 there were already 44 observatories equipped with a Schmidt camera [7]). The great luminous intensity and extensive field of vision of these instruments have been turned to good advantage.

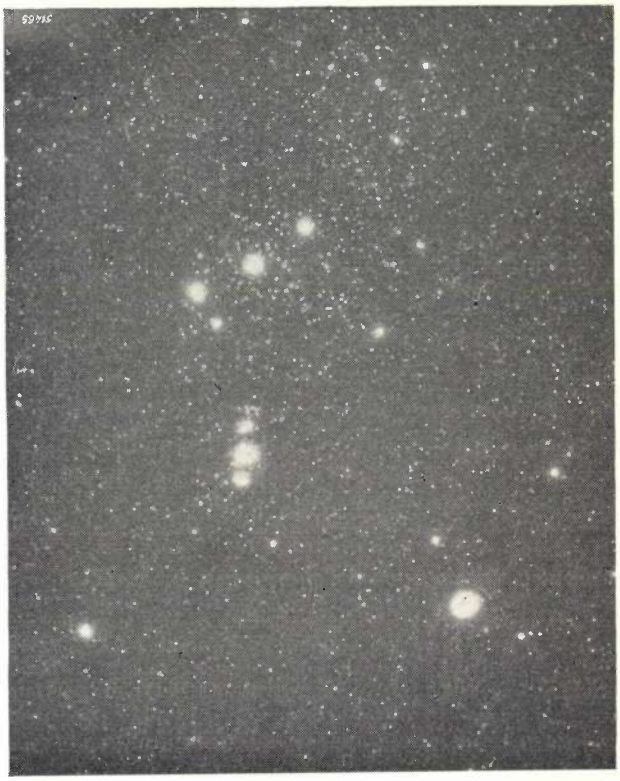


Fig. 7. Part of the Orion constellation as photographed in the Philips laboratory at Eindhoven when using a Schmidt mirror system (focal distance 8 cm) with a correction plate of gelatin. This photo demonstrates the large useful field of vision. Although its diameter is about 20° the image is sharp and free of aberrations right up to the edges of the photo.

The aperture ratio of these cameras — that is to say the ratio of the diameter $D$ of the objective to its focal distance $F$ — shows clearly enough how great the improvement is in luminous intensity. When $D/F$ is small one speaks of "slow" telescopes, because a long exposure time is required for making photographs. In the case of many previously built reflectors and refractors $D = F: 5$, $F : 6$ or $F : 8$, and sometimes still smaller. An exceptionally "rapid" telescope is the 200 inch telescope of Mount Palomar Observatory with $D = F : 3.3$.

But with the Schmidt optical systems the relative aperture is much greater. An astronomical camera with $D{=}F:1$ is quite a common type. Such a camera has a useful field of 25°, whilst the field of a parabolic telescope having an aperture equal to the focal distance would have to be measured in minutes. And cameras with $D = F : 1$ are still not

[6]) The importance of this can be understood from the following. As is known, a parabolic mirror can only be used for observing an infinitely remote object. If the object is at a finite distance then an elliptical mirror has to be used, with the object and the image in the two foci of the ellipsoid. When the distance of the object changes one therefore needs another elliptical mirror. The same applies to the Schmidt optical systems. When changing over from an infinitely remote object to one at a finite distance the correction plate has to be made "stronger". The required strength of the correction plate depends upon the distance of the object.

[7]) G. Z. Dimitroff and J. G. Baker, Telescopes and accessories, Blakiston Company, Philadelphia 1946, p. 292.

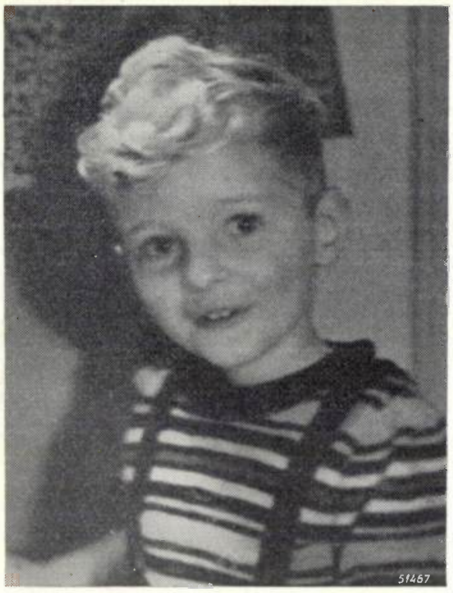


Fig. 8. Photograph taken with a camera made in the Philips laboratory at Eindhoven with a Schmidt optical system and a gelatin correction plate ($D = F : 0.7$). This optical system is of such a high power that the photo could be taken without any other illumination than that from a 25 W lamp and an exposure time of 1/10 sec.

the most rapid, some having been built even with $D = F : 0.6$.

The fact that good photographs of the constellations can be taken with the Schmidt optical system and simple accessories is demonstrated by *fig. 7*, where a photograph is reproduced, taken in the Philips laboratory at Eindhoven, of a part of the Orion constellation. A Schmidt mirror system with a focal distance of 8 cm was used. Although the field of vision has a diameter of no less than 20° the image is sharp and free of aberrations right up to the edges of the photograph.

The possibilities of application of the Schmidt optical systems is not confined to astronomy. This sytem has been applied in our laboratory in the construction of cameras of high optical power for experimental use (see *fig. 8*). Two other applications have already been mentioned in passing: cameras for photographing X-ray images on a fluorescent screen, and projectors as used in television receivers. It is the intention to describe the Schmidt optical systems designed for these purposes in a subsequent issue of this journal in due course.

# CIRCUIT FOR CONDENSER MICROPHONES WITH LOW NOISE LEVEL

by J. J. ZAALBERG van ZELST.

621.395.616: 621.395.822

Hitherto the condenser-microphone has generally been used in series with a direct voltage source and a resistor. The resistance of the latter must be of a high value, not only to obtain good reproduction of the low notes but also in view of the noise level. This high value of the resistance, however, makes it necessary to build in an amplifying valve close to the microphone. The same is also necessary with another method where the microphone forms part of the high-frequency oscillating circuit of an oscillator. There is a third method, dealt with more extensively here, which does not have this drawback and, moreover, is characterized by an exceptionally low noise level. With this method the microphone is incorporated in a bridge circuit which, as the diaphragm vibrates, supplies an amplitude-modulated high-frequency voltage. This voltage is amplified and detected. The bridge circuit (including the microphone but without valve) can be connected to the amplifier by means of a cable. This method is equally suitable for other capacitive vibration pick-ups, e.g. a capacitive gramophone pick-up.

## The condenser-microphone working on direct voltage

In the previous issue of this journal [1]) a type of condenser-microphone was described which excels by its uniform reproduction of the audio frequencies and at the same time has a high degree of sensitivity and is of small dimensions. However, when used in what might be called the "conventional" circuit this new design still has some drawbacks in common with the older types of condenser-microphones. In this circuit (*fig. 1*) the microphone works on direct voltage and is in series with a resistor $R$. This resistor is given a high value for two reasons:

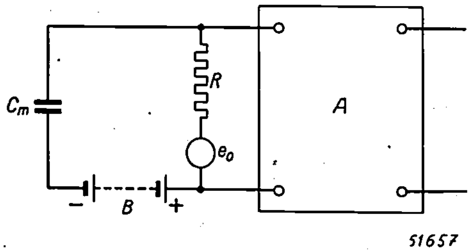


Fig. 1. Conventional circuit of a condenser-microphone ($C_m$). $B$ is a source of direct voltage, $R$ a resistor of high impedance, $A$ an amplifier; $e_0$ represents the noise voltage coming from $R$.

1) As explained in the article referred to [1]), for the sake of good reproduction of the low notes the charge of the condenser formed by the microphone must remain practically constant when the diaphragm vibrates. The condition required for this is that $2\pi f C_m R$ must be at least of the order of unity ($C_m$ = capacitance of the microphone), also for the lowest frequency $f = f_1$, which has to be reproduced well. With $f_1 = 30$ c/s and $C_m = 50$ pF this means that the resistance $R$ must be of the order of 100 megohms.

2) The second reason why $R$ has to have a high value is connected with the noise, the source of which lies in this resistor.

The average value $\overline{e}$ of the quadratic contribution to the noise voltage, in a frequency band $\Delta f$, amounts to [2]):

$$\overline{e_0^2} = 4\,kTR \cdot \Delta f, \quad . \quad . \quad . \quad . \quad . \quad (1)$$

in which $k$ represents Boltzmann's constant ($1.38 \cdot 10^{-23}$ Wsec/degree) and $T$ the absolute temperature of the resistor. In the case of the condenser-microphone, in an equivalent circuit one can imagine that noise-voltage source as being connected in series with the resistor and the microphone (fig. 1). What has to be considered is that part of this noise voltage which occurs across the input terminals of the amplifier, thus also across the capacitance $C_m$. The average quadratic contribution $\overline{e^2}$ to the noise voltage across $C_m$, in a frequency band $\Delta f$, is calculated from (1):

$$\overline{e^2} = 4\,kT \cdot \frac{R}{1 + (2\pi f\, C_m\, R)^2} \cdot \Delta f. \quad . \quad . \quad . \quad (2)$$

The total contribution in a frequency range extending from $f_1$ to $f_2$ is found by integrating (2) between the limits $f_1$ and $f_2$; thus, for the range of the audible notes, from $f_1$ = about 30 c/s to $f_2$ = about 14 000 c/s. For values of $R$ greater than $R_{\text{min}} = 1/(2\pi f_1 C_m)$ (i.e. approximately the minimum required for good reproduction of the low notes) the rule holds that in each frequency interval $\Delta f$ lying above $f_1$ the contribution to the noise becomes smaller as $R$ increases. This may be seen in *fig. 2*, where the fraction $R/[1 + (2\pi f C_m R)^2]$ occurring in

[1]) A. Rademakers, A condenser-microphone for stereophony, Philips Techn. Rev. 9, 330-338. 1947 (No. 11).

[2]) See *e.g.* Philips Techn. Rev. 2, 140, 1937 or 6, 130, 1941.

(2) is plotted as a function of $f$ for one value of $C_m$ and two values of $R$: curve (*1*) for $R = R_{\min}$, (*2*) for $R = 2R_{\min}$. To the right of $f_1$ curve (*2*) lies entirely underneath (*1*). A lower noise level is therefore obtained by choosing $R$ greater than $R_{\min}$.

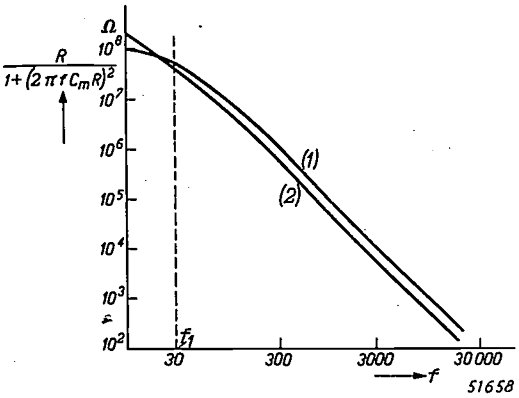


Fig. 2. The factor $R/[1 + (2\pi f C_m R)^2]$, represented here as a function of the frequency $f$ in c/s, in the diagram of fig. 1 is a measure for the contribution to the noise (in a frequency interval $\Delta f$) at the input of the amplifier. Curve (*1*) applies for the smallest value of $R$ which with a given value of $C_m$ yields a satisfactory reproduction of the lowest frequency $f_1$; curve (*2*) applies for twice as high a value of $R$. Curve (*2*) lies below curve (*1*) to the right of $f_1$. This diagram applies for $f_1 = 30$ c/s, $C_m = 50$ pF, $R = 100$ and 200 megohms.

The drawbacks mentioned above as being connected with the conventional circuit of the condenser-microphone are related to the high value of the resistance connected in series with the microphone. These drawbacks are the following:

a) The circuit in which the microphone is taken up is highly sensitive to low-frequency interference which may be induced in it, for instance by neighbouring a.c. mains.

b) The insulation of the microphone and of the grid of the first amplifying valve has to answer very high requirements.

c) The microphone cannot be connected directly to a cable, because the capacitance of the cable would then be in parallel with that of the microphone, so that variations in the cable capacitance — caused by bending, clinching, twisting, etc. of the cable — would induce interfering voltages. Consequently one is obliged to build in an amplifying valve, with accessories, near the microphone. In many cases this means an awkward complication.

The first-mentioned difficulty can be overcome by providing adequate screening. The second one can be partly met by employing a low series resistance; in order still to satisfy the requirement of $2\pi f_1 C_m R > 1$ it is necessary to increase $C_m$ proportionately, by connecting a fixed condenser in parallel to the microphone. In the article referred to in footnote [1]) it was stated, for instance, that $R$ had been reduced from 80 to 5 megohms and $C_m$ raised from 65 to 1000 pF. This, however, still does not overcome the necessity of having to build in an amplifier near the microphone. Moreover, yet another difficulty has to be taken into the bargain:

d) Both the absolute sensitivity decreases (because the relative capacitance variations — and thus also the resultant voltage variations — become just as many times smaller as $C_m$ is increased) and also the signal-to-noise ratio, because the noise voltage decreases only in proportion to $\sqrt{C_m}$.

All these difficulties are, as already stated, a direct result of the fact that the microphone is connected in series with a high resistance, such being unavoidable when working on direct voltage, and one does not wish to sacrifice anything in quality. For a long time already other circuits have been known which dispense with the high resistance. With such circuits a high-frequency alternating current flows through the microphone, whilst the high-frequency voltage on the microphone is modulated in some way or other by the speech oscillations. The high-frequency current is to be regarded as an auxiliary current with which the impedance of the microphone, varying at low frequencies, is "measured" or "scanned" as it were.

We shall consider two methods based on this principle.

