VOL. 11 No. 1, pp. 1-32 JULY 1949

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 "φ-DETECTOR", A DETECTOR VALVE FOR FREQUENCY MODULATION

by J. L. H. JONKER and A. J. W. M. van OVERBEEK.

621.396.622.7:621.396.619.13

With frequency modulation a much lower level of noise and other interference can be reached than with the method of amplitude modulation hitherto more commonly used. Frequency modulation is therefore being applied more and more, although it is limited to transmitters working on very short waves (at most a few metres). In the United States of America dozens of broadcasting transmitters are already working according to this method of modulation. It has even been laid down there that frequency modulation is to be used for the transmission of sound in television. A description is given of a new valve with which a frequency-modulation detector can be built possessing very good detection properties. Moreover with this valve the system can be greatly simplified, since it takes the place of several circuits and valves needed in other systems. Between the cathode and the anode this valve has seven grids.

An important advantage of telecommunication by means of carriers modulated in frequency instead of in amplitude lies in the fact that, given a well constructed transmitter and receiver, there is little interference in reception due to noise, to signals of other transmitters and to impulse-like interference.

As Armstrong showed in 1936[1]), in order to get this high degree of freedom from interference the transmitter should work with a frequency sweep which is great compared with the highest modulation frequency[2]). Moreover the receiver should be of such a construction that variations in the amplitude of the signal applied to the detector do not cause any appreciable a.f. voltage.

The circuits applied in good receivers for frequency modulation answer this requirement but are rather complicated. Here a new valve is described which allows of much simpler circuits and, as will be shown, moreover offers other advantages.

First of all we shall briefly consider the functioning of a detector for frequency modulation (FM-detector).

Functioning of an FM-detector

In *fig. 1* a block diagram is given representing a superheterodyne receiver for frequency modulation. The intermediate frequency f_i, obtained by mixing the (amplified) aerial signal with an oscillator voltage (constant frequency f_o), shows the same variations corresponding to the modulation as does the frequency f_t of the aerial signal. The central frequency at present usually chosen for the i.f. amplifier is about 10 Mc/s.

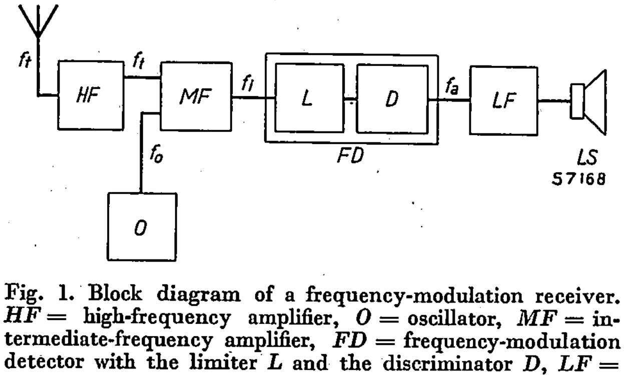


Fig. 1. Block diagram of a frequency-modulation receiver. *HF* = high-frequency amplifier, *O* = oscillator, *MF* = intermediate-frequency amplifier, *FD* = frequency-modulation detector with the limiter *L* and the discriminator *D*, *LF* = audio-frequency amplifier, *LS* = loudspeaker. f_t = frequency of the transmitter, f_o = oscillator frequency, $f_i = f_t - f_o$ = intermediate frequency, f_a = audio-frequency.

[1]) E. H. Armstrong, A method of reducing disturbances in radio-signaling by a system of frequency modulation, Proc. Inst. Rad. Engrs. **24**, 689-740, 1936.

[2]) By frequency sweep is understood the largest deviation from the central frequency, i.e. the frequency of the transmitter in the absence of modulation; see e.g. Th. J. Weijers, Frequency modulation, Philips Techn. Rev. **8**, 42-50, 1946.

The frequency detector has two functions: that of a limiter (performed by the part L, fig. 1) which renders the amplitude variations just mentioned innocuous, and that of fiequency discriminator (pait D). We shall first consider the latter.

Frequency discrimination

In the main two methods of frequency discrimination are known:

a) The fiequency fluctuations are converted into amplitude variations which are detected in the familiar manner (*fig. 2a*). Conversion is brought about with the aid of a network the impedance of which depends upon the frequency; detection (rectifying) is done by one or more diodes. This network may in principle consist of a simple LC circuit, but usually it is a system of two or three coupled circuits (band-pass filter). This gives a better approximation to a linear relation between the variations of frequency and amplitude, so that there is less distortion.

b) From the received signal two voltages (v_1 and v_2) are derived between which there is a phase difference φ varying with the instantaneous value of the frequency (fig. 2*b*). Detection takes place with the aid of a sort of mixing valve with two control grids to which the voltages v_1 and v_2 aie applied. This method, which hitherto has been employed much less than that under (a), will be reverted to in more detail farther on.

Limitation

The voltage at the output of the i.f. amplifier, thus the voltage that is applied to the frequency detector, would be expected to have a constant amplitude, since every endeavour is made in the transmitter to keep the signal transmitted as free of amplitude modulation as is possible. There aie, however, a number of causes that can be indicated whereby the said output voltage may be amplitude modulated to a fairly great depth. These causes are the following.

Even though a transmitter may be working with a perfectly constant amplitude, the signal received by the receiver aerial need not by any means be constant: inter-action between the waves reaching the receiver aerial along different paths and dispersion in the atmosphere may cause considerable amplitude modulation in the received signal, this modulation being found again at the output of the i.f. amplifier.

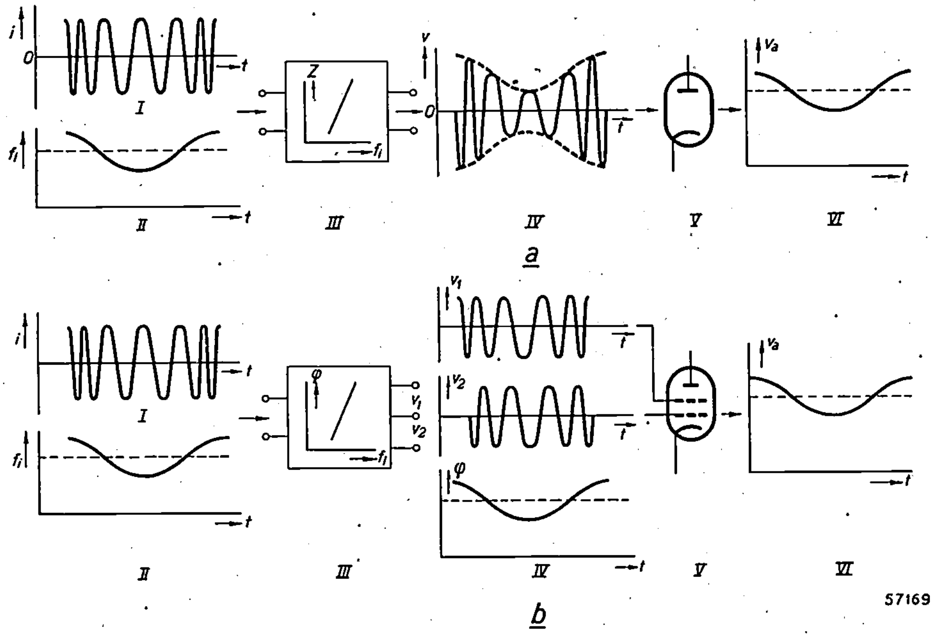


Fig. 2. Two methods of frequency discrimination.

a) At I a frequency-modulated current i and at II the instantaneous value of the frequency f_i are shown as functions of the time t. The current traverses a network (III), the impedance Z of which is linearly dependent upon f_i. Across the network there then arises an a.m. voltage v (IV) which with the aid of one or more diodes (V) is rectified to the a.f. voltage v_a (VI).

b) According to another method the current i (see I) is conducted through a network (III) which yields two voltages v_1 and v_2 between which there is a phase difference φ linearly dependent upon f_i. At IV v_1, v_2 and φ are shown as a function of t. v_1 and v_2 are each applied to a control grid of a kind of mixing valve (V) of which the output voltage v_a varies with the modulation frequency (VI).

Furthermore, interfering signals also cause amplitude variations. These may arise both from undesired stations and, for instance, from the ignition of passing motor vehicles, as well as from noise sources (principally the first valve in the receiver).

Finally, another source of amplitude modulation lies in the i.f. transformer at the output side of the i.f. amplifier. The ideal solution would be a band-pass filter with a rectangular frequency characteristic, i.e. a filter passing a frequency band of the desired width absolutely uniformly. In practice, however, one is content with a filter the characteristic of which has, for instance, the shape represented in *fig. 3*, curve *I*. The sweep of the intermediate frequency (curve *II*) results in the amplitude of the output voltage being modulated (curve *III*). As the illustration shows, this modulation is distorted; when detected, it would give rise to a more for less distorted reproduction.

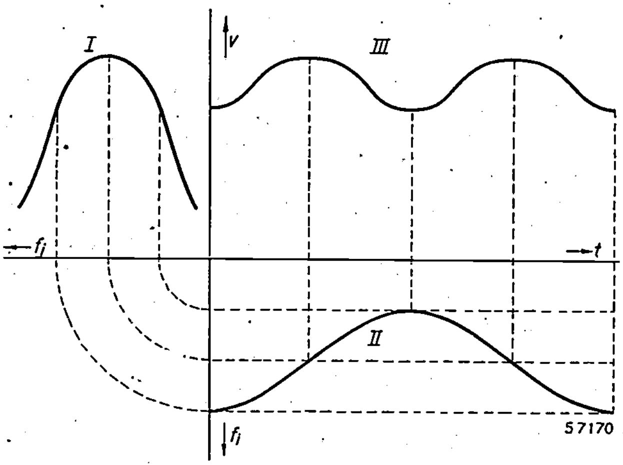


Fig. 3. I = resonance curve of an i.f. transformer: voltage V as a function of the frequency f_i. When f_i varies sinusoidally with the time (curve II) then V shows a distorted amplitude modulation (curve III).

It is now the task of the limiter to prevent the amplitude variations in the output voltage of the i.f. amplifier penetrating into the frequency discriminator. The use of a limiter not only suppresses noise and other interference but at the same time considerably attenuates the distortion just referred to.

Fig. 4 gives a diagrammatic representation of some known limiting systems with a brief explanation.

A frequently used circuit for frequency detection

Amplitude limitation can be combined with both the systems of frequency discrimination mentioned above under (a) and (b). *Fig. 5* represents a much used circuit where the limiter corresponds to that of fig. 4*a* and frequency discrimination is performed by the method given under (a). The frequency fluctuations here produce amplitude variations in a manner described by Foster and Seeley and already discussed in this journal [3]).

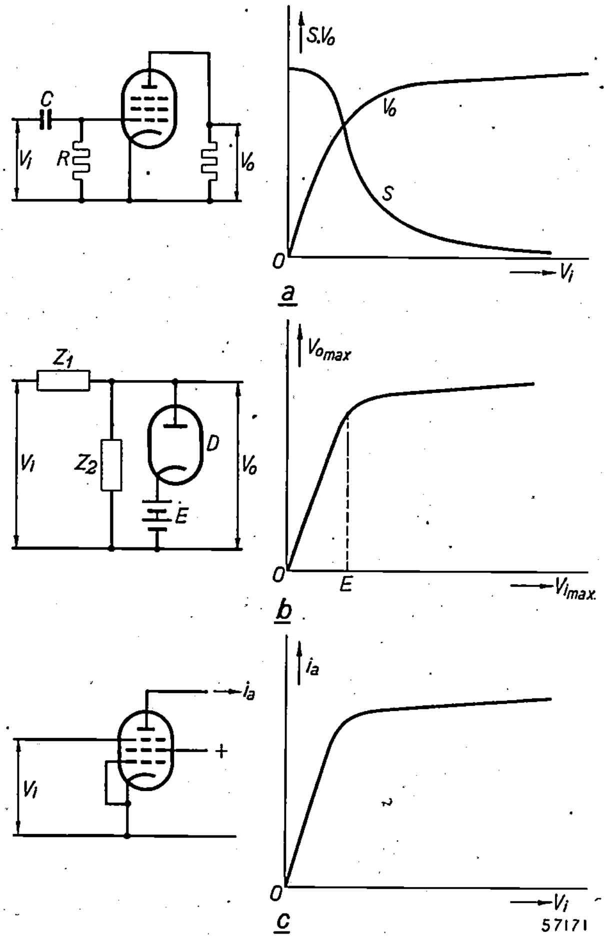


Fig. 4. Three methods of amplitude limitation.
a) Grid detection arises by the grid capacitor C and the leak resistor R, so that the grid voltage becomes more strongly negative as the signal V_i increases. At the same time, however, the slope S of the pentode drops, in such a way as to cause the output voltage V_o to remain practically constant (provided V_i is large enough).
b) Connected in parallel with the impedance Z_2 is a diode D in series with a threshold voltage E. When the amplitude $V_{i\max}$ of the input signal exceeds the value E the diode becomes conducting and the amplitude $V_{o\max}$ of the output voltage cannot rise much higher than E. The difference between V_i and V_o is taken up by the impedance Z_1.
c) The anode current i_a of a valve is practically independent of the voltage V_i at a control grid (provided V_i is not too small) if inside this control grid there are one or more grids at constant potential screening the control grid. If V_i is an alternating voltage with varying but continuously sufficient amplitude, then the amplitude of i_a remains practically constant.

The "φ-detector", the new valve that we are about to discuss, in its function as frequency dis-

[3]) D. E. Foster and S. W. Seeley, Proc. Inst. Rad. Engrs. **25**, 289-313, 1937. See also the article quoted in footnote [2]), in particular figs 10, 11 and 12 and accompanying text.

criminator comes under (b): detection of phase differences. At the same time it acts as amplitude limiter, according to the diagram of fig. 4c. Thus the two functions are combined in one valve.

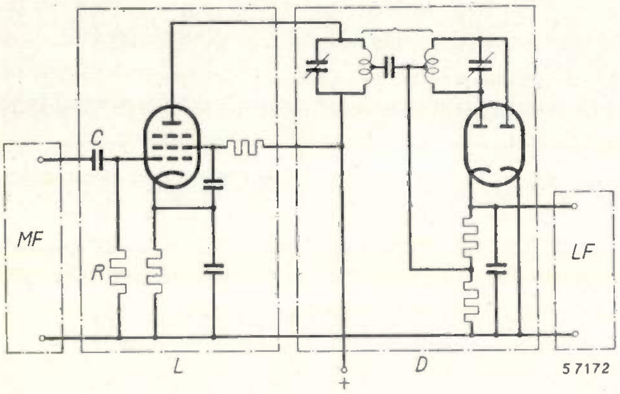


Fig. 5. A commonly used circuit for an FM detector. L = limiter according to fig. 4a, D = frequency discriminator according to fig. 2a. MF = intermediate-frequency amplifier, LF = audio-frequency amplifier.

Principle of the "φ-detector"

The "φ-detector" has seven grids, denoted in *fig. 6* by g_1-g_7 and shown separately in the photograph of *fig. 7*. The grids g_2, g_4 and g_6 are screen grids which provide a screening between the two control grids (g_3 and g_5) and between these grids and the other electrodes, and to which a constant, low, positive voltage (20 V) has to be applied. The innermost grid g_1 likewise has a constant potential, for instance that of the cathode. The grid g_7 is a suppressor grid counteracting secondary emission of the anode; it is connected to the cathode.

The anode is connected to a high positive voltage (275 V) via a resistor of high value.

The grids g_1 and g_2 are so constructed that the field of g_3 and that of the electrodes beyond g_3 are ineffective at the cathode. Just as is the case with a pentode for instance, the electron current emerging through the openings of g_2 is thus only dependent upon the voltage of g_1 and that of g_2. Since these voltages are constant (0 and 20 V respectively) the electron current referred to is likewise constant.

The voltage at the control grid g_3, however, does indeed influence what further happens to this current: if the voltage of g_3 is negative the electrons return to g_2 (g_2 then takes up practically the whole of the cathode current); if g_3 is positive the electrons continue on their path and pass through the openings of g_4. What then happens to them is determined by the voltage at g_5: if g_5 is negative the electrons have to return to g_4; if g_5 is positive they continue on their way, passing through the meshes of g_5, g_6 and g_7 and reaching the anode, which is at a high positive potential.

Briefly it therefore amounts to this, that anode current can only flow when g_3 and g_5 are positive simultaneously and that this anode current (I_a) is constant (about 1 mA). (The currents taken up by the grids through which the anode current is flowing are small compared with I_a.)

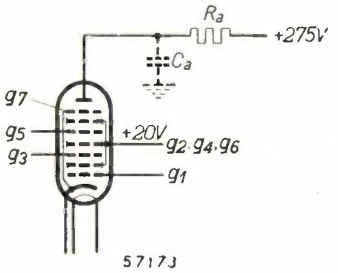


Fig. 6. Diagrammatic representation of the "φ-detector". g_1 = grid at approximately zero potential; g_2, g_4 and g_6 = screen grids at +20 V, g_3 and g_5 = control grids; g_7 = suppressor grid. R_a = anode resistance. C_a = anode stray capacitance.

If a sinusoidal alternating voltage is applied to each of the control grids and φ is the phase difference between these two voltages, then — according to the description given — anode current will only flow so long as both grids are positive. This is the case, in each cycle, during an interval of 180°—φ (*fig. 8*). Thus the anode current varies in block form between 0 and the constant value I_a. The mean value $\bar{i}_a$ of the anode current is therefore

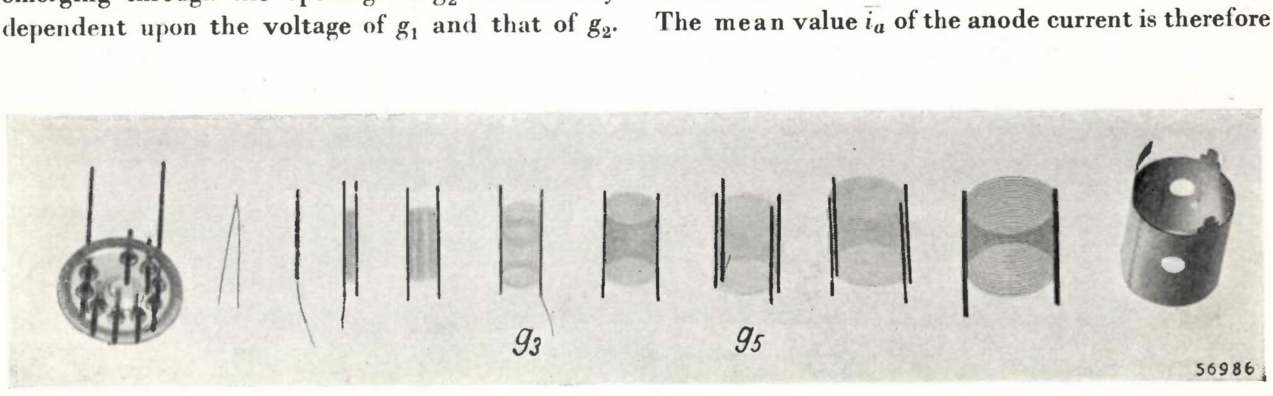


Fig. 7. Component parts of the "φ-detector". From left to right: the "Noval" base with 9 contacts pins, the filament, the cathode, the seven grids and the anode; g_3 and g_5 are the control grids. The illustration is about $^4/_5$ of the true size.

$$\bar{i}_a = \frac{180° - \varphi}{360°} \cdot I_a \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

and is thus a measure of the phase difference φ. Hence the name of the valve: "φ-detector".

When the frequency-modulated i.f. voltage in an FM receiver is converted with the aid of a suitable network into two alternating voltages with a mutual phase difference φ varying proportionately to the frequency deviation, and these voltages are applied to the grids g_3 and g_5, then, according to (1), the mean anode current will vary proportionately to φ, thus likewise in proportion to the frequency deviation, or in other words proportional to the a.f. signal to be transmitted.

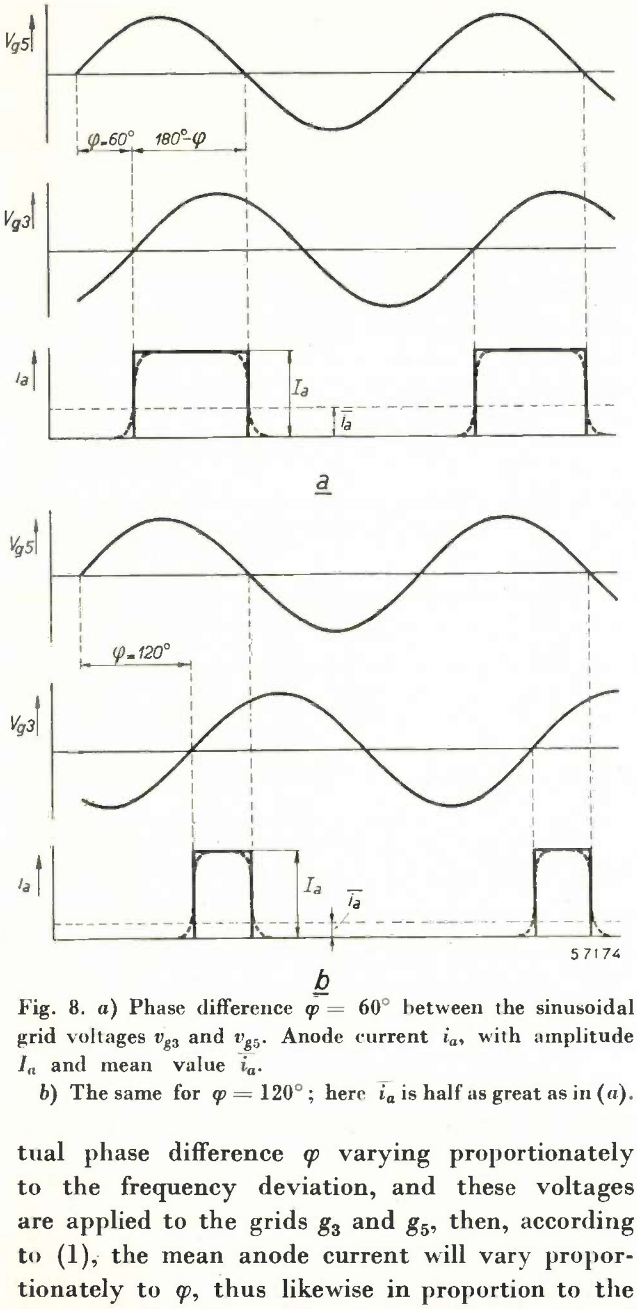


Fig. 8. *a*) Phase difference $\varphi = 60°$ between the sinusoidal grid voltages v_{g3} and v_{g5}. Anode current i_a, with amplitude I_a and mean value $\bar{i}_a$.

b) The same for $\varphi = 120°$; here $\bar{i}_a$ is half as great as in (*a*).

The rectangular pulses forming the anode current have the same periodicity as the intermediate frequency (about 10 Mc/s). The anode current therefore consists of an alternating current with this fundamental frequency (and its harmonics) superimposed on a direct current pulsating at the audio frequency. The anode resistor R_a (e.g. 0.5 MΩ) shunted by the stray capacitance of the anode C_a (about 25 pF) brings about an almost complete separation of these two components: the current changing according to the audio frequency flows through the resistor, whilst the i.f. current flows almost entirely through the capacitance. The internal resistance of the valve amounts to about 3 MΩ.

From (1) it is also to be seen that the "φ-detector" acts as a limiter: since, as we have seen, I_a is constant, also $\bar{i}_a$ is independent of the amplitude of the alternating voltages at g_3 and g_5.

For the sake of simplicity it has been assumed in the foregoing that the anode current changes discontinuously with the voltage applied to the control grids. Actually, however, a gradual transition takes place, about which more will be said farther on. As a result the rectangular shape of the anode current pulses is somewhat rounded off, as indicated in fig. 8. In essence the linear relation (1) between $\bar{i}_a$ and φ still holds.

Fig. 9 is an illustration of the "φ-detector" (type EQ 80). It is made in the same way as the "Rimlock" valves [4]) and, like these, has a diameter of 22 mm.

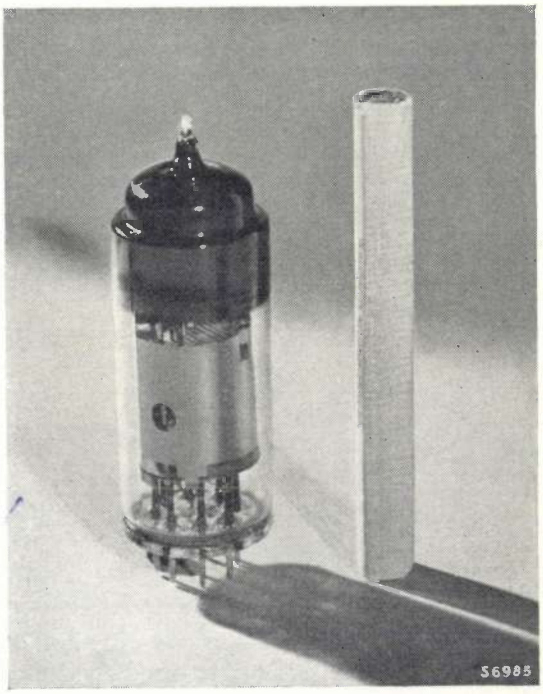


Fig. 9. The "φ-detector" EQ 80 with a cigarette for comparison of size.

[4]) G. Alma and F. Prakke, A new series of small radio valves, Philips Techn. Rev. **8**, 289-295, 1946. In contrast to the "Rimlock" valves, the EQ 80 has a "Noval" base with 9 contact pins placed at 9 corners of a regular decagon. The valve cannot be wrongly inserted in the holder because the latter has only 9 openings to match the pins. Since the EQ 80 requires only 8 leads, the cathode is connected to two contact pins.

Properties of the "φ-detector"

The "φ-detector" as frequency discriminator

Fig. 10 shows how the "φ-detector" can be used in a circuit. The i.f. transformer belonging to the i.f. amplifier, in which the first phase of the detection process takes place, namely the conversion of the frequency variations into φ-variations, consists for example of two coupled tuned circuits. Each of the circuits is connected to a control grid (g_3, g_5) of the "φ-detector". The amount of the phase difference φ between the voltages across the two circuits depends upon the instantaneous value f_i of the intermediate frequency. If the frequency deviation Δf_i is nil (as is the case if the transmitter is not modulated) then $\varphi = 90°$. When the transmitter is modulated f_i varies and φ shows a sweep around the value 90°. With the band-pass filter consisting of two tuned circuits as represented in fig. 10 the relation between φ and Δf_i is an arc cotan function (*fig. 11*) which in the vicinity of the point for $\Delta f_i = 0$ is only approximately linear. If at the maximum frequency sweep φ remains between 60° and 120° then the distortion due to the curvature of the arc cotan line amounts to 2.5% (curve D, fig. 11), which is to be regarded as the maximum permissible value. At the given frequency sweep, φ is kept between the said limits by giving the damping of the second band-pass filter circuit a suitable value. In this way reasonable linearity is obtained between Δf_i and φ.

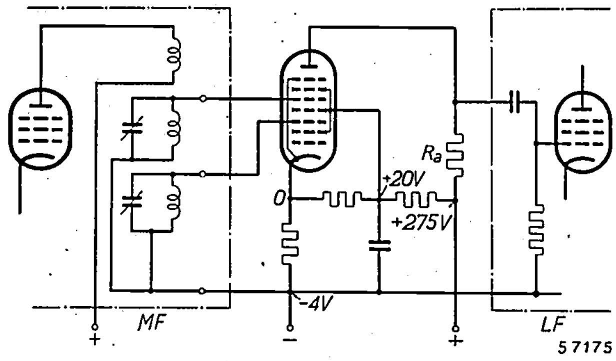


Fig. 10. Circuit for FM-detection with "φ-detector". MF = intermediate-frequency amplifier with a two-circuit band-pass filter, LF = audio-frequency amplifier.

It is to be noted that the distortion can be still further reduced at the extra cost of a more complicated filter. If a band-pass filter with three tuned circuits is used (*fig. 12*) under certain conditions the distortion is limited to 0.2% for φ = 60°-120° and to 1.2 % for φ = 50°-130°.

The conditions referred to are for example:

$$Q_3 = 0.45\ Q_2,$$
$$M_{23} = L_2/Q_2,$$

in which Q_2 is the quality factor of the circuit L_2-C_2 (fig. 12), Q_3 that of L_3-C_3 and M_{23} the mutual inductance between L_2 and L_3.

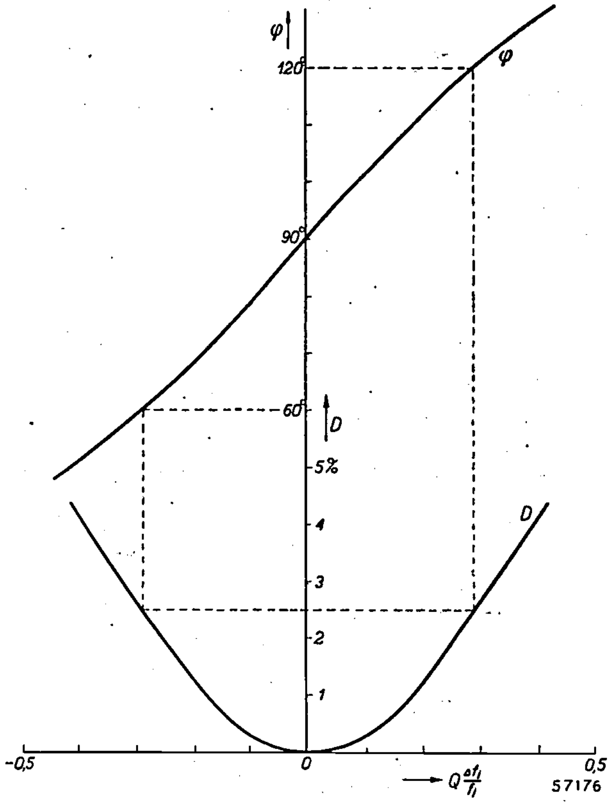


Fig. 11. Phase difference φ and distortion D as functions of the frequency deviation for an i.f. band-pass filter with two circuits. The distortion is a result of the φ-curve not being linear. As abscissa $Q.\ \Delta f_i/f_i$ is plotted, in which Q = quality factor of the second band-pass filter circuit and Δf_i = deviation from intermediate frequency f_i.

The second stage, conversion of the φ-variations into proportional current variations, is the task of the "φ-detector". How this is achieved can be judged from the characteristics represented in *fig. 13*, where $\bar{i_a}$ is plotted as a function of φ for several

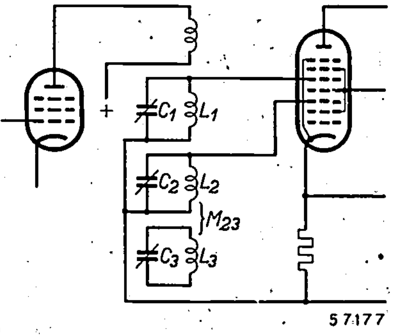


Fig. 12. Band-pass filter with three tuned circuits, by means of which the distortion can be kept smaller than with two circuits.

values of the alternating voltages $V_{g3} = V_{g5}$ at the control grids. It is seen that between $\varphi = 50°$ and 130° the relation between $\overline{i_a}$ and φ is practically linear. The slope in this range is 2.8 μA per degree phase difference. With a sweep of φ between 60° and 120° — limits within which φ is a practically linear function of Δf_i — the variation in the value of $\overline{i_a}$ thus amounts to $30 \times 2.8 = 84$ μA. The r.m.s. value of the a.c. anode current is then $84/\sqrt{2} = 60$ μA. This current flows through the anode resistor (0.5 MΩ) of the "φ-detector" shunted by the grid resistor (about 1 MΩ) of the a.f. amplifying valve, so that the voltage applied to the latter, at full modulation, amounts to about 20 V r.m.s.

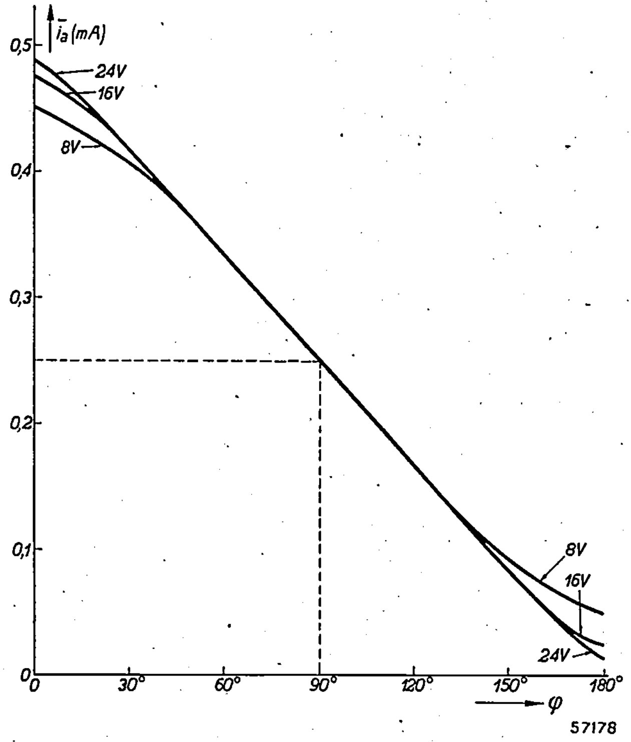


Fig. 13. Mean anode current $\overline{i_a}$ of a "φ-detector" as a function of the phase difference φ between the alternating voltages v_{g3} on the third grid and v_{g5} on the fifth grid, for $V_{g3} = V_{g5}$ respectively 8, 16 and 24 V r.m.s.

When a band-pass filter with three tuned circuits is used, the limits of the φ sweep can be chosen somewhat wider; the output voltage then rises to 25 V.

The "φ-detector" as limiter

Fig. 13 also shows that the "φ-detector", as already explained, acts as a limiter: within the above-mentioned limits of φ the curves $\overline{i_a} = f(\varphi)$ for various values of the alternating voltage $V_{g3} = V_{g5}$ at the control grids practically coincide, so that in the first instance amplitude modulation will not be detected. However, 8 V r.m.s. is the smallest value at which the curves sufficiently coincide and at which the linear part is sufficiently long. Therefore V_{g3} and V_{g5} must not drop below 8 V, a value which is rather higher than that required for some other FM-detectors but which will not as a rule be difficult to obtain.

In order to ascertain more accurately to what degree the "φ-detector" is insensitive to amplitude modulation, the following measurement has been taken. H.F. voltages not modulated in frequency but modulated in amplitude up to 30 % with a low frequency and having a constant phase difference of 90° were applied to the two control grids of the valve. The r.m.s. value I of the a.f. anode current was measured as a function of the amplitude of the control-grid voltage. If the "φ-detector" were absolutely insensitive to amplitude modulation — in other words, if the curves in fig. 13 were to coincide completely round about the point with abscissa $\varphi = 90°$ — then I would have to equal 0. The result of the measurement (*fig. 14*) shows that I is not quite zero but less than 1 μA, provided the control-grid voltages are greater than 8 V. (This value of I is to be compared with the value found above for the A.C. anode current at the full frequency sweep, viz. 60 μA.) The same applies for other values of φ between 50° and 130°. Thus all interference causing amplitude variations in the output voltage of the i.f. amplifier is effectively suppressed in the "φ-detector".