### The condenser-microphone as frequency-modulator of a high-frequency auxiliary current

Riegger [3]) has indicated a method where a condenser-microphone forms part of the high-frequency oscillation circuit of an oscillator, so that certain frequency changes correspond to the variations in capacitance. By means of suitable circuits an amplitude-modulated signal is obtained which corresponds to this frequency modulation and which after detection produces a low-frequency signal.

When this method was published in 1924 it did not meet with much success, partly because of the then inadequate means of detecting frequency-modulated signals. Since there are now better means available, Riegger's method could be applied more successfully, that is to say a fairly high sensitivity, a reasonably low noise level and little trouble from low-frequency induction interferences may be expected. It still has the drawback, however, that a valve — the oscillator valve — with accessories

[3]) Z. techn. Phys. 5, 579. 1924,

has to be built in near the microphone. Moreover, it is rather complicated and highly sensitive to variations in the values of various circuit elements.

## The condenser-microphone as amplitude modulator of a high-frequency auxiliary current

We shall go somewhat more closely into another method with high-frequency supply which has none of the abovementioned drawbacks attaching to it. According to this method a constant current $I$ is sent through the microphone with a likewise constant high frequency (the carrier-wave frequency, $= \omega_0/2\pi$). The voltage on the microphone is proportional to $1/C_m$ and thus varies in amplitude when the capacitance $C_m$ changes in consequence of the sound vibrations.

The relative variations in capacitance, however, are as a rule very small; in this manner one would then get only a small modulation depth. As a consequence the signal would bear an unsatisfactory relation to the noise, which now is not due to a resistor (for there is no resistor here) but partly to fluctuations in the amplitude of the high-frequency supply current and partly to fluctations of the carrier-wave frequency. The current drawn from an oscillator is unavoidably modulated with noise both in amplitude and in frequency. When the current modulated in frequency with noise flows through an impedance which is dependent on the frequency — and the condenser microphone is such an impedance — then a voltage arises which is modulated in amplitude with noise and which combines with the noise voltage coming directly from the amplitude-noise of the supply current.

Furthermore, there is still the abovementioned difficulty that when the microphone is connected to a cable the variations in capacitance of that cable make themselves felt as interferences.

From what follows it will be seen, however, that all these difficulties can be removed.

### *Tuning of the microphone circuit*

The first improvement to be made is to connect the microphone in series with a coil — for the present assumed to be free of loss — the self-inductance of which at the carrier-wave frequency is approximately in resonance with the microphone capacitance. The addition of the constant impedance of this coil does not affect the size of the absolute impedance variations occurring in the circuit when the diaphragm vibrates, but it affects the total impedance, which becomes very small. Thus the relative impedance variations increase considerably, and consequently so does the modulation depth.

Thanks to the fact that the impedance formed by the coil of the microphone in series is so low, the noise components of the current will not induce any noise voltage worth mentioning. At the same time another impedance — a cable — can be connected in parallel to this low impedance without any objection, so that the microphone can then indeed be connected direct to a cable (*fig. 3*). There is still the drawback, however, that the impedance is frequency-dependent; frequency fluctuations will therefore still give rise to noise.

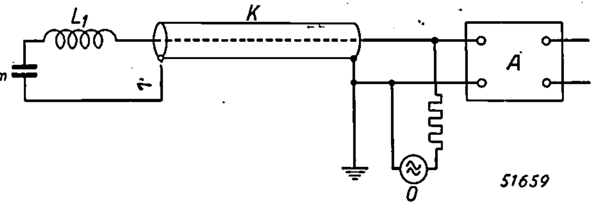


Fig. 3. The oscillator $O$ sends a constant current with a constant frequency through the condenser microphone $C_m$ and the coil $L_1$, which are tuned approximately to the frequency of the oscillator. Vibrations of the microphone diaphragm set up variations in capacitance, as a consequence of which the high-frequency voltage on the microphone is modulated in amplitude according to the sound vibrations. Owing to the low impedance of $L_1$-$C_m$ the connection to the microphone can safely be made by means of a cable $K$. $A$ is the amplifier, serving at the same time as detector.

Finally it has to be considered that the coil will actually have a certain loss resistance $r$. In the first place this resistance forms a source of noise — though it may be a feeble one — but it involves yet another difficulty of a different nature. Contrary to what would be the case with a loss-free coil, the output voltage has a carrier-wave component $Ir$ shifted 90° in phase with respect to the original carrier-wave component. This would lead to distortion in the detection, to which we shall revert presently.

### *Bridge circuit*

In order to circumvent also these difficulties we must extend our circuit somewhat. Instead of "measuring" the impedance of the series connection of microphone and coil direct, we compare it with that of a second branch consisting of an identical coil in series with a fixed condenser, the capacitance of which is just as large as that of the microphone in the state of rest (*fig. 4*). Equal currents are passed through the two branches, which together form a bridge; the difference $V$ of the voltages arising on the branches serves as output voltage.

In the state of rest apparently $V = 0$, even though the coils are not free of loss. When sound strikes the microphone then $V$ consists only of side-band components. The carrier-wave component is absent and therefore has to be added in the detector stage; we shall see presently how this can be done.

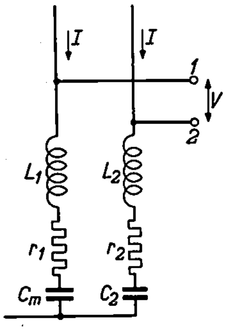


Fig. 4. Bridge circuit consisting of two branches $L_1$-$r_1$-$C_m$ and $L_2$-$r_2$-$C_2$ with a high-frequency current $I$ flowing through both. The self-inductances $L_1$ and $L_2$ are equal, as are also the loss resistances $r_1$ and $r_2$. The capacitance $C_2$ is equal to the microphone capacitance $C_m$ when the diaphragm is at rest. The voltage $V$ between the terminals *1* and *2* has no carrier-wave component, but consists of side bands corresponding to the sound striking the microphone.

How do matters stand now with the previously mentioned causes of noise which still remained, viz. the amplitude and frequency modulation of the current $I$ with noise, and the loss resistance of the coils?

The two causes of noise occurring in the current $I$ are now in principle rendered harmless, for with the diaphragm in the state of rest the bridge is in equilibrium for all frequencies, so that no noise voltage arising from the amplitude modulation of $I$ occurs across the output terminals. And as regards the frequency modulation of $I$ with noise, this too does not give rise to any voltage across the output when the diaphragm is in the state of rest, since the equal impedances of the bridge branches vary with the frequency in exactly the same way. When the equilibrium of the diaphragm is disturbed — owing to sound striking it — then the two noise components do indeed occur, but they do so in proportion to the relative variation in capacitance $(C_m—C_2)/C_2$, thus also proportionate to the sound pressure on the microphone.

Thus there remain as the only permanent source of noise the loss resistances of the coils. Actually we should add the mechanical vibration of the circuit elements and the wiring; this is accompanied by variations in capacitance or self-inductance, which may again constitute a source of noise.

Under this last heading are also the vibrations of the diaphragm when there is no "external" noise. A condenser microphone may also act as a telephone; there will thus be movement in the diaphragm when a noise-modulated current flows through the microphone. This is not the case with its opposite member in the bridge circuit — the condenser $C_2$ — so that the bridge is not in equilibrium and thus produces noise voltage. This effect, however, is found to be very small in comparison with the other noise effects.

*Fig. 5* shows how the two branches can be fed from one oscillator with the introduction of a pentode in between. In the anode circuit of this pentode is an oscillating circuit with self-inductance $L_a$ and a capacitance formed by the sum of the capacitances $C_3$ and $C_4$; the branches $L_1$-$C_m$ and $L_2$-$C_2$ are each tuned to the oscillator frequency (and may therefore be regarded as short-circuits), as is also the circuit $L_a$-$(C_3 + C_4)$. The currents $I$ (*fig. 4*) flowing through the two branches may amount, for instance, to 20-50 mA. The pentode need only yield a power of the order of 0.1 W, thanks to the impedance of the branches being so low. A small valve therefore suffices. If necessary a separate oscillator as indicated in fig. 5 can be dispensed with by causing the pentode itself to oscillate with the connected circuits.

It is now desirable to mention some numerical data. With sounds of a moderate strength the sound pressure (R.M.S. value of the alternating pressure) amounts to about 1 μbar. In the case of the microphone described in the article quoted in footnote [1]), this corresponds to a relative capacity change of about 1:30 000. The voltage $V$ of the side bands (fig. 5) will then be about 1/30 000 of the voltage on the microphone. Owing to the danger of sparking between the electrodes this latter voltage is limited to 100 V. Consequently with a sound pressure of 1 μbar $V$ will be about 3 mV.

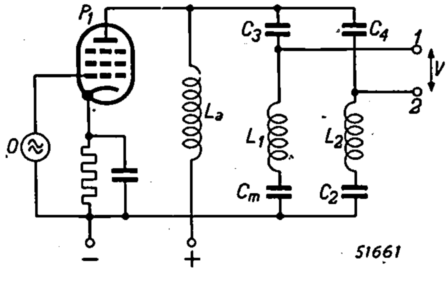


Fig. 5. System for feeding the two bridge branches ($L_1$-$C_m$ and $L_2$-$C_2$) of fig. 4. $O$ is an oscillator with the carrier-wave frequency, $P_1$ an amplifying valve (pentode), $C_3$ and $C_4$ are condensers in series with the two branches, $L_a$ is a coil in resistance with $C_3 + C_4$, *1* and *2* are the terminals across which the modulated voltage arises.

What is the level of the noise voltage between the terminals *1* and *2* (fig. 5)? This noise voltage will come mainly from the resistance between these terminals, thus from the resistance of the two coils together. A coil of good quality and having the required self-inductance has a resistance $r$ = approx. 15 Ohms. In the frequency range $\Delta f$ from 0 to 14 000 c/s the noise voltage is then equal to $\sqrt{4kT\cdot 2r\cdot \Delta f} = \sqrt{4\cdot 1.38\cdot 10^{-23}\cdot 300\cdot 2\cdot 15\cdot 14\,000} = 8.4\cdot 10^{-8}$ V $= 8.4\cdot 10^{-5}$ mV. For an equally large signal voltage a sound pressure of $8.4\cdot 10^{-5}/3 = 2.8\cdot 10^{-5}$ μbar would be necessary, *i.e.* 17 db below the international zero level ($10^{-16}$ W/cm$^2$)! This is indeed exceptionally low. In order to maintain a very low noise level when amplifying and detecting (see the next sub-heading) care has to be taken to ensure that no appreciable amount of noise is added. For certain reasons, however, it may be desirable not to aim at the lowest possible noise level. One has to take into account, for instance, the characteristic impedance of the cable, which has to be approximately equal to the resistance of the coils. In one practical example the only cable available had a characteristic impedance of 160 Ohms, so that the resistance of the coils had to be increased to that value. With a further 80-Ohms resistance of a transformer at the end of the cable, in that case the noise resistance was 2 × 160 + 80 = 400 Ohms, corresponding to a noise level of —6 db. Direct measurement showed the actual noise to be a few db stronger. This has to be ascribed to the variation of stray capacitances due to the vibration of the wiring as a result of mechanical disturbances in the surroundings; in this particular case no special precautions had been taken. It was also found necessary to use components of very good quality, *i.e.* components not subject to fluctuations in their values.

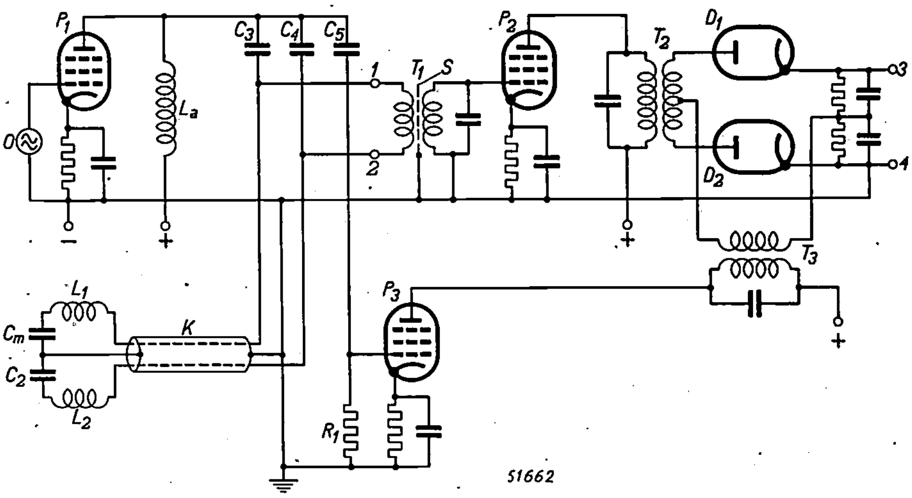


Fig. 6. Amplification and detection of a voltage between the terminals *1* and *2* (the part of the circuit to the left of these points corresponds to fig. 5). *K* is the microphone cable, $T_1$, $T_2$ and $T_3$ are transformers tuned to the carrier-wave frequency, $P_2$ and $P_3$ pentodes, $D_1$ and $D_2$ diodes for the detection. $C_5$ and $R_1$ form a phase-changing voltage divider. *3* and *4* are the output terminals for the low-frequency signal. Thanks to the diodes being in push-pull, no noise voltage from the carrier-wave channel ($C_5$-$R_1$-$P_3$-$T_3$) occurs on the output terminals. $S$ = screen in $T_1$.

Further, it is to be remarked that the system is much less sensitive for gradual variations in the values of the circuit elements than is the Riegger method.

### *Amplification and detection*

The signal voltage $V$ (fig. 5) is stepped up by a high-frequency transformer $T_1$ with a tuned secondary winding (*fig. 6*). Any low-frequency interferences that may be induced are short-circuited by the primary of $T_1$. The secondary voltage is amplified and detected.

Since the signal $V$ does not primarily contain any component with the carrier-wave frequency, such a component has to be added in the detection stage. This is done by means of a separate carrier-wave channel consisting of the components $C_5$-$R_1$-$P_3$-$T_3$ (see fig. 6).

Care has to be taken that no noise is added from this carrier-wave channel and, moreover, that the carrier-wave voltage has the right phase. As to the first point, this has been provided for by arranging for the detection to take place in a push-pull circuit with two diodes; if symmetry is ensured then any noise present in the secondary voltage from $T_3$ cannot reach the output terminals *3* and *4*.

The question of the right phase of the carrier-wave voltage is illustrated in *fig. 7*: fig. 7*a* represents the correct position, the carrier-wave voltage $V_0$ being in phase with the resultant $V_r$ of the two side band components $V_1$ and $V_2$, so that the result is a carrier-wave voltage purely modulated in amplitude. In fig. 7*b* it is assumed that $V_0$ is shifted 90° with respect to $V_r$. During one cycle of the modulation frequency the vector sum of $V_0$ and $V_r$ then varies considerably in phase but little in amplitude. The angular velocity of this vector is therefore no longer uniform; in other words, there is considerable frequency modulation. Moreover, the amplitude modulation takes place at a frequency twice that of the sound frequency. In the case of a different phase shift between $V_0$ and $V_r$ we have an intermediate state where again amplitude and frequency modulation occur simultaneously. This is highly undesirable, because it leads to great distortion in the detection. Care must be taken therefore that the situation represented in fig. 7*a* is actually obtained.

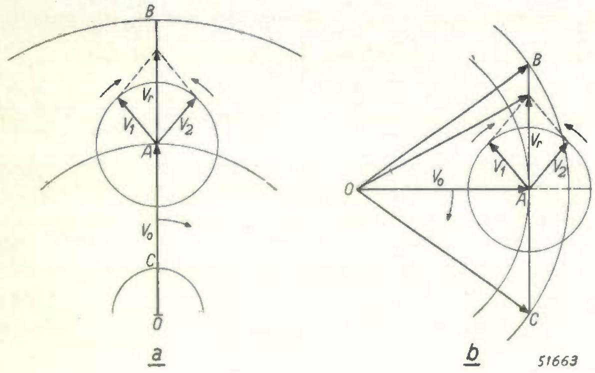


Fig. 7. The vector $OA = V_0$ represents the non-modulated carrier-wave; it rotates at an angular velocity corresponding to the carrier-wave frequency. The side bands are represented by the vectors $V_1$ and $V_2$, which turn about $A$ in opposite directions at an angular velocity corresponding to the signal frequency. The resultant of $V_1$ and $V_2$ is $V_r$, a vector whose maximum size is equal to $AB$ or $AC$. In fig. 7*a* $V_0$ has the right position with respect to $V_r$. The sum of $V_0$ and $V_r$ is a vector which rotates with the same constant angular velocity as $V_0$ and which varies in amplitude between $OB$ and $OC$. The result is pure amplitude modulation. In fig. 7*b* $V_0$ is turned with respect to $V_r$ (in this case 90°). The vector sum of $V_0$ and $V_r$ shows no constant angular velocity and varies only little in amplitude (from $OA$ to $OB = OC$). This means that mainly frequency modulation occurs and amplitude modulation only to a small extent. In a detection system arranged for amplitude modulation this would lead to a distorted low-frequency signal.

To this end it must be borne in mind that the output voltage from the secondary tuned transformer $T_1$ (fig. 6) is shifted 90° with respect to the primary voltage. Consequently it is necessary to provide for an equally large phase shift between the input and output voltages of the carrier-wave channel. This is done with the aid of the voltage so that the voltage on $R_1$ is shifted almost 90° with divider $C_5$-$R_1$, which is so designed that $R_1 \ll 1/\omega_0 C_5$, respect to the anode alternating voltage of the tube $P_1$.

An earthed screen between the coils of $T_1$ prevents capacitive transmission to the secondary coil of the voltage with the carrier-wave frequency present between the points *1-2* and earth.

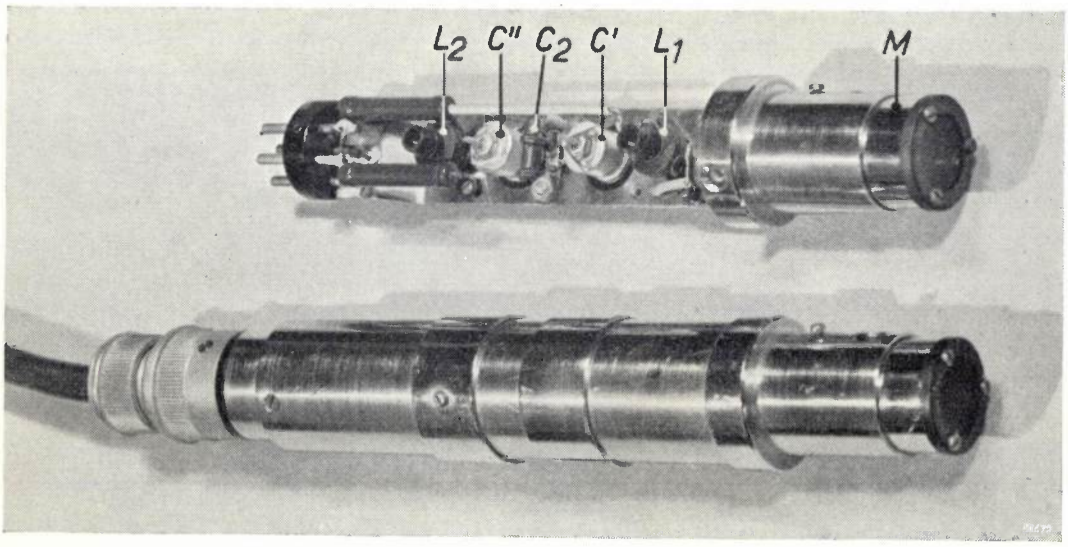


Fig. 8. Top: Opened condenser-microphone showing the component parts built in. $M$ is the microphone proper; $L_1$, $L_2$ and $C_2$ have the same meaning as in fig. 6; the small adjusting condensers (trimmers) $C'$ and $C''$ are connected in parallel to the microphone and to $C_2$ respectively. Bottom: The microphone encased and connected to a cable.

*Fig. 8* shows a complete microphone connected to a cable and also a decased microphone showing the components represented in the bottom left-hand corner of fig. 6.

It is to be added that the principle of the circuit described here is not new and is in fact even older than Riegger's system [4]). The great advantages it offers — low noise level and absence of an amplifier built in close to the microphone — have apparently not been realized hitherto and the methods had become more or less forgotten.

*Other applications*

In conclusion we would add that the method described here lends itself equally well for use in combination with a capacitive gramophone pick-up or any other capacitive pick-up for mechanical vibrations [5]). In principle it is also possible to modify the method for variable inductances instead of variable capacitances, thus, for instance, for electromagnetic vibration pick-ups (including microphones) reacting to the changes in self-inductance of a coil placed near to a vibrating iron membrane or armature.

[4]) See *e.g.* Swiss Patent 95 439 in the name of H. Vogt, J. Engl and J. Massolle.

[5]) P. J. Hagendoorn and M. F. Reynst, An electrical pressure indicator for internal combustion engines, Philips Techn. Rev. 5, 348-356, 1940, especially fig. 3.

# THE BLURRING OF X-RAY IMAGES

by H. A. KLASENS *)

537.531:778.33.022.8

Fluorescent screens cause blurring of the X-ray images. Here the factors governing this screen blurring are investigated. It is of importance that the magnitude of these factors should be expressed in the same terms as that of the geometrical and kinetic blurring. A method is described by means of which screen blurring is compared with kinetic blurring. It appears that the $S$-shaped curve which represents the distribution of intensity in the image of a sharp edge when using a fluorescent screen can be sufficiently characterized for this purpose by a straight line intersecting the $S$-curve at two points at a certain height. This method is applied for determining the screen blurring of a number of intensifying screens and viewing screens. It is also employed to find a formula for the total blurring when various kinds of blurring are combined.

In radiology fluorescent screens are used both for making the X-rays visible and for intensifying their photographic action. In an article published in the preceding number of this journal [1]) this conversion of X-ray energy into light energy was studied and the conditions with which X-ray screens have to comply in their various applications were discussed.

A drawback attaching to the employment of fluorescent screens is that they cause additional blurring of the image. In this article the cause of this screen blurring will be discussed and a method of expressing the magnitude of the effect numerically will be investigated. Further, we shall consider how its effect can be compared and combined with that of two other kinds of blurring, namely geometrical and kinetic blurring.

### The cause of screen blurring

*Fig. 1* gives a magnified representation of the cross section of part of an X-ray screen, in this case an intensifying screen. The fluorescent layer lies between a sheet of cardboard and the sensitized film. The arrows indicate the direction of the incident rays. This illustration applies to the case where the fluorescent layer is used as front screen; it may also be used as back screen, behind the film (as indicated in the article quoted in footnote [1])), which case will be considered farther on.

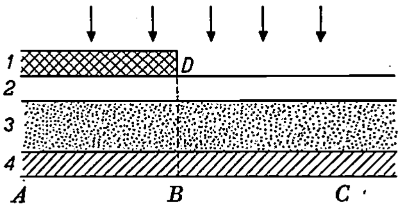


Fig. 1. Magnified cross section of a part of an X-ray screen (intensifying screen) with a lead plate having a sharp edge $D$ laid upon it. *1* is the lead plate, *2* the cardboard base of the fluorescent layer, *3* and *4* the film. The arrows indicate the direction of the incident X-rays.

It is assumed that the screen is partly covered with a sheet of lead of such a thickness as to be practically impermeable for X-rays. We shall now consider what image the edge of the sheet of lead leaves on the film.

The image of the edge will be perfectly sharp when no light falls on the film to the left of $B$. Owing to the finite thickness of the fluorescent layer, however, this is not the case. The crystals in the part of this layer to the right of $BD$ are brought to fluorescence by the X-rays. This light spreads out in both directions and consequently cannot be prevented from reaching that part of the film to the left of the line $BD$.

When we measure the distribution of light along the line $AC$ we obtain a curve as represented in *fig. 2*. The two curves, ($a$) relating to an intensifying screen and ($b$) relating to a viewing screen, are symmetrical with respect to the point $M$, corresponding to point $B$ in fig. 1.

The fact that the light distribution curves are symmetrical with respect to $M$ can be readily understood from the following. We imagine the screen as extending over a large distance to the left and right of the line $BD$. The point $B$ then receives light only from that part of the screen to the right of $BD$ and the luminous intensity at that point is thus half of what it would be if the sheet of lead were removed. Let us now consider the points $A$ and $C$ at equal distances to the left and right of $B$. In addition to the light rays from the right, $C$ receives light from the part above $BC$. The amount that $A$ receives is less than that falling on $B$, by just this quantity, viz. the light that would be generated in the part above $AB$ if the sheet of lead were removed.

*) Material Research Laboratory, Philips Electrical Ltd. London, England.

[1]) H. A. Klasens and W. de Groot, Light emission of X-ray screens, Philips Techn. Rev. 9, 321-329, 1947 (No. 11).

The luminous intensity at $A$ is therefore just as much below the value at $B$ as the intensity at $C$ is above it.

In order to observe more closely the spread of the light generated in the fluorescent screen, let us imagine that the sheet of lead is removed and also the whole of the screen except for one single crystal at a distance $a$ from the film (*fig. 3*).

When irradiated with X-rays this crystal will emit a light flux which we shall call $\Phi$. This causes an illumination at a point on the film of intensity

$$E = \frac{a\Phi}{4\pi\,(r^2 + a^2)^{3/2}},$$

where $r$ is the distance from the base of a perpendicular drawn from the crystal on the film to the point in question.

The curves in fig. 3 represent the variation in luminous intensity for two values of $a$. The larger the value of $a$, the flatter are the curves. The luminous crystal produces on the film a spot of light which becomes more diffused as the crystal is farther removed from the film.

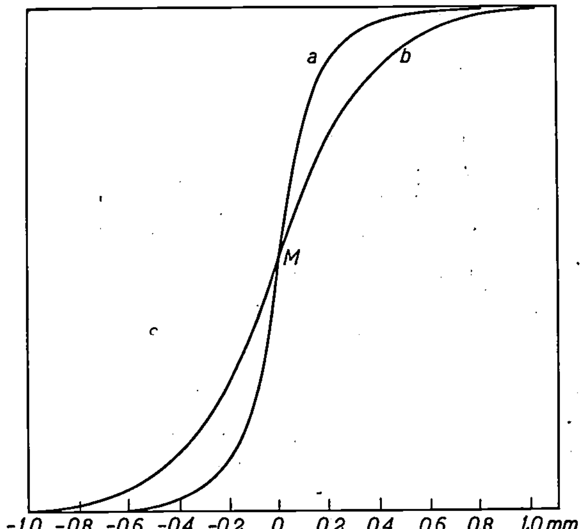


Fig. 2. The luminous intensity curve in the image of the edge of a metal plate on a film obtained with the aid of a fluorescent screen; *a*) when using an intensifying screen of fine-grained calcium tungstate; *b*) with a viewing screen of coarse-grained zinc-cadmium-sulphide. Both curves are symmetrical with respect to point $M$, corresponding to point $B$ in fig. 1.

When a light-absorbing medium is placed between the fluorescent crystal and the film we have a luminous intensity

$$E = \frac{a\Phi}{4\pi\,(r^2 + a^2)^{3/2}}\, e^{-\tau(r^2+a^2)^{1/2}},$$

where $\tau$ represents the absorption coefficient of the medium. The full curves then change into the dotted ones. It can be seen that not only does the luminous intensity decrease more strongly with increasing distance between the crystal and the film but also that the light does not spread out so far.