Contrary to the case with limiters having an *RC*-circuit (fig. 5), in the application of this detector as a limiter there is no other inertia than that of the electrons, so that the interferences are suppressed very quickly. Consequently, not only interference of a more or less continuous nature is suppressed

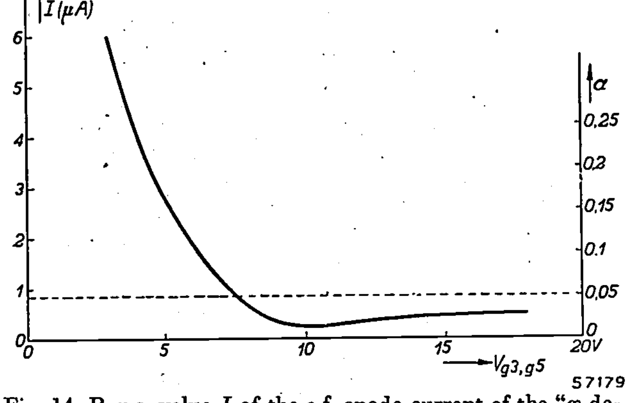


Fig. 14. R.m.s. value I of the a.f. anode current of the "φ-detector" as a function of the alternating voltage $V_{g3,g5}$ (r.m.s. value) at the control grids, in the case of 30% amplitude modulation and a constant phase difference $\varphi = 90°$ between V_{g3} and V_{g5}. When $V_{g3,g5} > 8$ V amplitude modulation is scarcely detected and the noise arising therefrom is less than that due to frequency modulation. (For the quantity α see the text in small type.)

(e.g. noise), but also short, impulse-like interference bursts such as may arise from transients due to transients in the mains or from the ignition in motor vehicles.

The extent to which a limiter should suppress amplitude modulation can be gathered from the following, taken partly from an article previously published in this journal[5]). Particular attention will now be paid to noise.

Suppose that at the output of an i.f. amplifier there is a signal with amplitude A, originating from the desired but non-modulated transmitter (interference is most troublesome during intervals in the modulation), and also a weak interfering signal with amplitude sA ($s \ll 1$), likewise non-modulated. Let the frequency difference between the two signals be f_a, a frequency in the audible range.

The combination of these two signals forms a signal the amplitude and frequency of which are both modulated, the amplitude with a modulation depth s, the frequency with the frequency sweep sf_a (see the first of the article quoted in footnote[5])). In both modulations the modulating frequency is f_a.

If this combined signal is applied to an amplitude detector then, roughly speaking, this yields an a.f. voltage with the amplitude sA and the audio-frequency f_a.

When, however, this combined signal, after all amplitude modulation has been removed in an ideal limiter, is applied to a frequency discriminator such as represented diagrammatically in fig. 2*a*, the discriminator yields a signal with the amplitude A modulated to a depth $sf_a/\Delta f_{\max}$ ($\Delta f_{\max}$ is the maximum frequency sweep allowed by the discriminator). Upon this amplitude modulation being detected, the result is an a.f. signal again with the frequency f_a but with an amplitude $sf_aA/f_{\max}$.

We shall now extend the single interfering signal from which we started to a noise, that is to say, to a continuous spectrum of interfering signals. We shall see what the result is a) after amplitude detection and b) after limitation and frequency detection.

a) After amplitude detection noise is produced, to the power of which, in any frequency interval df_a, the contribution is proportional to $s^2A^2df_a$. By integration over the range of the audio-frequencies ($f_a = 0$—$f_{a\,\max}$) we find for the power P_{AM} of the audible noise:

$$P_{\mathrm{AM}} = k \int_0^{f_{a\,\max}} s^2A^2 df_a = ks^2A^2f_{a\,\max},$$

in which k is a proportionality factor.

b) After complete limitation and frequency detection we find in a similar way for the power P_{FM} of the noise:

$$P_{\mathrm{FM}} = k \int_0^{f_{a\,\max}} s^2A^2f_a^2/(\Delta f_{\max})^2\, df_a = ks^2A^2 \cdot \frac{1}{3}\,\frac{f_a{}^3{}_{\max}}{(\Delta f_{\max})^2}.$$

Thus the two powers bear the relation of

$$\frac{P_{\mathrm{AM}}}{P_{\mathrm{FM}}} = 3\left(\frac{\Delta f_{\max}}{f_{a\,\max}}\right)^2, \text{ for which we write } = \frac{1}{a^2}.$$

A frequency detector without limiter would yield both the noise contributions (AM noise and FM noise) simultaneously. Of these two the AM noise is by far the stronger, as shown by the last formula given. If we substitute for $\Delta f_{\max}$ the usual value of 75 kc/s and for $f_{a\,\max}$ 12 kc/s, we obtain:

$$P_{\mathrm{AM}}/P_{\mathrm{FM}} = 3 \cdot (75/12)^2 = 117.$$

We now have to consider how well a limiter must function in order to suppress the AM noise so far as to make it just as weak as the FM noise (further suppression would be of little avail). We see from the above calculation that the amplitude variations of the input voltage of the discriminator have to be attenuated to $1/\sqrt{117} \approx 1/11$ of the original value.

In this rough calculation no account has been taken of the fact that in an FM transmitter the high notes are given a certain pre-emphasis, that is to say they are transmitted with a strength proportionately too great, this being compensated in the receiver by a de-emphasis filter. Taking into account the influence of this filter, the value of the limiting factor a where the two noise contributions are equal is $1/20 = 0.05$ instead of 1/11.

In fig. 14 a scale is given for a, which is proportional to I.

Following upon this consideration of the foremost properties of the "φ-detector" as discriminator and as limiter, we shall deal with some other features of this valve: the possibility it offers to suppress the side response occurring in a receiver that is not properly tuned, the a.f. amplification which has to be applied after detection, and the possibility of simplifying the circuit.

Suppression of side response

The reception of frequency-modulated signals is often disturbed by a troublesome noise occurring when the receiver is not properly tuned, i.e. when the superheterodyne oscillator does not yield the frequency which, together with the central frequency of the transmitter, produces the central intermediate frequency to which the i.f. transformer has been tuned. The fact is that if the oscillator is so tuned that the modulated intermediate frequency sweep is in the area *II* or *III* (*fig. 15*) instead of in the

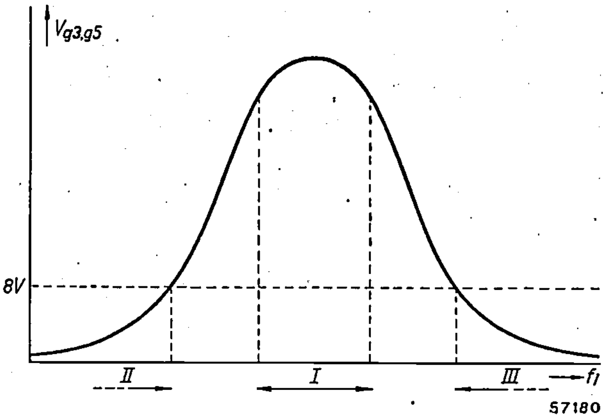


Fig. 15. Resonance curve of an i.f. transformer. With proper tuning the sweep of the intermediate frequency f_i lies in the area *I* and the output voltage $V_{g3,g5}$ is greater than the limit value of 8 V. In the case of incorrect tuning (area *II* or *III*) $V_{g3,g5}$ is too small

[5]) Th. J. Weijers, Comparison of frequency modulation and amplitude modulation, Philips Techn. Rev. 8, 89-96, 1946. For more detailed particulars see F. L. H. M. Stumpers, Theory of frequency-modulation noise, Proc. Inst. Rad. Engrs. 36, 1081-1092, 1948 (No. 9).

area I, the working point will lie in a part of the filter characteristic where strong amplitude modulation takes place, whilst the signal voltage remains below the limit at which the limiter comes into action. Thus the reproduction is distorted and accompanied by strong side response (interchannel noise) making it difficult to get the right tuning.

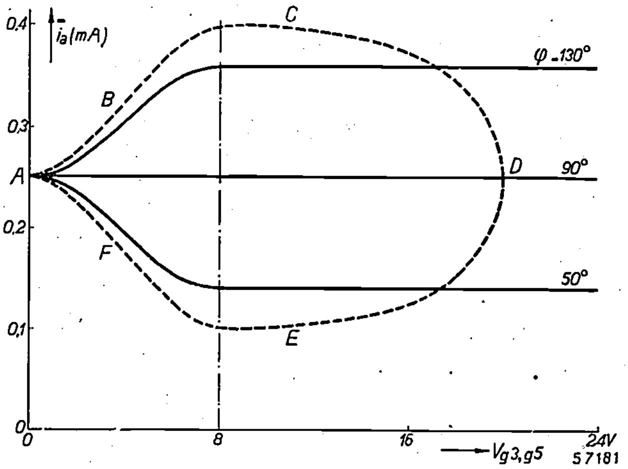


Fig. 16. Fully-drawn lines: mean anode current $\overline{i_a}$ of a "φ-detector" as a function of $V_{g3,g5}$ for several values of φ. Point A corresponds to a receiver very much out of tune, point D corresponding to a correctly tuned receiver. When turning the tuning knob the dashed line is followed.

We shall show how this side response is restricted as far as possible by means of the "φ-detector". It will also be seen that it can even be completely suppressed in a simple way.

In fig. 13 $\overline{i_a}$ is represented as a function of φ for some values of $V_{g3} = V_{g5}$. In order to see what happens when the receiver is detuned we prefer to choose the coördinates differently, plotting $\overline{i_a}$ against $V_{g3,g5}$ with φ as parameter. This gives curves as represented in *fig. 16* (continuous lines), from which it is also to be seen that if detection of amplitude modulation is to be avoided V_{g3} and V_{g5} must exceed a certain limit (8 V). If V_{g3} and V_{g5} are less than 8 V then $\overline{i_a}$ is dependent upon their magnitude, that is to say, the valve no longer functions as limiter.

The dashed line in fig. 16 denotes how $V_{g3,g5}$ and $\overline{i_a}$ vary when the frequency f_0 of the superheterodyne oscillator is gradually changed and the transmitter is not modulated. If there is still a great difference between f_0 and the transmitter frequency f_t then V_{g3} and V_{g5} are very nearly zero and in fig. 16 the position is thus represented at the point A. When f_0 is shifted towards f_t, so that $|f_t—f_0|$ approaches the frequency band passed by the i.f. circuits, V_{g3} and V_{g5} increase and the point indicating corresponding values of $\overline{i_a}$ and these voltages moves from A to B or F (according to the sign of $f_t—f_0$); let us suppose that it moves to the point B. Upon f_0 being further changed it passes through C to the point D, where correct tuning (φ = 90°) is reached. If f_0 were to be still further changed in the same direction the point would pass through E to F and finally, far beyond the tuning, reach A.

In the areas B and F, $\overline{i_a}$ depends upon the magnitude of V_{g3} and V_{g5}, so that the amplitude modulation of the signal is detected and the reproduction is distorted. The smaller the slope at B and F, the less this effect will however be. The shape of the curves given in fig. 16 is in fact much more favourable than that represented in *fig. 17*, where the slope is particularly steep at F. Such a state of affairs arises with a badly designed or improperly adjusted "φ-detector". The explanation of this is as follows.

So far we have been assuming for the sake of simplicity that the "φ-detector" is fully "open" at any positive value of the voltage at g_3 and g_5 and fully cut off at any negative value; in other words, that the characteristics $I_a = f(E_3)$ (with positive grid g_5) and $I_a = f(E_5)$ (with positive grid g_3) measured with a direct voltage (E) show the abrupt change represented in *fig. 18a*. Actually, however, these characteristics have a smooth slope (fig. 18*b* or *c*). If the characteristics are symmetrical with respect to a point P (fig. 18*b*) and this point

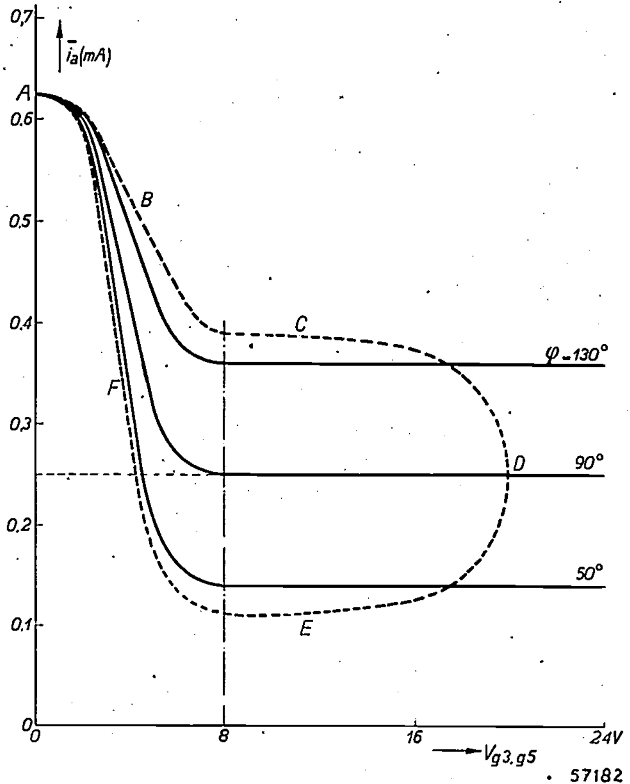


Fig. 17. As in fig. 16, but for a badly designed or wrongly adjusted "φ-detector". This functions properly at the correct tuning (D) but outside that point in the areas F and B of the dashed line, where $V_{g3,g5} < 8$ V, the sharp slope results in a greatly distorted reception and strong noise.

is, with the aid of a bias, chosen as working point, then the mean value of the anode current remains unchanged when an alternating voltage is applied to one of the control grids; the characteristics $\overline{i_a} = \mathrm{f}\,(V_{g3,g5})$ and $\varphi =$ constant then have the symmetrical shape of fig. 16. If, however, the bias is improperly adjusted (working point for instance P' fig. 18*b*) or if one bend of the static characteristic is much less pronounced than the other (fig. 18*c*), a rather small alternating voltage applied to one of the control grids will give rise to a rectifying effect, with the result that $\overline{i_a} = \mathrm{f}(V_{g3,g5})$ becomes asymmetrical in shape, the point A being situated either too high (fig. 17) or too low.

In the designing of the "φ-detector" particular care has been taken to ensure that the static characteristics are made symmetrical. By means of grids consisting of a helix with constant pitch, characteristics are obtained of the type indicated in fig. 18*c*, with a gentle upper bend and a sharp lower bend [6]). A known means of making the lower bend less sharp consists in giving the control grid a variable pitch; there are then, so to speak, a series of valves with mutually different characteristics connected in parallel. This means is commonly applied in variable-μ amplifying valves where the amplification is required to be gradually variable by changing the negative grid bias. This construction has been successfully applied also in the "φ-detector"[7]) in such a way that the characteristics obtained become symmetrical, as indicated in fig. 18*b*.

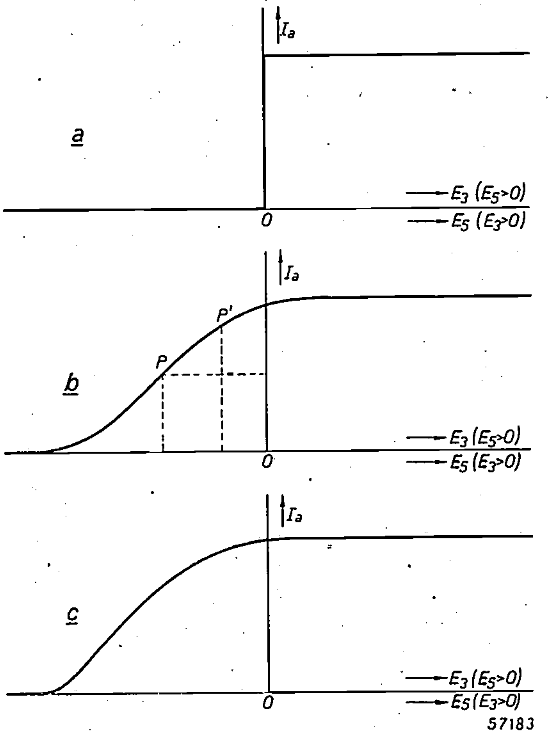


Fig. 18. Static "φ-detector" characteristics; anode current I_a as a function of the direct voltage E_3 on the third grid or E_5 on the fifth grid when E_5 respectively $E_3 > 0$.

a) Discontinuous characteristic, so far assumed for the sake of simplicity.

b) Continuous, symmetrical characteristic with P as the correct and P' as an incorrect working point.

c) Asymmetrical characteristic.

The slope of these characteristics has purposely been made rather small, so that inconstancy of the bias and shifting of the characteristic (contact potential!) have little effect.

The bias, which should be about —4 V on both grids, is taken from a potentiometer across the supply voltage, as shown in fig. 10.

Although, thanks to these measures, the "φ-detector" with such a circuit is already fairly favourable as regards the side response, one can easily go a step further and suppress interchannel noise entirely. This may be done by applying to g_1 an auxiliary voltage which cuts off the "φ-detector" cathode current so long as the voltage at the control grids g_3 and g_5 is less than 8 V. *Fig. 19* shows how this can be brought about: applied to g_1 are two direct voltages in series, a fixed negative voltage E_1 (taken from the potentiometer R_1-R_2-R_3) and a variable positive voltage E_1' (obtained from a diode rectifying one of the secondary voltages of the band-pass filter, and a smoothing capacitor). The "φ-detector" cannot come into action until the rectified voltage is high enough. A resistor R_4 prevents g_1 from becoming too strongly positive.

The diode can serve at the same time, in the known way, for automatic volume control and for controlling a tuning indicator.

Audio-frequency amplification

With the circuit shown in fig. 5 the maximum output voltage of the detector will scarcely reach 4 V. The "φ-detector", on the other hand, can give about 20 to 25 V, as already stated. This gain in output can be utilized in various ways.

With voltages of such amplitude it is possible, for instance, to control directly the output valve EL 41 [8]), which is already fully loaded with a control-grid voltage of 4 V; this leaves a factor

[6]) See J. L. H. Jonker, The control of the current distribution in electron tubes, Philips Res. Rep. 1, 331-338, 1946, in particular figs 4 and 5, or, by the same author, Current distribution in amplifying valves (Dutch), thesis Delft 1942, figs 68 and 69.

[7]) The irregular pitch of g_3 can be seen in fig. 7. This detail of g_5 has been lost in the reproduction.

[8]) Output pentode of the "Rimlock" construction (see footnote [4])) with a mutual conductance of 10 mA/V and an anode dissipation of 9 W.

of about 5 for the application of negative feedback, to counteract distortion by the output valve. This saves a stage of a.f. amplification.

Mostly, however, it is required that the output valve of a receiver can be fully loaded already at the average modulation depth of the transmitter, which is about 25 %. If, whilst the output valve is just fully loaded, the modulation goes deeper or the volume control is turned higher, then some distortion arises, but the impression that the sound makes is determined more by the high notes it contains than by this distortion. In the case of an A. M. receiver with a relatively narrow frequency band, when there is a moderate overloading, the sound is not so disturbing, but it is so in an F. M. receiver, which reproduces also the highest audio frequencies; at these frequencies there is only a small margin between the audible limit and the bearable limit of the sound, so that when the sound is too strong the reproduction becomes disagreeable. This is all the more reason why no A.F. stage should be used between the "φ-detector" and the output valve. If ,however, it should nevertheless be desired to keep this stage then advantage can be taken of the high output voltage of the "φ-detector" by applying an exceptionally heavy negative feedback (thus making the distortion exceptionally small).

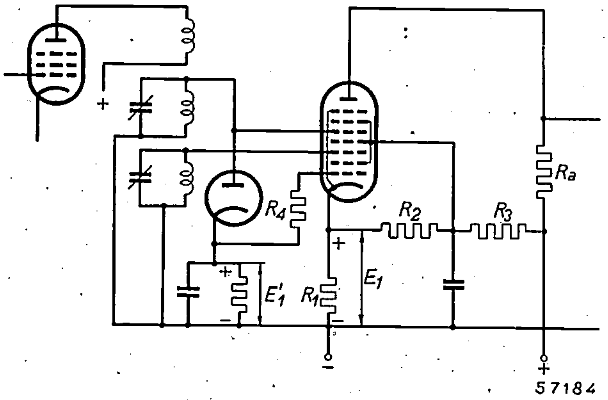


Fig. 19. The voltage at the first grid of the "φ-detector" amounts to E_1'—E_1, whereby E_1' is obtained by rectification of the voltage V_{g5} across one of the band-pass filter circuits. E_1 is so chosen that the "φ-detector" is cut off when $V_{g5} < 8$ V.

Simplicity of the circuit

When the circuit according to fig. 10 is compared with that according to fig. 5 it is seen that the "φ-detector" takes the place of the following components (without counting the less important ones): a pentode, a double diode and an i.f. transformer with two tuned circuits. One will particularly be glad to dispense with the i.f. transformer, which has to be so very carefully adjusted. There may be added the stage of a.f. amplification saved.

The circuit shown in fig. 19, where the "φ-detector" is blocked when the receiver is detuned, does not really entail any complication, because the diode, which as a rule is already present for tuning indication and automatic volume control, can be utilized for the object in view.

This ends the discussion of the properties of the "φ-detector". It has not been possible within the scope of this article to deal with a number of details, among others selectivity, which is particularly favourable when the "φ-detector" is used.

Summary. The "φ-detector" is a new valve for detecting frequency modulation. Of the seven grids which it contains, the second, fourth and sixth grids are screen-grids and the seventh is a suppressor grid. To each of the control grids (the third and fifth (an output voltage is applied from an i.f. transformer. The r.m.s. value of these voltages must be at least 8 V. The mean value of the anode current is a function of the phase-shift φ between the two control voltages, which phase-shift is in turn a function of the frequency deviation. Both functions are approximately linear when φ has a sweep between 60° and 120°. The amplitude of the anode current is not dependent upon the magnitude of the control voltages (provided they are greater than 8 V), so that the valve also acts as limiter. This renders certain sources of noise and of distortion harmless. Limitation takes place without any other inertia than that of the electrons, so that also short impulse-like interference bursts are limited. The "φ-detector" yields an a.f. voltage of about 20 to 25 V, so that an output valve can be fully loaded. With a "φ-detector" some of the components can be dispensed with, such as an i.f. transformer and two or three valves. The first grid can serve for blocking the cathode current in the event that the control voltages are not large enough, in this way suppressing the interchannel noise. The "φ-detector" EQ 80 is made in the same way as the "Rimlock" valves but has a "Noval" base (9 contact pins). The anode current (mean value) is only 0.25 mA.

AN APPARATUS FOR DETECTING SUPERFICIAL CRACKS IN WIRES

by P. ZIJLSTRA.

620.191.3:621.317.39:621.317.336

High-frequency alternating currents are not only used in radio and telecommunication engineering but can often also be employed for testing purposes, their high-frequency character being an advantage for some purposes. An example of such an application is the testing of wires for superficial cracks. The currents induced in a wire, when this is placed in a high-frequency magnetic field, flow almost exclusively along the surface, so that a superficial crack will greatly influence these currents and thus can easily be detected by a suitably chosen electrical measurement.

Introduction

For connecting the electrodes of electronic valves to other circuit elements use is made of wire or rod-shaped leads (e.g. lead pins). These have to be fused into the glass base of the valve in such a way as to ensure a vacuum-tight seal.

For good vacuum-tight sealing it is serious if there are fine cracks on the surface of the lead (as may easily be the case particulary with fairly thick wires) because in the fusing process the crack is not entirely filled with the glass and air can therefore leak in. This is particularly the case if the crack runs in the longitudinal direction of the wire or rod. In order to detect the presence of such cracks before a valve is mounted it is very convenient to have an apparatus with which the leads can be quickly tested.

The use of such an apparatus need not be confined to the testing of leads. For instance it is also highly important that in the manufacture of large electronic valves the thick filaments should be previously tested for the presence of fine cracks, since these are apt to lead to premature rupture of the wire.

In this paper an apparatus is described which has proved quite satisfactory for this purpose in practice. First we shall explain the principle of the method applied.

Principle of the method of testing

When a conducting wire or rod is placed in a coil through which a high-frequency alternating current is flowing, Foucault currents are induced in that wire or rod and at high frequencies these currents flow almost exclusively along the surface of the wire (the so-called skin effect).

Let us imagine that a round metal rod is placed inside a cylindrical coil with its axis coinciding with that of the cylinder. The combination of coil and rod may be regarded as a transformer with its secondary winding consisting of one single short-circuited turn, viz. the cylindrical surface layer of the rod through which the Foucault currents flow. The impedancc of this transformer, considered from the primary side, depends, inter alia, upon the resistance and self-inductance of the secondary winding and the coupling between the primary and secondary windings. This means that this impedance is related to the shape of the cross section of the rod.

If there is a crack in the surface of the rod then the impedance will differ from that obtained with a sound rod, particularly if the crack is not a gentle hollowing out but a sharp cleft in the practically circular profile of the cross section. It will be readily understood that the presence of such a crack increases the length of the path travelled by the induced high-frequency currents, this resulting in an increased resistance of the "secondary winding" of the transformer. This change in resistance is manifested on the primary side of the transformer by a change in the impedance of the coil. (To a first approximation the self-inductance will not change.) All that is now necessary is to connect the coil to a circuit with which any change in the impedance can easily be measured.

When a rod that has to be tested is drawn through the coil any change observed in the impedance gives an indication of the presence of a crack.

It is of importance to investigate what magnitude the change in impedance can be expected to assume. To that end we have to calculate the reaction of the rod upon the coil and see how a fine crack in the rod passed through the coil influences this effect. Assuming, for instance, that a molybdenum wire 1.5 mm in diameter and having a longitudinal crack 0.1 mm in depth is passed through the coil, then it will be found that (at the frequency used in the apparatus) the coil resistance increases by 1/3%. Even such a small variation can be observed quite well by the method applied by us.

Reaction of the rod upon the coil

We will consider a cylindrical coil with an average radius of r_1 metres [1]) and of a length that is great compared to its diameter. Placed concentrically inside the coil is a cylindrical rod with a radius of a metres. Let us suppose that the rod consists of a material having a specific resistance ϱ (ohm.metre) and a relative permeability μ_r, and that an H.F. current $I_1 = I_{1\max} \sin 2 \pi ft$ is flowing through the coil.

The magnetic alternating field H in the space between the coil and the rod has everywhere the same value H_1, which is given by the formula

$$H_1 = nI_1 \quad . . . . \quad (1)$$

in which n is the number of turns per metre of the coil.

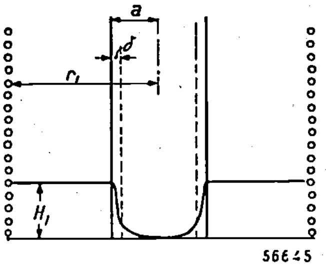


Fig. 1. Diagrammatic representation of the variation of the magnetic alternating field H inside the measuring coil (radius r_1) and in the rod (radius a). δ represents the penetration depth.

In the rod alternating currents are induced which, given a sufficiently high value of the frequency f, flow mainly along the surface of the rod. The current density is maximum at the surface of the rod; it decreases rapidly in the axial direction (whereby also the phase of the current changes continuously) and at a depth δ (the so-called penetration depth) it drops to $1/e$ times its greatest value (e = base of the natural logarithmic system). As the calculation shows, also the magnetic field rapidly decreases inside the rod: at the surface $H = H_1$ and at the depth δ H_1 has dropped to H_1/e (see *fig. 1*). The relation between the quantities δ, f, ϱ, and μ_r is given by the equation

$$\delta = \sqrt{\frac{\varrho}{\pi f \mu_0 \mu_r}} \text{ metres,} \quad . . . . . . \quad (2)$$

in which μ_0 = the vacuum permeability $= 4\pi \cdot 10^{-7}$ H/m.

The penetration depth δ is at the same time the thickness of an imaginary current envelope which, in the case of homogeneous distribution of the total current over this envelope, gives the same quantity of heat as the actual current. The Joule heat developed in the rod per metre length is given by the formula

$$P = H_{1\max}{}^2 \, \pi a \varrho / \delta \text{ watts,} \quad . . . . . . \quad (3)$$

which with the aid of (1) can also be written as

$$P = I_{1\max}{}^2 \, \pi \, n^2 a \varrho / \delta = I_{1\mathrm{eff}}{}^2 \, 2\pi \, n^2 \, a \varrho / \delta \quad . . . \quad (4)$$

Formula (4) shows that the heat development in the rod per metre length corresponds to that of a resistor connected in series with the coil:

$$R' = 2\pi \, n^2 a \, \varrho / \delta \text{ ohm} \quad . . . . \quad (5)$$

Thus this gives the reaction of the rod upon the impedance of the coil in so far as the resistance component is concerned.

The self-inductance of the coil is also changed by passing the rod through the coil. Before the rod was introduced the self-inductance per metre length of the coil was

$$L = \pi \mu_0 r_1{}^2 n^2 \text{ henry} \quad . . . . \quad (6)$$

After the rod has been inserted in the coil the magnetic field, roughly speaking, continues to exist only in the space between the coil and the rod, the cross-sectional area of which equals π ($r_1{}^2$-a^2). Originally there was the same magnetic field in the whole of the space enclosed by the coil (cross-sectional area $= \pi r_1{}^2$). Thus the reduction of the self-inductance of the coil due to the presence of the rod amounts per metre length to

$$L' = \pi \, \mu_0 \, a^2 n^2 \text{ henry} \quad . . . . \quad (7)$$

As an example we will consider the case of a coil having an average radius r_1 = 2.83 mm and with a number of turns per metre n = 26,000, this coil being fed with a high-frequency current of the frequency $f = 5.6 \times 10^6$ c/s. The quality factor $Q = 2\pi f L/R$ of such a coil (without the rod) proved to be 17.6.

From these data it follows that the high-frequency resistance R and the self-inductance L of the coil (without rod) per metre length amount respectively to:

$$R = 42{,}700 \text{ ohms,} \qquad L = 0.0214 \text{ henry.}$$

A molybdenum wire 1.5 mm in diameter is now inserted in this coil. The specific resistance of molybdenum is $\varrho = 5.7 \times 10^{-8}$ ohm·metre, whilst μ_r = 1. A calculation of the penetration depth with the aid of formula (2) gives $\delta = 50.8 \times 10^{-6}$ metres (thus about 50 microns). The increase in resistance of the coil due to the presence of this molybdenum wire is found directly from formula (5). Per metre length R' = 3576 ohms. Comparing this value with that of R it appears that the coil resistance has increased by a good 8%. Further, from (6) and (7) it is found that the self-inductance is reduced by about 7% of the original value.

What, now, is the effect of a crack in the surface of the rod upon the change in impedance of the coil? We shall investigate only the influence this has upon the resistance variation because it is mainly this that plays a part in the measurements taken with the apparatus which will be described farther on.

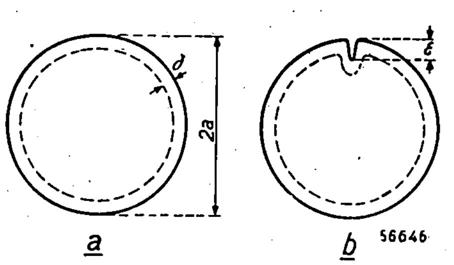


Fig. 2. Cross section of a rod where δ is the penetration depth of the induced Foucault currents. *a*) A sound rod or wire. *b*) A rod having a crack in the longitudinal direction to a depth ε.

As we have seen, the current in the wire is practically confined to a thin surface layer of the thickness δ (*fig. 2a*). The resistance of this thin layer (per metre length of the wire) is $2\pi a \varrho/\delta$ ohm, which expression is found in formula (5). If one narrow crack (of the depth ε) is present we may assume, as an approximation, that the flow of current will pass round the crack, so that its resistance will amount to $(2\pi a + 2\varepsilon)$ $\varrho \cdot \delta^{-1}$ ohm. It may therefore be expected that owing to the crack the resistance R' will increase in the proportion of $(\pi a + \varepsilon)/\pi a$.

[1]) In the calculations we shall use the Giorgi system of units.

Let us now take again a molybdenum wire with a radius of 0.75 mm as an example and imagine that it has a crack 0.1 mm in depth running along its entire length. The resistance R' will then increase to the extent of 152 ohms, i.e. by about 4%. Thus the total coil resistance formed by R and R' in series (42,700 ohms + 3576 ohms) increases by $^1/_3$%.

Further details of the method

It is now only a question of designing a circuit by means of which changes in the impedance of the coil of the given order of size can be measured with sufficient accuracy. For this purpose one can use, inter alia, a high-frequency bridge circuit or an oscillator circuit. We have chosen the latter and introduced the coil in the grid circuit of the oscillator, with a grid capacitor and leak resistor connected in the usual manner. Contrary to the methods already known, instead of the frequency change of the oscillator being taken as a measure for the change in the impedance of the coil, the influence of the impedance variation upon the control-grid current was taken as the measure. The principle of the circuit used is illustrated in *fig. 3.*

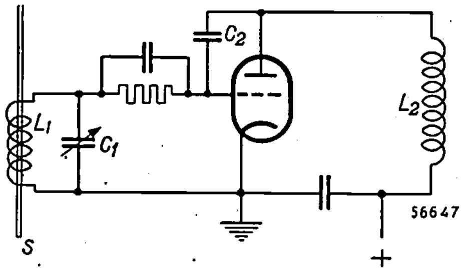


Fig. 3. Basic diagram of the apparatus for testing wires. The rod or wire S to be tested is passed through the coil L_1 which, together with a variable capacitor C_1, is included in the grid circuit of a radio valve. L_2 is the coil in the anode circuit and C_2 a feedback capacitor.

The grid circuit is formed by the coil L_1 (inside of which is the rod to be tested) and a capacitor C_1. The anode circuit comprises only a coil L_2, whilst the feedback via the grid-anode capacitance of the valve is augmented by an extra feedback capacitor C_2.

With such a circuit the magnitude of the control-grid current is to a first approximation proportional to the impedance $L_1/(R{+}R')C_1$ of the circuit. If $R{+}R'$ increases by $^1/_3$% then, as calculated above for a specific case, this results in a decrease of $^1/_3$% in the impedance and also in a reduction of $^1/_3$% in the grid current of the oscillator valve.

With the apparatus described in this article the grid current amounts to 2 mA. Thus, a crack in the rod resulting in a change of $^1/_3$% in the resistance causes the current to decrease by about 7 μA.

In order to observe such small current variations with sufficient accuracy a compensating circuit was used which reduces the current flowing through the measuring instrument from 2 mA to 100 μA.

With the aid of this compensating circuit the presence of cracks of the order of 0.1 mm depth can easily be detected. Where there are deeper cracks one will of course find a greater difference in the current intensity.

Description of the apparatus

The circuit diagram of the instrument is given in *fig. 4,* where L_1 is the measuring coil and C_6 the tuning capacitor of the grid circuit; R_4 and C_1 are respectively the grid leak and the grid capacitor, while C_2 and C_3 together form the feed-back capacitor. The coil L_2 is the anode choke. The anode and the grid d.c. circuits are decoupled by means of the by-pass capacitors C_4 and C_5.

The variations in current are read from the micro-ammeter, care being taken that the circuit oscillates adequately by adjusting the capacitors C_6 and C_3. The oscillator valve B_1 is a triode, the grid current of which is adjusted to about 2 mA. The compensation circuit already mentioned ensures that only a small part of this grid current (100 μA) passes through the micro-ammeter. The grid current flows from the grid via the cathode, micro-ammeter, selenium rectifier Se_1, through L_1 and R_4 back to the grid. If Se_1 were taken away the compensating current would flow from the positive side through R_1, R_2 and R_3 and back in the opposite direction through the micro-ammeter to the negative side. The difference of these two currents results in a current in the pass direction of the selenium rectifier.

The resistor R_3 is variable so that the compensating current can be adjusted to allow current readings to be taken from the micro-ammeter with rods of different diameters.

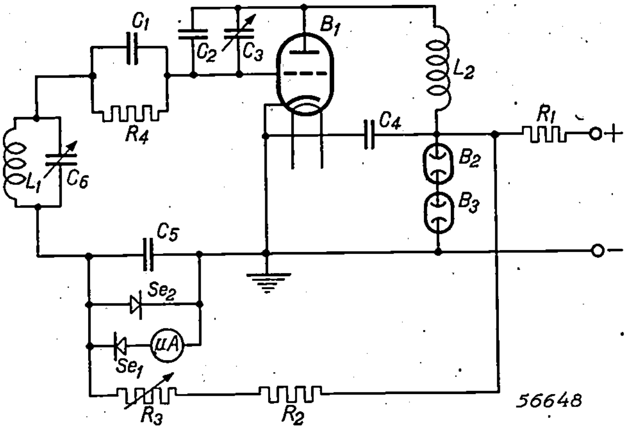


Fig. 4. Circuit diagram of the apparatus for testing wires. B_1 is the valve with the measuring coil L_1 in its grid circuit; B_2, B_3 are stabilizing valves, L_2 is the anode coil, Se_1 and Se_2 are selenium rectifiers.

Since the grid current is dependent not only upon the impedance of the grid circuit but also upon the direct voltage of the anode, the rectifier for the anode supply is provided with a voltage stabilizer consisting of two stabilizing valves B_2 and B_3 with a resistor R_1.

The valve acting as rectifier in the high-tension unit needs a longer time to warm up than B_1. This is of importance because it avoids the compensating circuit coming into action before the oscillator starts supplying currꝛnt; if that were to be the casc then the micro-ammeter might easily bc overloaded when switching on.

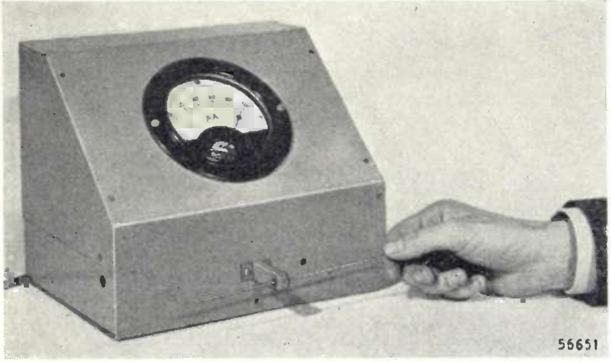


Fig. 5. Photograph of the apparatus described in this article being used for the testing of a wire.

The micro-ammeter is also safeguarded against overloading in another way. When the apparatus is switched off the grid current drops much quicker than the compensating current. The rectifying cell Se_1 then prevents the compensating current from flowing through the micro-ammeter, whilst the rectifying cell Se_2 forms an additional shunt for the leakage current from Se_1. *Fig. 5* shows the apparatus in operation.

Applications

As stated already in the introduction to this article, the apparatus described here has been designed for testing the lead pins of electronic valves. These are mostly made of molybdenum or tungsten in the form of rods with diameters of about 0.75 up to 2.5 mm. Not only these pins, already cut to size, but also the rods or wires from which they are made can be passed through the apparatus, thereby detecting any faulty parts and thus saving cutting and grinding costs. A larger measuring coil can be fitted in for testing thicker rods up to a diameter of 6 mm.

This apparatus has also been found very useful for the testing of tungsten filaments for large transmitting valves. It has been found that the cases of filament rupture that used to take place now and then were caused by long cracks in the wire. These wires are therefore now passed through the apparatus for testing before being cut to size.

Wires possessing ferromagnetic properties cannot be tested with this apparatus because local variations in permeability give rise to so many changes in the meter reading that those due to the presence of cracks cannot be distinguished.

Apart from its use for the detection of cracks, this apparatus is also suitable for detecting variations in the cross section of a rod or wire, provided this is accompanied by a variation in the length of the periphery. Thus it is possible to find local thickening or thinning in this manner; it is not necessary that the normal rod should have a circular profile.

Another application lies in the sorting of pins of equal dimensions but having divergent properties in their material. Pins not conforming to a specified composition can easily be picked out with this apparatus on account of the different value of their specific resistance.

Summary. For the proper functioning of electronic valves it is necessary that there should not be any fine cracks on the surface of the leads passing through the glass. In this article an apparatus is described with which the presence of such cracks can be detected. The rod or wire to be tested is placed in a measuring coil through which a high-frequency current is passed. If there is a crack in the wire then the impedance of the coil will vary from that obtained with a sound rod or wire. In order to determine this difference the coil is included in the grid circuit of an oscillator. The extent of the variation in impedance is derived from the influence that this has upon the control-grid current. The method is sufficiently accurate to detect cracks of about 0.1 mm depth. The use of this apparatus is not confined to the testing of valve leads; it can be used also in other cases for the testing of wires.

FLUORESCENT PIGMENTS AS AN ARTISTIC MEDIUM

by J. L. H. JONKER and S. GRADSTEIN

667.65:667.624.48:535.612

Tradition has it that in China about the year 1000 A.D. there was a picture by a Japanese artist representing an ox which was invisible by day but could be seen in the dark on account of light being emitted from it. Painting with luminescent materials is therefore in fact fairly old. That picture, however, must have been painted only for the mysterious effect of the afterglow. In the experiment described in this article there was no such intention, the object being to make use of the vivid colours produced by fluorescent substances under ultra-violet radiation. Fluorescent pigments have a selective emission, whereas the pigments commonly used in art painting are characterized by a selective absorption and reflection. The consequences of this difference, both for the practical realization of the experiment in question and for the possible artistic value of the new technique of painting, are explained in this article.

The wide use being made nowadays of fluorescent substances for various industrial products (fluorescent lamps, cathode-ray tubes, etc.) has stimulated the development of a number of new substances of this kind. The beautiful colours and the high light yield of many of these new substances have in turn led to the search for fuIther possible applications. One example that has aroused considerable public interest is the application of these substances for decorative purposes. Stage scenery, costumes, and all sorts of objects for attracting attention in exhibitions or in shop-windows are painted with fluorescent paints. When all visible light is excluded and the invisible ultra-violet radiation from a mercury lamp is projected onto the scenery or the object, astonishing effects are produced because the objects themselves are apparently emitting light in all sorts of vivid colours.

It is obviously only one step further to use fluorescent pigments for making a picture. By way of experiment one of the authors of this article (J. L. H. J.) ventured to take that step a few years ago (1946). *Fig. 1a* gives a coloured reproduction of the result, which was exhibited, inter alia, at a meeting of the "Commission Internationale de l'Eclairage" (International Commission of Illumination) held in Paris in June 1948. In what follows something will be said about the artistic aspects and the practical problems of this new technique in the art of painting.

The world that the artist aims at representing — presupposing that that is indeed his aim — is characterized from the physical point of view by the differences in colour and brightness of the objects perceived by the eye. In the classical techniques, such as painting in water-colours, pastel drawing, oil painting, the artist has a sufficient variety of paints on his palette to be able to represent faithfully the different colours occurring in nature. In trying to represent the brightness values occurring in a scene however he is at a disadvantage. Obviously one cannot expect a faithful imitation of the high brightness of all sorts of light sources; no one would blame the artist for a yellowish round spot on the canvas giving only a poor impression of the radiant sun it is supposed to represent. But even disregarding the high brightnesses of light sources, a faithful reproduction of the degree of brightness of most subjects remains an impossibility. It is

Table I. The measured brightness range (ratio of the greatest to the smallest local brightness) of different objects [1]).

Object	Lightest part	Darkest part	Brightness range
Open-air scene	White house in the sun	Black cat in the shade	90
Living-room	Light wall-paper in the sun	Black mantle-piece in shadow	5000
Living-room	Light wall-paper in diffuse daylight	Black mantle-piece in shadow	80
Cinema picture (for comparison)	—	—	25-80

[1]) Taken from General Electric Review 48, September 1945, page 18. A more extensive treatise, with figures, on the brightness ranges occurring in nature is to be found in E. Goldberg's "Der Aufbau des photographischen Bildes, I. Helligheitsdetails", publ. Knapp, Halle 1925. And further see L. A. Jones and H. R. Candit: The brightness scale of exterior scenes and the computation of correct photographic exposure, J. Opt. Soc. Amer. 31, 651-678, 1941.

not unusual for the "brightness range", i.e. the ratio of the greatest and smallest brightnesses occurring, of an indoor subject to amount to a factor of 1000 or still more; in the open air, owing to the equalizing effect of the contribution of diffuse light from the vault of the heavens, the brightness range is generally smaller (see *table I*), but in the presence of glass or water, owing to ıeflections from their surfaces, the brightness range in the open air may also be far more than 1000. The greatest difference that the artist can produce between light and dark however is only a factor of 20 to 50, according to the technique chosen (see *table II*).

The artist is fully conscious of this shortcoming and tries to meet it in various ways. To lend more brightness to a patch of colour he chooses contrasting colours for the adjoining parts of the picture (blue shadows against yellow lights) or tempers their brightness. He enhances the effect of reflections by flattening with a palette knife the pure white that he uses for these effects, thereby avoiding cast shadows from the uneven surface; etc. If one speaks of the "simulation of light" by famous painters such as Leonardo da Vinci, Rembrandt or one of the great impressionists, it must be realized that each case is in fact a special example of the art of manipulating the limited scale of brightness in order to produce the greatest artistic effects.

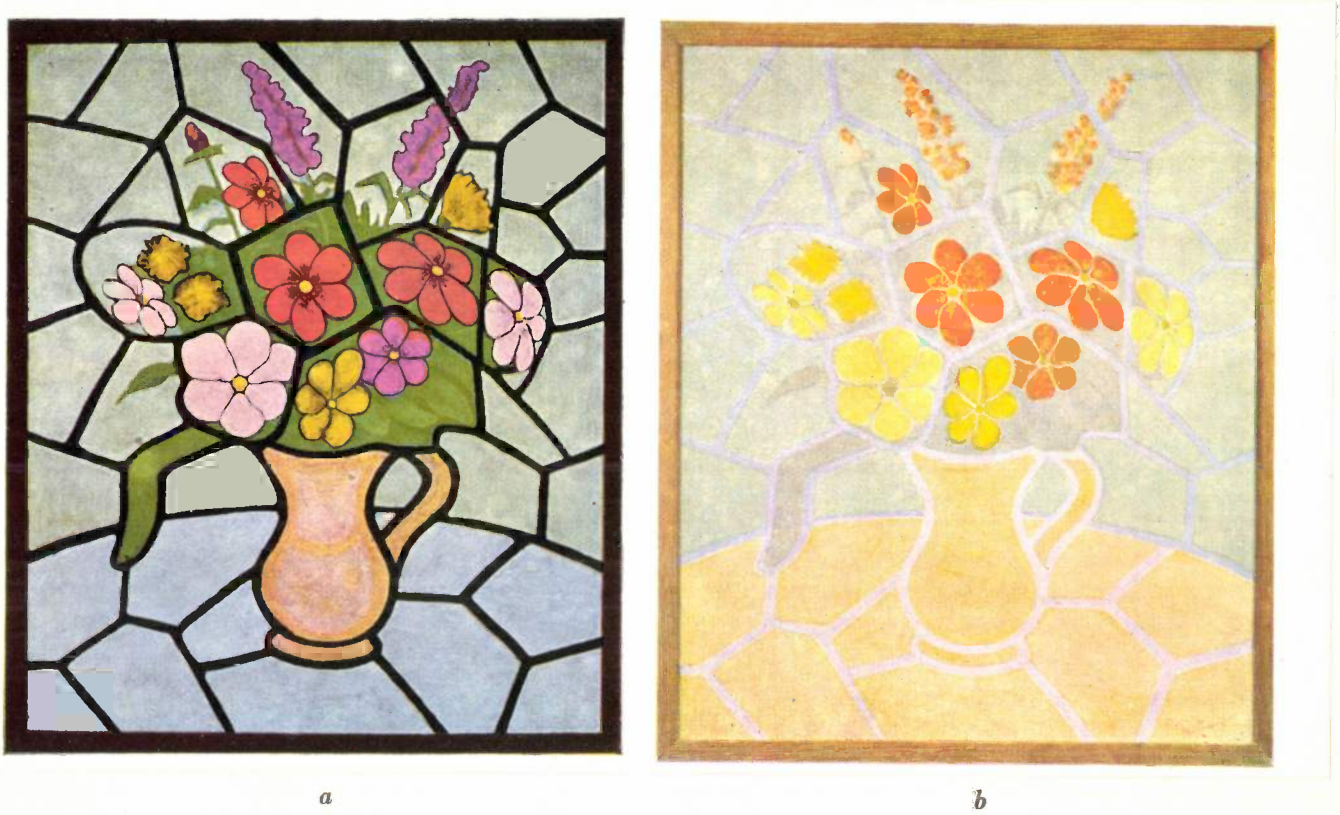


Fig. 1. Reproduction of the fluorescent painting
a) when exposed to ultra-violet radiation,
b) when exposed to daylight.
In the case *a*) two "Biosol" mercury lamps were used for irradiation. Filters of nickel-oxide glass absorbing practically all the visible rays were placed in front of the lamps. In front of the lens of the camera a filter was used which absorbs the ultra-violet rays. The latter was necessary because the painting reflects a large part of the incident ultra-violet light, to which the photographic plate is sensitive.

The specific impression that the fluorescent painting under ultra-violet irradiation makes upon the observer cannot, fundamentally, be fully represented by the reproduction; the reason for this is explained in the text.

Table II. The brightness range in various techniques of painting. The brightness was measured, relatively, of the clearest white and the darkest black of each technique (with the same intensity of illumination for all four techniques). Each colour was applied in a well-covering layer on an area of about 10 cm × 10 cm (4″ × 4″).

Technique	Base	Brightness of white	Brightness of black	Brightness range
Water-colour	white paper	4.58	0.211	22
Pastel	grey paper	4.35	0.112	39
Gouache	white paper	5.35	0.130	45
Oil painting	painters' canvas	4.45	0.096	46

There has been no lack of attempts to raise the brightness range of a painting. Perhaps one of the oldest known cases is that of the painter Mariotto Albertinelli, who worked at Florence about 1500. Vasari[2]) tells of the efforts of Albertinelli to produce a highly plastic effect by means of deep shadows while still bringing out details among the shadows. When this proved unsuccessful, even after transposing the piece to higher or lower brightnesses, it occurred to him that he would have to try to find a brighter white for the high lights, and so he attempted to increase the purity of the white-lead paint used at that time, though without success.

Some painters have tried to solve the problem by gluing small mirrors of glass or metal onto the canvas where the high lights occur and partly covering these with paint. It is obvious that the effect of this artifice depends very much upon the direction of the incident light and consequently cannot, in general, give satisfaction.

Rather than attempt to get a brighter white, it is better to try to increase the brightness range by making the black darker. On the white side, where the reflection is 80-90 %, at most a factor of 1.1-1.2 can indeed be gained. The shortcoming of the normal techniques of painting lies rather in the fact that the "black" is never quite so black, corresponding mostly to a reflection coefficient of 2-5 % (from which then follows the abovementioned limit of 50 for the brightness range). In principle, therefore, much more is to be gained on this side. Several methods of getting a deeper black are described in a patent granted to F. Blau[3]). These methods amount to this, that, at the places where no reflection towards the eye of the observer is desired, either the light is allowed to pass through the picture to be absorbed in a "black body" behind it, or the light is caused to be reflected in directions other than those of the observer by applying directed lighting and a reflecting background in the picture. A difficulty then, however, is to get intermediate tints; for this purpose a sort of pointillism technique has been proposed, among others.

It may be surprising that the problem of brightnesses should find a satisfactory solution, without any further artifices, in the application of fluorescent pigments. Putting it paradoxically, we might say that this is due to the non-fluorescing parts in and around the painting! If no other light whatever is allowed to fall on the painting and the visible light from the mercury lamp used for irradiation is filtered out, these non-fluorescent parts can be almost absolutely black. The brightness range of the fluorescent picture may then theoretically reach the value of infinity!

The brightness range measured on the fluorescent painting was about 130. The fact that the infinite value is not reached is due to two causes acting in conjunction. In the first place the filter used in front of the mercury lamp (a nickeloxide glass) allows a little red and violet light to pass through it, and furthermore the human eye is not quite insensitive to the ultra-violet mercury line with wavelength 3650 Å by which the fluorescence is excited. In the second place the backing of the picture (and the frame, etc.), though it does not fluoresce at all, will always reflect to a certain extent the ultra-violet and the violet and red, and will thus possess a definite though very small brightness. In our case the situation in this respect was as unfavourable as possible, because normal linen with a white chalk base was used as a backing for the picture. In the filtered mercury light the white chalk, owing to the reflection, is much less "black" than one would expect. Nevertheless, according to the values mentioned, a factor of 3 to 4 has already been gained in brightness range compared with the normal techniques of painting. An experiment with a backing that was black in the visible light (and also non-fluorescent) showed that a brightness range of more than 1400 could thereby easily be reached (the brightness of the "black" under ultra-violet irradiation fell below the sensitivity of the measuring apparatus employed).

In addition to the large brightness range of the fluorescent picture there is yet another point. In the normal techniques of painting there is no objection to translating "brightness range" by the more familiar "difference between the brightest white and the darkest black". When painting with fluorescent pigments, however, this translation would be incorrect, for here we have the remarkable fact that some colours are brighter than the brightest possible white! This fact, which from the artistic point of view may of course also be of great importance, can be explained as follows.

A normal pigment owes its colour to the absorption of a specific part of the spectrum of the light falling upon it. The more saturated a colour is, the narrower is the part of the spectrum that the pigment reflects (or allows to pass through). Consequently the total reflection coefficient of a pigment, and with that also its brightness under a given intensity of light, will necessarily decrease

[2]) G. Vasari, Le vite de' più eccellenti architetti, pittori e scultori, Florence, Torrentino 1550.

[3]) See Z. techn. Phys. 6, 279-280, 1925 (Dr. Blau - Festschrift).

according as the colour is more saturated [4]); see *fig. 2*. The brightest possible colour is that having a satuıation zero, and that is white.

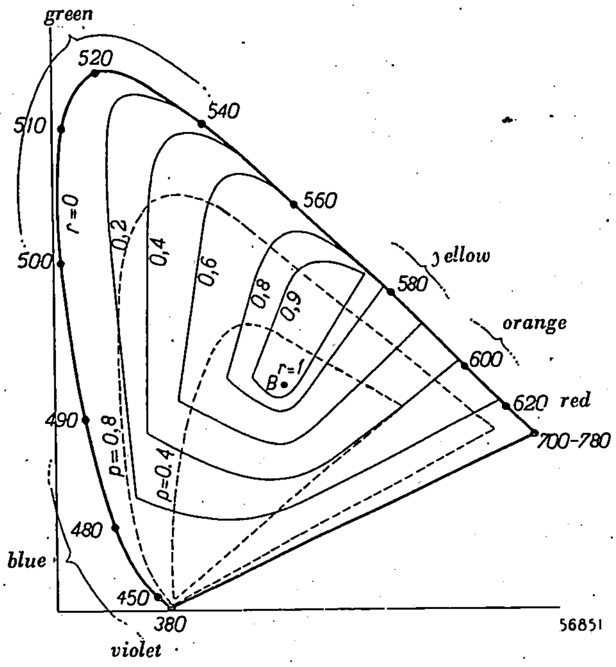


Fig. 2. In the chromaticity triangle, for each point characterized by the hue and the degree of saturation there is a maximum total reflection factor *r* that can be theoretically reached. (The pigment with that colour point must possess a spectral reflection curve of a certain shape to obtain the maximum reflection factor.) In the diagram, which applies for the case where the pigments are exposed to sunlight (so-called standard illuminant *B*), each fully-drawn line connects all points having a certain value of *r*. Two dotted lines have been drawn for constant degree of saturation ($p = 0.4$ and $p = 0.8$, according to the definition of the C. I. E. 1948). It is seen that in a normal painting the saturated colours (colour points near the curve of the spectral colours) can only be obtained with lower brightnesses than white (saturation 0, colour point *B*).

A fluorescent pigment, on the other hand, owes its colour to the fact that it re-emits the absorbed radiant energy with a specific spectrum. The efficiency of the energy conversion, thanks to the progress made in preparative chemistry, is nowadays very high (quantum efficiency about 80 %) and practically independent of the spectrum of the fluorescent light. With a given intensity of ultra-violet light a pigment with a highly saturated fluorescent colour (narrow emission spectrum) will theıefore, as far as energy is concerned, roughly speaking be able to emit just as much as a white fluorescent pigment having an emission spectıum covering the whole of the visible wavelength range or large parts thereof. If the narrow emission band of the first-mentioned pigment lies near the maximum of the relative visibility curve (5550 Å, in the yellow-green) then the energy supplied produces a very high brightness, whereas the white fluorescent pigment must of necessity emit part of the available energy in less favourably situated spectral ranges and thus lose in brightness. This is clearly demonstrated by the spectral distribution curves given in *fig. 3* for three oil pigments and three roughly corresponding fluorescent pigments.

It is fully in line with this reasoning that the value of 130 mentioned above for the brightness range of the fluorescent picture was found by comparing the darkest black with the brightest (not so very saturated) green. A comparison between black and white led to the smaller factor of 100. These figures confirm what the eye immediately sees when looking at the picture: it is the "radiant" saturated colours, together with the great range of brightness, which lend the typical character to the picture.

After what has been said above it is hardly necessary to point out that the reproduction in fig. 1*a* is incapable of fully representing the impression produced by the fluorescent picture. The brightness range of such a reproduction is in fact

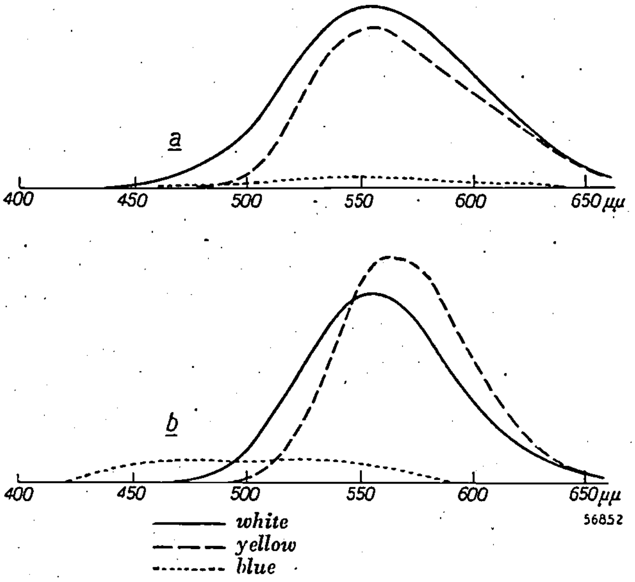


Fig. 3. *a*) The spectral distribution of the reflection has been measured for three normal oil paints, white, yellow and blue. For each wavelength the product of the spectral reflection coefficient and the relative visibility factor has been plotted. Thus the area underneath each curve gives a direct measure for the brightness (for exposure under a light source with uniform energy distribution, standard illuminant *E*). The curve for "white" is everywhere higher than the other curves and thus also has the largest area. The differences in the trend of the curves are but slightly pronounced, the colours not being highly saturated.

b) The same for three fluorescent pigments with about the same colours, white, yellow and blue. The product of the spectral emission coefficient and relative visibility factor has been plotted. Both the curve for yellow and that for blue lie in part higher than that for white. The colours are more highly saturated than in (*a*), but at the same time they are brighter in proportion to the white. The curve for yellow has even a larger area than that of white.

[4]) A clear explanation of this is given in P. J. Bouma, Physical aspects of colour, Philips Techn. Library, publ. Meulenhoff, Amsterdam 1946; see sections 46 and 47.

still less than that of a normal painting, usually not more than 10 to 12. As regards the high relative brightness of the saturated colours, this property can be found to a certain extent in the reproduction when all the "natural" white (and also all shining objects or surfaces) are covered over or removed from the surroundings of the reproduction. If this is not done, then by comparison with the natural white the brightest "white" in the reproduction will give a greyish impression and the characteristic feature of the saturated colours will escape notice.

By taking the same precautions the effect of a bright, saturated colour can be obtained also in the normal techniques of painting: all the white must be purposely tempered. The effect of this can be very striking in the art of glass-painting, where conditions are favourable for avoiding the undesired impression mentioned, namely that the observer notices only the tempering of the white: the light source is behind the stained-glass window and, provided the window does not contain any non-painted or too bright panes, the surroundings are automatically sufficiently tempered. Moreover, tempering of the white in the painting, which with the normal painting techniques would constitute a further unpleasant limitation for the already small range of brightness, is very easily acceptable in glass paintings, as transparent pictures have the property that their brightness range may be much greater than that of pictures viewed in reflected light [5]).

This two-fold similarity, of great brightness range and, in the case of some windows, radiant deep colours, accounts for the fact that when viewing the fluorescent painting most observers remark that the effect reminds one very much of that of a stained-glass window. The broad black lines dividing up the plane, intended to demonstrate the great contrast obtainable between light and dark, undoubtedly help to create the suggestion of a window with leaded panes.

Let us now turn to the technique of painting with fluorescent paints.

In *table III* is a list of the fluorescent powders that were used in the painting of the picture concerned. They were all chosen from the zinc-cadmium-sulphide system with silver as activator, that is of the type [xZn.(1—x) Cd] S-Ag, with different proportions x:/(1− x) of zinc and cadmium. The spectral distribution of the fluorescent light from this system shows a fairly sharp maximum, the position of which (wavelength $\lambda_{\max}$) can be varied within wide limits by the selection of x, as may be seen in the table.

The number of basic colours used may seem to be small compared with the enormous variety of paints available to the painter for instance in the normal oil-painting technique. This limitation, however, was quite voluntary. By a further variation of x and the use of still other fluorescent systems the number of basic colours could be increased at will and it would be easy to compete with the wealth of oil paints. But even the largest possible variety of basic colours will not relieve the painter of the necessity to produce intermediate tints by mixing the paints. Only then can he hope to reproduce the infinitely varying tints occurring in nature to make every stroke of the brush harmonize perfectly with the whole and to create that "vibration" of the colour, so indispensable to give life to every part of the picture. Since, therefore, mixing is unavoidable, the number of basic colours to be used is mainly a question of convenience, of custom and perhaps of individual style. Few artists use more than 20 or 30 oil paints, whilst some purposely use no more than 10 or even less. For a first acquaintance with the new technique the small number of fluorescent pigments given in our list proved to be quite sufficient.

Table III. Fluorescent pigments used for painting the picture. All are of the system [xZn.(1—x) Cd] S-Ag and each pigment is thus characterized by the value of x. The peak of the spectral distribution of the fluorescence of each pigment lies at the wavelength $\lambda_{\max}$.

x	Colour of the fluorescence		Colour in daylight
	$\lambda_{\max}$ (mμ)	colour	
1.00	460	blue	white
0.85	492	green-blue	white
0.77^5	514	green	white
0.68	541	yellow-green	white
0.58	570	yellow	light-yellow
0.50	600	orange	medium yellow
0.39	630	orange-red	yellow
0.33	655	red	orange

Although the mixing of paints thus is a normal routine to the artist, when he comes to mix fluorescent paints he finds there are several unexpected peculiarities. In the first place the mixing is not subtractive, as is the case with all other techniques of painting, but additive.

Of course this fact in itself is not new, but in this connection it is perhaps worth while considering it more closely.

In water-colour painting the paints to be mixed are applied one after the other in thin transparent layers at the desired place. The incident light passes through all the layers in succession and is then reflected back through the layers again (*fig. 4a*). In each layer the light undergoes the selective absorption which determines the visible colour of that layer, so that ultimately only those parts of the spectrum in the emerging light remain which are not appreciably absorbed by any of the layers. This is subtractive mixing. With pastel colours conditions are slightly different. Different coloured chalks (grains of chalk with adsorbed grains of pigment) are applied to the paper and thoroughly mixed by rubbing, for instance with the finger. A ray of light falling on the drawing is reflected (scattered) several times in the microscopic

[5]) This fact is well-known in photography, where one can count on a maximum brightness range of about 30 for positive prints on paper and one of 100, or in some cases even 1000, for transparencies.

accumulation of coloured grains on a particular spot before emerging from the paper again (fig. 4*b*). With each grain the spectrum is weakened in a certain part by the selective reflection determining the colour of that grain, and the emerging light contains only those parts of the spectrum that have not been absorbed by any of the kinds of grain which it has encountered. Thus here, too, we have subtractive mixing, each component contributing towards the ultimate colour by subtracting certain parts of the spectrum. In mixing oil paints the position is rather similar to that existing with pastel colours, except that the mixing is more complete owing to the suspension of the extremely small grains of pigment in the linseed oil; the mixed colour may therefore be more homogeneous.

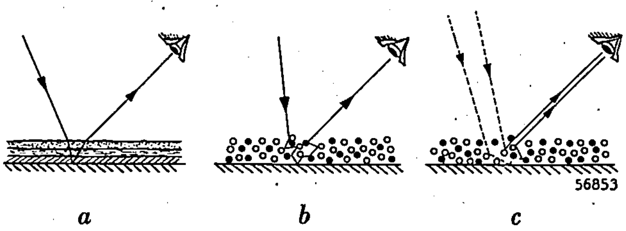


Fig. 4. *a*) In water-colour painting a mixed colour is obtained by applying different transparent layers of paint one on top of the other, each of which reduces a specific part of the spectrum in the transmission of the light: subtractive mixing. (Each layer consists in fact of innumerable grains of pigment; for the sake of clarity this is not indicated here.)

b) In pastel drawing one likewise has subtractive mixing: each ray of light undergoes a number of reflections from the grains of the chalks mixed together, each kind of grain weakening a specific part of the spectrum. The situation is somewhat similar also in oil painting; the pigment grains themselves are mixed in a suspension, the light is scattered in the grains and thereby selectively attenuated before diffusing outwards.

c) In the case of painting with fluorescent paints a mixed colour arises, owing to the radiation emitted from grains of a different kind impinging upon the same element of the retina: additive mixing. The ultra-violet rays falling upon the layer of paint are indicated by dotted lines.

The situation in the case of a mixture of fluorescent pigments resembles that of oil-paints. There is a microscopic accumulation of pigment grains, in this case with different colours of fluorescence (fig. 4*c*). The incident ultra-violet rays are absorbed by the grains, each one of which then radiates its own specific colour, and after being scattered again by other grains these coloured light rays emerge. With many fluorescent substances the grains are "white" (see the first lines of table *III*), that is to say their absorption of the visible light is not selective. The coloured light rays then emerge unaltered and where different kinds of light reach the same spot of the retina the eye adds up all the individual spectral contributions. This is the essence of additive mixing. If the pigments are not completely white (see the last lines of table *III*) we get a sort of blending of additive and subtractive mixture.

In the case of additive mixing the colour point of the resultant colour in the chromaticity triangle can easily be predicted from the colour points of the components; *fig. 5a* serves to remind us of this. For subtractive mixing there is no such simple law (except in a few special cases; cf. fig. 5*b*). This does not mean to say, however, that additive mixing would be easier for the artist than subtractive mixing. The contrary is true. For mixing, such as he is accustomed to, the artist does not rely upon any physical law but works according to a series of rough rules, the product of his own experience and that of others. These rules of course do not hold for additive mixing. Thus, the painter is astonished when in the mixing of say a yellow and a blue fluorescent paint he does not get green, as would be the case with normal paints, but a highly unsaturated, perhaps pink or even white tint (see fig. 5*a*). Mixing will therefore continuously give rise to surprising results, at least until one gets accustomed to the mixing of these paints.

We may mention a second unexpected technical problem that occurred in painting the fluorescent picture. In order to apply the pigments durably to the canvas they must be dispersed in a binding agent. This binding agent must not be fluorescent, it must not cause any chemical reaction in the fluorescent substances and, furthermore, it must possess all the various properties required of a paint vehicle. More or less by chance, normal linseed oil as used in oil-painting also proves to fulfil the former conditions fairly well, and this was therefore chosen as the binding agent. Now, in the usual oil-painting technique, if a pigment is coarse-grained it is first ground very fine and then mixed with the linseed oil to a paste. The very fine grinding gives normal paints a good covering power and at the same time two or more suspensions of paint obtained in this way can very easily be mixed, resulting in a highly uniform tint. Fluorescent pigments of the system used must not, however, be ground to a fine grain, because this results in a loss of the light yield! Consequently, in order not

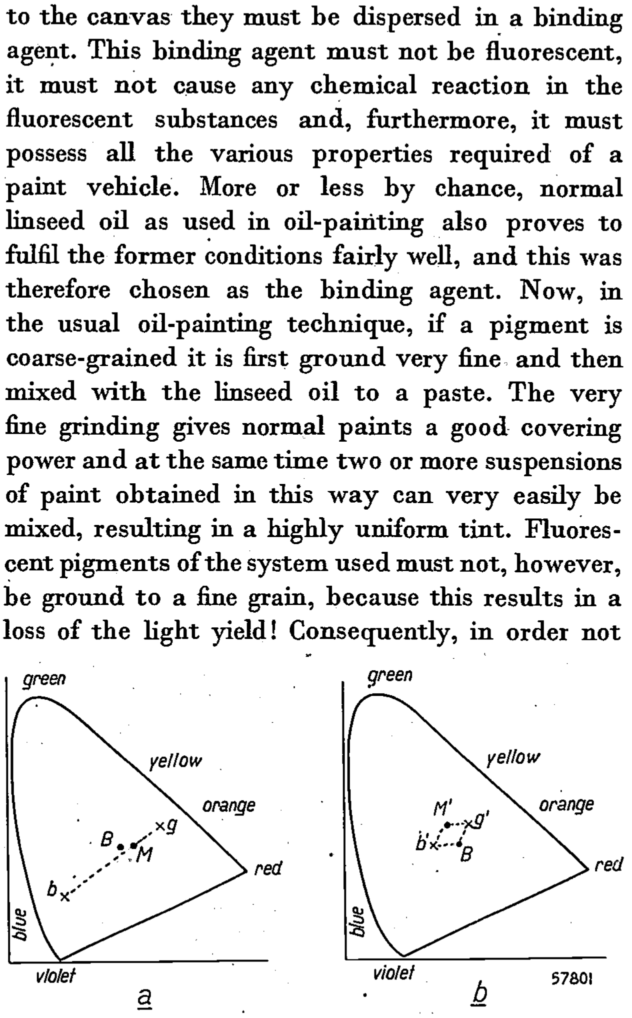


Fig. 5. *a*) Additive mixing of two colours represented in the chromaticity triangle by the points *g* and *b*. The colour point *M* of the mixture lies on the line through *g* and *b*; in the case drawn here (*g* a yellow, *b* a blue colour) by mixing in suitable proportions one can get the mixed colour near the white point *B*.

b) There is no general law for subtractive mixing. For certain simple cases there are approximate rules. If, for instance, two fairly unsaturated colours are mixed, such as *g'* and *b'*, the colour point *M'* of the mixed colour is found approximately as the fourth corner of the parallelogram *Bg'b'M'* (cf. H. E. J. Neugebauer, thesis Dresden 1935).

to sacrifice too much of the beautiful effect of fluorescent paints, one has to work with suspensions of fairly coarse grains, and in mixing such suspensions special precautions must be taken: it is advisable first to sort the powders accoiding to the size of crystal and to mix only those paints that have practically equal crystal sizes. If this precaution is not taken then, probably owing to differences in the rates of sedimentation of the particles of different size, when the paint is being applied to the canvas the mixed components have a tendency to separate.