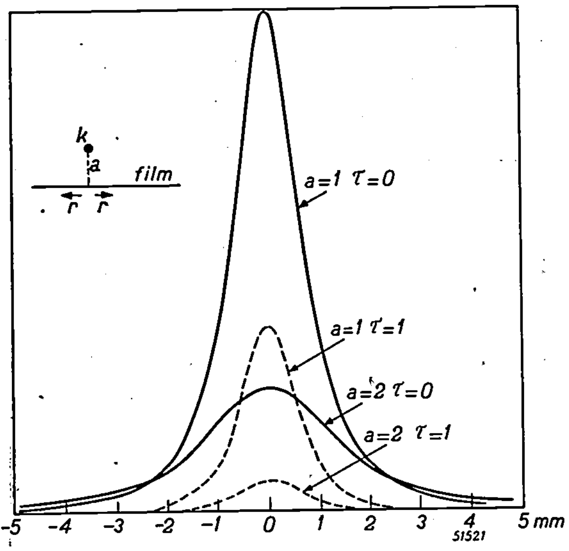


Fig. 3. Distribution of the luminous intensity on the screen resulting from the irradiation of a single crystal at a distance of $a$ cm from the film. The full curves relate to the cases where $a$ is 1 and 2 cm respectively and the absorption coefficient $\tau = 0$. The dotted curves apply to the cases where $a$ is 1 and 2 cm and $\tau = 1$. In the inset $k$ represents the crystal and $r$ the distance from the point of the film in question to the base of the perpendicular drawn from $k$ onto the film.

If we concentrate upon the spreading of the light from one crystal then, to a first approximation, we may regard the action of the rest of the screen as being that of an absorbent medium. Actually, however, the light is not only absorbed but also repeatedly reflected (scattered). The result is, on the one hand, that the actual length of the path travelled by the rays of light in the fluorescent layer is lengthened by repeated reflection, thereby intensifying the action of absorption; on the other hand however other crystals adjacent to the one originally considered act as if they were new sources of light, thus resulting in expansion of the spot of light. The question which of these two influences predominates is not easily decided and, as far as we know, has not been sufficiently investigated theoretically. It has been established in practice that as a rule increased scattering has a favourable effect on the blurring.

The blurring of the image as expressed in the curves of fig. 2 is caused by the spreading of the light of all the irradiated crystals together. Further expanding our theory for the single crystals, we arrive at the following conclusions:

1) The crystals in the screen farthest away from the film contribute most to the blurring. Thus the blurring diminishes the thinner the layer.
2) Blurring can be reduced by increasing absorption in the screen. This can be achieved by:
   a) increasing the actual absorption, for instance by adding a dyestuff to the binder,
   b) intensifying the scattering by using smaller crystals.

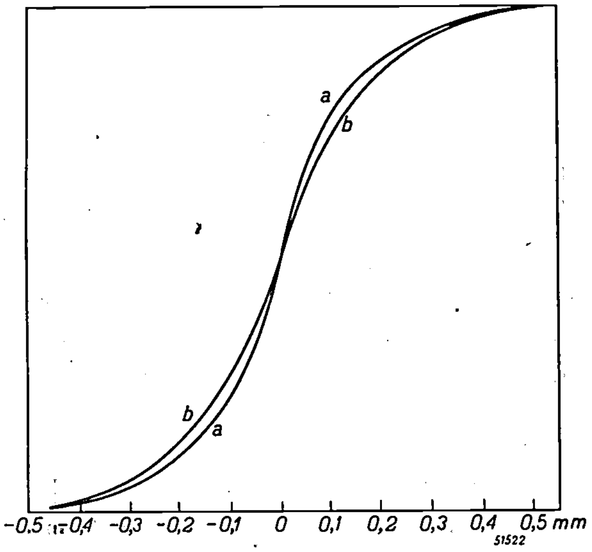


Fig. 4. The luminous intensity distribution in the image of a metal edge on a film obtained with an intensifying screen; *a*) when the screen is used as back screen, *b*) when the same screen is used as front screen. In the former case there is less screen blurring.

All these factors tending to reduce blurring have at the same time the effect of reducing the brightness of the screen, so that any gain in sharpness is always accompanied by a loss in brightness.

Screen blurring depends not only upon the nature of the screen but also upon the manner in which the light is generated in the screen. In the screen the intensity of the X-rays decreases exponentially. Consequently the greater part of the light is generated on that side of the fluorescent layer which is turned towards the X-ray tube. When the screen is used as back screen then that side is in contact with the film. Most of the light therefore comes from a layer which contributes least towards the blurring. When, however, the same screen is used as front screen the reverse is the case, most of the light then coming from a layer farther removed from the film. *Fig. 4* illustrates how this difference affects the light distribution curve of an edge image.

## Measuring screen blurring

Apart from screen blurring, in radiology we encounter geometrical and kinetic blurring [2]). It is therefore of importance to know what happens when all these kinds of blurring occur together. It is above all necessary to be able to express these three blurrings by one and the same measure.

Geometrical blurring is caused by the finite width of the focus. Let us consider again the image cast on a photographic film by an object with a sharp edge. The distribution of energy from the X-rays along a line perpendicular to the shadow of the edge of the object is represented in *fig. 5a.*

A similar energy distribution is also obtained when the object is in motion during the exposure if the geometrical blurring in the radiograph can be ignored, for instance by making the focus small or holding the object close to the film. This is shown in fig. 5*b*. Kinetic blurring is particularly important when photographing the lungs, the heart and the stomach of the human body.

In both cases the energy reaching the screen or the film increases linearly from the minimum value in the shadow corresponding to the irradiation passing through the object to the maximum value in the area freely exposed. It is obvious to define these blurrings by the distance or the transition from the minimum to the maximum energy (*AB* in fig. 5).

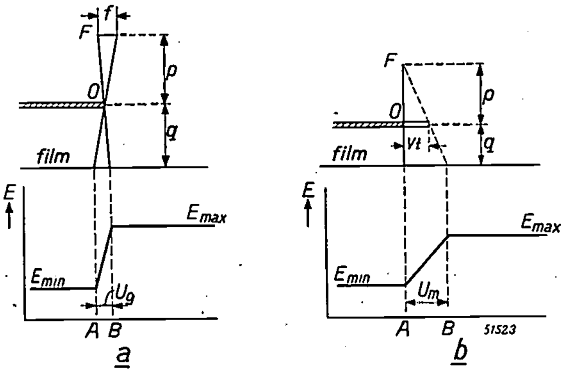


Fig. 5. *a*) Distribution of the energy *E* of the X-rays in the shadow of a stationary object *O*; *f* is the width of the focus *F*, $AB = u_g$ is the geometrical blurring.
*b*) Distribution of the X-ray energy in the shadow of a moving object. $AB = u_m$ is the kinetic blurring. Since the focus is considered as a point, in this case the geometrical blurring is zero.

Given a focus width of *f* mm, a distance *p* from object to focus and a distance *q* from object to film,

[2]) These three kinds of blurring have already been discussed in previous articles in this journal. Cf. Philips Techn. Rev. 5, 270-275, 1940; 8, 32;-329, 1946 (No. 11).

the geometrical blurring is represented by

$$u_g = \frac{q}{p} \cdot f \quad . \; . \; . \; . \; . \; . \; . \; . \quad (1)$$

and the kinetic blurring by

$$u_m = \frac{p+q}{p} \cdot vt \, , \quad . \; . \; . \; . \; . \; . \quad (2)$$

in which $v$ is the speed of motion in a direction perpendicular to the edge of the object and $t$ the exposure time. The geometrical and kinetic blurring can therefore be deduced directly from the conditions under which the photograph is taken.

A clearly defined specification of screen blurring should conform to the same requirements. This is often overlooked. Screen blurring is frequently judged according to the clearness of details in the photograph, but there are many other factors playing a part, for instance the exposure time and the contrast. A photograph with a certain geometrical blurring may appear to be sharper when the contrast is intensified. A difference in the exposure time may entirely alter the character of an image. Spiegler has given some striking examples of this [3]). In order to avoid these difficulties any definition of screen blurring should be based solely upon the intensity distribution of the light in the screen image.

To arrive at such a definition it is best to start once more from the image of the edge of a plate. We then have to develop a method which makes it possible to arrive at a definition of screen blurring which is directly comparable with a blurring having a linear intensity variation.

Nemet, Cox and Walker [4]) have studied the blackening in an edge image on a film obtained with the aid of a fluorescent screen. The geometrical blurring was negligible and there was no kinetic blurring. From the blackening observed they deduced the luminous intensity distribution in the edge image on the screen. This was done while determining at the same time the blackening curve of the film, i.e. the curve indicating the relation between the blackening and the amount of light absorbed. The luminous intensity distribution $E$ found on the screen was represented by an $S$-shaped curve (just as in the figs. 2 and 4) and this has been reproduced in *fig. 6*.

In order to arrive at a definition of the screen blurring these investigators measure the area of the shaded parts $I$ and $II$ in fig. 6. If we had to do with a geometrical or kinetic blurring $u$ then the intensity distribution would not be represented by an $S$-shaped curve but by a straight line. If we indicate the area of $I + II$ by $A$ then in that case the formula applying would be:

$$A = {}^1/_4 \, u \, (E_{\max} - E_{\min})$$

$$\text{or } u = \frac{4A}{E_{\max} - E_{\min}} \, , \quad . \; . \; . \; . \; . \; . \; . \; . \quad (3)$$

where $E_{\max}$ and $E_{\min}$ represent respectively the maximum and the minimum blackening.

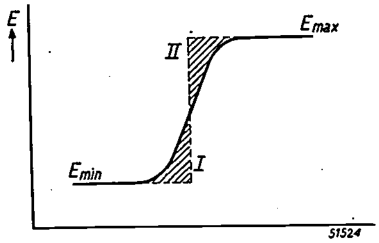


Fig. 6. The luminous intensity $E$ in the image of an edge is distributed, owing to screen blurring, according to an $S$-shaped curve. Nemet, Cox and Walker take the sum $A$ of the areas $I$ and $II$ for defining the screen blurring, putting $u_s = 4A/(E_{\max} - E_{\min})$, where $E_{\max}$ and $E_{\min}$ are respectively the maximum and the minimum luminous intensities in an edge image on a film.

Nemet, Cox and Walker now define the screen blurring $u_s$, in accordance with (3), by the equation:

$$u_s = \frac{4A}{E_{\max} - E_{\min}} \quad . \; . \; . \; . \; . \; . \; . \; . \quad (4)$$

This definition makes it possible to express the screen blurring in the same measure as the geometrical and kinetic blurring. It is not a priori certain, however, that screen blurring and geometrical or kinetic blurring, represented according to this method by the same number, will make the same impression upon the eye.

We have employed an experimental method for comparing screen blurring with kinetic blurring [5]), again using an edge image. The object used was a copper plate 0.2 mm thick, brought into contact with a casette containing the screen to be examined and a panchromatic film, in such a way that the edge of the copper plate was projected onto the middle of the film. The conditions of exposure (the voltage was 70 kV) were so chosen that the density on the film was about 0.4 in the shadow and 0.8

[3]) G. Spiegler, Photogr. J. 83, 410-413, 1943.

[4]) A. Nemet, W. F. Cox and G. B. Walker, Brit. J. Radiology 19, 257-271, 1946 (No. 223). The method followed by these investigators was developed on the basis of a method previously applied by Nitka, who, however, did not first convert the blackening curve into a luminious intensity curve; cf. H. Nitka Phys. Z. 39, 436-439, 1938.

[5]) See also Philips Research Reports 1, 241-249, 1946 (No. 4).

in the freely exposed part, it having been found that with the particular film used the blackening in that range varied practically linearly with the luminous intensity, so that the blackening curve indicated directly the variation of the luminous intensity.

Another piece of the same metal plate was mounted on an iron rod in the manner shown in *fig. 7.*

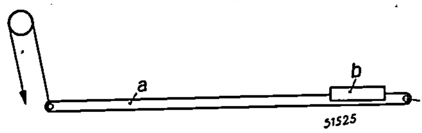


Fig. 7. Device employed to obtain a series of varied kinetic blurrings. $a$ = iron rod, $b$ = copper plate, $c$ = hinge.

This rod was hinged at one end. Thus it was possible to move the plate during exposure over a distance varying from 0.3 mm at one end of the copper plate to 1.2 mm at the other end. No intensifying screen was used, so that there was no screen blurring. By holding the casette with the film close to the copper plate we obtained an image of the edge of the plate without any geometrical blurring and with a kinetic blurring which increased from 0.3 mm at one end to 1.2 mm at the other end. The conditions of exposure were so chosen that the image showed the same maximum and minimum blackening as that of the image which had only screen blurring.

The edge image with screen blurring and that with kinetic blurring were then compared with each other in a viewing lantern. Owing to the difference in the character of the two blurrings this comparison was rather difficult. For instance, when looking at a photograph with kinetic blurring under a very bright light one had the illusion of observing bright and dark lines parallel to the edge. In order to avoid this the brightness of the light in the viewing lantern had to be reduced to below the usual level.

The question to be decided was what parts of the image with kinetic blurring gave the same impression of blurring as the image with screen blurring. Ten observers were asked to give their impressions and from these an average was taken. This investigation was carried out with a blue intensifying screen and a yellow viewing screen, and it was found that the screen blurrings corresponded to kinetic blurring of 0.46 and 0.92 mm repectively.

The $S$-curve in *fig. 8* represents the energy distribution observed in the case of the edge image with a screen blurring of 0.92 mm. The straight line represents the energy distribution in an image with a kinetic blurring of 0.92 mm. A similar drawing has also been made for the screen and kinetic blurrings of 0.46 mm. The $S$-curve and the straight line intersect in both cases close to the points where $E = E_{\min} + 0.16\ (E_{\max}\text{-}E_{\min})$ and $E = E_{\min} + 0.84\ (E_{\max}\text{-}E_{\min})$.

This having been established, it is possible to determine the blurring of other screens by a much quicker method than that outlined above. First the energy distribution in an edge image is determined and plotted for any arbitrary screen. Then a straight line is drawn through the points where $E\text{-}E_{\min} = 0.16\ (E_{\max}\text{-}E_{\min})$ and $0.84\ (E_{\max}\text{-}E_{\min})$. The distance between the points where according to this linear distribution the energy would be exactly equal to $E_{\max}$ and $E_{\min}$ (the points $A$ and $B$ in fig. 8) is then taken as a measure for the screen blurring.

We have applied this method to determine the blurring of a number of commonly used screens. In the case of combinations of intensifying screens (i.e. a front and a back screen) the results appeared to lie between 0.25 and 0.50 mm, and in the case of viewing screens between 0.4 and 1.0 mm.

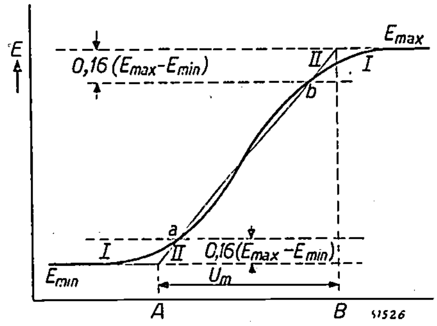


Fig. 8. The $S$-shaped curve $I$ represents the energy distribution in an edge image with a screen blurring of the same magnitude as the kinetic blurring $u_m$ to which the energy distribution according to the straight line $II$ corresponds. The straight line passes through the points $a$ and $b$ for which $E\text{—}E_{\min} = 0.16\ (E_{\max}\text{—}E_{\min})$ and $0.84\ (E_{\max}\text{—}E_{\min})$ respectively. The distance $AB$, which in this case is 0.92 mm, is a measure of the screen blurring.

Thus we have another method for deriving a measure for the blurring direct from the blackening or luminous intensity curve. This method, compared with that of Nemet, Cox and Walker, offers greater certainty of the adequacy of the measure, it having been established experimentally that two photographs having equal blurring according to the present definition, e.g. screen blurring and geometrical blurring, make the same impression upon the eye.

The Nemet, Cox and Walker method would have yielded slightly higher values for the blurring of the screens referred to, but the difference is of no great importance.

### Combination of screen blurring with other kinds of blurring

Screen blurring can easily be combined with geometrical and kinetic blurring by holding the copper plate a certain distance away from the casette containing the film and screen and moving it during exposure by means of the device shown in fig. 7. The total blurring then resulting can be measured by the method described. In this manner a formula can be deduced for determining the resulting blurring when various kinds of blurring are combined.

For screens with little blurring, such as those usually employed in normal contact photography, the blurring observed can well be represented by the formula

$$u = (u_g^3 + u_m^3 + u_s^3)^{1/3}, \quad . \quad . \quad . \quad . \quad (5)$$

where $u_g$ is the geometrical blurring
$u_m$ the kinetic blurring and
$u_s$ the screen blurring.

The same formula applies also for a combination of geometrical and kinetic blurring ($u_s = 0$).

In the case of viewing screens with a blurring of the order of 0.7 mm the value observed is usually slightly higher than that obtained by calculation from equation (5), the average deviation being of the order of 5%. In practice such differences are scarcely noticeable.

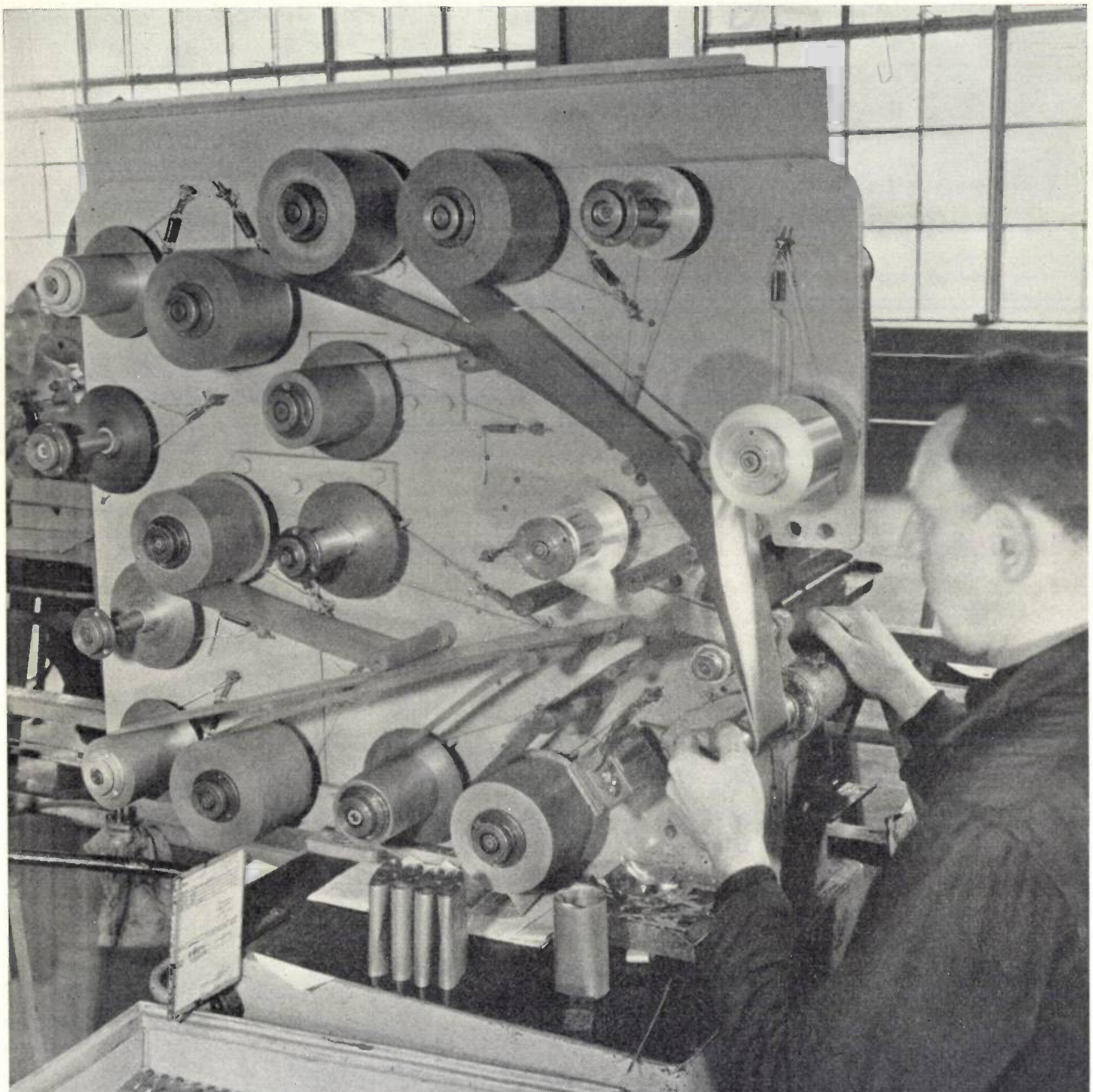
We have also investigated whether the formula used by Newell [6]), viz.

$$u = (u_g^2 + u_m^2 + u_s^2)^{1/2} \quad . \quad . \quad . \quad . \quad (6)$$

properly represents what is actually observed. It appeared that the resulting blurring is always less than the square root of the sum of the squares of the component blurrings. The average deviation, however, is not greater than 10%. This means that the formula (6) likewise gives a reasonable approximation of the edge blurring, although the deviation is much greater than that of the formula (5).

[6]) R. R. Newell, Radiology 30, 493-499, 1938.

# THE MANUFACTURE OF PAPER CONDENSERS



This photograph shows how paper condensers are wound. The outer covering of alumiuium foil is fed from the two polished rollers. The dielectric consists of layers of paper, which are fed from the other rollers; the number of layers and the thickness of each depend upon the voltages for which the condensers are made. Here five layers are being wound. A second set of layers is required to prevent the coverings from making contact with each other. The circuiting contacts are made of strips of copper foil, seen underneath the left hand of the winder. When the counter (to the left of the man's left hand) shows the number of turns required for the particular capacity, the coil is taken off the pin and pressed flat, giving it the shape seen in the tray at the bottom. In the next stage the coils are provided with the electrical connections and placed in a metal box, which is then inpregnated under vacuum and sealed by soldering.

# MEASURING REVERBERATION TIME BY THE METHOD OF EXPONENTIALLY INCREASING AMPLIFICATION

by W. TAK

534.844.1:621.317.351:621.314.3

The reverberation time of an enclosed space can be measured by producing in that space periodical sound impulses and conducting the voltage from a microphone set up within the space to an oscillograph *via* an exponential amplifier. The amplification begins every time with a low value but increases exponentially with the time by a certain amount, say 60 db, and then drops back to the initial low value. The rhythm of the sound impulses and of the exponential variation in amplification is determined by the generator of the time-base voltage of the oscillograph. The frequency of this voltage is regulated in such a way that the oscillogram shows as far as possible a constant amplitude; when that is the case then the period corresponding to that frequency is approximately equal to the reverberation time sought. Various parts of the apparatus and some measuring results are described. At the same time it is demonstrated experimentally that when a sound is produced it excites mainly the characteristic vibrations of the space with the adjacent frequencies and that beats arise between those frequencies.

## Introduction

The phenomenon of reverberation is of great importance when judging the acoustics of halls and rooms. The theory of this phenomenon and of the measuring of the reverberation time has been fully discussed in an article [1]) especially devoted to that subject. Two methods of measuring have been particularly discussed, each of which allows of a fairly complete quantitative investigation of reverberation. In the meantime one of these methods, *viz.* that of exponential amplification, has been further developed and employed in a large number of tests. The present article describes the apparatus used and deals with some results obtained. First of all, we shall briefly outline once more the origin of reverberation, giving also the definition of reverberation time and dealing with the principle of the method of exponentially increasing amplification for measuring that time.

Let us imagine that a source of sound is active in one single frequency in an enclosed space during an indefinite length of time. A state of stability will then obtain. By this is understood the following: the sound pressure in the whole of the space as a function of time will be given by a sine function with the frequency of the source; the distribution of the sound pressure in space is a superposition of standing waves corresponding to the characteristic vibrations of the space in question. The shape and dimensions of the space determine what standing waves may occur in a given space. Which of these standing waves will be represented in the said superposition (in other words, which characteristic vibrations will be excited) and to what extent they will occur depends upon the frequency, the position and the shape of the source of sound. There will certainly not be present the standing waves which at the place where the sound source is situated (if this may be regarded as a point) would show a node. The fact that among the characteristic vibrations represented there will always be several yielding a considerable contribution towards the superposition is evident when one remembers (see the article quoted in footnote [1])) that the successive characteristic frequencies show very small differences.

After the source of sound is switched off the intensity of the sound of each of the characteristic vibrations excited will die out exponentially owing to absorption by the walls. This sound heard after switching off the source is called the reverberation. The sound pressure $p$ will then be given as function of the time $t$ by the equation:

$$p = p_0 \cdot e^{-k't} \cdot \sin(2\pi\nu_e t + \varphi); \quad . \quad . \quad (1)$$

where $p_0$ is the amplitude of the sound pressure before the sound source is switched off, $k'$ is a constant which increases with the absorption by the walls, $\nu_e$ is the characteristic frequency in question and $\varphi$ is a phase angle the magnitude of which depends upon the moment of switching off.

The reverberation time $t_{60}$ is defined as the time it takes for the sound intensity to drop 60 db, *i.e.* by a factor $10^6$. Now this intensity is proportional to the square of the amplitude of the sound pressure. From equation (1) it therefore follows that

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

or

$$t_{60} = \frac{6.9}{k'} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

[1]) W. Tak, The measurement of reverberation, Philips Techn. Rev. 8, 82-88, 1946 (No. 3).

One of the methods for measuring reverberation time, viz. that of exponentially increasing amplification, is based on the following principle. The voltage induced by the reverberation by means of a microphone is amplified by an amplification which increases with time according to $e^{at}$, where $a$ has a known and variable positive value. The amplified voltage is observed with a cathode-ray oscillograph, on the screen of which a sinusoidal line is to be seen whose amplitude increases or decreases according as $a > k'$ or $a < k'$ (see fig. 8 of the article quoted in footnote [1])). By adjusting $a$ so as to keep the said amplitude constant, thus by making $a$ equal to $k'$, one finds from equation (2) the reverberation time $t_{60}$. Generally the constant $k'$ is a function of the frequency, as is also the reverberation time. In order to determine the acoustics of a space it is therefore necessary to carry out successive measurements with a series of frequencies in the audible range.

The foregoing applies for the hypothesis assumed in the beginning where a source of sound is acting with one single frequency for an indefinite length of time. Actually such a condition will seldom occur with speech or music. Rather one has to do with series of sound impulses varying in duration and frequency, these impulses being so short as to preclude a state of stability. In order to approximate the condition existing with speech or music, in such a way that reproducible results may be expected, one can proceed in various ways. For instance the frequency of the sound can be modulated, or else the sound can be made to consist of short impulses of a certain tone. Since we consider the latter to be the best imitation of music or speech we have employed sound impulses when carrying out our measurements. Owing to the large number of characteristic vibrations excited by every impulse and the beats between these vibrations, the shape of the oscillograms becomes very complicated. We shall revert to this in the discussion of the measuring results.