In principle this difficulty can be overcome by selecting or specially preparing fluorescent substances which without grinding already have the desired very fine structure found with normal paints (size of grain, for instance, 20 μ). The effect of grinding upon the light yield, which causes the said complication, can, on the other hand, be turned to advantage for shading down colour spots where such is desired. In the normal techniques this can be done by admixing some black (or other dark colour) and for a fluorescent picture one could similarly dilute the paint suspension with "black", i.e. non-fluorescent substances, such as for instance common white (!) chalk. Tempering by using a pigment that has been ground, however, gives a more homogeneous impression.

Finally a few words must be said about other practical questions which arise in the new technique. In order to avoid photo-chemical dissociation, as a result of which the "luminosity" of the picture would in course of time diminish, it is important that the crystals should be well protected against moisture. The embedding of the crystals in the linseed oil vehicle is favourable in this respect, and in addition the completed picture can be sprayed with a covering layei of zapon lacquer (which is also non-fluorescent). The backing, as already stated, was ordinary artist's canvas treated with white chalk (again non-fluorescent materials). In daylight the white fluorescent pigments thus stand out only slightly against the background and the form of what is represented in the picture can only be seen faintly (cf. fig. 1*b*). Consequently when the source of ultra-violet radiation is switched on the metamorphosis is all the more surprising, but on the other hand there is the drawback (in addition to the limitation of the brightness range already referred to) that some of the effect is very soon lost if any visible light falls on the picture. If the presence of a tempered visible lighting is to be allowed for, it is better to use a backing that is black (i.e. black also in daylight).

It almost goes without saying that while one is painting the canvas may exclusively be exposed to the light from the mercury lamp (with its filter). The subject, on the other hand, — if one is painting from nature — must be normally lighted. To keep the lighting of the canvas separate from the lighting of the subject is a somewhat unusual problem for the artist and one that is certainly not always easy to solve.

The filter in front of the mercury lamp suppresses not only the visible light but also the short wavelength ultra-violet, i.e. the strong mercuiy line of wavelength 2537 Å. It is almost exclusively the mercury line of wavelength 3650 Å that is responsible for exciting the fluorescence. Irradiation with this wavelength has practically no effect upon the skin, so that there is no fear of suffering from erythema on the hands even when working on the canvas for long periods under the mercury lamp.

May the expeiiment described and the experience communicated here induce painteis to tiy out themselves the possibilities of the new technique.

Summary. One of the authors (J. H. L. J.) has painted a picture with fluorescent paints. The pigments — 8 different ones chosen from the system of zinc-cadmium-sulphide with silver as activator — were mixed with linseed oil; the canvas was ordinary artist's canvas treated with white chalk; the finished picture was sprayed with zapon lacquer. All these materials fulfil the requirement that they themselves should not show any fluorescence. When preparing intermediate colours from two or more fluorescent pigments one is dealing with additive mixing; the rough mixing rules which the artist usually applies and which are based on subtractive mixing do not hold here. The fluorescent powders used in this experiment presented some difficulty in the mixing because they had to be used in the form of rather coarse grains; grinding these pigments would cause a loss of light yield. — The artistic value of painting with fluorescent pigments lies, in the opinion of the authors, in the great brightness range thereby attainable; this is very much greater than the maximum obtainable by the normal techniques of painting, as is proved by actual measurements. Moreover, the fluorescent picture shows exceptionally bright saturated colours, some even brighter than the brightest white, a property due to the fact that the yield of light conversion is practically the same for all fluorescent pigments. A large brightness range and "radiant" colours are likewise found in some stained-glass windows, to which the fluorescent painting therefore shows a striking resemblance.

THE MEASUREMENT OF CHANGES IN LENGTH WITH THE AID OF STRAIN GAUGES

by A. L. BIERMASZ and H. HOEKSTRA.

621.317.39:630.172.222

For some decades the technique of measuring electrical quantities has been ahead of that for the measurement of other, e.g. mechanical, quantities, as far as sensitivity and convenience are concerned. This advance has been further augmented by the development of amplifiers, cathode-ray oscillographs and the like. It is therefore obvious that this advanced technique should also be made available in other fields of measuring by means of instruments designed for converting non-electrical into electrical quantities. An example of such an instrument is the strain gauge dealt with in this article, which shows differences in resistance when its length is changed, which can be measured electrically. In engineering there are wide fields of application for strain gauges in combination with a measuring bridge and/or a cathode-ray oscillograph.

A fundamental problem in the dimensioning of any mechanical construction is the distribution of stresses arising under different conditions of loading. Only with the knowledge of this distribution of stresses — togethei with the necessary knowledge of the properties of materials — can one arrive at the ideal construction which is nowhere too weak nor anywhere unnecessarily strong.

However, if the construction is complicated great difficulties are encountered in calculating the distribution of stresses. One must therefore often resort to practical measurements on existing constructions (or models). Usually measurements can only be taken on the surfaces, but it is just there that the stresses are as a rule greatest. As a measure for the stress in for instance a rod or bar under a tensile load one takes the change in length of the rod, for as long as the deformation is elastic, it is proportional to the stress. In more complicated cases the stress can be deduced from the changes in length measured in two or three directions.

For carrying out these measurements there are extensometers with which it is possible to determine the changes taking place in the distance between two suitably chosen points of the construction when the load acting upon it varies. In order to measure this change in distance with sufficient accuracy either mechanical or optical extensometers are employed which magnify the movement with the aid of a set of levers or by optical means before the actual measurement is made.

For simple cases with a static load these extensometers are quite practicable, but they are of no use with more complicated constructions, in places difficult to get at or where vibrations that are not of a very low frequency must be accounted for. We have only to take as an example aircraft construction, where variations in length have to be measured during the flight, often at some hundreds of points at the same time, on the wings, fuselage and tail planes, whilst moreover large vibrations may be superimposed on the static load. In such cases the solution of the problem is offered by what are known as strain gauges, with which variations in length are measured by electrical means, this method possessing, moreover, advantages over extensometers.

Principle and advantages of strain gauges

It has been known for some time that the electrical resistance of a wire changes with the mechanical stress in the wire; Lord Kelvin described this phenomenon as far back as 1856. This forms the fundamental principle of strain gauges, since these consist of a wire which is glued onto the constructional element under test in such a way that it undergoes all the changes in length taking place in that part of the structure. The measured resistance variations of the wire then give a measure of the change in length. It is only during the last ten years, however, that in several countries strain gauges have been developed with a sufficient degree of reliability.

The main reasons why the method of measuring with strain gauges has come to be so widely applied are the following:

1. The gauges are only a few centimetres long and can thus be applied in places which would otherwise be almost inaccessible.
2. The measuring instrument proper is set up at a distance and if desired can be connected successively to a large number of strain gauges.
3. By this method also dynamic loads (vibrations) can be measured.

4. In combination with an oscillograph it is possible
   a) to record the observations photographically,
   b) to observe and record simultaneously the resistance variations of two or more strain gauges.

Construction, components and use of strain gauges

Fig. 1a is a sketch of the larger of the two types of strain gauges made by Philips. It consists of a resistance wire of constantan (55 % Cu, 45% Ni) bent in zig-zag fashion and glued onto a paper carrier. By means of this cairier, when glued onto the object to be tested, the variations in length of the object are faithfully translated in the wiie, whilst at the same time the carrier provides electrical insulation. Connected to each end of the constantan wire by a soldered joint is a copper lead which is also partly glued onto the paper carrier, thereby providing sufficient safeguard against any strain on the leads being transmitted to the thin constantan wire. The resistance of this gauge is 600 ohms.

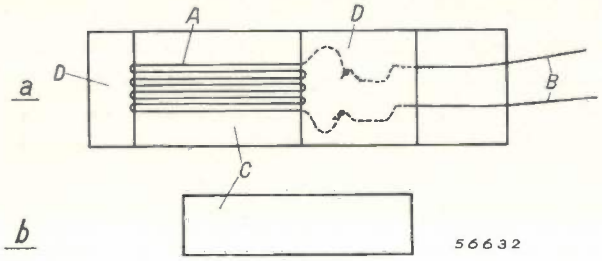


Fig. 1. *a*) Sketch of the strain gauge GM 4472. *A* = resistance wire, *B* = leads, *C* = paper carrier (actual size), *D* = paper covering strips. *b*) Carrier of the strain gauge GM 4473 (actual size).

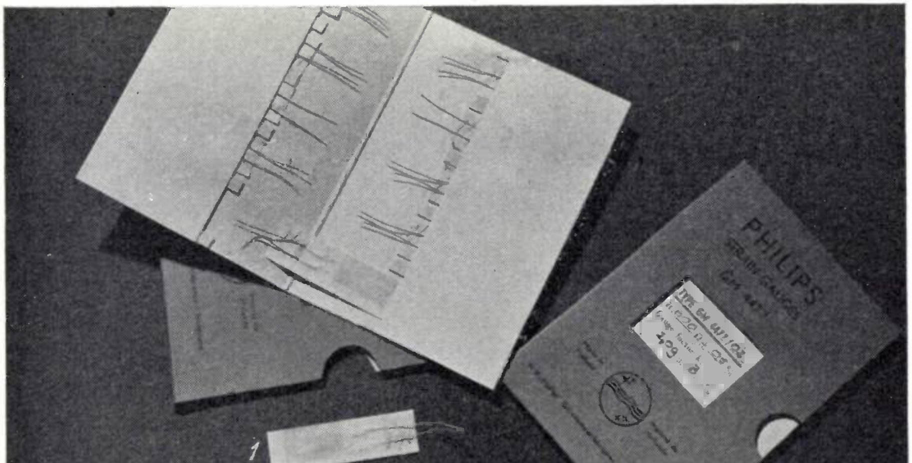
The smaller type (GM 4473, fig. 1*b*) is of a similar construction but has a resistance of 120 ohms. Owing to its smaller dimensions it can be applied in even less accessible places more easily than the larger type (GM 4472), which, however, with its larger surface, can dissipate a greater power, thus allowing of more sensitive measuring.

Fig. 2 shows a specimen of each of the two types and the cases in which the strain gauges aie supplied in packets of 10. The frequency range of the vibrations covered by the strain gauge extends to above 10,000 c/s.

We shall now discuss first the principal components of the strain gauge: the resistance wire, the glue and the paper carrier, and then the method of gluing.

The resistance wire

Since the relative resistance variation, even with the greatest elongation occurring, is only small (in the order of 0.1%), one has to guard against the effect of accidental resistance variations that may occur for instance in switches oi in the long connections often necessary between the stiain gauge and the measuring instruments. It is therefore essential that the resistance of the strain gauge should be at the least in the order of 100 ohms. In order to get



Figl 2. *1* = strain gauge GM 4472, *2* = strain gauge GM 4473; also the two cases in which the strain gauges are sold.

such a resistance in the strain gauge only a few centimetres long with a reasonably robust wire, a material has to be used which possesses a high specific resistance. This excludes the pure metals and leaves only alloys like constantan and chromium nickel, and carbon.

In the second place the choice of material for the wire is determined by the behaviour of the gauge factor k, that is the ratio of the relative variation $\Delta R/R$ of the resistance to the relative variation $\Delta l/l$ of the length:

$$k = \frac{\Delta R}{R} : \frac{\Delta l}{l} = \frac{\Delta R}{R} \cdot \frac{1}{e}.$$

($\Delta l/l = e$ is also called the elongation.) The greater the value of k (absolute), the higher is the sensitivity of the measurement, but a linear relation between $\Delta R/R$ and e (i.e. a gauge factor which is independent of the amount of the elongation) is even more important.

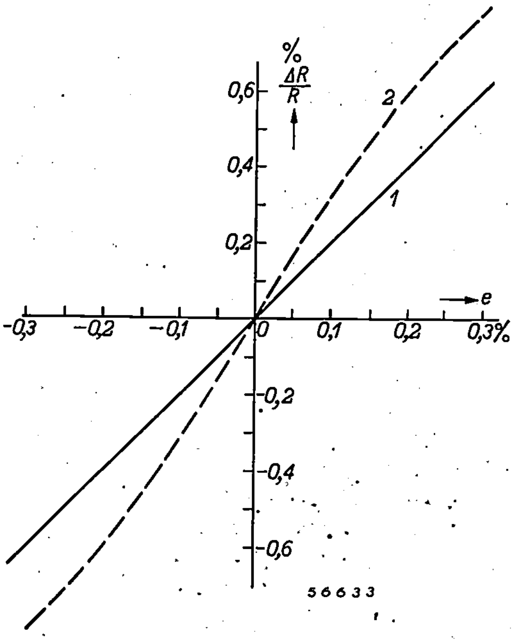


Fig. 3. The relative resistance variations $\Delta R/R$ as function of the elongation e, (*1*) for constantan, (*2*) approximate curve of a non-linear relation as found with chromium nickel.

The high k value found with carbon (about 20), which is due to the changeability of the transfer resistance between the particles of carbon, makes it tempting to use this material in strain gauges [1]), but the results in the long run are not sufficiently reproducible and as a consequence carbon is not used as a material for the resistance element any more.

The relation between $\Delta R/R$ and e for constantan is represented by the continuous line in *fig. 3.* This relation is practically linear, with $k \approx 2$ both in the elastic area (e between $\pm 0.2\%$) and outside it. With other alloys having a high specific resistance this relation is a non-linear one: chromium nickel for instance gives a value of $k \approx 3$ for a small elongation, this value dropping to about 2 in the plastic area, as roughly indicated by the broken curve in fig. 3. This non-linear characteristic is a great disadvantage in practice and explains why constantan is preferred.

The resistance of a conductor varies not only with the elongation but also with the temperature. Where only a rapidly changing elongation has to be measured (vibrations) gradual variations in resistance due to temperature fluctuations are not disturbing, but in the case of static loads they may well be so. We shall presently discuss a method whereby the influence of temperature is compensated, but it is nevertheless desirable that the temperature coefficient (a) of the strain gauge should be as small as possible. In this respect, too, constantan is highly favourable, a being in the order of 1×10^{-6} compared with about 100×10^{-6} for chromium nickel.

In *table I* the quantities discussed are listed for constantan, chromium nickel, carbon and, for comparison, a pure metal (iron).

Table I. Approximate values of the specific resistance ϱ, the gauge factor $k = (\Delta R/R)$: e (within the limits of elasticity) and the temperature coefficient a of the resistance of some materials.

Material	ϱ $\mu\Omega\cdot$m	k	a 10^{-6} (°C)$^{-1}$
Iron	0.1	—4	5000
Constantan	0.5	2	1-3
Chromium nickel	1	3	100
Carbon	70	20	—500

Here we can go rather more deeply into the value of the gauge factor k of metals.

The resistance of a wire with diameter D is:

$$R = \frac{l\varrho}{\frac{\pi}{4} D^2}. \quad \ldots \ldots \ldots \quad (1)$$

When the wire is stretched (or compressed) not only l changes but also D and ϱ. As regards the change in diameter Poisson's relation applies:

$$-\frac{dD}{D} : \frac{dl}{l} = \mu,$$

in which for most metals, in the elastic area, $\mu \approx 0.30$-0.35. As regards the specific resistance ϱ it is known that when a metal is subjected to compressive forces from all sides ϱ changes with the density, thus with the volume V. In the case

[1]) See for instance S. L. de Bruin, The investigation of rapidly changing mechanical stresses with the cathode-ray oscillograph, Philips Techn. Rev. 5, 26-28, 1940.

of constantan we then have the linear relation

$$\frac{\mathrm{d}\varrho}{\varrho}:\frac{\mathrm{d}V}{V}=c, \quad \dots\dots\dots \quad (2)$$

with $c = 1.13$. We shall assume, as Opechowski does, that this relation also holds for extension and contraction.

Since $V = {}^1\!/_4\pi D^2 l$ equation (1) can be written as follows:

$$R = \frac{\varrho l^2}{V}.$$

By differentation we arrive at

$$\frac{\mathrm{d}R}{R}=\frac{\mathrm{d}\varrho}{\varrho}+2\,\frac{\mathrm{d}l}{l}-\frac{\mathrm{d}V}{V}=2e+(c-1)\,\frac{\mathrm{d}V}{V}. \quad (3)$$

Plastic deformation is characterized by cons.ancy of the volume, hence for all materials (even if eq. (2) does not hold) the rule applies that

$$\frac{\mathrm{d}R}{R}=2e,$$

or

$$k = 2,$$

in accordance with experiment.

With elastic deformation on the other hand

$$\frac{\mathrm{d}V}{V}=2\,\frac{\mathrm{d}D}{D}+\frac{\mathrm{d}l}{l}=(1-2\mu)\,e,$$

so that (3) becomes

$$\frac{\mathrm{d}R}{R}=2e+(c-1)\,(1-2\mu)\,e,$$

or

$$k = 2+(c-1)\,(1-2\mu)\,.$$

For constantan, with $c = 1.13$ and $\mu = 0.325$, this works out to

$$k = 2 + 0.13 \times 0.35 = 2.05,$$

which falls within the extreme values measured (2.09 ± 0.06). It must be said however that with other materials the agreement is much less satisfactory.

The glue and the paper

The paper carrier and the glue used for fixing the constantan wire on the carrier and the latter to the object to be tested have to translate the deformations of the object faithfully to the wire. Consequently the adhesive strength of the glue must be exceptionally high. The best kinds of glue are those having a cellulose base and consisting of long molecules with numerous polar atom groups. It is the latter which result in the strong adhesion (both to paper and to metal) and, moreover, show practically no after-effects (creep).

After-effects in a strain gauge constitute a most undesirable property, because then the unambiguous relation between resistance variation and deformation is lost. Both in the wire as well as in the glue and the paper after-effects must therefore be kept small in comparison with other inaccuracies in the measuring method. This is a requirement that is particularly difficult to meet as regards the glue, but in the manufacture of the Philips strain gauges this has been satisfactorily met. With an elongation of 0.1% for instance the permanent relative resistance change is normally less than 10^{-5}.

The long molecules of cellulose glue have the property of adapting themselves somewhat to the direction of the elongation, thus lending great toughness to the adhesive layer. Other kinds of glue on the other hand become more or less brittle (for instance polysterene, which moreover is less polarized than cellulose glue).

Finally the glue must not take too long to dry; this we will refer to farther on when dealing with the method of gluing .The thinner and more porous the paper, the quicker does the glue dry, and for this reason the thinnest possible paper is used for the Philips strain gauges, having regard to insulation and the necessity for the gauge to remain flat.

For gluing strain gauges onto the objects to be tested, Philips have placed on the market tubes of glue (GM 4479) which can be used at temperatures up to 60 °C.

Gluing on the strain gauges

In order to ensure proper adhesion of the strain gauge to the object being tested, at the place where it is to be applied the metal must be thoroughly cleaned first with emery paper and then with some grease-remover, for instance pure acetone. The cleaned surface must not then be touched with the hands. A thin layer of glue (GM 4479) is then spread out over it and when this is dry the back of the strain gauge is likewise covered with a thin layer of glue and pressed down on top of the other layer, taking care that the longitudinal axis of the strain gauge coincides exactly with the direction of the deformation to be measured.

After about 6 hours the glue is dry enough and the adhesion is sufficient for taking rough measurements with the strain gauge. For accurate measurements, however, particularly when they are to be of a long duration, the glue should be allowed to dry a few days, preferably while heating to 70 °C. Even after that every possible precaution should be taken to exclude moisture, since moisture has a serious effect upon the measuring results, due to two causes: variable insulation resistance and variable mechanical stress both in the glue and in the paper. As regards the insulation resistance, if for instance with a strain gauge of 600 ohms it is desired to measure resistance variations of 0.1% with no greater error than 1% then the insulation resistance (in parallel with the strain gauge) would answer high demands of constancy if it were only in the order of 60 megohms. If, on the other hand, it is of the order of 1000 megohms, as is to be reached with the above-mentioned measures, then great changes in the insulation resistance are harmless.

In the second place moisture causes local swelling of the glue and of the fibres of the paper. Even in a glued-on strain gauge, which is not free to expand, such deformations are apt to arise and cause perceptible, irregular resistance changes.

Precision measuiements of long duration should not be begun until the insulation resistance has reached 500 MΩ. The gauge must then be shut off from the air either with a layer of wax or with a sheet of rubber glued round the gauge and enclosing at the same time a moisture-absorbing substance (silica gel). This latter method has been developed in the Laboratory for Applied Mechanics of the Foundation for Technical-Industrial Research ("T.N.O.") at Delft (Netherlands), with which institution we are fruitfully cooperating in this field. It is a method that has proved to be very satisfactory, even when taking measurements over a long period in the open air in rainy weather.

The measuring apparatus

We shall now proceed to discuss the apparatus employed for measuring small resistance variations. A Wheatstone bridge where one branch is formed by the strain gauge (G_1 in *fig. 4*) is suitable for this, though in some cases, as will be shown presently, a simpler system may also answer the purpose. The bridge circuit, however, is to be preferred whenever the influence of temperature has to be excluded. The following will show what this effect is.

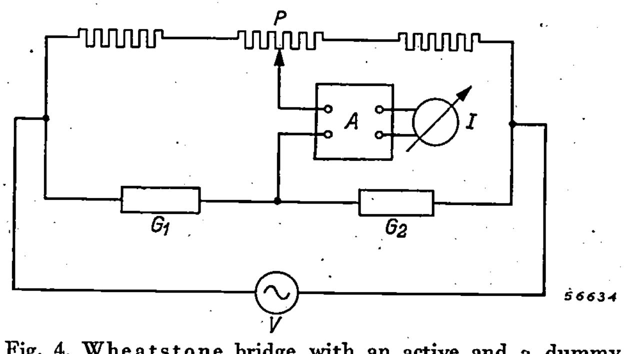


Fig. 4. Wheatstone bridge with an active and a dummy gauge (G_1 and G_2). V = alternating voltage supply, P = potentiometer for balancing the bridge, A = amplifier, I = indicator.

The effect of temperature

When the temperature of the object to which the strain gauge is affixed changes while the measurements are being taken this may lead to two errors:

1) The resistance variation observed in the strain gauge is partly due to the temperature coefficient of constantan, small as it may be; this part of the variation must not be interpreted as a deformation.
2) Even in the absence of stresses the object undergoes deformation, namely thermal expansion, which it is not as a rule desired to measure but which shows itself as a variation of the resistance of the strain gauge because of the usually different expansion coefficient.

Both these errors can for the greater part be avoided by using a compensating or dummy strain gauge in the bridge circuit. This is a strain gauge of the same type as the "active" or primary gauge G_1 (fig. 4); it forms an adjacent branch of the bridge (G_2). It is glued onto the same material as G_1 but this piece of material is not mechanically loaded, and the dummy gauge has to be placed close enough to the actual measuring point so that it undergoes the same temperature variations as the active strain gauge. Thus the active and the dummy gauges are subject to the same thermal influences. The resistance variations resulting cancel each other in the bridge circuit, so that only the variations which correspond to mechanical stresses in the object are measured.

Often it is possible to fix the dummy gauge in such a way that it not only compensates the thermal effect but also contributes towards the unbalance of the bridge, so that the measuring system becomes more sensitive. In such a case the compensating gauge is fixed at a place on the object where the material is subject to contraction while the "primary" gauge undergoes an elongation, or vice versa. We shall come across some examples of this farther on.

The two types of measuring bridge developed for use with strain gauges have both been designed for working with a compensating gauge.

We shall now consider the measuring bridges more closely.

Measuring bridge for static load

The diagram of fig. 4 represents the principle of both types of measuring bridge. Before the active gauge (G_1) is deformed, the bridge is balanced with the aid of the potentiometer P, that is to say, the deflection of the indicator I is set to zero. In order to increase the sensitivity the indicator is preceded by an amplifier. So as not to complicate the latter the bridge is not fed with D.C. but with an alternating voltage (derived from a valve oscillator incorporated in the apparatus).

Owing to an alternating voltage being used, the bridge must also be balanced capacitively. The measures taken for this are not indicated in fig. 4 and we shall not go into them here.

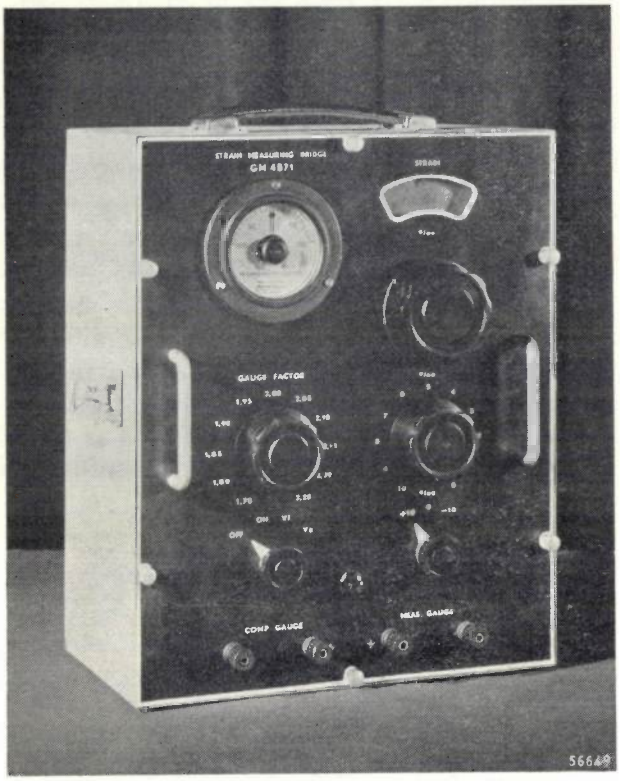


Fig. 5. Measuring bridge (GM 4571) using the zero method for measuring static loads. *To the right*, one above the other, three knobs for balancing the bridge; *to the left*, from top to bottom, the indicator, the correcting knob for the gauge factor of the strain gauge, and the knob for switching on and off, acting at the same time as the switch by means of which the indicator can be used for checking the batteries.

Deformation of the active strain gauge will throw the bridge out of balance. As is known, one can then proceed in two ways:

a) rebalance the bridge with the potentiometer *P* and determine the resistance variations from the two positions of *P*;

b) use the deflection of the indicator as a direct measure of the resistance variation of G_1.

The measuring bridge of the type GM 4571 (*fig.* 5) works on the zero method mentioned sub (*a*). This has the advantage that the result is independent of the degree of amplification, but on the other hand it is of course limited to static load only.

The potentiometer is provided with a scale from which the elongation can be read directly. In order to allow for the fact that there is a slight difference in the gauge factor k as between one strain gauge and another, a correction knob is provided which has to be previously set in the position corresponding to the gauge factor of the strain gauge used. This factor is shown on the strain gauge case (fig. 2).

As indicator a moving-coil meter is used, which is connected via selenium rectifiers [2]) to the output of the amplifier, and the indicator can be used at the same time as a voltmeter for checking the batteries feeding the oscillator and the amplifier. These batteries make the bridge particularly suitable for taking measurements in places where no A.C. mains are available.

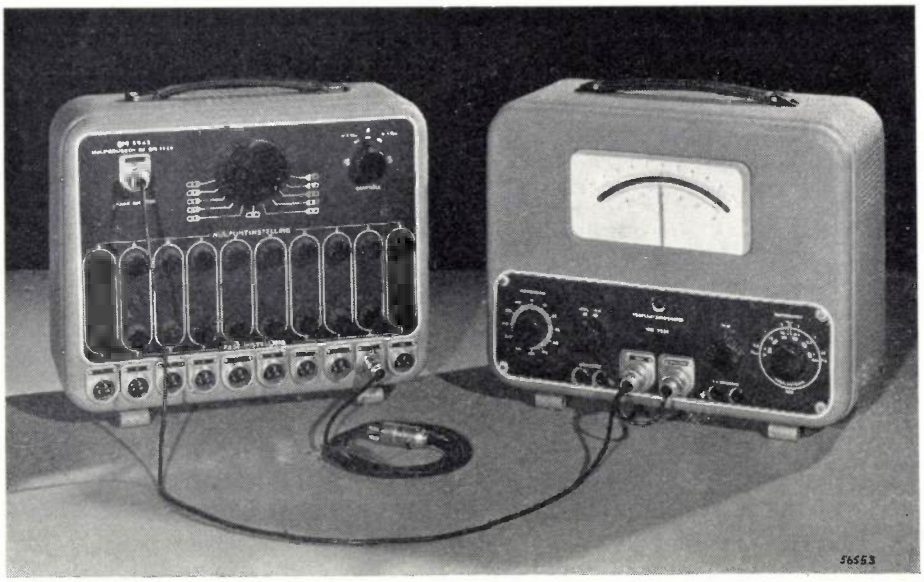


Fig. 6. *On the right* the measuring bridge GM 5536 for measuring static and dynamic loads. The elongation can be read directly from the meter. The amplified output voltage from the bridge can be applied to an oscillograph. *On the left* the switch GM 5545 with which the bridge can be connected successively to any one of 10 pairs of strain gauges.

[2]) J. J. A. Ploos van Amstel, Small selenium rectifiers, Philips Techn. Rev. 9, 267-276, 1947.

This bridge is suitable for use with strain gauges of more than 100 ohms resistance.

Measuring bridge for static and dynamic loads

The measuring bridge of the type GM 5536 (*fig. 6*) is used in the manner indicated above under (*b*).

This bridge has to be balanced before starting to take measurements. When the balance is disturbed it produces an output voltage which in the case of dynamic load is a modulated alternating voltage; the "carrier" of the oscillator feeding the bridge is modulated with the frequencies of the mechanical vibrations. With the aid of a ring modulator [3]) consisting of four selenium rectifiers, the output voltage is demodulated and converted into a rectified voltage which is fed to a moving-coil meter acting as indicator. Static load is to be regarded as a special form of dynamic loading where the frequency of the vibrations is zero. Both for static and for dynamic loads the deflection of the meter is a measure of the resistance variation of the strain gauge.

Instead of the meter, a cathode-ray oscillograph can be used as indicator. By this means the vibrations can be visualized and if necessary recorded photographically. In order to avoid distortion an oscillograph should be used which has an amplifier suitable for the very low frequencies that may occur with mechanical vibrations. Such an oscillograph is for instance the type GM 3156 [4]), which is suitable for frequencies down to 1 cycle per second.

As an accessory for the measuring bridge GM 5536, which is intended for use with strain gauges of 600 ohms (and if need be higher), a change-over switch is provided (fig. 6), by means of which the bridge can be quickly connected in succession to 10 different combinations of an active and a dummy strain gauge. The change-over switch is provided with a device by means of which the meter or the oscillograph can be calibrated.

A cathode-ray oscillograph offers the possibility of visualizing simultaneously the resistance variations of two or more strain gauges, so that time and phase differences of the vibrations can be determined. For that purpose one or more electronic switches [5]) can be used.

Circuit for an exclusively dynamic load

If measurements have to be taken only under a dynamic load then no account need be taken of gradual variations due to temperature fluctations. In that case a compensating strain gauge and a bridge circuit are superfluous and one can manage quite well with the simple arrangement illustrated in *fig. 7*, where the strain gauge is connected to a battery via a resistor. Resistance variations of the strain gauge set up across its terminals voltage variations which are applied to an oscillograph via a pre-amplifier.

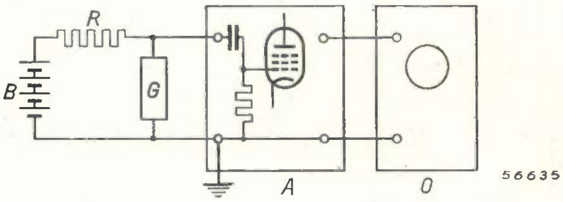


Fig. 7. Circuit for measuring dynamic loads. B = battery (45 V), R = series resistor, G = strain gauge, A = pre-amplifier, O = oscillograph.

Fig. 8. The pre-amplifier GM 4570, which is suitable for the arrangement according to fig. 7. Gain 4 times or 20 times.

As pre-amplifier use can be made of the unit illustrated in *fig. 8*.

The sensitivity of the oscillograph can be calibrated with a known alternating voltage.

Applications

The applications of strain gauges in engineering can be divided into direct and indirect applications. Under the former are to be understood cases where strain gauges are used for measuring stresses in all sorts of mechanical constructions, whereas indirect applications cover their use with accessory apparatus required for taking various measurements.

[3]) See e.g. Philips Techn. Rev. 7, 86, fig. 6, 1942.

[4]) S. L. de Bruin and C. Dorsman, A cathode-ray oscillograph for use in tool making Philips Techn. Rev. 5, 277-285, 1940.

[5]) See e.g. E. E. Carpentier, An electronic switch with variable commutating frequency, Philips Techn. Rev. 9, 340-346, 1947.

Direct applications

An extensive field for the use of strain gauges lies in the measuring of stresses in bridges, cranes, ships, aircraft, rolling mills, etc. As remarked in the introduction, better knowledge of the stresses occurring leads to more rational constructions with more reasonable margins of safety.

Of the numerous examples of the successful direct application of strain gauges we shall mention here only three.

In the first place there are the measurements taken on the hull of a ship at the time of launching, when a very complicated state of stress may occur. In order to measure the changes taking place in the stresses at a certain point of the ship's hull three strain gauges are applied at that spot close together at angles of 120°. The state of stress can then be easily calculated from the resistance variations of the three gauges.

A second example is the measuring of torque in a shaft. Two strain gauges are glued onto the shaft with an angle of 90° between them and at an angle of 45° with the centre line of the shaft (*fig. 9*). When torque arises one of the gauges is stretched and the other contracted. The torsion is calculated from the resistance variations of the two gauges.

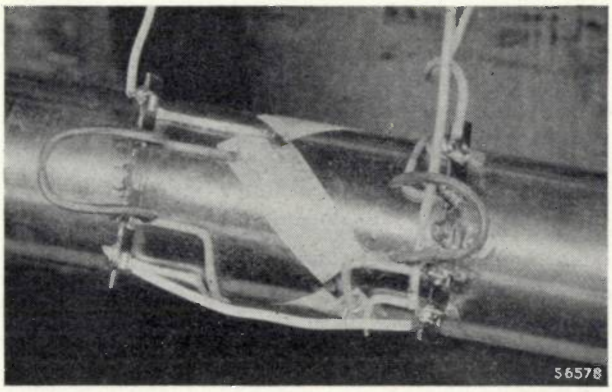


Fig. 9. Torsional vibrations in a shaft can be measured with two strain gauges at right angles to each other and making an angle of 45° with the centre line of the shaft. (In the experiment illustrated here four gauges were used, connected in parallel in pairs.)