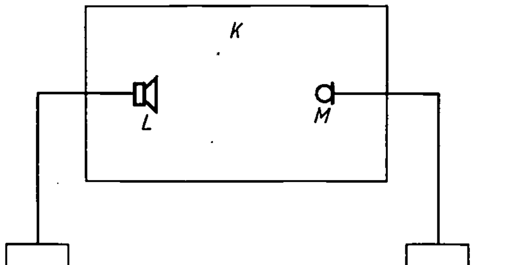


Fig. 1. In an enclosed space $K$ where the reverberation time is to be measured a loudspeaker $L$ and a microphone $M$ are set up. The loudspeaker is connected *via* an amplifier $A_1$ and an impulse exciter $I$ to a tone generator $G$. The microphone is connected to an oscillograph $O$ *via* a pre-amplifier $A_2$ and an amplifier $A_{\text{exp}}$ with exponentially increasing amplification. The generator $T$ of the time-base voltage for the oscillograph also controls the impulse exciter and the exponential amplifier.

### Description of the measuring apparatus

*Fig. 1* shows diagrammatically of what components the measuring apparatus consists and how these are interconnected.

The microphone, set up in the space whose reverberation time is to be measured, is connected to a cathode-ray oscillograph via a pre-amplifier and the exponential amplifier to be discussed below. As source of sound a loudspeaker is used which is connected to a generator via an output amplifier and an impulse exciter, the latter serving to switch the loudspeaker on and off periodically so as to givc a series of sound impulses.

Each sound impulse has to be synchronized with the beginning of the exponential amplification and the beginning of the movement of the luminous spot across the screen of the oscillograph from left to right. As soon as the amplification has reached its final value it has to return as quickly as possible to its initial value and the spot from the right side of the screen to the left side to start a new cycle.

What takes place is illustrated in *figs. 2a-d,* showing successively the sound intensity produced ($I_L$) and the sound intensity received ($I_M$), the logarithm of the amplification $g$ of the exponential amplifier, and the horizontal deflection $x$ on the oscillograph, all as functions of the time $t$. When the ratio of the final value $g_T$ of the amplification to the initial value $g_0$ has been definitely adjusted so as to correspond to 60 db and the duration $T$ of the cycle is so chosen that *the amplitude $y$ of the oscillogram* (fig. 2*e*) *remains constant,* then $T$ is the reverberation time $t_{60}$ that is sought. Bearing in mind that the regulation of $T$ practically means giving the right slope to the line $\ln g = \mathrm{f}(t)$ (fig. 2*c*), it is evident that the result is independent of both the duration of the impulse $\tau$ (fig. 2*a*) and the time $t'$ (fig. 2*b*) the sound takes to travel the shortest path from the loudspeaker to the microphone.

If, for instance, $t'$ is increased to $t''$ (fig. 2*b*) by setting up the microphone farther away from the loudspeaker, then the oscillogram will begin at $t''$ but the same cycle $T$ will be found during which the amplitude remains constant.

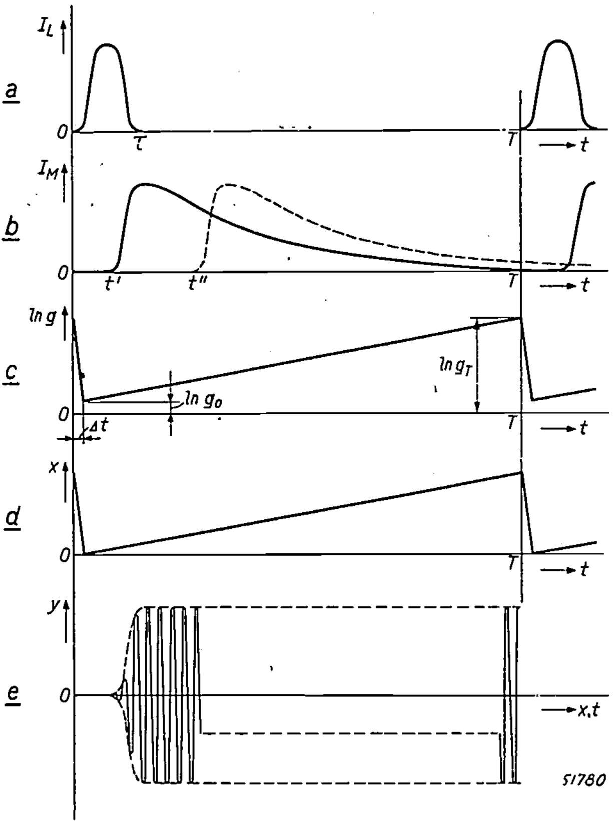


Fig.2.
*a*) The sound intensity $I_L$ at the loudspeaker,
*b*) the sound intensity $I_M$ at the microphone,
*c*) the logarithm of the amplification $g$ of the exponential amplifier,
*d*) the horizontal deflection $x$ and
*e*) the vertical deflection $y$ on the oscillograph, all plotted as function of the time $t$.
$T$ = duration of the cycle, $\tau$ = duration of the sound impulse, $t'$, $t''$ = time required for the direct sound to travel from the loudspeaker to the microphone, $\Delta t$ = time taken for the amplification of $A_{\text{exp}}$ to return from the final value $g_T$ to the initial value $g_0$.

Only the interval $\Delta t$ (fig. 2*c*) required by the amplifier to return from the final value to the initial value of amplification should be deducted from $T$, but this correction is so insignificant as to be negligible.

The aforementioned synchronization between the various phenomena is obtained by controlling from one point both the impulse exciter and the exponential amplifier and the horizontal deflection on the oscillograph. This central control point is the generator of the linear time-base voltage ($T$ in fig. 1).

We shall now consider in succession this time-base voltage generator, the exponential amplifier and the impulse exciter.

### *The time-base voltage generator*

In essence this consists of a capacitor gradually charged via a resistor from a direct voltage source $E$ and rapidly discharging via a relay valve, the charging and discharging alternating periodically. The wiring diagram is given in *fig. 3* and explained in the text below it [2]). If the resistor through which the charge flows were an ordinary resistor then the capacitor voltage $v_C$ would rise exponentially with $t$, but by using a pentode as resistor and causing it to work in the range where the anode current $I_0$ is independent of the anode voltage a constant charging current is obtained, so that $v_C = I_0t/C_0$ changes linearly with $t$ ($C_0$ = capacitance of the capacitor). The same applies for the voltage $v_P$ on the pentode, $= E—I_0t/C_0$.

The maximum amplitude of $v_C$ is limited to a value which is a certain amount smaller than $E$; it depends in fact upon the striking voltage of the relay valve. With the aid of a variable direct voltage on the grid of the relay valve (see fig. 3) the striking voltage is so adjusted as to make the amplitude of $v_C$ correspond to the voltage required to carry the luminous spot across the whole width of the screen of the oscillograph.

The period of the relaxation oscillation of the system described must, as we have seen, be equal to the reverberation time of the space under investigation. This reverberation time may be anything between about 0.1 sec (small, strongly damped room) and a number of seconds (large, cavernous spaces, *e.g.* a church.) The frequency of the relaxation oscillation must therefore be variable. Its cycle consists mainly of the charging interval of the capacitor, which is charged all the quicker

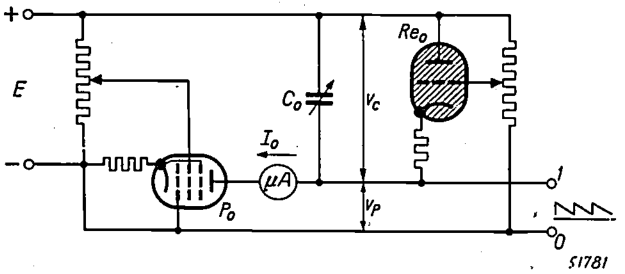


Fig. 3. Circuit for generating a linear time-base voltage ($v_P$). The capacitor $C_0$ is charged *via* the pentode $P_0$ with the anode current $I_0$, which is independent of the anode voltage, so that the voltage $v_C$ on $C_0$ increases linearly with time. When $v_C$ reaches the striking voltage of the relay valve $Re_0$ then $C_0$ discharges across $Re_0$ in a short time, after which it is charged again, and so on. The voltage $v_P = E—v_C$ is taken from the terminals *0-1*.

[2]) The same circuit is used for the time-base of the cathode-ray oscillograph type GM 3156 (see Philips Techn. Rev. 5, 289-297, 1940). The lowest frequency obtainable with this oscillograph (1/4 c/s) is in many cases still too high for our purpose.

according as $C_0$ is smaller or $I_0$ is greater. In the practical construction of the apparatus these two quantities have been made adjustable: $C_0$ in coarse and fine stages, and $I_0$ continuously with the aid of the screen grid voltage of the pentode (see fig. 3). Before measuring is started the screen grid voltage is adjusted so that $I_0$ has a value fixed once for all. The actual measuring, i.e. making the vertical amplitude on the oscillograph constant, is done by regulating $C_0$.

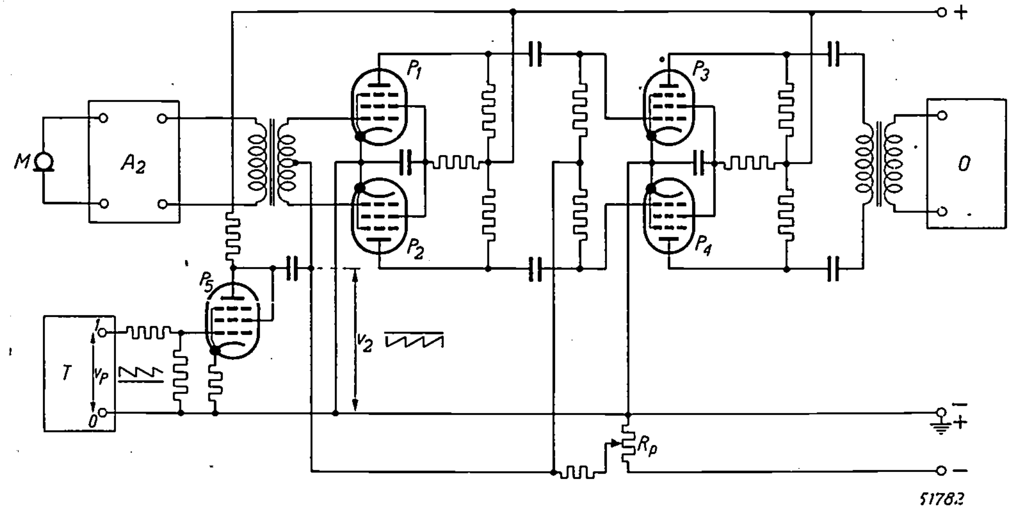


Fig. 4. Diagram of the exponential amplifier. The control grids of the two push-pull stages with the pentodes $P_1$-$P_2$ and $P_3$-$P_4$ receive a variable bias $v_2$ consisting of a fixed negative direct voltage derived from the potentiomenter $R_p$ and a "sawtooth" voltage supplied by the pentode $P_5$. The control grid circuit of $P_5$ is connected via a voltage divider to the terminals *0-1* of the time-base voltage generator (fig. 3). The rest of the symbols have the same meaning as in fig. 1.

### *The exponential amplifier*

In the exponential amplifier pentodes are employed whose anode current in a certain range is an exponential function of the control grid voltage. When a certain range of this voltage is traversed linearly with time then the anode current — and likewise the mutual conductance — changes with time according to an exponential law. The same applies for the amplification factor of an amplifier in which such pentodes are used. Pentodes possessing this property are, for instance, those of the types EF 5 and EF 41.

Measurements taken on a push-pull amplifier with two stages equipped with these pentodes have shown that when the control grid voltage is changed simultaneously in both stages the limits of the range just mentioned correspond to values of the amplification lying about 50 db apart. The definition of reverberation time ($t_{60}$) is, it is true, based upon a fall of 60 db in the sound level, which should be compensated by an equally large increase in the amplification, but, as will be seen from the following, this is neither necessary nor desirable.

Starting from the highest practicable intensity of the sound impulses, one finds in practice that after having been attenuated about 40 db the reverberation has already reached a level at which various interferences become troublesome. These interferences consist partly of the noise of the amplifier and partly of sounds having nothing to do with the measurement. For this reason it is not desirable that the amplification should increase by more than 40 db. And this is in fact sufficient, since the time $t_{40}$ in which the reverberation decreases 40 db in intensity is correlated in a simple manner to the time $t_{60}$ that is sought:

$$t_{40} = \frac{2}{3}\, t_{60}.$$

(This may be seen at once, for instance, from fig. 7 of the article quoted in footnote [1]).) On the other hand it is not desirable to let the reverberation drop much less than 40 db, because then in the conversion to $t_{60}$ the inevitable measuring errors would be increased by too large a factor.

We therefore decided to work upon a drop of 40 db in the reverberation and consequently an increase of 40 db in the amplification. With the abovementioned two-stage push-pull amplifier this increase of 40 db is obtained by causing the negative grid voltage to drop by about 32 V. To ensure that the amplification starts and stops at the right moments the grid voltage variation is brought about by means of the time-base voltage generator. How this is done is shown in *fig. 4*. This generator supplies to terminal *1* a voltage which is highly positive with respect to terminal *0* (earth) and which gradually drops to a lower positive value. On the control grids of the pentodes $P_1 \dots P_4$ in the exponential amplifier. on the other hand a voltage is required ($v_2$, fig. 4) which rises from a highly negative to a less negative value. In order to derive the second voltage variation from the first one ($v_P$, fig. 4) a valve $P_5$ is introduced whose output voltage is superposed upon a direct voltage drawn from a potentiometer $R_P$; this sum ($v_2$) serves as a grid bias for $P_1 \dots P_4$. The amplitude of $v_2$ and the direct voltage referred to are so chosen as to give the amplification variation of 40 db; to make sure that this remains so, the feeding voltages have to be carefully stabilized.

It is to be added that the frequency characteristic of the amplifier is a horizontal straight line in the audible range of about 50-10 000 c/s.