As third example *fig. 10* shows the set-up used for measurements which were taken in cooperation with the Netherlands Railways to investigate the stresses arising in one of the arches carrying the overhead electric conductors in the event of rupture of the overhead line or the cable from which it is suspended. Strain gauges were glued onto one of the uprights at certain points and connected to a

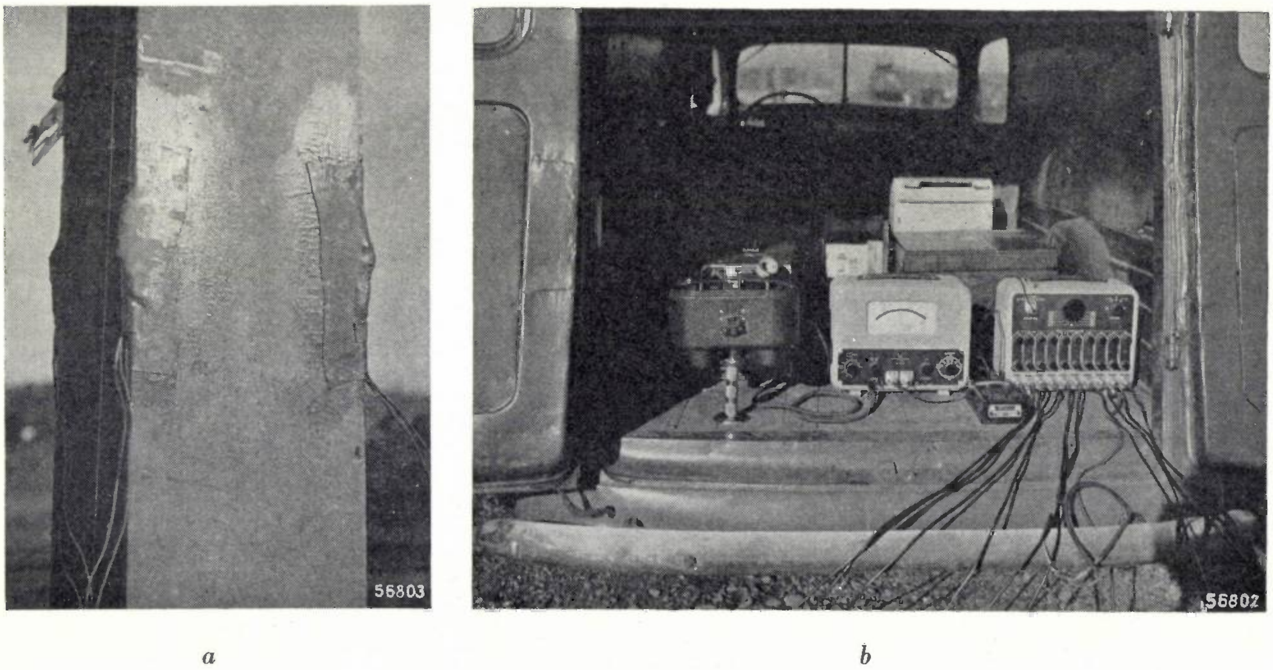


Fig. 10. *a*) Upright of an arch carrying the overhead power line of an electric railway, with four strain gauges for measuring the static load variations arising in the upright in the event of rupture of the overhead cable. Two of the strain gauges undergoing variations in length in the opposite sense to each other form a pair. They are complementary to each other as regards the elongation to be measured, but mutually compensate the effect of temperature variations. In order to exclude moisture the strain gauges are covered with a sheet of rubber enclosing a quantity of silica gel.

b) For measuring vibrations the strain gauges glued onto the upright could also be connected to a measuring bridge for dynamic loads. In the back of the car are to be seen from right to left the measuring bridge (GM 5536,) the circuit-selecting unit (GM 5545) and the oscillograph with camera set up in front of it.

measuring bridge GM 5536 (fig. 6). Upon the overhead power line being cut through, vibrations are set up which modulate the output voltage of the ,bridge, as may be seen from the oscillogram in

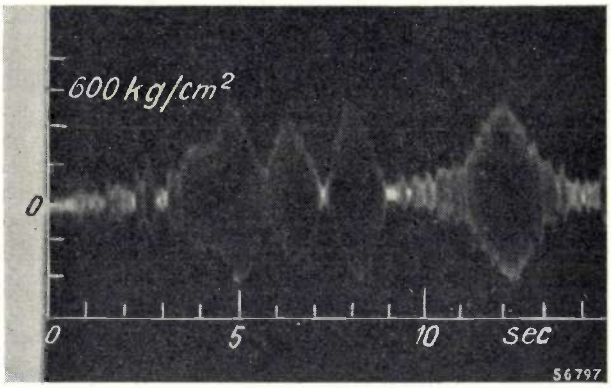


Fig. 11. Oscillogram (produced with the apparatus illustrated in fig. 10*b*) of the stress arising in the arch upright when the overhead power line is cut through close to the upright. This sets up vibrations which are propagated along the power line and the carrier cable and which are reflected partly in the next suspension points. This oscillogram gives a picture of the resultant interferences. The highest peak corresponds to a stress of about 600 kg/cm^2 at the point where the strain gauges are applied.

fig. 11. From the measured amplitude it was deduced that at the point where the strain gauge was applied stresses occurred amounting to 600 kg/cm^2; for further details see the caption of fig. 11.

Indirect applications

Of the indirect applications of strain gauges we shall likewise confine ourselves to only a few examples which, in so far as they lie in the field of mechanics, concern apparatus developed by the "T.N.O." (Delft).

The ground-pressure and ground-water-pressure meters so far used have the drawback, among others, that they need a long adjustment time and often do not give any high degree of accuracy. Strain gauges provide a great improvement in this respect. The ground pressure or water pressure is transmitted to a diaphragm onto which a strain gauge has been fixed. The extent to which the diaphragm gives is a measure of the pressure, which can be made directly readable without any noticeable inertia and with a high degree of accuracy.

In *fig. 12* a dynamometer is illustrated with which forces of 1 to 10 tons can be accurately measured. The force to be measured is transmitted to a steel ring to which strain gauges have been applied for determining the elongation. A calibrating table gives the force as a function of the elongation measured.

As last example we would mention an entirely different field of application, namely physiology, where strain gauges are used for oscillographing the heart beat, respiration and suchlike [6]).

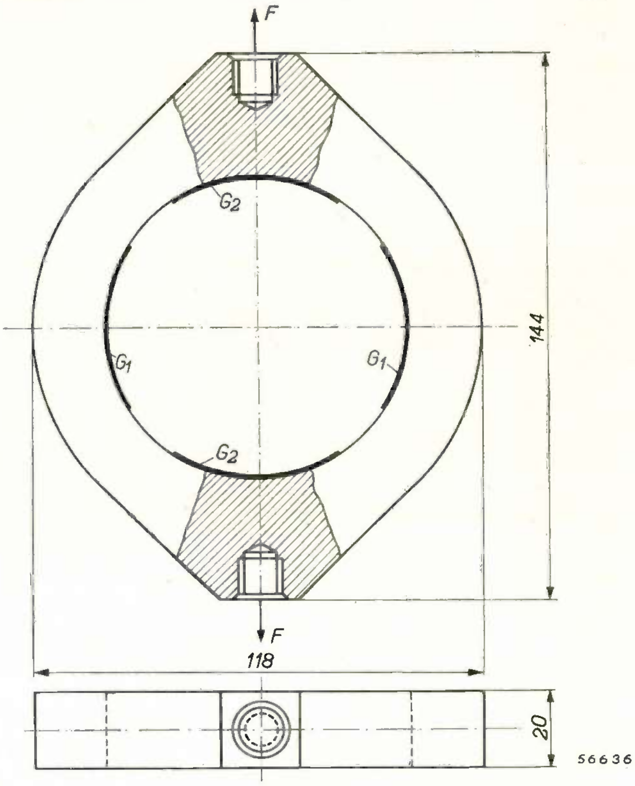


Fig. 12. Dynamometer (force meter) consisting of a steel ring which is elongated or contracted by the force F to be measured. The deformation of the ring is a measure of the force. It is measured by means of the two active strain gauges G_1. The gauges G_2 provide for temperature compensation and also act somewhat in the active sense since they change in length in the opposite sense to that of G_1. The dimensions given are in millimetres.

Summary. Strain gauges are being used more and more as a means for determining changes in length. In their present form they consist of a wire bent in zig-zag fashion and glued onto a paper-carrier. The two types manufactured by Philips are 5.5 cm and 3 cm long (GM 4472, 600 ohms, and GM 4473, 120 ohms). The wire is of constantan, which has a constant gauge factor k (about 2.1), i.e. the ratio of the resistance variation to the elongation. The glue used for fixing the wire onto the carrier and the latter onto the workpiece is a strongly polar cellulose glue with great adhesive strength and showing very little "creep". This glue is sold in tubes. Two measuring bridges are described for determining the resistance variations: one for static loads only and the other for dynamic loads (vibrations) as well. Both are suitable for applying a second, compensating strain gauge for neutralizing the effect of temperature. This temperature compensation is not needed when only vibrations are to be measured; one strain gauge, a pre-amplifier and an oscillograph are then sufficient. The frequency range of the vibrations to which a strain gauge can respond extends to over 10,000 c/s. Finally a number of applications are discussed, both indirect (with force meters, pressure meters and suchlike) and direct, for determining the mechanical stresses in all sorts of constructional elements, particularly where other measuring methods are made impossible owing to the presence of vibrations or to inaccessibility.

[6]) See e. g. H. R. Bierman, A device for measuring physiologic phenomena, using the bonded electrical wire resistance strain gauge, Rev. sci. Instr. **19**, 707-710, 1948 (No. 10).

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

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

1810*: J. L. Meijering: Hardening of metals by internal oxidation (Physical Society, Bristol Conference Report, 140-151, 1948).

Certain alloys of Ag, Cu and Ni with a few atomic percent of a homogeneously dissolved metal having a sufficient affinity for oxygen (Mg, Al) can be hardened by diffusing oxygen into them. The greater the affinity for oxygen of the basis metal, the greater must be the affinity of the solute to produce oxide which is not too coarse, as the conglomeration will take place by way of the atoms. This is worked out in a tentative thermodynamic scheme. X-ray and electrical resistivity measurements show that the MgO and Al_2O_3 particles which harden silver are very small indeed. The mechanical properties are not much affected by long annealing, especially when the dissolved free oxygen is not removed after the oxidation. Recrystallization is slowed down considerably. A drawback is the intercrystalline brittleness of these materials, which is less serious when somewhat smaller hardnesses are aimed at; single crystals are completely ductile. In some alloys complications arise at higher concentrations from the formation of inner oxide films, which slow down the penetration of oxygen (See Philips Res. Rep. **2**, 81-102 and 260-280, 1947).

1811: R. Vermeulen: Melodic scales (J. Acoust. Soc. Amer. **20**, 545-549, 1948, No. 4).

In this article an attempt is made to give a derivation of musical scales on a melodic rather than harmonic principle. Starting from the hypothesis that the scale shall contain, at most, two different intervals and shall contain a reasonably perfect fifth, all possible scales have been investigated. The customary scale appears here in the form of the Pythagorean scale. Apart from this certain other possibilities are found, which partly coincide with other existing scales.

1812: J. M. Stevels: Some chemical aspects of opacification of vitreous enamels (J. Sheet Metal Industries **25**, 2234-2237, 2240, 1948).

A vitreous enamel generally consists of a vitreous phase in which another phase is dispersed. For this reason vitreous enamels show opacification. The dispersed phase may be formed: 1) by crystals formed from the melt (devitrification; not used in practice), 2) by introduction of crystals insoluble in the melt into the batch, 3) by two unmixible vitreous phases. The influence on the opacification of particle size and particle shape is discussed extensively. Finally the influence of the chemical composition of the vitreous and the dispersed phase is discussed.

1813: H. B. C. Casimir: On the attraction between two perfectly conducting plates (Proc. Kon. Ned. Akad. Wetensch. Amsterdam **51**, 793-795, 1948, No. 7).

It is proved that between two parallel metallic plates an attractive force exists, which is independent of the material of the plates as long as the distance is such that for waves of a wavelength comparable with the distance the depth of penetration is small compared with the distance. This force may be interpreted as a (negative) pressure due to the zero point energy of electromagnetic waves. Similar effects come to the fore in calculating the influence of retardation on the Van der Waals-London attraction between two atoms (see these abstracts, No. **1793**).

1814: C. J. Bouwkamp: A note on Mathieu functions (Proc. Kon. Ned. Akad. Wetensch. Amsterdam **51**, 891-893, 1948, No. 7).

A short proof of the theorem stating that the Mathieu equation $y'' + (a - 2q\cos 2z)\,y = 0$ does not possess two linearly independent periodic solutions of period 2π, unless $q = 0$ and $a = n^2$ (n integer). A conjecture of Mac Lachlan relating to the characteristic values for imaginary q is disproved.

1815: K. F. Niessen: Nodal planes in a perturbed cavity resonator, I (Appl. sci. Res. 's-Gravenhage, **B1**, 187-194, 1948, No. 3).

In this paper a cavity resonator is considered which is generated from a prismatic cavity with square cross section by rotation of one of the walls through a small angle δ about its edge. The electric field is supposed to be parallel to that edge. The change in resonance frequency and the distortion of the electromagnetic field due to the perturbation are calculated.

VOL. 11 No. 2, pp. 33-64 AUGUST 1949

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

PLASTICS AND THEIR APPLICATION IN THE ELECTROTECHNICAL INDUSTRY

by J. C. DERKSEN and M. STEL.

679.56:679.57

Plastics is a name with which almost every one is familiar now that fancy goods and objects of domestic use are being made from these materials. Plastics were first used in the electrotechnical industry more than 20 years ago but the rapid developments of recent years in this field have created all kinds of new possibilities. All plastics are essentially comprised of macro-molecules, each consisting of a very large number of atoms and formed from much simpler molecules by a repeated chemical binding process. There is a very wide choice of available basic materials, and so a great variety of plastics can be made with a surprising range of properties.

Since the beginning of the nineteenth century chemists have developed a number of new materials for industrial uses and for the community at large. From natural rubber has been derived, by chemical treatment with sulphur, what is now known as vulcanized rubber, this having better technical properties than natural rubber. Cellulose has been extracted from vegetable fibres and is converted with nitric acid into nitrocellulose, from which are derived celluloid (a substitute for ivory) and rayon (artificial silk). Casein has been used for making artificial horn.

During the last 20 years these developments have taken the world by storm: the artificial products industry has advanced by leaps and bounds. This industry however has reached a new stage of development, in that whereas formerly the aim was to imitate natural materials now many synthetic products are being produced which have exceptional properties not found in nature and which, therefore, can no longer be said to be artificial. "Philite" and "organic glass" are examples of such products, all of which have been given the name of "plastics". By this term is understood any material the main component of which is a macro-molecular substance, usually organic, and which in some stage of the processing is either plastic or liquid and solidifies in a later stage.

In almost every domain we come across articles made from plastic materials. These materials play an important part also in the electrotechnical industry: the development of radio, television and radar, to mention only a few examples, would certainly have been impossible if no plastics had been available as raw materials.

The fact that the electrotechnical industry finds a great use for plastics is due in the first place to the valuable electrical properties which most of these materials are found to possess, such as high break-down strength, low dielectric losses and high "non-tracking" quality [1]). Polystyrene and polyethylene for instance have very low dielectric losses, and moulding materials of urea-formaldehyde are highly proof against "tracking".

But these favourable electrical properties are not the only reasons, because other materials like porcelain and mica possess them too. The outstanding property which makes plastics so attractive, and by reason of which they excel over the other materials just mentioned, is their easy mouldability, or plasticity. The electrotechnical industry is one of mass-production, for which the technique of the processing of plastics is highly suitable.

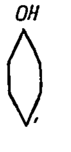
To this is to be added the fact that it is quite easy to give a glossy appearance, coloured if necessary, to objects made from plastics.

[1]) By this it is meant that the material suffers little from electrical discharges between electrodes placed on the surface, especially when the surface is covered more or less with an electrolyte.

These last two properties, plasticity of the material and good finish of the product, are reasons why plastics are also used where the electrical properties are not of such great importance, as in the manufacture of radio cabinets, telephones, door-knobs, lamp fittings, etc.

In this article something will first be said about the molecular structure of plastics, this being followed by a survey of the principal moulding materials and their specific properties. We shall then briefly discuss the methods employed in manufacturing articles and components from plastics. In conclusion a number of examples are mentioned of the use of such articles, mainly in the electrotechnical industry.

Macro-molecular substances



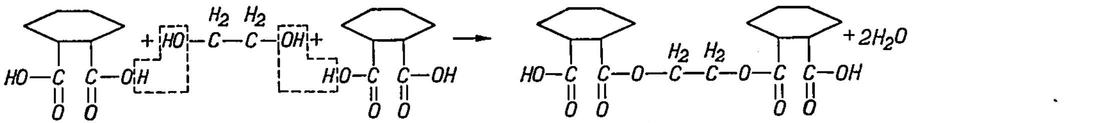
A substance that plays an important part in the preparation of plastics is phenol, the structural formula of which is given here in a simplified form. Compared with that of the substances dealt with below, the phenol molecule is simple in structure. It is a micro-molecular substance with a molecule consisting of 13 atoms. Plastics, on the other hand, are all macro-molecular, being built up in very large molecules each consisting of 1000 or more chemically bound atoms. As an example the structure of a phenol-formaldehyde molecule is shown in *fig. 1*, indicating at the same time by what reaction this substance is formed. Plastics have recently been developed having molecules containing 10,000 and even 100,000 atoms.

In the synthetic production of plastics one starts from micro-molecular substances, and in the choice of these it is necessary to consider the functionality of the reactants, by which is understood the number of hydrogen atoms or reactive groups of the original elements taking part in the formation of macro-molecular compounds. The reactants must be chosen so that after the first reaction has ceased there is still sufficient functionality left. A couple of examples may serve to explain what is meant.

In the esterification of acetic acid and ethyl alcohol each reactant has the functionality 1 (it is a socalled 1:1 reaction) and when the ethyl acetate is formed the reaction is completed; macro-molecules cannot arise in this way.

When glycol and acetic acid, respectively bifunctional and monofunctional, are brought together, the reaction comes to an end by the formation of the di-acetate.

If, however, the acid is also bifunctional, possibilities are then opened for further reactions. An example of such a 2:2 reaction is that between phthalic acid and glycol, which can be represented by the following formula:



When two molecules of phthalic acid have reacted upon one molecule of glycol the result is a molecule which has two reactive groups. Two of these new molecules will then in turn react upon a glycol molecule. This goes on and on and leads to the formation of polymers. The original phthalic acid molecules are joined together by CH_2 links.

This manner of producing plastics is called the condensation method, where, in the bonding of two different molecules, a simple separation product (in this case water) is eliminated.

Another form of condensation is obtained when a trifunctional substance is caused to react upon a bifunctional substance, for instance glycerine with phthalic acid. We can imagine that here again polymers are formed as in the 2 : 2 reaction, but in these chains alcohol groups are left which have not yet been esterified. If now there is sufficient phthalic acid present, two hydroxyl groups of different chains will be esterified by one molecule of phthalic acid. In this manner the polymers are coupled together to form three-dimensional struc-

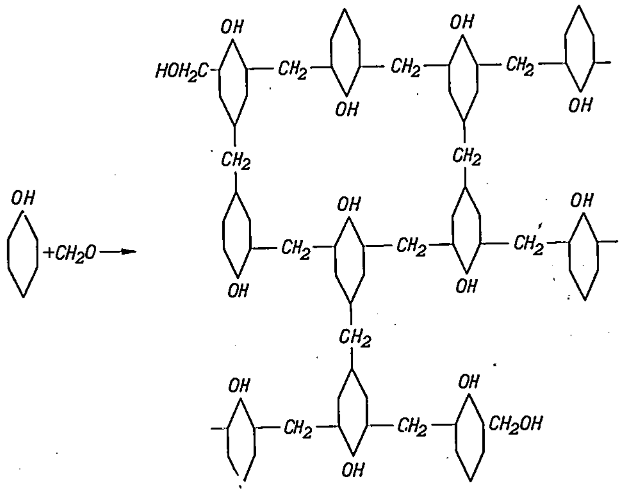


Fig. 1. The formation of phenal-formaldehyde resin. This representation is only diagrammatic. One must imagine the benzene nuclei as being three-dimensional and coupled together in irregular order.

tures. This transition is characteristic for the 3 : 2 reaction.

A second method of producing entirely synthetic macro-molecular substances is the polymerization method. Here small molecules of the same kind are joined together and without elimination of by-products. To give an example of this we will take the preparation of polystyrene, a product of great importance nowadays. The formulae relating to this case are given in *fig. 2*. The basic material for the macro-molecule is styrene (*a*). This molecule is activated either by means of catalysts or by irradiation or heating, as a result of which it is strongly agitated and the double bond opens out as indicated in fig. 2*b*. When this activated molecule is brought together with another molecule a reaction takes place (*c*). The activated molecule then created can react in turn with a new molecule of styrene. The ultimate result is a polymer (*d*). In this way one can obtain from styrene the macro-molecular polystyrene; many other substances possessing a $-C^{H} = CH_2$ group are also capable of polymerization.

It is particularly in the field of polymerization that great advances have been made during the last 20 years, resulting in the development of a number of valuable products such as polyvinyl acetate, polyvinyl chloride, polyvinylidene chloride, polyvinyl alcohol, polyethylene, polyacrylic acid and methacrylic acid esters.

An important property of macro-molecular substances is their strong cohesion, much stronger than that found with micro-molecular compounds. This may be illustrated by comparing micro- and macro-molecules with grains of sand and wads of cotton wool. Grains of sand packed together show very little cohesion, whereas wads of cotton wool are not easily parted.

The molecules of a substance attract each other through forces of an electrical nature, called van der Waal's forces, which increase in strength as the molecules become larger; these give to macro-molecules a mutual cohesion much stronger than that of micro-molecules.

Many plastics are made in the form of fibres (e.g. nylon yarns). It appears that the strongest fibres are obtained when the linear polymer molecules are orientated longitudinally in the fibre direction. This is in fact always the case with natural fibres.

The two kinds of macro-molecular substances — the two-dimensional (polymer) molecules and the three-dimensional molecules — behave very differently when heated. The first category usually becomes plastic when the temperature is raised sufficiently and solidifies again upon cooling down, apparently no chemical conversion taking place in the plasticizing. This group of plastics are called thermoplastics. On the other hand the substances formed from three-dimensional macro-molecules belong to the group of thermosetting plastics. These materials first become plastic upon being heated but after a time set hard and remain so after cooling down. It is not possible to make them plastic again by further heating.

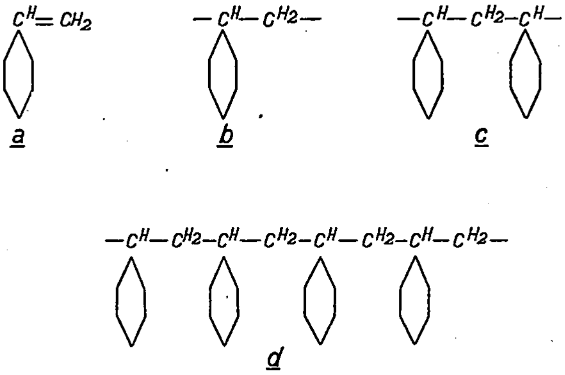


Fig. 2. The formation of polystyrene. *a*) Styrene. *b*) Activated styrene molecule. *c*) Two activated styrene molecules joined together. *d*) Polystyrene.

The difference just mentioned between two-dimensional and three-dimensional macro-molecular structures is not an invariable rule. If the mutual cohesive force of two-dimensional macro-molecules is very strong then the temperature at which the material can be made plastic is so high that disintegration takes place before that temperature is reached. Cellulose is a known example of this.

The principal moulding materials and their specific properties

When an article is to be made from a plastic the material has to be put into a mould or press to give it the right shape. The material prepared for this is called the moulding powder. The moulding powders most used are:

1) the thermosetting powders: phenolic resins, amino-plastics, polyesters and some silicones;
2) the thermoplastic powders: various cellulose derivatives, polyamides, polystyrene, polyethylene, polydichlorstyrene, polyvinylidene chloride, polyvinyl acetate, polyvinyl butyrate, vinyl chloride and various mixed polymers, acrylate resins and silicones.

These differ greatly in their physical and chemical properties. Furthermore it is possible to vary these properties by the addition of suitable fillers.

In order to give a survey of the properties of these plastics we shall classify them under several

Table I. Properties of some plastics.

		Specific weight	Shock-bending test N·m/m^2 [2]	Notch toughness N·m/m^2 [2]	Distortion tempera-ture in °C.	Insulating resistance ohm·cm	Dielectric constant at 1000 c/s	$10^4\times\tan\delta$ at 1000 c/s	Disruptive strength in kV/mm
Thermo-setting plastics	Phenol plastics with various fillers	1.35-1.8	$7\text{-}30\times10^3$	$1.5\text{-}25\times10^3$	150-170	$10^9\text{-}10^{12}$	5	100-500	10-30
	Amino-plastics	1.5	7×10^3	1.7×10^3	120	10^{12}	5-7	500	15
Thermoplastics	Polystyrene	1.05	20×10^3	5×10^3	65-100	$>10^{12}$	2.5-2.7	0.5-5	20-28
	Polyethylene	0.92	—	—	50	$>10^{12}$	2.25-2.3	3-5	16-19
	Polyvinyl chloride	1.38	175×10^3	5×10^3	65	10^{12}	4.9-5.6	1000	7-16
	Polymethacrylic acid esters	1.18	25×10^3	—	65-100	$>10^{12}$	3.2-3.4	500-600	20
	Polyamides	1.08-1.14	$>150\times10^3$	10×10^3	65	10^{12}	4-5	200-500	14-16
	Ethyl cellulose	1.08-1.18	—	—	40-90	$>10^{12}$	2.5-3.5	170-360	16-24

[2] 1 N (newton) = 10^5 dyne, 1 N·m/m^2 = 1.02×10^{-3} kg cm/cm^2.

groups. In *table I* some numerical data are tabulated for the most important moulding materials.

Phenolic resins

Phenolic resins (for example "Philite") are synthetically composed from raw materials obtained from coke and tar. The admixed fillers (wood-flour, mica, asbestos and suchlike) often constitute half the total weight of the moulding powders and considerably influence their properties. Powders of such a composition are widely used for making radio cabinets, telephones, door-handles, etc.

Phenolic resins without fillers are much used in glues, varnishes, etc.

Amino-plastics

Among the amino-plastics are urea- and melamine formaldehyde, which are often used mixed together to give a material that has good electrical properties and which can be made in a great variety of reasonably stable colours. This material is used instead of phenoplast in cases where a good non-tracking quality is required or where the products have to be made with bright colours.

Cellulose derivatives

The kinds of cellulose widely used in industry are: cellulose nitrate (made from wood cellulose or cotton fibres treated with sulphuric acid and nitric acid and softened with camphor), cellulose acetate (prepared from cotton fibres with sulphuric acid, acetic acid and acetic anhydride), cellulose acetobutyrate (made from cotton fibres treated with acetic acid, acetic anhydride and butyric acid) and ethyl cellulose (likewise made from cotton fibres treated with soda and ethyl chloride).

Plastics with a cellulose base are exceptionally suitable for mass production and are used for the manufacture of articles of a widely divergent nature. Cellulose nitrate is best known under the name of celluloid.

Vinyl compounds

The best known vinyl compounds are: polyvinyl acetate, polyvinyl chloride, polyvinylidene chloride and polyvinyl alcohol.

Polyvinyl acetate serves for the cementing of glass, paper and other material.

Polyvinyl chloride serves as a synthetic rubber; since softeners are used in the preparation of this substance, materials of different hardness can be made.

Under the polyvinyl compounds there are also polystyrene, polyethylene and acrylate resin.

Polystyrene is as clear as glass. It is used, inter alia, for all sorts of cheap jewellery. In the electrotechnical industry it is used because it is an exceptionally good insulating material with a very low loss factor and possessing, moreover, the favourable property of absorbing very little water in humid surroundings.

A relatively new, somewhat leathery, flexible and transparent material is polyethylene. Owing to its very good electrical properties it is used instead of guttapercha in the covering of submarine cables. For the same reason it is much used in radar technique.

Acrylate resin (polymethyl acrylic acid ester) is an important kind of resin. This material in itself is completely colourless and perfectly transparent, more so than any other material. In recent years it has been used more and more for making windows in aircraft. Further, it is used for optical glasses, artificial eyes and teeth, etc.

Polyamides

It was not until 1940 that polyamides came to the fore. They form a very strong and tough material which so far has found its place on the market mainly as a fibre material (nylon yarns).

Silicones

Some silicones are thermosetting materials and others are thermoplastic. They form a series of materials (liquids, greases, rubber-like and resinous substances) all of which have this in common that although they are not organic substances in their structure they nevertheless show a great resemblance to them. The main links of the molecule chains are formed by silicon and oxygen atoms. The side chains, derived from the silicon atoms, bear methyl groups or other groups containing carbon atoms. It is also possible to link via an oxygen atom a silicon atom from one chain with a silicon atom from another chain.

The most striking property of all these compounds is their resistance to high temperatures, for instance up to 300-400 °C, at which temperatures organic substances are sure to be decomposed. This is highly important because such is not the case with any of the plastics so far described; at high temperatures, owing to chemical disintegration accompanied by the separation of carbon, they are transformed into brownish black and useless masses [3]). Silicones are therefore especially employed in cases where high temperatures occur: the liquids and greases for lubricating parts of machinery which get very hot, the rubber silicones for the insulation of wires exposed to high temperatures and the resins as insulating material in engines. In the lacquer industry these resins are used for making lacquers for painting ovens, flues, radiators and suchlike.

The silicones have good electrical properties. It is also of importance that they are water-repellent; an object that has been immersed in a silicon solution will not be wetted by contact with water.

Moulding of the articles

It has already been pointed out that many applications of plastics have been made possible owing to the ease with which the products made therefrom can be given any desired shape. Use is thereby made of the plastic properties of the material, which show to advantage when it is heated. The mass is pressed at a high temperature when in a more or less liquid state, to give it a certain shape.

The shaping of articles from thermosetting masses is done in moulds under pressure. We need not go into this in detail here because it has already been discussed in two articles in this journal dealing with "Philite" [4]).

The mould is usually of steel and made in two parts. A certain weighed quantity of the mass is placed between the two halves of the mould, which are then heated to about 160 °C and pressed together, this being mostly done with the aid of hydraulic presses. Some products turned out in this manner are illustrated in *fig. 3*.

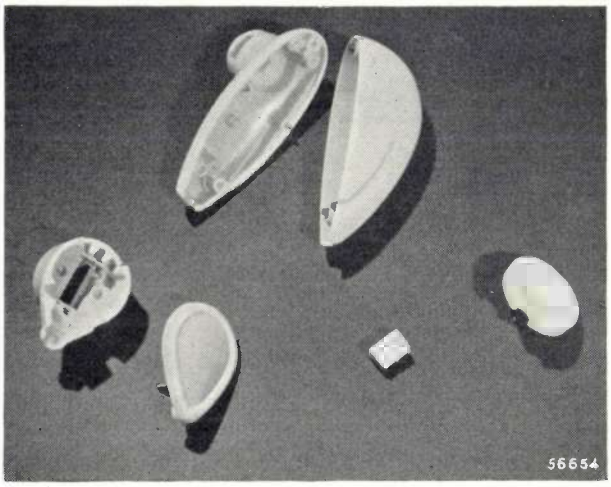


Fig. 3. Parts of a "Philishave" dryshaver made of urea-formaldehyde. This material allows of the objects being given an attractive light colour.

Obviously one cannot in this way make articles of every desired shape. Products made from thermosetting moulding masses must be capable of being taken out of the mould; after the shaping, the top and bottom halves of the mould have to be drawn away from the product in opposite directions. Sometimes this can be arranged by deviating from the conventional shape of the article (compare for instance the handle of a teacup made from a plastic material with that of a china cup). A product

[3]) During the second World War in the U.S.A. another plastic material was developed, called "Teflon" (polytetrafluoro ethylene), which can be heated up to 400° C without any noticeable decline in its properties. This material, which resembles stiff leather, has very remarkable properties, the most important of which is its chemical inertness. Thus it is not affected by any organic liquid nor by any corrosive reactants such as, for instance, aqua regia. This material opens up new possibilities of great importance in technical engineering.

[4]) R. Houwink, Properties and application of artificial resin products, Philips Techn. Rev. 1, 257-263, 1936; L. L. C. Polis, "Philite" as a structural material, Philips Techn. Rev. 3, 9-16, 1938.

that does not come away immediately but which can still be moulded in a mould consisting of two halves is one that has a single screw thread, such as a jampot lid; when the mould is opened the product is left in the half of the mould which shapes the thread, but it can be removed by unscrewing it.

Laminated materials take a special place in the moulding technique. In a certain respect these are comparable to sheets of plywood. In the manufacture of these materials paper, tissue or wood veneer is used in the form of sheets of about 1 m × 1 m (about 3′ × 3′). These sheets are saturated with resin and pressed together between hot plates, thus producing a resinous mass interspersed with layers of paper, tissue or wood.

In practice this is done by passing a strip of paper or fabric first through a bath of resin in solution and then through a long oven to remove the solvents and partly harden the resin. The strip is then cut into pieces which are stacked one upon the other and then pressed between flat plates. The materials produced in this manner from phenolic plastics are called resin-bounded papers or sheets.

In the moulding of articles from thermoplastic masses different methods of manufacture are employed, but these too are based upon the plastic properties of the material at a high temperature. Account is taken of the fact that owing to the nature of their macro-molecules these materials do not set hard when heated. With these materials the products are shaped by injection, extrusion, pressing and blowing.

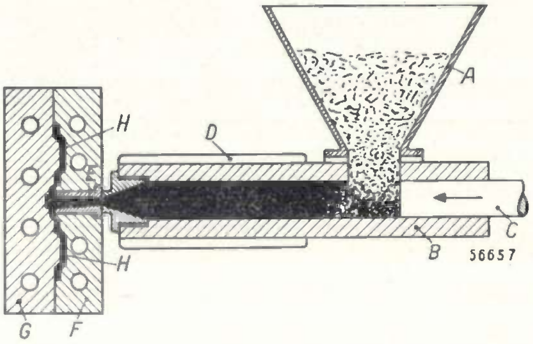


Fig. 4. The process of injecting thermoplastic material. *A* hopper charged with the moulding mass, *B* cylinder, *C* plunger, *D* heating element, *E* injection channel, *F* and *G* dies of the mould with cooling channels, *H* two products (small trays).

The process of injection is illustrated in *fig. 4*, a diagrammatic representation of a machine used in this process. The moulding material is fed into a heated cylinder with a constricted opening at one end. This cylinder is heated to such a temperature as to soften the material. The softened material is injected into the mould by means of a plunger fitting into the cylinder.

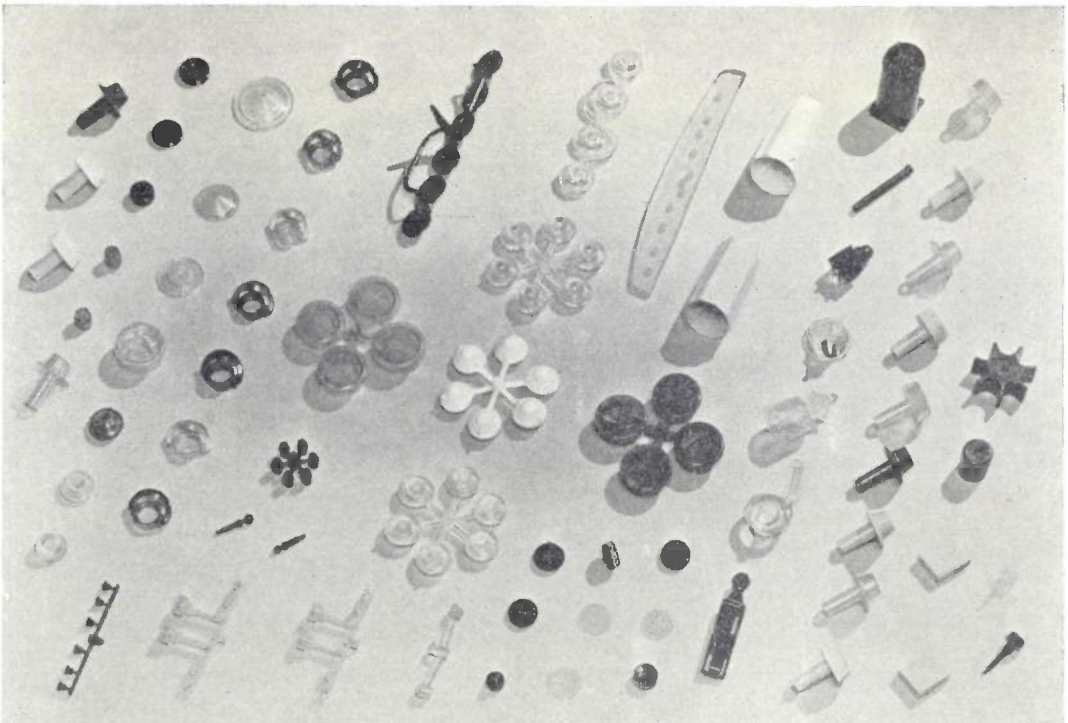


Fig. 5. A number of articles produced by the injection method. The raw material used for some of these is polystyrene and for others polymethacrylic acid esters ("Plexiglas").

Often the mould is cooled. When the cavity in the mould is filled the plunger is stopped. *Fig. 5* shows a number of articles produced by the injection method.

Rods, tubes, etc. are made by the method of extrusion or squirting. This method is illustrated in *fig. 6*. The material is placed in a cold hopper and transported through a cylinder by means of a worm spindle. The cylinder is heated higher and higher as the material passes through it. In order to get a homogeneous mass the softened material is often pressed through one or two sieves. The end of the cylinder has a constricted opening. The soft mass passes through the opening into the air and sets hard.

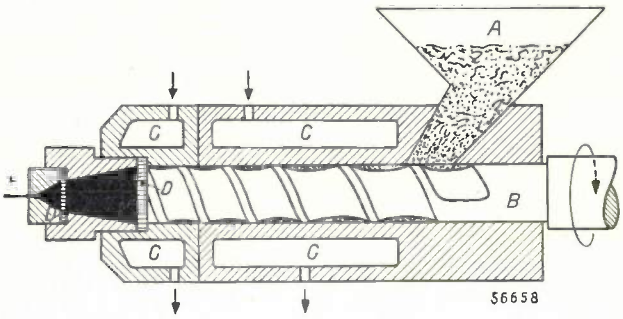


Fig. 6. The extrusion of thermoplastic material. *A* hopper charged with the moulding mass, *B* worm spindle, *C* heating elements, *D* sieves, *E* extruded product.

The shape of the outlet determines whether rods (round opening) or tapes and plates (rectangular opening) or tubes (annular opening) are produced. It is also possible to apply a coating of the material to copper wire for use as flex or cable.

When the pressing process is applied with thermoplastic masses it is done in much the same way as with thermosetting material. Bottom and top dies are used, both heated. Since at the elevated temperature the mass remains soft, it is necessary to cool the mould while it is still closed. Pressing is applied, inter alia, in the manufacture of plates, although these are usually made by the injection method.

In the blowing method one starts with plates or tubes in the form which they have already been

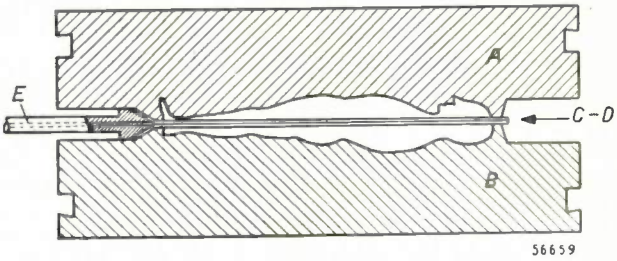


Fig. 7. The blowing method used in the moulding of thermoplastic material. *A* and *B* top and bottom dies, *C* and *D* two heated plates of the material (not yet moulded), *E* channel through which air is forced in between the plates.

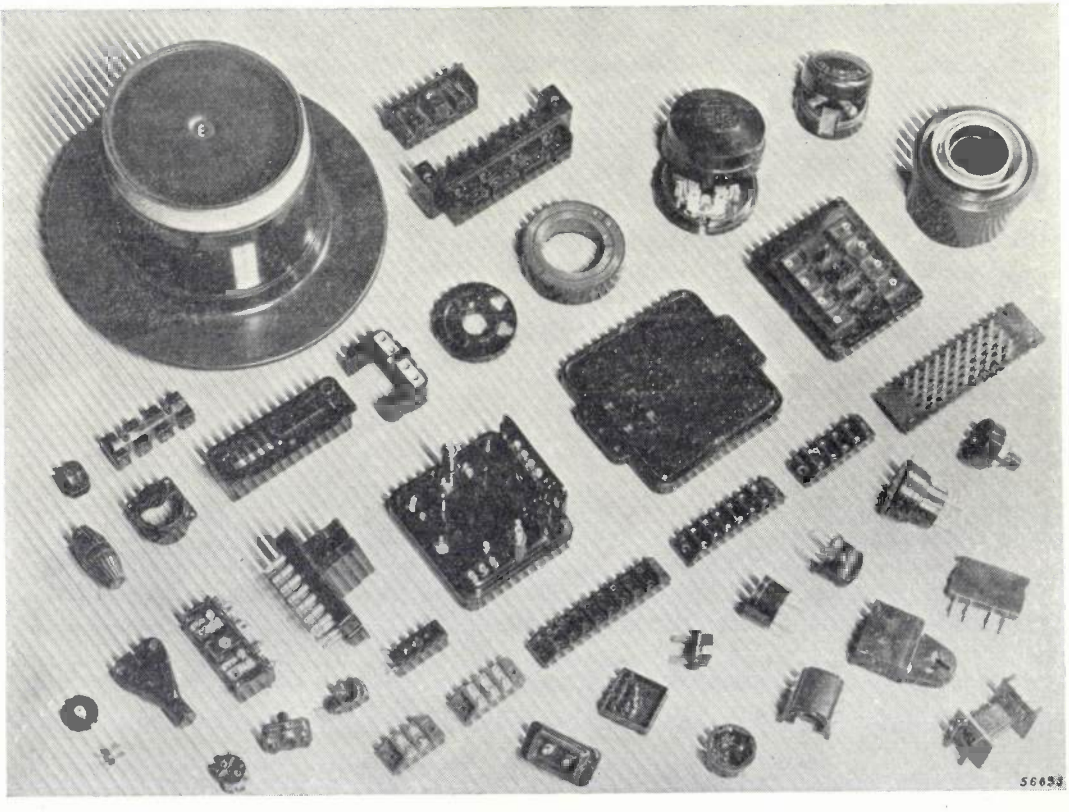


Fig. 8. Component parts of electrotechnical apparatus, all press-moulded from phenol-formaldehyde ("Philite").

given by injection, extrusion or pressing. This is illustrated in *fig. 7*. For instance two hot plates of material are placed between the top and bottom dies of a mould. Between the plates a hollow needle is inserted through which air is blown. The plates are thus blown apart and given the shape of the mould. The manufacture of celluloid dolls is an example of the application of the blowing method. An advantage of this method of working is that no steel dies are needed because only low pressures have to be applied.

In conclusion some examples are given of the use of plastics as constructional material for electrotechnical apparatus.

Applications

From table I it is seen that in the thermosetting category there are materials with greatly divergent properties. Among the thermoplastic materials included in this table are those most used in the manufacture of electrotechnical apparatus; as these, too, show widely divergent mechanical and electrical properties, care is needed in the selection of the most suitable material for each application.

In *fig. 8* a number of component parts of electrotechnical apparatus are illustrated, all moulded from phenol-formaldehyde ("Philite"). The moulding masses used for these products differ considerably in composition. Some of these articles, for instance, are used in X-ray apparatus and for that purpose they are often required to have a high dielectric strength (30 kV/mm at 90 °C), while at the same time they must be resistant to the effects of hot oil. Furthermore, the material must be easily mouldable in view of the often complicated and precise form of the parts of such apparatus. These articles are therefore made from a kind of "Philite" which has a high dielectric strength, which fully answers these requirements and has a much higher breakdown voltage than the normal "Philite".

Polyethylene and polystyrene are particularly noticeable in the table on account of their low dielectric losses. This explains why these two materials are so frequently used in the electrotechnical industry. Polyethylene is much used as a covering for cables carrying high-frequency currents, whilst polystyrene is used for all sorts of apparatus or components having to answer high electrical requirements.

The low softening point of polystyrene is a drawback, but by using substituted styrene the chemical industry has already succeeded in producing polymers with higher softening points and still retaining the good electrical properties.

In *figs. 9* and *10* some coil bodies with adjusting screws are illustrated. Polystyrene was chosen as the material for these articles on account of its excellent electrical properties. Often, however, this material is chosen only because it is easily shaped by the injection or extrusion processes.

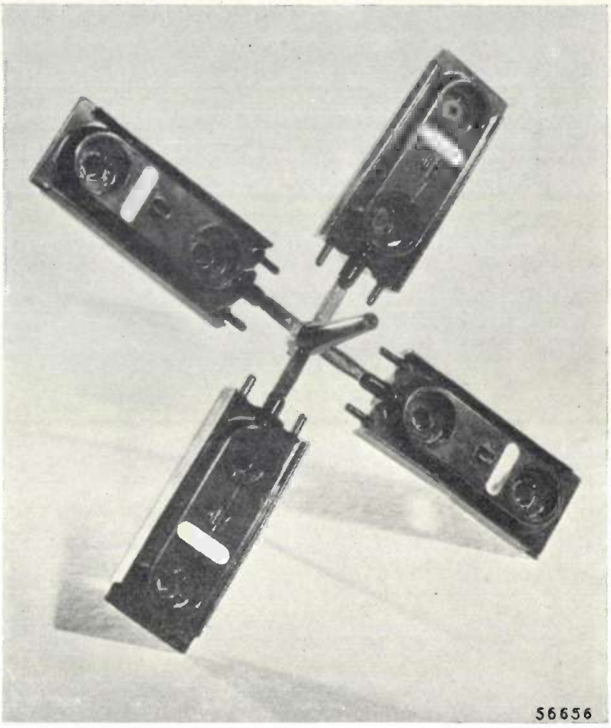


Fig. 9. Coil bodies of polystyrene. This shows the workpiece as produced by the injection method. It consists of four coil bodies which are subsequently pinched off the common holder.

Polyvinyl chloride is used on a large scale for spraying wire used for the manufacture of flexes and flexible connections, examples of which are illustrated in *fig. 11*.

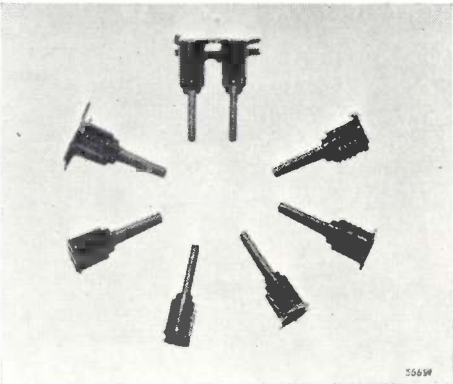


Fig. 10. Adjusting screws for the coil bodies illustrated in fig. 9. These screws (in the illustration they are $^3/_4$ of the actual size) comprise a core of "Ferroxcube" with a holder of polystyrene; they serve for adjusting the self-inductance of the coils, ultimately mounted in the coil bodies. These screws, too, are produced by the injection method.

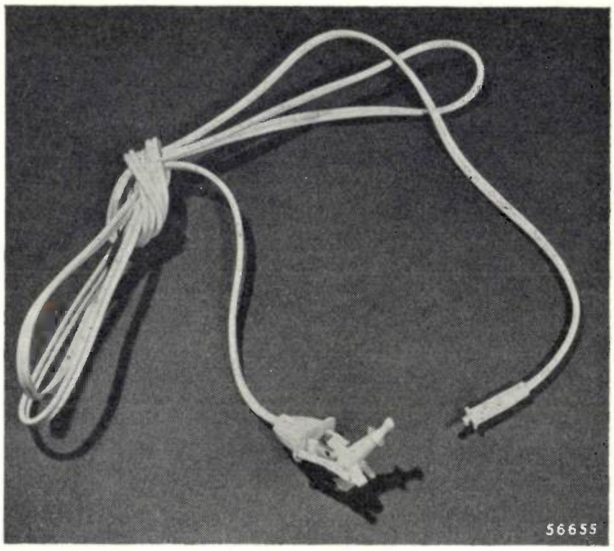


Fig. 11. A flex made of polyvinyl chloride. The flex itself is extruded, whilst the sleeve and the plug are injected. This flex belongs to the "Philishave" dryshaver, some parts of which are illustrated in fig. 3.

The optical properties of polymethacrylic acid esters are utilized for instance in the manufacture of lenses and prisms. The high index of refraction and the small light absorption make this material highly suitable for this purpose. Some of the articles illustrated in fig. 5 are made from this material, whilst *fig. 12* shows an example of a product made from this material in the electrotechnical industry.

The examples given are confined to some of those belonging to the field of mass production, but this is only a selection taken at random from a multiplicity of applications. For the sake of completeness it is to be mentioned that in the electrotechnical industry plastics are used not only as raw meaterials for mass-produced articles but also for all kinds of individual purposes where it is not so much a matter of easy processing as of certain combinations of properties. Thus it often happens that for one particular component of a single apparatus use is made of one of the typical characteristics offered by the family of plastics (including silicons and "Teflon") in such great variety.

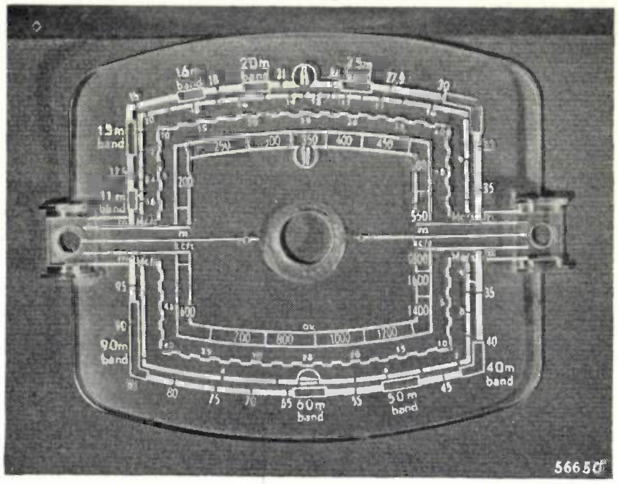
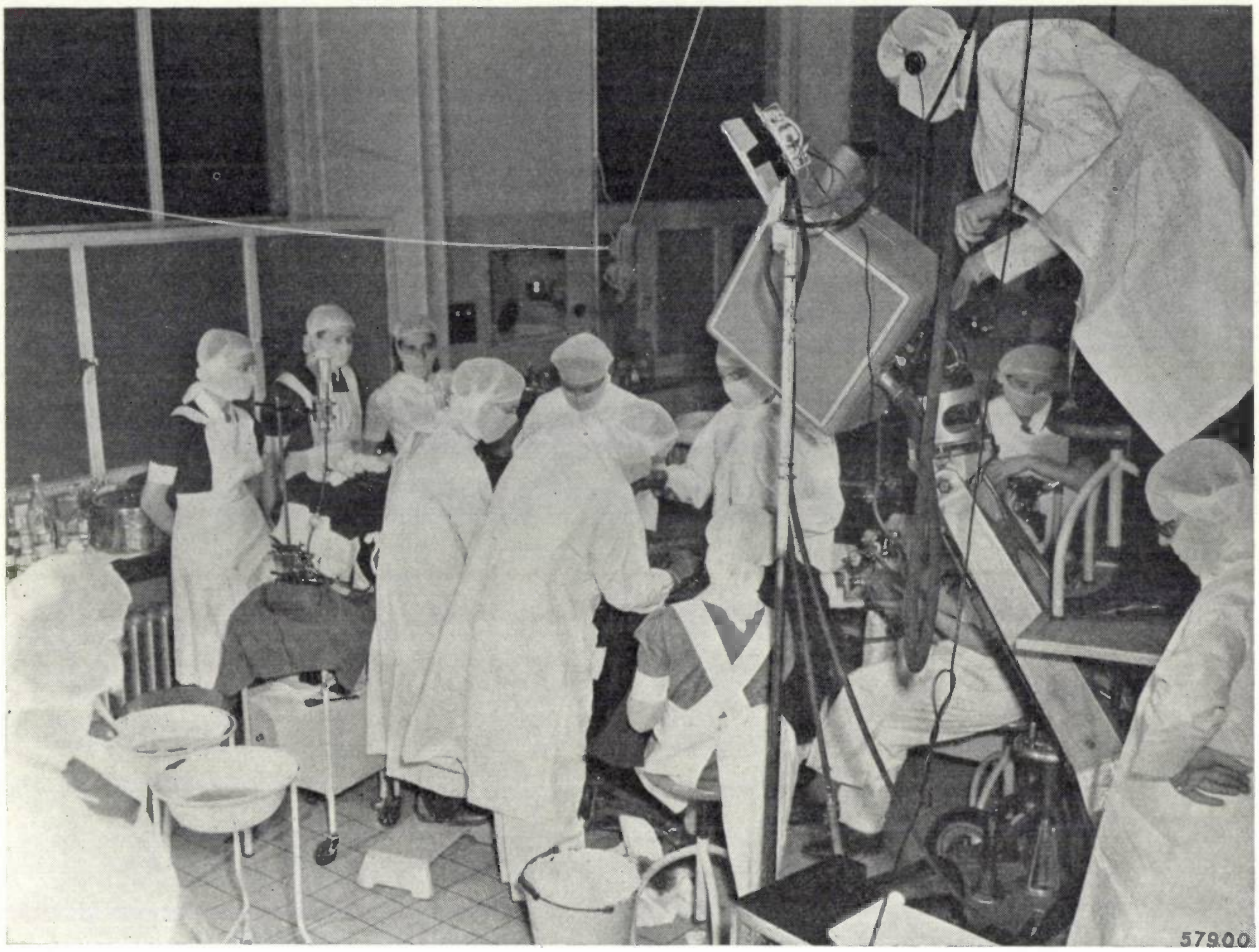


Fig. 12. A station dial for a radio set, made from polymethacrylic acid ester ("Plexiglas") according to the injection method.

Summary. Just as in many other fields, plastics are extensively applied in the electrotechnical industry as constructional materials for all kinds of apparatus. Often they are used for these purposes on account of their good electrical properties, whilst in other cases it is their easy mouldability and the good finish that can be given to the products which count most. It is very often possible to synthesizc these materials with the properties required for a particular purpose. Of the methods employed in the formation of the macro-molecules characterizing all plastics, the condensation and polymerization methods are discussed. Use is made not only of thermosetting moulding materials but also, and to an increasing extent, of thermoplastics materials. The most important materials and their specific properties are dealt with briefly. A table gives an idea of the great diversity of these properties. In the moulding of articles from thermoplastic material four methods are employed: injection, extrusion, pressing and blowing. Finally a number of applications of plastics in the electrotechnical field are discussed.

AN OPERATION WITNESSED BY 200 SPECTATORS



At the occasion of the *Dies* Celebration at the University of Leiden (Netherlands) a demonstration was given on February 5th 1949 in order to enable a large audience to witness the televising of a surgical operation. The above photograph was taken during this operation, that was carried out by Prof. Dr. W. F. Suermondt, of the above-mentioned University, and his main assistant Dr. J. Kweekel. As can be seen from the photograph the pick-up device — an iconoscope — is situated next to the surgeon and directed towards the patient. In front of the iconoscope one of the two stands, each carrying three mercury lamps, can be seen. These lamps are water-cooled so that no trouble is experienced with the heat developed. In addition, incandescent lamps fixed to the ceiling were used for the general illumination. — The amplified video signals from the iconoscope were transmitted over a cable to a lecture hall where the image was projected onto two screens of 1 m × 1.3 m (about 3′ × 4′). Each half of the audience, consisting in total of over 200 spectators, were easily able to follow the operation. At the receiving end two units comprising a cathode-ray tube and an optical projection system as described in this Review [1]) were used.

[1]) Projection-television receiver, I. The optical system for the projection, by P. M. van Alphen and H. Rinia, and II. The cathode-ray tube, by J. de Gier, Philips Techn. Rev. **10**, 69-78, 1948 (No. 3) and 97-104, 1948 (No. 4).

INFLUENCE OF LIGHT UPON PLANTS

by R. van der VEEN.

581.1.035

Plants need light! By and large this is undoubtedly true, but a closer investigation shows that this cannot be accepted without reserve. The most favourable intensity and composition of the light for the feeding, the shaping and the flowering of a plant is not always the same, and as regards the duration of the exposure to light it appears that dark periods are often just as essential for flowering as light periods. All such factors should be considered in order to arrive at a method of irradiating plants with artificial light that is most suitable for a certain object.

The fact that light is one of the most important factors in the life of a plant is commonly recognized. In a non-tropical climate the intensity of the light received from the sun during the winter months is, for many plants, inadequate for a strong development. It is therefore understandable that both the nurseryman and the lover of plants often tries to promote the growth and development of his plants by means of artificial light.

Light, however, has many different effects upon plants. Artificial lighting can therefore only be applied successfully when one first makes quite sure what result is to be expected and knows how it should be applied to attain the object in view. Attention has to be paid both to the intensity and duration of the irradiation as well as to the composition of the light used.

In this article some of the effects of light upon the life of a plant will be discussed, namely those of:
1) light for the feeding,
2) light for the shaping,
3) light for the flowering.

Light for assimilation or photosynthesis

One of the first to carry out systematic research into the processes of plant life was Jan Ingen Housz, a Dutchman. His principal book "Experiments upon vegetables", the result of some hundreds of experiments, was published in 1779 while he was in England. There is it stated that green plants correct bad air when they receive sunlight. Later on it appeared that this purification consists in the absorption of carbon dioxide and the excretion of oxygen.

Through his experiments Ingen Housz became the discoverer of the photosynthesis of plants. This photosynthesis is to be regarded as the feeding process of the plants. Disregarding water, only an extremely small fraction of the material from which a plant is built up comes from the soil. By far the greater part is assimilated in the form of carbon dioxide and converted into carbohydrates. At the same time, however, as Ingen Housz demonstrated, another process is taking place, the respiration of the plants. In this process the substances formed by photosynthesis are burnt, carbon dioxide being excreted and oxygen absorbed.

Photosynthesis is a strongly endothermic reaction. The energy required for this is supplied by the light. Investigations have shown that only the visible part of the spectrum can act as the source of energy. Ultra-violet and infra-red rays are both injurious to plants. The various colours of the visible spectrum do not have the same effect upon photosynthesis. Taking assimilation of CO_2 in red light ($\lambda = 6500$ Å) as 100, according to Gabrielsen that in yellow light ($\lambda = 5450$ Å) is about 60 and that in blue light ($\lambda = 4400$ Å) only 38 [1]). The absorption of light energy takes place through the chlorophyll granules, the corpuscles which give a green colour to the stems and leaves.

In addition to the composition of light, the intensity of the radiation also plays a great part in photosynthesis. As a general rule the greater the intensity the greater is the photosynthesis, but there are a number of factors restricting the validity of this rule.

The first of these factors is the temperature. The manner in which the growth of a plant is affected by the intensitiy of radiation at any temperature can be found experimentally, somewhat different results being obtained according to the kind of plant. In *fig. 1* some curves are given taken from experiments carried out by Bolas with tomatoes [2]). It is seen that the growth of a plant reaches a maximum at a certain intensity of irradiation, this

[1]) The fact that the strongest photosynthesis takes place under red light accounts also for the relatively favourable effect of neon lamps for irradiating plants. The wavelength of the radiation from these lamps lies between 6000 and and 7000 Å. See J. W. M. Roodenburg and G. Zecher, Irradiation of plants with neon light, Philips Techn. Rev. 1, 193-191, 1936.

[2]) B. D. Bolas, Cheshunt Annual Report 19, 84, 1933.

optimum intensity increasing as the temperature rises. With an intensity of 2000 lux a temperature of 18 °C is most favourable, whilst with intensities of 7500 and 10,500 lux the most favourable temperatures are respectively 25 and 30 °C.

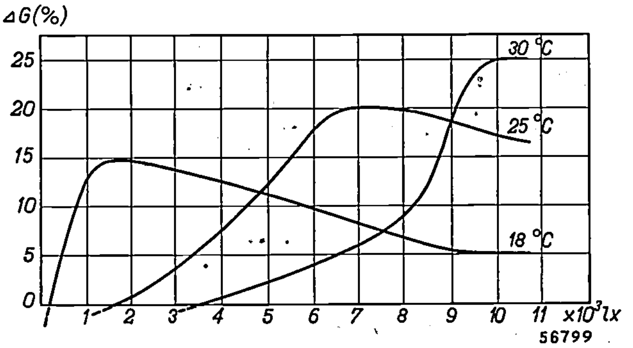


Fig. 1. The growth of a plant as function of the intensity of light at three temperatures (taken from Bolas). On the vertical axis the increase ΔG of the "dry weight" of the plant in 7 hours is set out as a percentage. This test was taken with tomatoes.

In such investigations one has to bear in mind that both photosynthesis and respiration play a part in the growth of a plant. It is however possible to separate these two factors, since if a plant is placed in the dark, or given a small dose of chloroform, photosynthesis ceases while respiration continues. When thus determining separately the degree of photosynthesis it appears that at any temperatures only a certain maximum photosynthesis is possible. Once this has been reached photosynthesis is not increased by a greater intensity of light. The higher the temperature, the greater the maximum photosynthesis, within certain limits. Further, it has been found that as the temperature is raised also respiration increases.

Fig. 2 shows the relation between photosynthesis and the intensity of radiation for different levels of temperature. In these graphs the numerical values, which differ according to the plant species, have been omitted. From the trend of the curves it is to be seen that no purpose is served by giving a plant more than a certain amount of light at a certain temperature. But it is also undesirable to raise the temperature higher than that which is suitable for a certain intensity of light, because then, as already stated, respiration is intensified and this is a process which counteracts photosynthesis, as far as the increase of weight of the plant is concerned.

By irradiating with certain, rather small intensities one can by experiment find a point where respiration and photosynthesis are balanced. This is called the compensation point. Fig. 1 shows with what intensity of light, for three given temperatures, this compensation point is reached in the case of the tomato.

If plants are to be made to grow, the irradiation of light should be such that the difference between photosynthesis and respiration is positive. In many cases irradiation can be applied only during a part of the 24 hours. The intensity of the light must then lie far above the compensation point, since respiration is continuous.

The higher the temperature, the higher is the compensation point. More light is then required to compensate respiration, since this increases at higher temperatures.

At a temperature of 20 °C and with a light intensity of 500 lux one is, in most cases, above the compensation point. The plants then certainly grow but often very slowly. If rapid growth is required, a light intensity of 1500 to 3000 lux is necessary for most plants. Such an intensity is easily reached with high-power lamps placed close above the plants.

If, however, incandescent lamps are used for this purpose, the heat from the radiation will as a rule be so great as to injure the plants. This can be avoided by using a water filter between the lamps and the plants, this filter holding back the infra-red rays. This method, however, has so many objections that it is preferable to use lamps which radiate very little heat and which can be placed close above the plants without having to take any special precautions. Fluorescent lamps (MCF/U lamps) are highly suitable for this purpose.

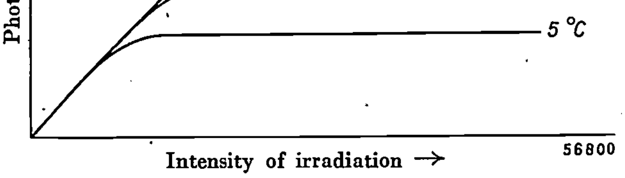


Fig. 2. Graphs representing the relation between photosynthesis and the intensity of irradiation at various temperatures.

Owing to the relatively low power of these lamps a large number are required (about 5 per square metre) and this makes the installation rather expensive. Nevertheless it is undoubtedly of advantage for many nurserymen to apply this method for intensively lighting some of the space in their glasshouses. Seeds can then be sown already in the winter in this part of the glasshouses and as a result the nurseryman can begin in the spring with better developed plants. Most plants can easily be raised a month in advance in this way,

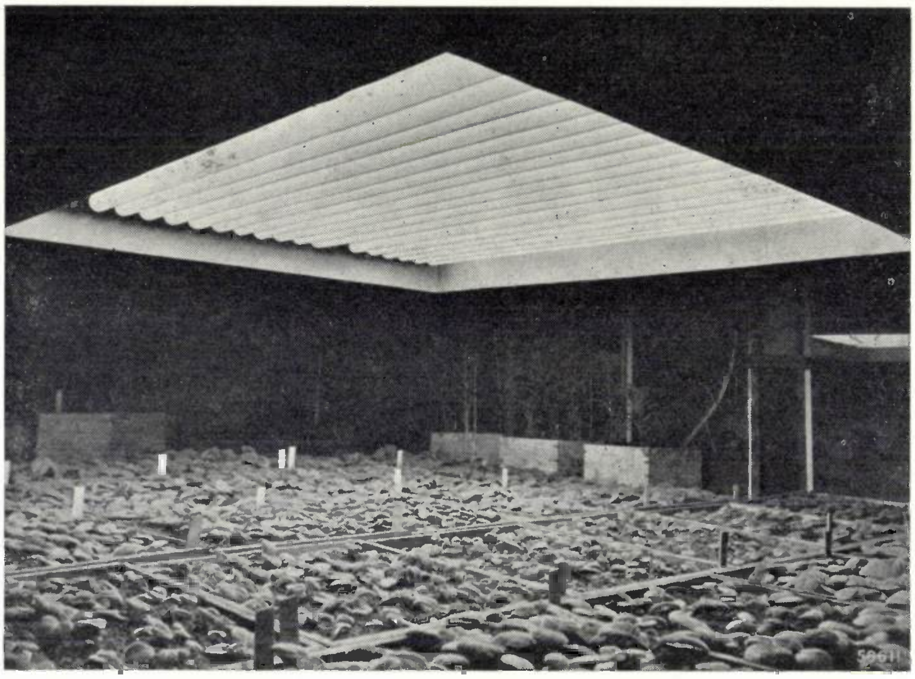


Fig. 3. Artificial lighting of a glasshouse in which Gloxinias are being raised. The lighting is from 40 W MCF/U lamps of the "daylight" type.

Fig. 3 shows the lighting installation of a glasshouse in which gloxinia plants have been raised. *Fig. 4* demonstrates the difference between boxes of plants raised without artificial light, with moderate lighting and under a strong light.

This method of lighting also makes it possible to raise young plants during the winter months in a shed or room instead of in a glasshouse. For instance, cornflowers sown on 15th January in a celler heated to 21 °C were already flowering on 25th February. In such cases the saving in fuel for heating a glasshouse will usually not outweigh the cost of lighting and heating the cellar, so that this method is still expensive.

Light for shaping

Light has also a considerable effect upon the shape in which a plant grows [3]). This is a subject, however, that cannot be discussed in general terms because in this respect the various kinds of plants react to light in a different manner. We can therefore only make a few observations and give an example.

In the dark, plants grow tall and spindly, the foliage remains small and the colour yellowish. The plants are then said to be etiolated [4]). Further, under red and yellow light many plants shoot upwards and these are likewise said to be etiolated although they turn green. Since the light from incandescent lamps is mainly composed of yellow and red rays, with regard to plants raised under this light one speaks of incandescent-light etiolation.

Generally speaking one finds in a plant two substances which absorb light of a certain wavelength and use the absorbed energy for the growth and development of the plant, namely chlorophyll and carotene. Chloropyll, as already stated, serves especially for photosynthesis and absorbs mainly red rays. Carotene affects the shape of the plant (length, foliage, etc.) and absorbs particularly blue rays.

Plants raised under blue light are usually shorter but stronger than those grown under white light. For normal development all plants need a certain amount of blue light, more in some cases than in others. Mirabilis is one of the plants needing a

[3]) This influence of light upon the development of plants has already been discussed in two previous articles. See R. van der Veen, Forcing tulips with artificial light, Philips Techn. Rev. **10**, 282-285, 1949 (No. 9); Storing seed potatoes in artificially-lighted cellars, Philips Techn. Rev. **10**, 318-322, 1949 (No. 10).

[4]) etiolate = to become pale.

great deal of blue radiation. *Fig. 5* shows some specimens of this plant raised under light of a different composition, one under red light with about 10% blue, another under yellow light with about 10% blue, and a third specimen under blue light with about 20% yellow and red, all irradiated with practically the same total energy. Only the last specimen was well developed.

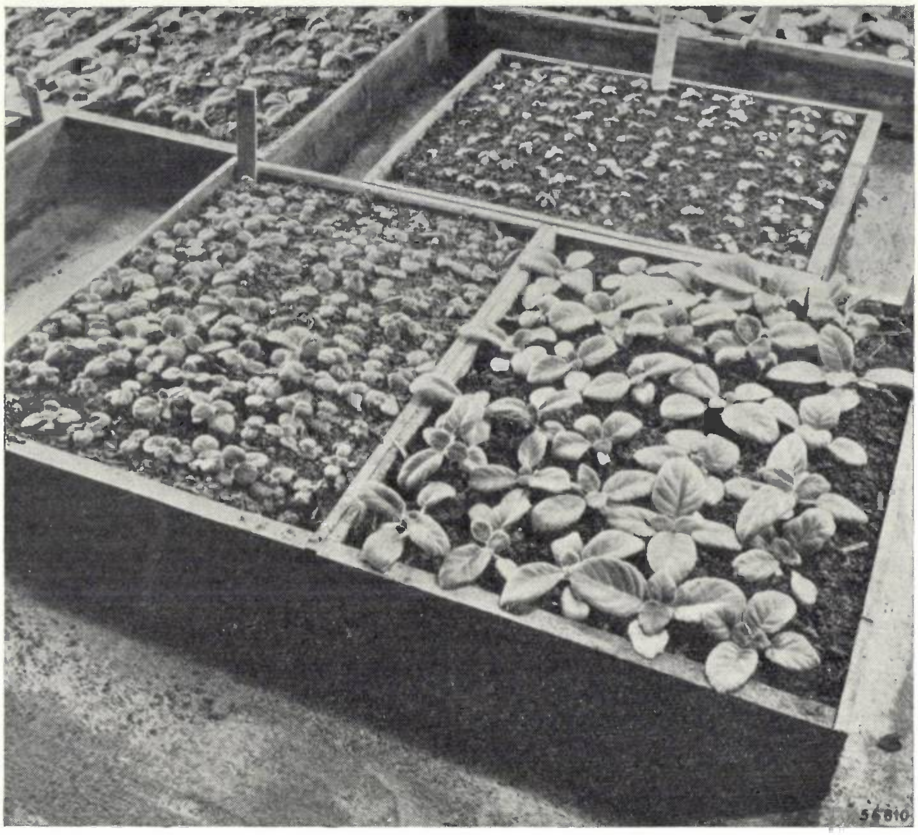


Fig. 4. The effect of lighting upon Gloxinias. On the three boxes of plants in the foreground the rearmost was exposed only to sunlight. The left-hand one was given additional weak artificial light during November and December; the right-hand one was irradiated with strong artificial light during those two months. The photo was taken in the beginning of February.

This example goes to show how great a difference there can be in the effect of different kinds of light upon the development of a plant.

Light for flowering

The flowering of several kinds of plants depends entirely upon the time during which they have been exposed to light. As regards flowering, the intensity of the light, so important for growth, is of minor importance.

With respect to the need of light, the varieties of plants can be placed in three groups. There are plants which only flower when exposed to light during a certain number of short days (10 hours light or less per day) and these are called short-day plants. Others will never flower if they get only short days of light, these requiring long periods (14 hours or more per day); they are called long-day plants. Finally there is a group of plants that are indifferent to the length of daylight. The whole of this phenomenon is termed photoperiodicity. In the field of photoperiodism there are a multitude of contradictions; as regards the mechanism of the actions of light in that repect one is still entirely in the dark [5]). We shall, therefore, not go into this too deeply and will confine our considerations to a few aspects of this problem.