### *The impulse exciter*

The impulse exciter is in fact an amplifying stage introduced between the tone generator and the final-stage amplifier (see fig. 1) and biassed in such a way as to be alternately either blocked or caused to act as an amplifier, so that the loudspeaker produces a series of sound impulses.

This varying bias is likewise drawn from the time-base voltage generator, so that the rhythm of the sound impulses corresponds to the frequency of the time-base voltage. This time-base voltage is first subjected to a certain distortion so as to satisfy the requirement of being able to regulate the duration of each impulse ($\tau$, fig. 2*a*) within certain limits, and preferably without changing the maximum sound intensity. This requirement is made for the following reason.

A long impulse (of the order of the reverberation time) is undesirable, because during that time no reverberation can be observed and the width on the oscillograph screen corresponding to $\tau$ is therefore lost; $\tau$ must consequently be kept small with respect to the reverberation time. On the other hand very short impulses would not give sufficient volume of sound, particularly in large halls. The most favourable compromise must therefore be found.

The circuit applied by us, which answers these requirements, is represented in *fig. 5*. The resistor $R_1$, the capacitor $C_1$ and the relay valve $Re_1$ form a circuit which under the influence of the time-base voltage $v_P$ produces a relaxation oscillation with the same frequency as that of $v_P$. The voltage on $C_1$ has a trend analogous to that of $v_P$ but in the reverse sense (obliquely rising instead of falling "saw-teeth") and controls a pentode $P_8$. In the anode circuit of $P_8$ is a filter consisting of a system

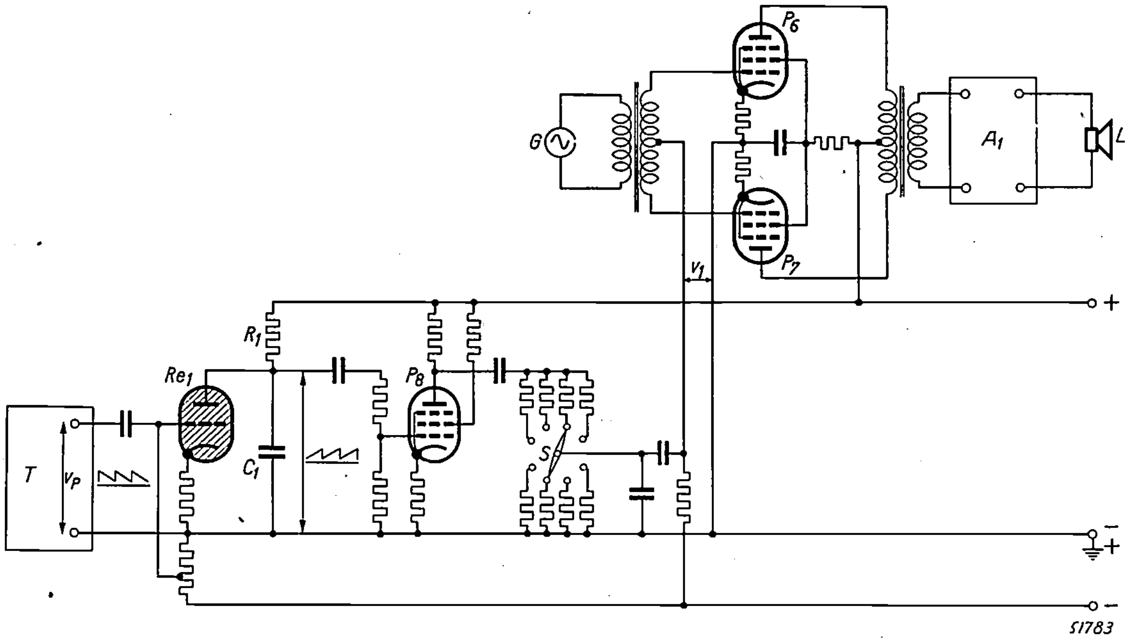


Fig. 5. Circuit diagram of the impulse exciter. The voltage $v_P$ of the time-base voltage generator $T$ (fig. 3) controls a relaxation oscillation in the circuit formed by the resistor $R_1$, the capacitor $C_1$ and the relay valve $Re_1$. The sawtooth voltage on $C_1$ is amplified by the pentode $P_8$ and distorted in a filter of capacitors and resistors into a periodical impulse which together with a fixed negative direct voltage forms the bias $v_1$ of a push-pull amplifier with the pentodes $P_6$ and $P_7$. The impulse width and thus also the time during which the amplifier is active are regulated by means of the switch $S$. For the rest of the symbols see fig. 1.

of capacitors and (partly variable) resistors. The output voltage of this filter system together with a fixed direct voltage forms the bias $v_1$ of the impulse exciter proper. The variation of $v_1$ with time is shown in *fig. 6* for the two extreme positions of the switch $S$ indicated in fig. 5. It is seen that a more or less rounded off voltage impulse arises which during a time variable between $\tau_{\min}$ and $\tau_{\max}$ exceeds the voltage $V_0$ at which the impulse exciter is blocked. The amplitude of the amount by which the impulse $v_1$ exceeds the value $V_0$ is practically independent of the position of the switch; this amount determines the amplification of the impulse exciter and thus ultimately the maximum sound intensity. This satisfies the requirement that sound impulses should be produced with a variable duration without affecting the maximum intensity.

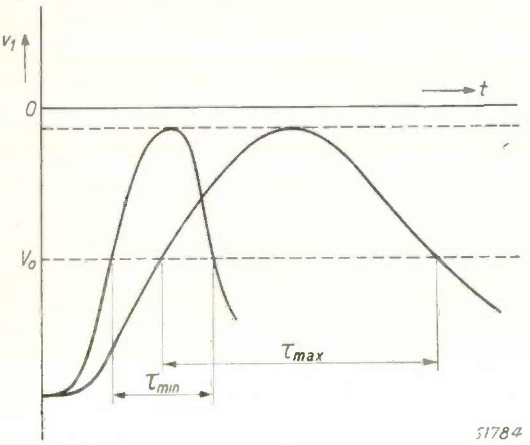


Fig. 6. Trend of the bias $v_1$ generated in the system of fig. 5 The time during which this bias exceeds the grid voltage $V_0$ at which the valves $P_6$ and $P_7$ (fig. 5) are blocked is variable between $\tau_{\min}$ and $\tau_{\max}$.

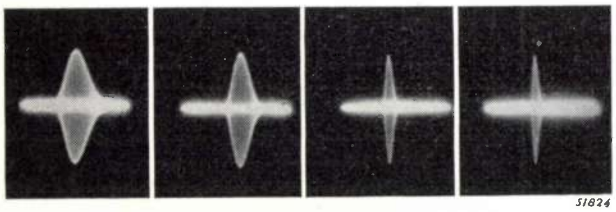


Fig. 7. Oscillograms of sound impulses obtained with the system according to fig. 5. The duration of the impulses is respectively 0.12, 0.07, 0.035 and 0.02 seconds.

*Fig. 7.* shows the oscillograms of sound impulses with a duration of 0.12, 0.07, 0.035 and 0.02 seconds.

### *Practical construction of the apparatus*

The arrangement given diagrammatically in fig. 1 may consist partly of a number of normal apparatuses, viz. the tone generator, the final-stage amplifier, the oscillograph, the pre-amplifier and further the loudspeaker and the microphone. Except for the two last-mentioned components and the pre-amplifier, the others are mounted on a mobile rack (*fig. 8*) which also carries a cabinet containing the time-base voltage generator, the exponential amplifier and the impulse exciter. A separate cabinet contains the apparatus required for supplying the various components with stabilized direct voltages [3]). Finally there is an auxiliary oscillograph on which a Lissajous figure can be produced by the voltage from the tone generator and the a.c. mains. This makes it easy to adjust the tone frequency exactly to a multiple of the mains frequency, and if the latter may be regarded as constant then one is sure that the same frequency has been used when taking the various measurements, so that the results may be more comparable.

A measurment is carried out in the following way. When the apparatus described has been started up and the tone frequency given the desired value, the capacitance $C_0$ in the time-base voltage generator (fig. 3) is regulated so as to keep the amplitude of the oscillogram as nearly constant as possible. When this has been done the duration of the reverberation can be read from the position of the controls for regulating $C_0$.

### Measuring results

Obviously when it comes to practical application matters will not be so simple as they have just been outlined above, considering that an enclosed space has a large number of characteristic vibrations all of which are more or less excited by the sound impulses and beats will arise between those vibrations. (Experimental confirmation of this will be referred to later on.) The picture on the oscillograph will never have such a simple shape as that shown in fig. 2*e*, but will always show a number of peaks. It will be seldom that an adjustment can be found to give peaks of equal size, but after some experience there will be no doubt about what is the best adjustment. For instance *figs 9a, b* and *c* are recordings taken under identical conditions with $T$ values of 1.3, 1.1 and 0.9 second respectively, and there the best adjustment is $T = 1.1$ second.

Of course it is more difficult in cases where one has to do with two or more average reverberation times, as occurs for instance when a "soft" room (with short reverberation) is in communication *via* an open door with a "hard" corridor (long

[3]) H. J. Lindenhovius and H. Rinia, A direct current supply apparatus with stabilised voltage, Philips Techn. Rev. 6, 54-61, 1941.

reverberation). *Fig. 10* shows two recordings made in such a case.

Finally a description is given of an experiment confirming the fact that sound impulses excite a series of characteristic vibrations of the room in which they are produced and that beats arise between those vibrations, as mentioned in the foregoing and in the article quoted in footnote [1]). The apparatus used in this experiment was the same as that previously described.

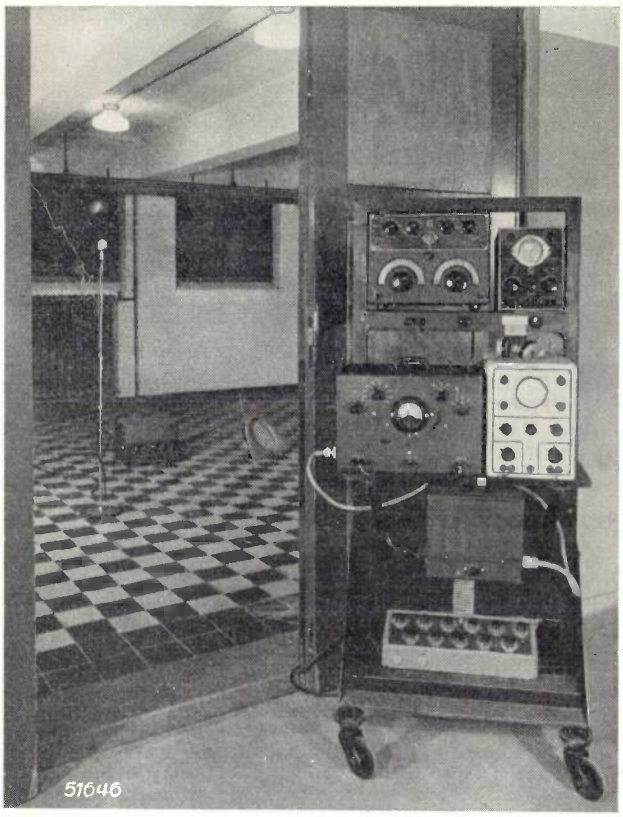


Fig. 8. Arrangement of the measuring apparatus. The mobile rack carries the following apparatus:

*Top row:* to the left the tone generator (type GM 2307), to the right an auxiliary oscillograph (type GM 3153) for adjusting the tone frequency to a multiple of the mains frequency.

*Second row:* to the left a cabinet containing the time-base voltage generator, the exponential amplifier and the impulse exciter; to the right the oscillograph (type GM 3159); underneath this the supply apparatus for the stabilized direct voltages.

*At the bottom* : the 24 W output amplifier (type 2843).

In the background in the room where the reverberation time is being measured are the microphone, the pre-amplifier (type 2843) and the loudspeaker.

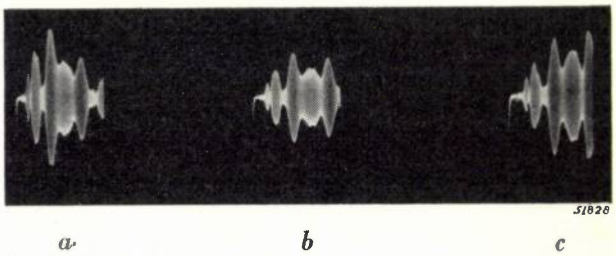


Fig. 9. Since there are always a number of characteristic vibrations excited simultaneously, the oscillogram does not show the simple form of fig. 2*e*, but has a number of peaks and valleys which make it difficult to find the right adjustment. Recordings have been made, under otherwise the same conditions, with $T$ = (*a*) 1.3 sec, (*b*) 1.1 sec, (*c*) 0.9 sec. The adjustment (*b*) is the best.

First the impulse exciter was put out of action, so that the loudspeaker produced an uninterrupted note. Also the exponential amplifier was temporarily out of action; a constant amplification was applied between the microphone and the oscillograph tube. This tube received no voltage for the horizontal deflection, so that the image consisted only of a small vertical line the length of which was proportional to the sound pressure at the microphone. This line was photographed on a continuously moving film, whilst the frequency of the tone generator was gradually changed. The result was a picture of the sound pressure (at the microphone) as a function of the frequency (*fig. 11*). The frequency scale was fixed by interrupting the pick-up circuit every time a certain frequency was passed.

Fig. 11 shows that in the room where the experiment was carried out and with a given arrangement of loudspeaker and microphone few pronounced characteristic frequencies occur in the range between 1020 and about 1065 c/s, but that they do decidedly occur between about 1065 and 1120 c/s.

What conclusions are therefore to be drawn from this regarding the reverberation that will occur when sound impulses are produced in this room with different pitches within the frequency range traversed in this experiment and with different durations of the impulses?