As already remarked, the influence of the length of daylight is not bound directly to high intensities of light. Consequently a short winter day for a

[5]) The results of some experiments would seem to show that under the influence of light a certain hormone may be formed which stimulates flowering and without which the plant cannot flower. The symptoms would indicate that this hormone is formed in the leaves. So long as a short-day plant is exposed to a long day it will develop vegetatively. When, however, a leaf is taken from a plant exposed to a short day and grafted upon such a plant the latter will flower. Apparently the hormone referred to is transported from the grafted leaf to the point of growth on the main stem of the plant.

plant can often be turned into a long day by adding a few extra hours of artificial lighting of 50 lux.

Winter-flowering Begonias give blooms when the days are short, whereas when the days are long the buds only develop vegetatively, so that big but not flowering plants are formed. When incandescent lamps are kept burning during the winter months in begonia glasshouses flowering is checked and the buds develop vegetatively, which is just what the nurseryman desires, because in the winter season he wants plenty of material for cuttings. In the same way Poincettias and Euphorbias are raised in the autumn and winter for cuttings.

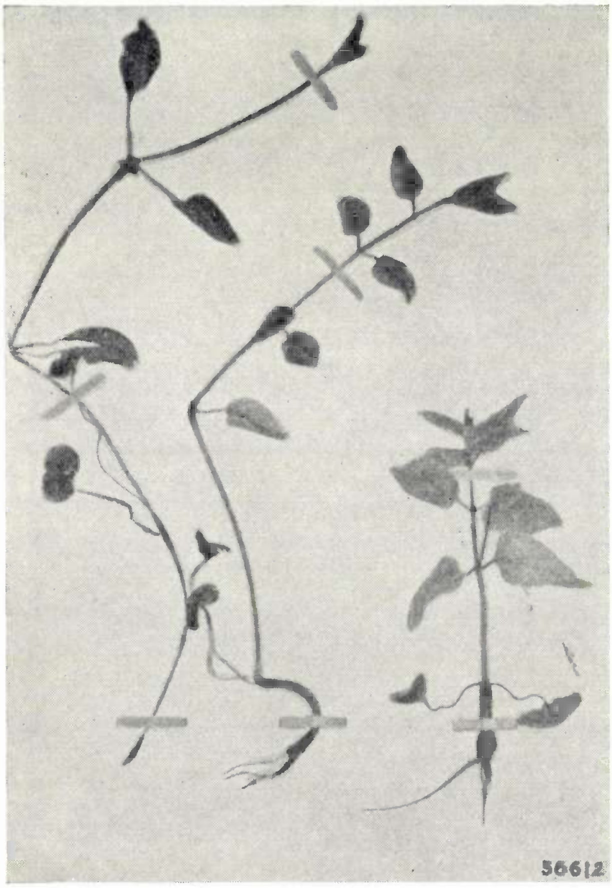


Fig. 5. Three Mirabilis plants raised respectively under red, yellow and blue light. Only the last one developed properly.

The results of various experiments show that what takes place in the dark period is something quite different from that taking place during the period of light. Hammer [6]) distinguishes the factor A, the process during exposure to light, from the factor B, the process taking place in the dark. These together yield the factor C which causes the plants to flower.

The factor A probably has something to do with the photosynthesis of the plant. The flower-inducing effect of the length of day presumably takes place in the chlorophyll. Several established facts point to this.

The soya bean, a typical short-day plant, needs, to start flowering, a dark period of at least 10 hours with a light period of at least 4 and at most 14 hours, with a decided optimum of 10 hours. The minimum intensity of light that has to be given during the light period is 1000 lux. Given this intensity of light for 5 hours daily, the plant yields a number of blooms. With higher intensities, thus promoting photosynthesis, the number of blooms increases up to a certain maximum. The duration of the exposure to light must, however, exceed the lower limit of 4 hours, regardless of the intensity used.

The minimum intensity of 1000 lux is understandable when one bears in mind that (with the optimum temperature for growth of the soya bean) this is about the limit at which an excess of photosynthesis begins. The effect that assimilation during the period of light has upon flowering is evident from the smaller number of blooms formed when photosynthesis is checked during the period of exposure to light by reducing the carbon dioxide content of the air. Further, Borthwick and Parker [7]) have succeeded in getting twice as many blooms from a soya bean by increasing the carbon dioxide content of the air from about 0.03% to 1% [8]).

Observations made with other short-day plants agree, on the whole, with what has been found in the case of the soya bean.

Remarkable discoveries have also been made from a study of long-day plants. Went [9]) found that with a period of 10 hours light the Baeria plant is just stopped from flowering, regardless whether the intensity of the light is 2000 or 15,000 lux, whereas any additional exposure, with weak or strong light, beyond these 10 hours causes the plant to flower equally well [10]).

Went has also investigated the effect of the

[6]) K. C. B. Hamner, Bot. Gaz. **101**, 658, 1940.

[7]) H. A. Borthwick and M. W. Parker, Bot. Gaz. **102**, 256, 1940.

[8]) The photosynthesis of a plant can also be checked by reducing the number of leaves. When three fourths of the number of leaves on a soya bean plant are removed still just as many blooms develop. From this it is to be concluded that Hamner's factor A is dependent upon the photosynthesis per surface unit and that the total photosynthesis matters comparatively little.

[9]) F. W. Went, J. Bot. **32**, 1, 1945.

[10]) By weak is meant here an intensity of 2000 lux. This is still a fairly high intensity of light compared with that which interferes with the effect of the period of darkness.

composition of the light to which long-day plants are exposed. He found that red light greatly stimulates bloom development, whilst yellow and blue light do so to a less degree and green light has no effect at all at an intensity of 2500 lux. This is yet another indication that the effect of daylight upon these plants has something to do with the photosynthesis, at least with the chlorophyll.

Other investigations have also shown that the colour of the light influences the flowering of a plant.

Klebs found that blue light checks the flowering of Sempervivum, whilst on the other hand red light promotes this process. Lettuce flowers quickly under red light, whereas under blue light nice firm heads are formed. Cornflowers come into bloom under red, yellow and blue light, but under green light they bloom very slowly. In all these cases the blooms differed in colour, red light giving small blooms with a faded colour whereas blue light gave blooms of the brightest blue.

Quite frequently irradiation with blue light results in a strong formation of anthocyanin. The stems and the leaf veins then turn red, owing to the red colour of the anthocyanin predominating over the green. Flowers which owe their colour to anthocyanin therefore often have a deeper colour when cultivated under blue light.

Most long-day plants also flower when exposed to continuous light; they do not need a dark period at all. In the case of short-day plants, however, the length of the dark period is just as essential for flowering as the period of light, and one might therefore just as well call them long-night plants.

Short-day plants need an undisturbed night. If a night of 16 hours is interrupted in the middle by a quarter of an hour's light this has the effect of breaking it up into two nights of 8 hours, with the result that the plant does not flower. The extent of this effect depends upon the total amount of light given during that interruption. As a rule, within certain limits, two minutes with 10,000 lux has the same effect as 20 minutes with 1000 lux or 200 minutes with 100 lux.

Even light of a very low intensity (100 lux) thus has its effect. Advantage is taken of this in practice, in order to prevent short-day plants flowering, by switching on a number of low-power incandescent lamps in the glasshouse during the dark period. With the low intensity of the light from these lamps there can be no question of excessive photosynthesis. The factor B is sensitive to light but, in contrast with factor A, has nothing to do with photosynthesis.

The effect of the period of light upon plants is not confined to the development of blooms, for there are also other phenomena influenced by the length of day. As an example may be mentioned the growth of the stem. But although stem growth and flowering are both influenced by the length of day, there are probably two quite independent reactions at play; various observations point in that direction.

In the case of Rudbeckia for instance a long day (14 hours light) is beneficial for stem growth and flowering, whereas when the day is short the plant does not flower and bears only leaf rosettes. When young plants are exposed to light for a number of long days their stems begin to grow at once. Upon this being followed by short-day exposure stem growth stops and a rosette is formed at the top of the stem. The remarkable fact is that blooms then begin to develop on this rosette. Apparently bloom development is induced by the long-day treatment and the effect of this continues even after the stem growth has been brought to a standstill.

The effect that long days have upon stem growth is also to be noticed in the case of the strawberry, which is a short-day plant. The flowering process starts during the short days of the autumn but it is not until the long days arrive that the plant blooms and bears fruit. During the short days the leaf stems are very short, but as the days lengthen they grow longer and ultimately the flowers open. But during the long days no new flowering is started. Thus in the case of this plant it can also be said that the influence of the length of day upon flowering is quite different from that upon the stems.

The foregoing serves to show in broad lines and very briefly the stage which has been reached in the problem of the photoperiodism of plants. The general picture is still very confused. The theories formed by some investigators to explain the mechanism of the action of light upon plants are not very convincing. Much more research will be needed before a satisfactory insight into this matter can be obtained.

These investigations are of importance for the biologist studying the laws applying in natural life, but they are also of importance for the nurseryman, who can only utilize natural and artificial light successfully when he knows how various kinds of plants react to such light.

Since the various kinds of plants make entirely different demands upon the nature of the light it is difficult to find a lamp of universal use for their irradiation. Mercury lamps, which give mainly blue light, and neon and sodium lamps with their predominantly red and yellow radiation respectively, do not by any means provide the best solution in all cases. In this respect fluorescent lamps offer more possibilities. With these lamps there is a choice between the "warm white" type with its spectrum mainly in the red and yellow, and the "daylight" type in which much blue occurs. By a combination of "warm white" and "daylight" lamps it is also possible to obtain intermediate colours.

For good plant development irradiation with exclusively red and blue light would probably suffice, the red light for promoting photosynthesis and the blue light for the shaping of the plant. Fluorescent lamps with a mixture of cadmium borate and magnesium tungstate would be the most suitable, but it would have to be investigated what ratio of these two phosphors offers the most economical solution for each kind of plant.

Summary: During the winter months, for many plants, there is too little sunlight for strong development. Attempts are then made to stimulate growth by means of artificial light. Intensification of the light, however, is not favourable under all circumstances. It is necessary for a nurseryman to know exactly the various effects that light has upon plants. In this article the importance of light is discussed in connection with the feeding, the shaping and the flowering of the plants. Attention has to be paid both to the intensity and duration of the irradiation and also to the composition of the light. Different kinds of plants react differently. The effect upon flowering is determined for the greater part by the photoperiodicity, i.e. the ratio of the periods of alternating light and darkness. It is clearly shown that the solution of the problems discussed is still only in an initial stage.

AN AUTOMATIC BRAKING DEVICE FOR X-RAY APPARATUS

by J. M. CONSTABLE *).

621.313.334.07-59;621.386.14

The operation of modern X-ray equipment for mass chest survey must in most respects be made automatic in order to enable the examination of hundreds of persons per hour. An interesting problem in the design of such equipment is the provision of means for rapid adjustment to the individual requirements of each examinee. Differences in chest size necessitate different exposure times. An article published earlier in this Review described the automatic timer which solved this problem. Differences in height require an adjustment of the vertical position of the apparatus for each examinee. The present article describes how in performing this adjustment by means of an electric motor the movement of the apparatus can be rapidly interrupted in order to avoid overshooting the required position.

The problem of decelerating or rapidly interrupting a movement is rather old. Many braking devices have been designed for the purpose in order to meet the requirements of every special case.

In vehicles it is a common practice to use the motor itself as a brake, i.e. to make it produce a torque opposite to the actual rotation of the wheels. This method is especially useful when applied to electric motors, as their torque can be reversed in a simple way. Moreover, in this case the kinetic energy to be dissipated need not be converted into heat but in some cases may be recuperated in the form of electric energy supplied back to the power line. The economic importance of this feature is evident in the case of electric trains travelling over mountainous routes. Other merits of the method as applied to electric motors are the reliability and good controllability of the braking action. These features account for its frequent use in elevators and cranes.

As a further advantage of the principle we may mention that no mechanical braking equipment is required. A limitation, however, will in many cases be imposed by the fact that the braking torque produced by the motor cannot exceed the starting torque to any significant extent.

After this short prelude, we can turn to the special case we have in mind.

The apparatus for series production of miniature X-ray photographs, described in this Review some time ago [1]), is provided with a mechanism allowing vertical adjustment of the X-ray tube and photographic camera in accordance with the height of the examinee. The weight of all moving parts is approximately counterbalanced by springs. Nevertheless, the vertical adjustment of the apparatus, if performed by hand, in addition to being tedious and time-consuming, may draw heavily on the physical strength of the operating personnel, who in many cases must cope with 200 or more examinees per hour. Therefore an electric motor which can be switched on and off simply by pressing and releasing a lever actuating one of two micro-switches is provided to move the apparatus up and down.

Rapid starting and stopping is of prime importance in this adjustment in order that too much time is not lost with every examinee. Since the torque requirement for stopping is essentially the same as that for starting, using the motor itself as a brake is a logical solution to the braking problem. Moreover, a previously mentioned advantage of the method — that no special parts adding weight to the apparatus are required — was particularly important for our purpose, as every effort was made to cut down the weight of the apparatus in order to make it easily transportable.

It is evident that the interruption of the movement in this case must be made automatic. The motor and the moving parts must come to a stop as soon as possible after the operating lever is released, without interference from the personnel and regardless of whether the movement is "up" or "down".

Fig. 1 shows in schematic fashion the solution applied for our apparatus. The motor is of the single-phase induction type. As is well known, the stator winding of such a motor does not produce a starting torque but will make the rotor run in either direction when once started. For starting, an auxiliary winding on the stator is provided, which is placed at an angle of 90° with respect to the main winding and is energized with a phase difference of approx. 90° with respect to this winding. When the rotor has reached full speed the auxiliary winding is cut out. The starting torque may be reversed by simply reversing the polarity of either the auxiliary

*) Formerly North American Philips Co., Inc., New York, N.Y.

[1]) H. J. Di Giovanni, W. Kes and K. Lowitzsch, A transportable X-ray apparatus for mass chest survey, Philips Techn. Review **10**, 105-113, 1948/49 (No. 4).

winding or the main winding. After this brief recapitulation, the wiring diagram in fig. 1 will readily be understood.

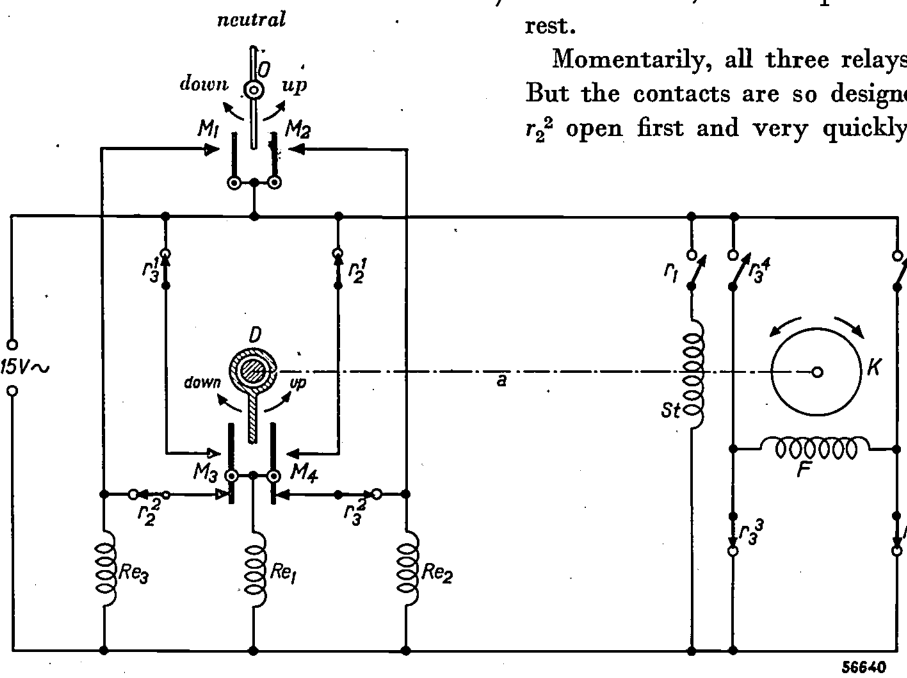


Fig. 1. Schematic circuit diagram of the braking device. K single-phase induction motor, with main winding F and auxiliary (starting) winding St; D magnetic drag mounted on motor shaft a; O operating lever; M_1-M_4 micro-switches; Re_1-Re_3 coils of relays; r_1 contact actuated by Re_1; r_2^1, r_2^2, r_2^3, r_2^4 contacts actuated by Re_2; r_3^1, r_3^2, r_3^3, r_3^4 contacts actuated by Re_3. All contacts are indicated in their initial position (coil of relay de-energized).

Re_1, Re_2. Re_3 are relays with a number of contacts r_1, r_2^1, r_2^2, . . .; r_3^1, . . . , all indicated in the initial position (coil of relay de-energized). M_1 and M_2 are two micro-switches actuated by the operating lever. M_3 and M_4 are two additional micro-switches actuated by the magnetic drag (D) on the shaft of the rotor. In *fig. 2* the drag is shown separately. The shaft carries a small permanent magnet (A); a copper ring supported by a ball bearing is placed around this magnet with a small separation distance. When the shaft is rotating eddy currents are induced in the ring, tending to make it rotate with the shaft. Hence, the switching lever fixed to the ring is turned over in much the same way as by a friction coupling between ring and shaft, but with the significant difference that this "friction" disappears when the motor comes to a stop: the switching lever then falls back to its initial position. These properties of the magnetic drag make it possible not only to energize the motor automatically in an opposite sense after the operating lever is released, but also to interrupt this energizing current as soon as the movement is stopped, thus preventing the motor from actually starting rotation in the opposite sense [2]).

Let us now consider four different situations.

I) The operating lever O on "neutral", rotor at rest.

All coils of relays and windings of motor are de-energized.

II) O on "down", hence M_1 closed, rotor still at rest.

Momentarily, all three relays are energized. But the contacts are so designed that r_3^2 and r_2^2 open first and very quickly; hence Re_2 is

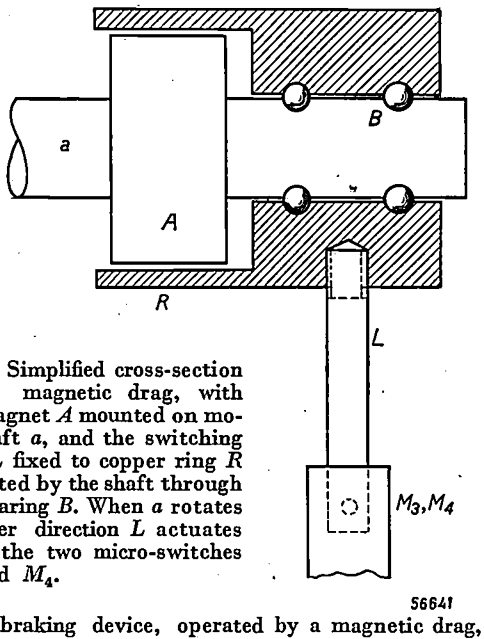


Fig. 2. Simplified cross-section of the magnetic drag, with ring magnet A mounted on motor shaft a, and the switching lever L fixed to copper ring R supported by the shaft through ball bearing B. When a rotates in either direction L actuates one of the two micro-switches M_3 and M_4.

[2]) A similar braking device, operated by a magnetic drag, has been designed earlier for various machine tools, as e.g. punching and shearing machines.

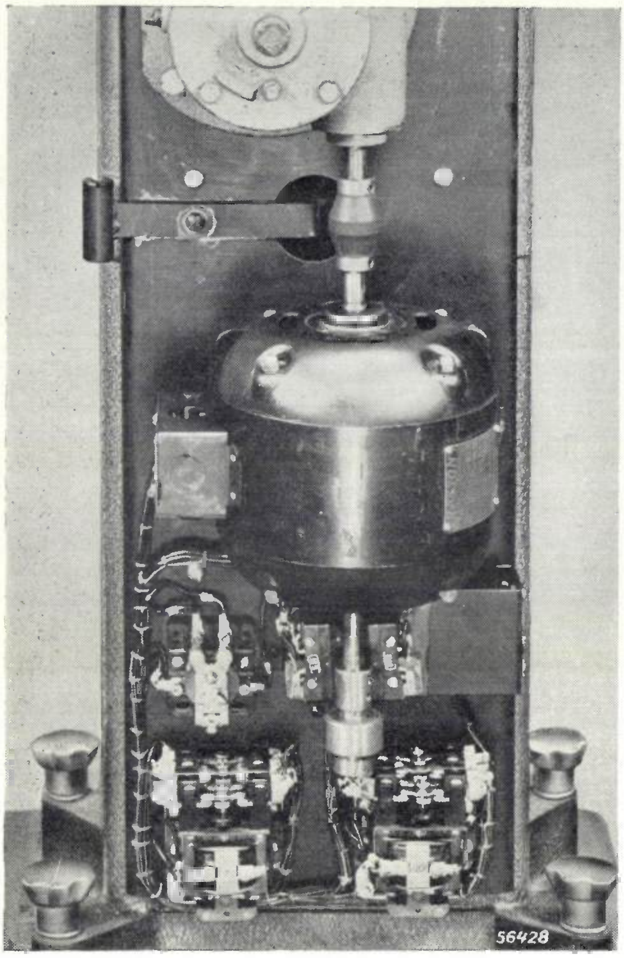


Fig. 3. Showing the motor and relay assembly in one of the pedestals of the X-ray apparatus (cf. the article quoted in footnote [1]). The ring of the magnetic drag and the two micro-switches are clearly visible on and next to the lower end of the motor shaft.

immediately de-energized and r_2^2 closes again (while r_3^2 remains open). Now Re_1 is energized via r_2^2 and M_3, hence the starting winding St is energized. As the main winding F is energized via r_2^3 and r_3^4, a torque is produced in the "down" sense.

III) O held on "down", rotor rotating in "down" sense. The magnetic drag is turned over, so M_3 is actuated, Re_1 is de-energized (r_3^1 being open) and St is cut out; Re_3 remains energized, so does F; rotor continues running "down".

IV) O again on "neutral", whereby M_1 is opened, rotor still running "down", M_3 remains actuated. Re_3 de-energized; Re_2 energized via r_3^2, M_4, M_3, r_3^1, hence main winding F energized via r_3^3 and r_2^4, i.e. in the sense opposite to situation II. Moreover, Re_1 is energized via M_3, r_3^1, hence the starting winding St is energized also and the desired torque in the "up" sense is produced, strongly braking the still continuing "down" rotation.

When the rotor comes to a stop the magnetic drag is released, M_3 returns to its initial position and all coils and windings are de-energized; situation as in I.

Analogous considerations apply to the case where O is first set on "up" and again on "neutral" after a sufficient upward displacement of the apparatus has been obtained.

The micro-switches M_1 and M_2 are closed by the operating lever O through a limit bar assembly which ensures returning the lever to "neutral" when the upper or lower limits of travel of the apparatus have been reached.

The total weight of all moving parts is 200 lbs. In a position of medium height this weight is exactly counterbalanced by the springs, in higher positions it is not completely balanced and in lower positions it is overbalanced. The motor, which rotates at 1750 revolutions per minute, is capable of lifting the apparatus from every position at a rate of 1 inch in 1 second. Without any braking device, i.e. relying only on the friction of the moving parts, in general the motor would not come to rest before about 20-30 revolutions after the operating lever is released, thus in many cases overshooting the desired position of the apparatus. Under the most unfavourable circumstances — when the apparatus moves downward from the extreme upper position — it would in some cases (when friction is low) even continue down due to its incompletely balanced weight until it reached some medium position. With the braking device described above, however, regardless of the position of the apparatus when the operating lever is released, the motor comes to a stop within 2 or 3 revolutions, corresponding to an apparatus displacement of not more than 0.1 inch.

Fig. 3 shows a photograph of the combination of relays mounted on the motor frame.

Summary. In the Philips apparatus for series production of miniature X-ray photographs the height of the X-ray tube and camera is adjusted by means of an electric motor in accordance with the height of the examinee. The motor is operated with a single control lever for up and down movement. When the lever is returned to neutral one of a pair of micro-switches, actuated by a magnetic drag mounted on the motor shaft, causes a set of relays to change the energizing of the motor in such a way that a torque opposite to the still continuing motion is produced. This braking torque brings the motor to a stop within 2 or 3 revolutions, thus avoiding the risk of overshooting the desired position of the apparatus. The magnetic drag returns to its initial position as soon as the rotor comes to a standstill and prevents the rotor from actually starting rotation in the opposite direction.

DARK-ROOM LIGHTING

771.24:535.736.14:771.524.53

To speak of dark-room lighting, as the photographer desires it, seems paradoxical; how can a room be illuminated so that it still remains dark? The paradox disappears when it is borne in mind that the conceptions of "dark" and "light" are used here for two different organs of perception. The room has to be dark, or at least sufficiently so for the photographic material, and it has to be light, or again at least sufficiently so for the human eye. These two requirements can indeed be compatible one with the other provided there is sufficient difference in the spectral sensitivity of the two organs. This article deals with the fundamental possibilities for dark-room lighting existing by reason of these considerations, and their practical realization is discussed for various purposes.

The problem of dark-room lighting in photography amounts to this, that it is desired to give the eyes sufficient light for performing a certain task without that light causing any noticeable blackening of the photographic material.

If the general sensitivity of the photographic material is relatively small the problem can easily be solved by a suitable choice of the level of illumination. Otherwise the object in view can still be reached provided there is sufficient difference between the spectral sensitivity of the photographic material and that of the eye, in which case a suitable spectral distribution of the light used must be chosen.

As a matter of fact all practical systems of dark-room lighting are based on these principles. When we come to consider, however, normal photographic negative material in particular, disregarding for [a moment positive papers, X-ray films and suchlike, then we find that the lighting of the dark room has become more and more difficult. The photographic industry is aiming at making the negative emulsions more and more sensitive so that good photographs can be taken with little light or with very short exposures; at the same time attempts are being made to make the spectral sensitivity curve of the emulsions approximate as far as possible to that of the eye, so as to reproduce the various colours with the correct shades of grey. As a result the differences between the eye and the negative will gradually disappear and we may therefore say somewhat paradoxically that the aim of the modern photographic industry in regard to negative material seems to be directed towards rendering dark-room lighting impossible.

In the case of some "panchromatic" emulsions this object has already been fairly well attained. Such emulsions have to be developed in almost complete darkness (unless prior to or during the development the sensitivity of the emulsion is reduced by means of a so-called desensitizer, which however also has its disadvantages). Now in order to avoid having to sit in the dark during the whole developing process the method of tank developing is used for panchromatic films. The developer is contained in a tank, which can be closed in a light-proof manner, and in which the film is placed while the room is in darkness or only very weakly illuminated, the film then being taken out again after a certain time has expired; once the tank has been closed the photographer can turn on the lights and carry on with some other work. But this solution of the problem of dark-room lighting is not really what is desired, for it is just the developing process itself that the photographer most desires to watch, and to do so he has to examine the film or plate several times during that process, thus of necessity exposing it to some light. Something similar applies in the photographic industry where these emulsions are made: if the continuous processes of manufacture are to be checked at all then the emulsion must be exposed to a certain, though very small, amount of light.

In this article we shall consider what possibilities still exist for the lighting of a dark room in this sense while developing various modern emulsions. We shall then deal with the practical realization of these possibilities with different kinds of dark-room safe-lights, and also briefly discuss the dark-room lighting for handling positive papers (which is fundamentally simpler).

From the foregoing it appears that the investigation into the possibilities of dark-room lighting will amount to an investigation of the sensitivy of the eye for light on the one hand and of the photographic material, in this case the negative emulsion, on the other. We shall have to ask ourselves the questions:

1) What amount of light does the eye need in the dark room?
2) How much light can the emulsion safely stand?

The fundamental difficulty outlined above can be avoided by applying a so-called image converter. By means of infra-red rays an image of the object to be viewed is cast upon a screen covered with a certain phosphorescent substance or with an electron-emitting layer. In the first case the screen, which must previously have been activated by irradiation with a radio-active preparation, produces a visible "fluorescent picture" directly, whilst in the second case the electrons emitted produce a visible picture on a normal fluorescent screen placed farther away [1]).

When a film or plate with a panchromatic emulsion, which is insensitive to the infra-red rays, is being examined in this way no trouble whatever is experienced from the fact that its spectral sensitivity in the visible zone so strongly resembles that of the eye, and one can raise the "illumination" to a high level.

As far as we know this method has not yet been applied in the practice of photography.

What the eye requires

In order to decide how much light the eye needs we must first establish what the eye has to do and under what circumstances it has to perform that duty.

To exercise control over the developing process, which we take to be the purpose of the dark-room lighting, the photographer examines the partly developed negative in transmitted light to see whether there is sufficient contrast of the details in the light and dark parts. Since the contrast sensitivity as a rule diminishes with the level of illumination, when working in the dark room the photographer will certainly not be able to observe just as fine contrasts in the negative as can be seen in daylight. But this is not necessary. It has been found that the process of development can be followed quite well if, for instance, one can perceive the difference between two densities of the order of 1.0 and 1.1 [2]).

There are, however, also parts in the negative having a very much smaller density, say 0.1, and since these parts let ten times as much light through as those where the density is 1.1, there will be a certain amount of glare, so that the observer's contrast sensitivity in the darker parts is adversely affected.

The quicker the negative can be examined, the better it will be protected against fogging. The time that the photographer must allow himself to examine and judge the negative can be taken as 7 seconds, to which is to be added 3 seconds before that for the eye to adapt itself to the mean level of brightness behind the negative.

Van Kreveld and Van Liempt [3]) have determined experimentally the brightness that an observer needs under these conditions — thus with the glare referred to and a total observation time of 10 seconds — to be just able to distinguish the two densities 1.0 and 1.1, likewise the two densities 1.0 and 1.2 and also 1.0 and 1.3 The experiments were carried out with different kinds of monochromatic light (wavelengths $\lambda = 5086, 5890, 6438$ and > 6800 Å). The results, averaged over a large number of observations by different persons, are represented in *table I* by the required radiance in W/m²·steradian at the part of the negative where the density is 1.0. The corresponding irradiance in W/m² at the back of the negative is shown in the next column. It is assumed that the negative emulsion fully diffuses the transmitted light, and account is taken of the light-absorption of the non-reduced silver bromide (not yet removed by fixing), which averages for various kinds of plates and films about 60 %.

For the complete control of the development of the negative the photographer will normally have to examine it twice for contrast of details.

Table I. The minimum illumination required to permit the distinction of various contrasts (first column) in a photographic negative under the conditions given, in the case of monochromatic light of different wavelengths.

Densities to be distinguished	Wavelength in	Required radiance (at the place with density 1.0)	Irradiance on back of negative	Quantity of radiation when examining negative twice
	Å	10^{-3} W/m²·sterad.	10^{-3} W/m²	10^{-3} Wsec/m²
1.0-1.1	5086	0.14	1.1	22
	5890	0.18	1.4	28
	6438	1.0	8	160
	>6800	7.7	60	1200
1.0-1.2	5086	0.06	0.45	9
	5890	0.02	0.15	3
	6438	0.25	2	40
	>6800	1.8	14	280
1.0-1.3	5086	0.002^{5}	0.02	0.4
	5890	0.005	0.04	0.8
	6438	0.013	0.1	2
	>6800	(0.4)	(3)	(60)

[1]) G. Holst, J. H. de Boer, M. C. Teves and C. F. Veenemans, Physica 1, 297-305, 1934.

[2]) When light with an intensity I_0 is thrown upon an exposed plate and behind the plate the intensity is only I, the density is defined as $D = \log I_0/I$. In what follows, the density above the so-called natural fogging of the negative is meant. In this case I_0 means the intensity behind a part of the negative on which no light has fallen when the photograph was taken.

[3]) A. van Kreveld and J. A. M. van Liempt, Measurements on dark-room illumination, Physica 5, 345-373, 1938.

It is assumed that during the rest of the time the negative is not exposed to any light in the dark room (the developing bath being covered over for instance). Altogether, therefore, the negative will have to be exposed for 20 seconds to the irradiance quoted in table I. The number of Wsec/m² (quantity of radiation) calculated in this way is indicated in the last column of the table and plotted in *fig. 1* as a function of the wavelength for the three different contrast sensitivities in the first column of table I.

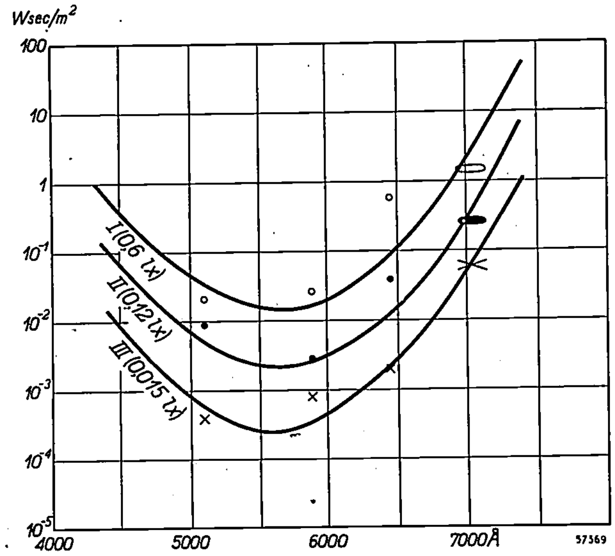


Fig. 1 To distinguish the difference between two densities under monochromatic light one needs for each wavelength a certain irradiance in W/m² on the back of the negative. Curves have been plotted representing the energy in Wsec/m² corresponding thereto with an exposure of 20 seconds for distinguishing the densities 1.0 and 1.1 (curve *I*), 1.0 and 1.2 (curve *II*) and 1.0 and 1.3 (curve *III*). (Measured in lux the necessary irradiance is practically the same for all wavelengths; this is indicated against the curves.)

The experimental results can also be expressed in the required brightness (c/m²), illumination (lux) and exposure (luxseconds). The figures given in table I, each for a certain wavelength, then have to be multiplied by the relative luminosity factor corresponding to the wavelength (and by the mechanical equivalent of light 683 lm/W). We have not introduced here the visual quantities because for comparison with the effect of the radiation upon the emulsions, which will be our next step, the eye sensitivity has in any case to be eliminated from the figures. It is to be observed here, however, that if the visual quantities are used (employing the normal eye-sensitivity curve for high brightnesses) one will arrive at the unforeseen result that the requisite numbers of c/m², lux and luxseconds are roughly independent of the wavelength of the light used. For the three contrast observations considered in table I it appears that about 0.6, 0.12 and 0.015 lux respectively are required.

This independency is to be so interpreted that the contrast observation and the perception of brightness depend upon the wavelength in much about the same way: the points of measurement in fig. 1 lie on curves showing about the same trend as the reciprocal relative luminosity factor. Since this similarity is not likely to be a mere matter of chance, the relative luminosity curve has therefore in fact been taken as being representative for the rather strongly scattered points of measurements.

As regards the difference between the relative luminosity curves with low and with high levels of brightness, something will be said about this farther on.

What the emulsion can stand

What happens in the negative when it is exposed to light while being developed in the dark room? Following the line of thought of Van Kreveld and Van Liempt [3]), let us consider the density curve of the negative. This gives the relation between the quantity of radiation H that has fallen upon a point of the film and the resultant density D at that point. When the log H is plotted along the abscissa this relation, which also depends upon the kind of emulsion and the manner of developing, is represented approximately by a straight line, such as line *1* in *fig. 2*. The slope of this line, thus $\mathrm{d}D/\mathrm{d}\log H$, provides a measure for the richness of contrast for the negative, namely the difference in density that will result from a certain difference in brightness between two parts of the object photographed.