Let us first give the tone a frequency equal to a characteristic frequency of the room lying roughly midway between two adjacent characteristic frequencies, for example 1072 c/s (see fig. 11). It may

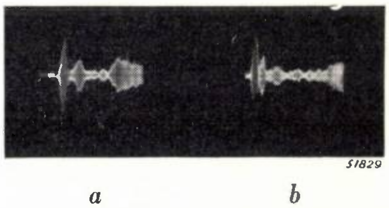


Fig. 10. Reverberation in a room communicating with a corridor through an open door. (*a*) was recorded with $T$ = 0.9 sec, (*b*) under otherwise the same conditions but with $T$ = 2.3 sec. In the case of (*a*) the speed of the amplification practically balances with the expiration of the reverberation in the room but not with the slower expiration in the corridor. It is seen that after some time the deflection increases, due to predominance of the reverberation in the corridor. In the case of (*b*) $T$ was so chosen as to obtain a practically constant amplitude of the reverberation in the corridor. Here, too, there is an increase of amplitude towards the end, but this is due to the noise of the amplifier.

then be expected that this characteristic frequency as well as the two adjacent ones (1069 and 1076 c/s) will be strongly excited, whilst more remote vibrations (*e.g.* 1085 c/s) will be much less excited, and that beats will arise with a frequency that is the difference between the note frequency and that of the first-mentioned vibrations (3-4 c/s).

This is confirmed by *fig. 12a*, recorded with the whole of the measuring apparatus described in this article in action, thus including impulse exciter, exponential amplifier and time-base. Here we see a more or less sinusoidal pressure amplitude varying with time and showing exactly two cycles in the time-base period. In this experiment the time-base period was 0.54 second, so that the frequency of the amplitude variation was 2/0.54 = 3.7 c/s, which corresponds well with the frequency differences between the pitch used (1072 c/s) and the two adjacent characteristic frequencies. The phenomenon represented in fig. 12*a* may therefore be ascribed to beats between these frequencies.

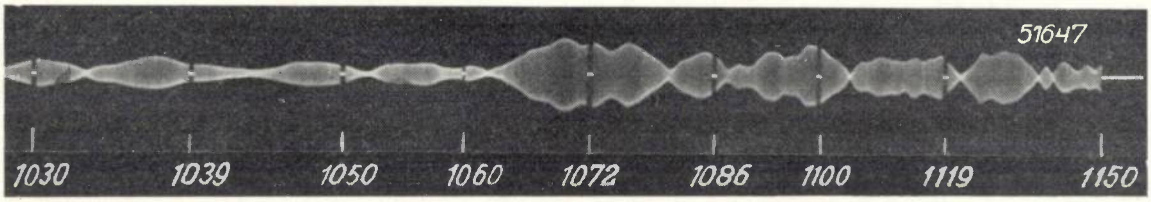


Fig. 11. Amplitude $p_0$ of the sound pressure at a point in a room as function of the frequency of a continuous note (in cycles per second).

If we now choose 1060 c/s as tone frequency then, as may be seen from fig. 11, the nearest pronounced characteristic frequencies are the same three as just mentioned (1069, 1072 and 1076 c/s). In the oscillogram of the reverberation (*fig. 13a*) we now recognize — though less clearly than in fig. 12*a* — a frequency of about 14 c/s, which again corresponds, within the accuracy of measurement of the frequencies, to the difference between the pitch used and the adjacent characteristic frequencies. The same is the case with a tone of 1086 c/s (*fig. 14a*), where according to fig. 11 one would expect excitation not only of the group around 1072 c/s but also of the characteristic frequencies at 1093 and 1100 c/s, and therefore, in the image of the reverberation, beat frequencies of 7 and 17 c/s, among others. These two interferences are indeed to be recognized in fig. 14*a*.

Figs 12*a*, 13*a* and 14*a* have all been recorded with a fairly long duration of the sound impulses, $\tau$ = 0.12 sec, so that each impulse covered more than 100 cycles. The images obtained with shorter impulses are given in the same figures above *b*, *c* and *d* where the values of $\tau$ were respectively 0.07, 0.035 and 0.02 sec. The shorter the duration of the impulses, the more complicated is the structure of the oscillograms, which points to a larger number of excited characteristic frequencies. This is not surprising considering that in the Fourier spectrum of the sound impulse more terms are involved according as the impulse is shorter.

A change of the impulse duration $\tau$ can therefore contribute towards giving the oscillogram such a shape as to make it easier to find the right adjustment of the period *T*. This was in fact another reason why $\tau$ was made variable.

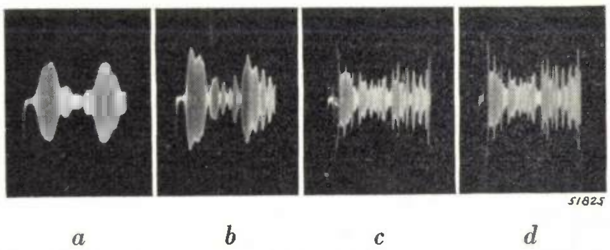


Fig. 12. Oscillograms recorded with the apparatus according to fig. 8. Duration of the sound impulses: (*a*) 0.12 sec., (*b*) 0.07 sec., (*c*) 0.035 sec., (*d*) 0.02 sec. Pitch 1072 c/s. The width of the oscillogram corresponds to 0.54 sec. Since it was not necessary in this case to adjust as accurately as possible the constant amplitude, this was only done approximately.

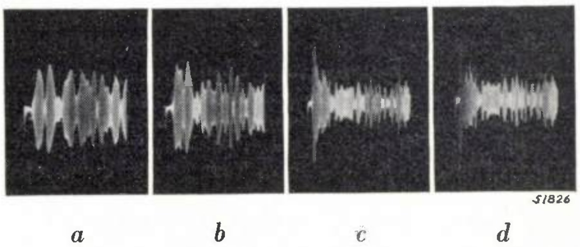


Fig. 13. The same as fig. 12 but with a tone of 1060 c/s.

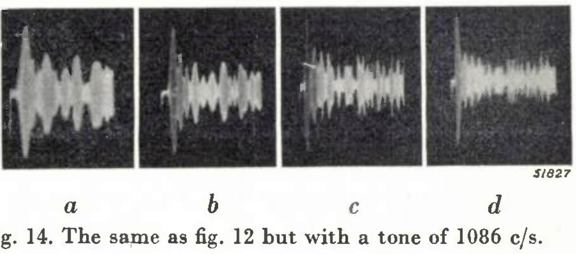


Fig. 14. The same as fig. 12 but with a tone of 1086 c/s.

# 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 asterik.

**R 32:** B. D. H. Tellegen: Coupled Circuits (Philips Res. Rep. 2, 1-19, 1947 No. 1)

The theory of coupled circuits leads, in the determination of the frequencies and dampings of the free oscillations, to an equation of the fourth degree. If we confine ourselves to cases in which the resonance curve has a small relative width, that equation can be reduced to one of the second degree which can be resolved into factors in a simple manner. Each of the factors determines one of the free oscillations. Inductive or capacitive coupling leads to a real coupling factor $k$, resistance coupling to an imaginary $k$, mixed coupling to a complex $k$. If the circuits are equally damped and equally tuned, then, when $k$ is real, the factors have çqual dampings but different characteristic frequencies; when $k$ is imaginary they have equal characteristic frequencies but different dampings; when $k$ is complex they have both different characteristic frequencies and different dampings. If the circuits are equally damped but differently tuned the effect on the shape of the resonance curve is the same as that of an increase of $k^2$. If the circuits are equally tuned but unequally damped $k^2$ is apparently diminished. If the circuits are differently tuned and unequally damped $k^2$ is apparently altered by a complex amount. All symmetrical resonance curves can be drawn in a single family. The height of the resonance curve is also investigated. Finally, systems are considered in which the circuits are coupled over an arbitrary four-terminal network which may also contain an amplifying valve.

**R 33:** W. Elenbaas: On the excitation temperature, the gas temperature, and the electron temperature in the high-pressure mercury discharge. (Philips Res. Rep. 2, 20-41, 1947, No. 1)

The excitation temperature of a high-pressure discharge, the temperature of which was formerly determined from the intensity decrease, perpendicularly to the axis, of spectral lines having different initial levels, is determined from the absolute value of the intensity of the yellow lines (5770/5791 Å) combined with the transition probability of these lines, as measured by Schouten and Smit. The temperature found by the last method is considerably lower than that previously determined. The cross section of the Hg atom which must be used in the Langevin equation in order to arrive at the right current (the density of the electrons being determined by the Saha equation), as well as the vapour pressure, are computed by assuming the excitation temperature to coincide a) with the gas temperature and b) with the electron temperature. In case b) we determine the gas temperature by assuming that the electrons deliver energy to the atoms through elastic collisions just sufficiently to compensate the loss by conduction of heat. In both cases we find for the cross section of the Hg atom a reasonable value, viz. $4.10^{-15}$ cm$^2$. The pressure as computed from the gas temperature and the mean vapour density agrees better with the measured value in the case b) than in the case a) (differences 2.5 and 8 per cent, respectively). In case b) one has to apply a correction to the temperature as formerly determined from the intensity decrease perpendicularly to the axis, resulting in a satisfactory agreement between the two methods. Finally, the excitation probability of the 3d and 3D levels from the ground level was estimated. The excitation by electron impact is much more frequent than the excitation by mutual collisions of the atoms, but yet it is far too small to explain temperature equilibrium.

**R 34:** H. B. G. Casimir: On the theory of eddy currents in ferromagnetic material (Philips Res. Rep. 2, 42-54, 1947, No. 1)

The theory of eddy currents in ferromagnetic material is developed for the limiting case that the depth of penetration $d$ is small, although the product $\mu d$ is not necessarily small.

First the rigorous solution for the sphere is discussed for this limiting case. Next it is shown that the solution can be obtained from Laplace's equation with a new type of boundary condition. This boundary condition is then applied to a discussion of eddy currents in spheroids. Explicit formulae are found both for small and for very large values of $\mu d/R$

Special attention is given to the limits for very long and very flat spheroids.

**R 35:** H. C. Hamaker: Radiation and heat conduction in light-scattering material. I. Reflection and transmission. (Philips Res. Rep. **2**, 55-67, 1947, No. 1)

On the basis of a set of simultaneous differential equations originally due to Schuster the transmission and reflection of light in light-scattering layers is discussed. Formulae previously developed by Kubelka and Munk are briefly recapitulated; they are extended so as to describe the luminescence of fluorescent screens excited by X-rays or electron bombardment. Likewise formulae are derived that include temperature radiation.

**R 36:** H. A. Klasens: The light emission from fluorescent screens irradiated by X-rays. (Philips Res. Rep. **2**, 68-78, 1947, No. 1)

Applying Schuster's theory as extended by Hamaker for the scattering and absorption of light, general equations are deduced for the amount of light emitted by fluorescent screens irradiated by X-rays. Some commercial screens are examined to measure the "absorption" coefficient of the fluorescent light. Several means of increasing the brightness of a screen are discussed. (see Philips techn. Rev. 9, 364-370, 1947, No. 2)

**R 37:** J. L. Meyering and M. J. Druyvesteyn: Hardening of metals by internal oxidation I. (Philips Res. Rep. **2**, 81-102, 1947, No. 2)

Alloys of silver, copper and nickel can be dispersion-hardened by diffusing oxygen into an alloy with e.g. 1—2 atomic % of an element having a sufficient affinity for O. Too small an affinity leads to a coarser distribution of the oxide formed, because conglomeration must take place via the atoms, and dissociation occurs more frequently when the oxide is not very stable. Thermodynamical considerations are given, and calculations of the penetration of the reaction front.

**R 38:** H. C. Hamaker: Radiation and heat conduction in light-scattering material. II. General equations including heat conduction. (Philips Res. Rep. **2**, 103-111, 1947, No. 2)

The equations of Schuster describing the transmission of radiation through light-scattering material are now extended so as to include temperature radiation and heat conduction. Linearization leads to a system of three simultaneous linear differential equations, amongst which one of the second order. These are resolved and certain general features of the solution are discussed. In particular an expression is deduced for the energy transport due to radiation.

**R 39:** H. C. Hamaker: Radiation and heat conduction in light-scattering material. III Application of the theory. (Philips Res. Rep. **2**, 112-125, 1947, No. 2).

The set of equations developed in the foregoing paper and describing the combined transmission of energy by radiation and by thermal conduction in light-scattering material are applied to practical problems. In the first place the boundary conditions at the surface of a thick layer are investigated. Near the surface the temperature is no longer a linear function of the distance from the surface. Formulae for the deviation from linearity are deduced and the order of magnitude of these deviations is indicated. On the basis of these theories it is pointed out that observations of the thermal conductivity of oxide cathodes by various authors are unreliable since they are founded on unjustifiable assumptions.

Application of the present theory to layers of finite thickness are only briefly discussed. On the whole the theory is a very crude one, so that it is not to be expected that detailed applications will be of great practical value.

**R 40:** A. van Weel: An improved method for coupling valves for ultrashort waves. (Philips Res. Rep. **2**, 126-135, 1947, No. 2)

A method for coupling two electron valves or one valve with an antenna is described, by which method the difficulties due to the finite inductance of the internal electrode leads of a valve can be eliminated up to very high frequencies. In addition to this the new system provides a very simple way to realize matching of the valve impedances.

**R 41:** F. L. H. M. Stumpers: Interference problems in frequency modulation. (Philips Res. Rep. **2**, 136-160, 1947, No. 2)

After a survey of definitions, the general problem of interference with frequency-modulated signals is treated. Special attention is paid to the pauses of the desired signal. The case of equal amplitudes gives rise to some interesting mathematical relations. The loudness level of disturbances is computed. In the last two sections the interference caused by synchronized transmitters (or by two-path transmission of one signal) is extensively dealt with. Many numerical examples illustrate the theory.