Suppose that line *1* in fig. 2 is the density curve $D(H)$ of a negative developed in absolute darkness. If instead of being developed in the dark the negative receives a small dose of light H_0 equal over the whole surface then the density at each point will increase to a value

$$D_1(H) = D(H + H_0)\; . \; . \; . \; . \; . \quad (1)$$

Thus the negative gets a new (apparent) density curve which can be constructed from line *1* in fig. 2 and is represented by line *2*. It is seen that the slope of line *2* at each value of H is smaller

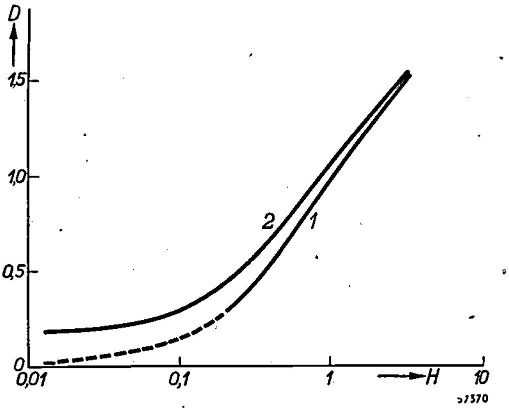


Fig. 2. Line *1* is the density curve $D(H)$ of a normal negative developed in absolute darkness. Line *2* is the "apparent" density curve $D_1(H)$ of the same negative fogged during developing by a small extra exposure (H_0).

than that of line *1*. In other words, the extra exposure of the negative results in a deterioration of the contrasts over the whole line. The relative change x of the contrast is

$$x = \frac{\mathrm{d}D_1/\mathrm{d}\log H - \mathrm{d}D/\mathrm{d}\log H}{\mathrm{d}D/\mathrm{d}\log H} \quad . \; . \; . \quad (2)$$

With the aid of (1) this can be written as

$$x = H_0 \frac{\mathrm{d}^2D}{\mathrm{d}H^2} \bigg/ \frac{\mathrm{d}D}{\mathrm{d}H}.$$

Experience teaches that in the case of a normal negative a contrast reduction x of maximum 10 % can be allowed without the negative suffering any appreciable loss in quality. The question is now with what value of the extra exposure H_0 one can still be sure that the permissible value of x is not exceeded.

A simple answer can be given thanks to the fact that x has a maximum value for a certain density. This can easily be understood if one studies the density curve (line *1*) in fig. 2. In the straight part this has an equation of the form

$$D = a \log H + b \quad . \; . \; . \; . \; . \quad (2)$$

(a and b are constants). From this it follows that:

$$\frac{\mathrm{d}^2D}{\mathrm{d}H^2} \bigg/ \frac{\mathrm{d}D}{\mathrm{d}H} = -\frac{1}{H},$$

$$x = -H_0/H.$$

As H increases, the contrast reduction thus approaches zero. On the other hand, for very small values of H, where the density curve deviates considerably from the straight line, the curve can be represented by an equation of the form:

$$D = cH + d \quad . \; . \; . \; . \; . \; . \quad (3)$$

(c and d are again constants). From this it follows that $\mathrm{d}^2D/\mathrm{d}H^2 = 0$, so that in this part of the density curve the contrast reduction likewise equals nil. The value of x must therefore pass through a maximum between the very small and the very large values of H.

With the density curves usually found in practice this maximum mostly lies round about the density 0.3. Since equation (2) already holds to a good approximation here, the maximum contrast reduction taking place in the negative is

$$x_{\max} = H_0/H_{0.3},$$

where $H_{0.3}$ signifies the exposure required to blacken the film to a density of 0.3 on the emulsion concerned. Substituting for $x_{\max}$ the above-mentioned permissible value of 10%, we see that the extra exposure (H_0) of the negative while it is being developed is limited by the condition

$$H_0 \leqq 0.1\, H_{0.3}. \quad . \; . \; . \; . \; . \; . \quad (4)$$

So far we have been disregarding the possibility that the light under which the photograph is taken (H) and that to which the negative is exposed while being developed (H_0) may have a different spectral composition. In practice this is actually the only case that occurs. In the above hypothesis the starting point that the density will everywhere increase owing to the extra exposure H_0 still holds. The extent to which this takes place, however, cannot be expressed off-hand by equation (1) because as a rule the sensitivity of the emulsion for the two kinds of light differs. Equal quantities of radiation of the one kind of light and of the other will not result in the same density, and even if this were so for one particular value of the quantity of radiation, it would not necessarily apply for all values. In other words, for different kinds of light the density curve might assume a different form, having a different slope in the straight part. Fortunately this proves to be not the case, for in general the slope of the density curve is to a sufficient approximation the same for all wavelengths. Further a law of addition[4]) is found to apply for the photographic emulsion just as for the human eye: for a given emulsion each wavelength has a certain effect that can be expressed by a number and after multiplication by the respective number two quantities of radiation of different wavelengths can simply be added, this sum then determining the density that will be obtained by the action of the two quantities together. Thus we can simply convert the extra exposure H_0 taking place in the dark room into an equivalent exposure of the same spectral composition as that which acted upon the photographic emulsion while the photograph was being taken (e.g. daylight), or vice versa. If we now regard the symbols H_0, H and $H_{0.3}$ used above as indications for such converted and, as regards density, equivalent quantities of radiation, the whole deduction of equation (4) remains valid. .

When writing down equation (1) we have also tacitly assumed, for simplification, that the sensitivity of the emulsion is the same whether wet or dry and that the density is determined only by the quantity of radiation and not partly also by the time of the exposure (absence of the Schwarzschild effect).

In order to deduce from the foregoing the fundamental possibilities for the illumination of a dark room let us first assume that we are using a monochromatic light source. We then have to find experimentally what number of Wsec/m² of that wavelength is required to obtain a density of 0.3. This test has to be repeated with various monochromatic light sources of different wavelengths. Such measurements — in essence amounting to the determination of the spectral sensitivity curve of the emulsion in question — have been

[4]) A. van Kreveld, thesis, Utrecht 1933, and Physica 1, 60, 1933.

taken by Van Kreveld and Van Liempt[5]). When the values found, multiplied by a factor 0.1 according to equation (4), are plotted as function of the wavelength one obtains for each kind of film a certain curve, such as represented in *fig. 3a* for two orthochromatic emulsions and in fig. 3*b* for three panchromatic emulsions.

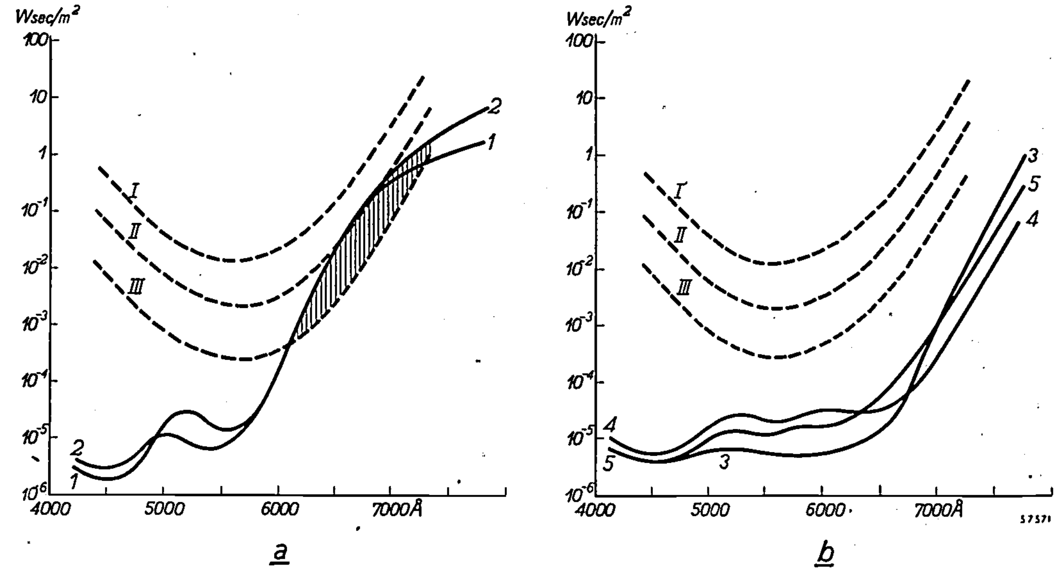


Fig. 3. One tenth part of the energy in Wsec/m² required to obtain a density 0.3 on the film, plotted as function of the wavelength,
a) for two orthochromatic emulsions: 1. Ilford Double X-Press, 2. Agfa Isochrom;
b) for three panchromatic emulsions: 3. Ilford Hyper Pan, 4. Avia Argus Pan, 5. Agfa Isopan S.S.
The dotted curves are the curves of fig. 1.

These curves give an upper limit for the dark-room lighting if a noticeable contrast reduction due to fogging of the negative is to be avoided. On the other hand the three curves of fig. 1, reproduced by dotted lines in figs 3*a* and *b*, each give a lower limit for the dark-room lighting if the density differences 1.0-1.1, 1.0-1.2 or 1.0-1.3 respectively are to be distinguishable. It is therefore only possible to solve the problem of dark-room lighting by using those wavelengths for which the corresponding dotted curve lies below the full-line curves.

Figs 3*a* and *b* show at once that we have already gone too far by stipulating the condition that it should be possible to observe a difference in density between 1.0 and 1.1. Neither with panchromatic nor with orthochromatic emulsions is there a wavelength making this possible without harm to the negative. With panchromatic emulsions this is not even possible if we reduce our demands to the perception of the difference in density of 1.0 to 1.2 or 1.0 to 1.3; the corresopnding curves still do not intersect each other. There is, however, an intersection of the curves in the case of orthochromatic emulsions (see the shaded part in fig. 3*a*): it appears possible to observe a difference from 1.0 to 1.2 (illumination 0.12 lux) when using wavelengths between 6600 and 6800 Å; to perceive a difference between 1.0 and 1.3 (illumination 0.015 lux) we can use the wavelength range from about 6200 up to 7300 Å.

Dark-room safe-lights for negative emulsions

Orthochromatic emulsions

Within the shaded part of fig. 3*a* all wavelengths are not equal. The best result will apparently be obtained with a monochromatic light source having the wavelength at which the dotted curve extends farthest below the full-line curve. In practice, however, it is not possible to make a monochromatic light source for any arbitrary wavelength (only the sodium lamp can be considered as monochromatic light source suitable for common use, but its wavelength of 5890 Å falls outside the area found suitable for orthochromatic emulsions). Most dark-room lamps are therefore incandescent lamps with a bulb of coloured or

[5]) See the article quoted in footnote [3]). Actually the quantities of radiation were not measured for a density 0.3 but for 0.1, since this is as a rule taken as starting point for determining the sensitivity of emulsions. For most emulsions $H_{0.3} = 2.5\ H_{0.1}$ applies to a good approximation; the curves in fig. 3 have been drawn with the aid of this equation.

lacquered glass intercepting certain parts of the continuous spectrum of the incandescent filament. For developing orthochromatic emulsions a dark red lamp is being made which gives light only with wavelengths greater than 6400 Å; see the spectral distribution given by curve *a* of *fig. 4.*

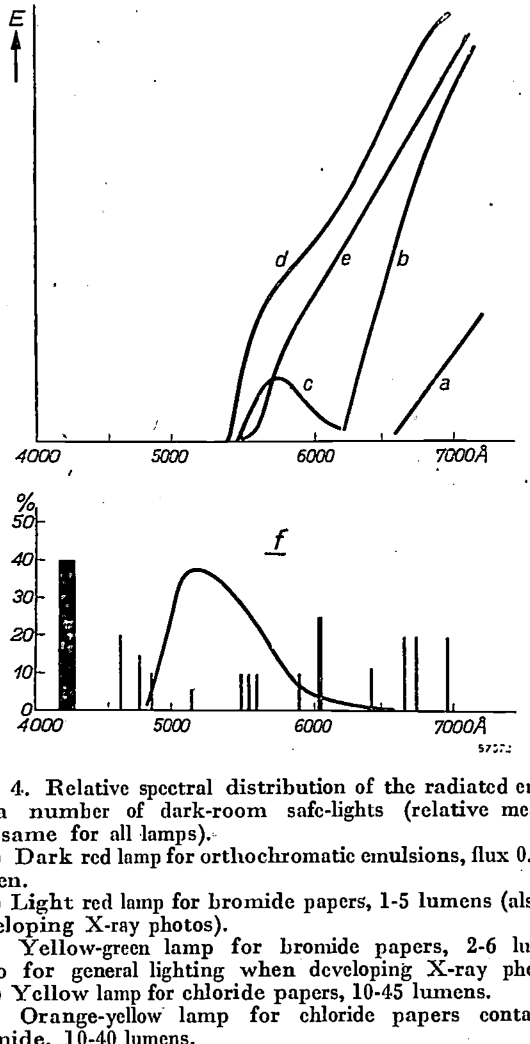


Fig. 4. Relative spectral distribution of the radiated energy of a number of dark-room safe-lights (relative measure the same for all lamps).

a) Dark red lamp for orthochromatic emulsions, flux 0.3-0.5 lumen.

b) Light red lamp for bromide papers, 1-5 lumens (also for developing X-ray photos).

c) Yellow-green lamp for bromide papers, 2-6 lumens (also for general lighting when developing X-ray photos).

d) Yellow lamp for chloride papers, 10-45 lumens.

e) Orange-yellow lamp for chloride papers containing bromide, 10-40 lumens.

In the case of the green lamp, serving for the general lighting when developing panchromatic emulsions, the radiated energy is so small that it cannot well be drawn in the same relative measure as for the other lamps. Moreover, the light from this lamp has a line spectrum and not a continuous one. In the separate diagram (*f*) the spectral position is indicated with the estimated ratio of intensity of the most important argon lines of interest to us, as also the transparency curve of the green lacquer used.

For such a more or less continuous spectrum, which in principle could be allowed to extend even farther than the intersection area of the curves in fig. 3, we have to reverse, as it were, the formulation of our problem. With the aid of the "relative sensitivity curve" of the emulsion (fig. 3) and the spectral energy distribution of the given lamp we determine the density contribution of each wavelength and by integration over the whole spectrum we can calculate what irradiance is needed with that lamp in order to obtain in 20 seconds a density of 0.3. From equation (4) we then find the permissible irradiance (which can also be expressed in lux if desired) and with that we can find in table I by interpolation what contrasts can just be made perceptible with the spectrum in question. In the case of a wide spectrum the result will of course be the less satisfactory the more rays are emitted with wavelengths outside the intersection area. With the spectrum of the dark red lamp the illumination for an orthochromatic emulsion may amount to about 0.1 lux, with which somewhat smaller contrasts than 1.0 to 1.3 can be observed; with some routine it is thus possible to follow the developing process.

Assuming that when being examined during the developing the negative is held at a distance of about 50 cm from the lamp, it is calculated that the lamp may have a luminous intensity of about $0.1 \cdot 0.5^2 = 0.025$ candle and therefore with the normal light distribution may emit a light of a few tenths of a lumen. The dark red lamp emits this light with a filament power of 15 W. The "efficiency" is therefore very small, about 99.7% of the light being absorbed in the coloured bulb but with this low power the efficiency does not of course matter much.

Panchromatic emulsions

As regards panchromatic emulsions we have already come to the conclusion that no kind of light exists which makes it possible for a photographer to follow the developing process in the manner described above. All that can be allowed is a very weak general lighting of the dark room.

Instead of the criterion of perception of contrast which we had taken as starting point, it is now simply the perception of brightness that determines the performance of the eye. If the spectral sensitivity curves of the eye and of the panchromatic emulsion were in entire agreement — which is in fact the object of the "pan"-chromatic emulsion — then there would be no optimum spectral area to be indicated for the safe-light in the dark room. All panchromatic films, however, as is to be seen in fig. 3*b*, are still relatively sensitive in the blue, so that these rays are definitely unfavourable.

But also the red rays are not likely to be suitable. Agreement between panchromatic emulsion and the eye is desired and it is in fact approximated as for as possible for the normal levels of brightness in daylight; at low levels of brightness (illuminations from about 10 to 0 001 lux) the maximum of the relative luminosity curve is gradually shifted towards the shorter waves, owing to the visual task of the eye being transferred from the cones to the rods in the retina (Purkinje effect). For the brightnesses that one could have in the dark room we therefore have to reckon with a reduced sensitivity of the eye for the red [6]).

It appears that for general lighting when dealing with panchromatic emulsions it is best to choose a spectral area near the maximum of the shifted relative luminosity curve, i.e. near 5050 Å. That is why for this purpose a green dark-room safe-light has been made, the maximum radiation lying around 5600 Å. The maximum cannot be brought much closer to the said optimum wavelength because the spectrum must inevitably extend towards still shorter waves and there the sensitivity of the emulsion (in comparison to that of the eye) begins to increase considerably. The flux of this lamp is much smaller than 0.1 lumen.

Formerly this lamp was made in the form of a 5 W incandescent lamp with a bulb of coloured glass ("natural glass"). Now, like the dark red and other lamps still to be mentioned, the bulb is covered with a coating of coloured lacquer. This simplifies manufacture and closer tolerances can be prescribed for the transparency. It proved to be difficult, however, to get the desired small flux for the green lamp with a filament in a lacquered bulb, since the layer of lacquer has to be so thick that it cannot safely dissipate the power it has to absorb from the filament; in course of time it would peel off. For this reason instead of the incandescent lamp a glow lamp filled with argon is now used. The argon spectrum contains lines having wavelengths favourable for the purpose; the undesired spectral lines are intercepted by the green lacquer. This can be seen in fig. 4*f*. The power of this lamp is about only 1 W, so that the difficulty described above does not occur. The electrodes are made in the form of two semicircular discs lying in one plane perpendicular to the axis; see *fig. 5*. Thus the scanty light is emitted mainly in the axial direction of the lamp, that is to say downwards when the lamp is mounted in the normal position. In order to get some light at the side — the lamp has to serve as "beacon" for orientation of the not yet adapted eye — the bulb is frosted on the inside.

Notwithstanding the very small flux, this light has to be used with care, as will be understood from the foregoing. The panchromatic film must be exposed to it as little as possible, certainly less than 20 seconds at such a distance from the lamp that the illumination is 0.0005 lux (about 2 metres when the flux is 0.01 lumen). By way of comparison it is to be noted that the illumination of the earth on a clear starlit night with no moon averages about 0.0003 lux [7]).

The developing of positive paper

As explained in the introduction, dark-room lighting is fundamentally difficult for emulsions which while having high general sensitivity tend to approximate the relative lunimosity curve of the human eye. In the case of positive papers neither one nor the other is essential. In principle the general sensitivity need not be high, since during the exposure there is no movement and the paper can therefore be exposed as long as desired. Further, the colour sensitivity can quite well be limited to a small part of the visible spectrum, say violet and blue, since the negative is in any case colourless. The general sensitivity should, it is true, not be too

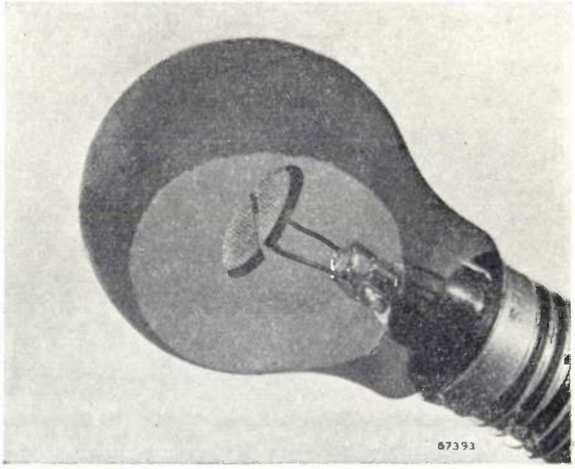


Fig. 5. Construction of the green dark-room safe-light for general lighting when developing panchromatic emulsions. It is a green-lacquered argon glow lamp with disc-shaped electrodes perpendicular to the axis of the lamp; the lacquer has been partly removed to show the electrodes. The bulb is normally frosted on the inside.

[6]) It may be surprising that in the experiments regarding the perception of contrasts in the negative a sensitivity curve was found corresponding approximately to the relative luminosity curve of the eye at high levels of brightness (fig. 1), although the level of illumination in those experiments was already in the Purkinje range. A possible explanation for this is that when one is looking sharply at an object it is particularly the central part of the retina (the fovea), where there are almost exclusively cones, that is brought into action. The sensitivity of the eye does not then depend upon the level of brightness. H. Arens and J. Eggert (Z. wiss. Phot. 24, 229-248, 1926) have made a study of the consequences of the Purkinje effect for dark-room illumination, but without differentiating between the criteria for the general lighting and those for the examination of the negative. A separation of the lighting for these two objects, making allowance for the Purkinje effect, has been proposed specially for the developing of X-ray films by F. Luft and M. Biltz, Trans. Ill. Eng. Soc. Japan 25, 101-112, 1941.

[7]) P. J. Bouma, Philips Techn. Rev. 5, 297, 1940.

low, so that reasonable times of exposure can be allowed for when making enlargements. Nevertheless, it is much easier to provide a useful lighting of the dark room here than is the case for the developing of negatives.

Bromide papers

The matter is relatively most difficult with the sensitive bromide papers, the relative sensitive curve of which extends from the violet to the yellow and orange. The best light to be used in the dark room is that with wavelengths above about 6200 Å, like that for orthochromatic negative emulsions, only with a much greater luminous intensity. The bulb of the light red lamp made for these papers is coated with a lacquer similar to that used for the ruby lamp but with a greater transparency. The spectral distribution of this lamp is given in fig. 4*b* in the same relative measure as the other curves. The flux is 1 to 5 lumens, thus ten times as high as that of the dark red lamp.

When developing positives under this light a remarkable difficulty may arise. When looking at a positive under red light in the dark room we have the impression that contrasts in the picture are much more pronounced than they appear in white light (given the same or greater illumination). Consequently, when we see the picture in daylight the contrasts are disappointing and the positive is judged to be too flat. This phenomenon, which also occurs in road lighting with sodium lamps [8]), is related to the already mentioned Purkinje effect and can be explained as follows. When a small surface is illuminated with a few lux of a long wave kind of light then with diminishing reflection coefficient the "subjective" brightness drops more than proportionately, owing to the fact that with diminishing brightness the maximum of the relative luminosity curve of the eye is shifted towards the short waves (in other words the eye becomes less sensitive for long waves). When we plot as function of the wavelength of the light the relation between the "subjective" contrast and the contrast by daylight for two areas the reflection coefficients of which are in the ratio of 1:10, graphs are obtained as shown in *fig. 6*. Each curve represents this relation for a certain brightness of the darker area. The phenomenon of contrast magnification owing to the Purkinje effect appears to be strongest when this brightness amounts to about 0.01 c/m^2 and it is still perceptible at 0.3 c/m^2, e.g. the range of brightnesses to be considered for developing bromide papers [9]).

An experienced photographer will of course know how to allow for this effect by developing somewhat longer. It can, however, be entirely avoided by not using red light in the dark room. Also with yellow or yellowish-green light a useful illumination is permissible for many kinds of bromide paper, though somewhat less than in the case of red light. Fig. 6 shows that the disturbing effect hardly occurs at all when using a light with wavelengths between 5200 and 5500 Å. At 5200 Å, however, there is a rather great risk of fogging the paper. A lamp has therefore been designed with a maximum in the spectral distribution at about 5700 Å, whilst the spectrum extends on the short-wave side to 5400 Å and on the long-wave side

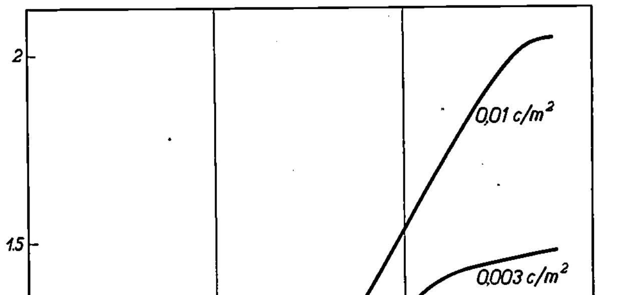


Fig. 6. Relation between a contrast observed under light with wavelength λ and the contrast observed under light of the composition of daylight, as function of λ. Two areas were observed the reflection coefficients of which bear the ratio of 1:10. Each curve applies for a different level of illumination; the brightness of the darker of the two areas is used as parameter. (Taken from the article by Van Liempt quoted in footnote [8])).

[8]) P. J. Bouma, Philips Techn. Rev. **1**, 166, 1936. — For dark-room illumination the effect has been further investigated by J. A. M. van Liempt, Zur Physik des Dunkelkammerlichts für Schwarz-weiss-Positive. Physica **10**, 645-660. 1943, where a closer definition is given of the conception "relation between contrast perceptions" used in this article.

[9]) Even without differences in colour, when examining a positive in the dark room and in daylight there would be a difference in the judgement of the contrasts owing to the difference in the level of illumination. This effect, however, is of less consequence. When developing negatives one is not concerned with the problem as a whole, partly because it is not the judgement of the magnitude of the contrasts but only their appearance that serves as criterion, and also partly because this work has to be done with still lower levels of illumination where the effect is again less felt (see fig. 6).

there is an unavoidable gradual drop to about 6500 Å; see fig. 4 curve *c*. The flux of this yellow-green dark-room lamp, having a power of 15 W, amounts to 2-6 lumens.

This yellow-green lamp proves to be very useful also for the general lighting of dark rooms for X-ray films; close to a developing bath, however, the light red lamp has to be used (see footnote [6])).

Chloride papers

In the case of chloride papers the spectral sensitivity curve extends from the violet no farther than about 5000 Å. Since in this spectral range the incandescent lamp gives little radiation, when making enlargements with incandescent lamps very long exposure is required, the more so since in enlargement apparatus only lamps of a limited power can be used in view of the heat developed. Owing to their greater sensitivity range bromide papers are more satisfactory in this respect and are therefore preferred by photographers although the chloride papers as a rule give more contrast. Nowadays, however, in the high-pressure mercury lamp (e.g. the HP 80 W) we have a light source yielding a relatively large amount of radiation just in the spectral range in which the chloride paper is sensitive. By using a mercury lamp in the enlargement apparatus it is therefore possible to make enlargements very quickly also on chloride paper and, owing to the limited spectral sensitivity range of chloride paper, it is then possible to provide ample illumination in the dark room[10]). For the latter purpose one can use for instance an incandescent lamp with a yellow bulb which emits only light in the wavelengths above 5000 Å and, in the case of a 15 W lamp, yields a flux of 10-45 lumens (for the spectral distribution see curve *d* in fig. 4). Still better, however, is the sodium lamp. The sensitivity of the chloride paper, provided it really contains no bromide, is so small at the wavelength of the yellow sodium light that one can safely use a 45 W sodium lamp, yielding 2700 lumens. In that case the dark room no longer does honour to its name, for it is then much more brightly illuminated than most living rooms.

In the spectrum of the sodium lamps some green and bluish-green lines occur (originating from the sodium and from the neon always present in the sodium lamp) for which the chloride paper is sensitive. In order to intercept this radiation the sodium lamp can be fitted with an orange-coloured vacuum jacket instead of the normal plain vacuum jacket.

For chloride papers containing bromide an incandescent lamp is being made coated with a lacquer the colour of which is between yellow and light red. The spectral distribution of this orange-yellow lamp (see curve *e* in fig. 4) lies as far as possible towards the short-wave side so as to avoid trouble from the above-mentioned effect of increased contrast perception.

Although, as we have seen from the foregoing, entirely different kinds of light are required for the various jobs that have to be done in the dark room, if necessary it is possible to manage with one single incandescent lamp by placing in front of it a filter of a certain colour according to the work in hand. This is a method frequently applied but it does not constitute any fundamental simplification for the photographer. In particular it lacks the advantage of being able to switch over easily from one kind of light to another when a number of different lamps are installed. This may be of importance for instance when it is desired to combine a very weak general lighting with a stronger and possibly differently coloured light to be switched on momentarily for the examination of a negative or a positive.

As regards the general lighting it may be added that if desired this can be indirect, one or more lamps of the type indicated being directed towards the ceiling of the dark room, given a light colour for that purpose. In the case of positive papers, where the picture is viewed in incident light, such a diffuse lighting (of a suitable level) can quite well be used also for checking the process of development itself. We shall not go further into the advantages and disadvantages of various lighting systems here.

Compiled by G. D. RIECK.

Summary. The fundamental possibilities for the lighting of a dark room while developing photographic negatives have been investigated by Van Kreveld and Van Liempt. The results lead to the determination of the most favourable spectral ranges for the light emitted by dark-room safe-lights. In the case of panchromatic emulsions it appears that only an extremely weak general lighting, preferably of a green colour, can be allowed; in the case of orthochromatic emulsions, by limiting the spectrum of the safe-lights to wavelengths above about 6400 Å the illumination can be increased to about 0.1 lux, under which light it is indeed possible to follow the developing process. For positive papers the problem of dark-room lighting is fundamentally simpler. Some dark-room lamps are mentioned which are suitable for bromide and for chloride papers respectively. The most favourable is the combination of chloride paper and a high-pressure mercury lamp in the enlargement apparatus and a sodium lamp as dark-room lamp. A sea of light is then available in the "dark-room".

[10]) See Philips Techn. Rev. **3**, 91, 1938.

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

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

1816: W. G. Perdok and H. van Suchtelen: A sensitive apparatus for qualitative testing of crystals on piezo-electricity (Appl. sci. Res. 's-Gravenhage **B1**, 195-204, 1948, No. 3).

A modern apparatus for qualitative testing of crystals on piezo-electricity, based on a principle announced by Giebe and Scheibe, is described. It is easily built and functions very constantly and reliably.

1817: J. Volger: On estimation of tenths (Appl. sci. Res. 's-Gravenhage **A1**, 215-218, 1948, No. 3).

A test is described which was made to check the possibility of estimating to one-tenth of a scale unit the position of a pointer on the scale of an instrument. It is proved that systematic errors arise in this sense that a general and outspoken tendency exists to read 0.3 and 0.4 one-tenth too low and 0.6 and 0.7 one-tenth too high.

1818: H. C. Hamaker: De invoering van moderne statistische methoden en opvattingen in het massa-producerend bedrijf (T. Efficientie en Documentatie **18**, 266-269, 1948, No. 11). (The introduction of modern statistical methods and ideas in mass-production; in Dutch.)

General considerations on decisions to be taken on the basis of essentially inaccurate data and on the use of sampling.

1819: H. C. Hamaker, J. J. Taudin Chabot and F. G. Willemze: De tolerantiekeuringen van gehele partijen (T. Efficientie en Documentatie **18**, 235-241, 1948, No. 11.) (Tolerance testing of whole lots; in Dutch.)

General considerations on inspection by sampling. Simple and double sampling. Description of a sampling table in use in the Philips Works (Netherlands).

1820: H. C. Hamaker: Foutentheorie en wiskundige statistiek (Statistica **2**, 177-205, 1948, No. 5/6). (Theory of errors and mathematical statistics; in Dutch.)

The writer advocates the use of standard deviation as a means of indicating the inaccuracy of observations. The mean and the standard deviations are subject to simple mathematical relations, independent of the form of the frequency curve. These relations are discussed and employed for a thorough treatment of the degree of rounding off permissible in practice. The maximum possible rounding interval is either 1/2 of the standard deviation or 1/6 of the range computed from 5 to 10 observations or 1/6 of the maximum mutual difference observed in ten pairs of observations. The rounding interval should at least be 1/5 of the maximum just specified.

1821: F. A. Kröger: The incorporation of uranium in calcium fluoride (Physica **14**, 488, 1948, No. 7).

Strongly fluorescent CaF^2-U is known to be obtained when CaF_2 with 0.004-0.5 mol % UO_3 is heated together with 6-25 mol % CaO; if no CaO is added the fluorescence is considerably weaker. The writer supposes that uranium is incorporated in such a way that the excess of charge of U (6+) over Ca (2+) is compensated by a simultaneous replacement of four F^--ions by four O^{2-}-ions. The result, therefore, can be described as a solid solution of Ca_2F_4 with $(CaU)O_4$. The centre of fluorescence is a uranium ion, surrounded by four O^{2-}-ions and four F^--ions.

1822: F. A. Kröger and W. Hoogenstraaten: Decay and quenching of fluorescence in willemite (Physica **14**, 425-441, 1948, No. 7).

With the aid of measurements of the efficiency and the decay of the fluorescence of willemite, it is shown that fluorescence excited by short-wave ultra-violet radiation is quenched by two kinds of radiation-less processes. One starts from the high excited state reached in the excitation process; its rate is governed by a linear process which, competing with the quadratic process leading to fluorescence, causes the efficiency of fluorescence to be dependent on the exciting intensity. This process is increased by iron and also by ball-milling. The other radiation-less process starts from an excited state of the manganese centre. It runs parallel to the slow fluorescence transition and thus influences the decay of the fluorescence. It increases strongly with temperature; the variation can be satisfactorily described with the aid af an activation energy. For excitation by long-wave ultra-violet only the latter process plays a part.

1823: Balth. van der Pol and H. Bremmer: Modern operational calculus based on the two-sided Laplace integral (Proc. Kon. Ned. Akad. Wetensch. Amsterdam **51**, 1005-1012 and 1125-1136, 1948, No. 8 and 9).

After having made some general remarks on Laplace transformations the writers claim the superiority of the two-sided form (integration from $-\infty$ to $+\infty$) over the one-sided form (integration from 0 to ∞). This is illustrated by examples: δ-function, linear differential equations with constant and with variable coefficients, originals having arguments of exponential character, operational identities, generating functions.

1824: C. J. Bouwkamp: On the mutual inductance of two parallel coaxial circles of circular cross-section (Proc. Kon. Ned. Akad. Wetensch. **51**, 1280-1290, 1948, No. 10).

Calculation of the mutual inductance of two parallel circular loops of circular cross-section (coaxial toroids). A series development is given. The coefficients are expressible in terms of a simple integral involving Bessel functions, which can be readily evaluated at least for the lower order terms. Finally approximate expressions for the mutual inductance are given.

1825: Joh. Hoekstra and H. A. W. Nijveld: The determination of the hardness of organic films (Rec. Trav. chim. Pays-Bas **67**, 685-689, 1948, No. 11).

An apparatus is described for the determination of the indentation hardness of organic films. The deepness of the indentation can be measured with an accuracy of 0.1 μ. The measurement is performed during the application of the force.

A few results are given, indicating how the hardness of some lacquer films changes with time and thickness of the layer.

1826: N. W. H. Addink: A general method for quantitative spectrochemical analysis (Rec. Trav. chim. Pays-Bas **67**, 690-696, 1948, No. 11).

Up till now Gerlach's internal standard method has been applied to all kinds of quantitative spectrochemical analysis. According to this method it is necessary to prepare for each individual material to be analysed a series of samples in order to set up working curves relating concentration and intensity ratio of lines of the element sought for and the standard element. Harvey (1947) uses the background near a line of the element as a kind of internal standard and his book contains many useful data for two-component systems. However, materials containing many major constituents such as mixtures of salts, oxides or metals cannot be analysed with great accuracy according to Harvey's method. In the following paper a general method derived form Harvey's is described; it can be divided into two parts, one for the rough estimation of the element concentration and a second part for the exact determination, obtained by successive additions of the element sought for. If the straight portion of the characteristic curve of the photographic emulsion is used a simple evaluation of the concentration is given, whereas the knowledge of the accurate shape of the characteristic curve is not essential.

1827: F. de Boer: On the use of Evjen's method in calculating Madelung potentials (Rec. Trav. chim. Pays-Bas **67**, 697-702, 1948, No. 11).

Evjen noticed that, in evaluating the sum $\Sigma e/r$, a reasonable accuracy may be obtained by summing up the contributions of the ions of a finite part of the lattice if this part is, as a whole, electrically neutral. The complications arising if this condition is not fulfilled are investigated for a simple lattice (CsCl, bounded by cubic faces). A correction is calculated and applied to the spinel lattice.

1828: A. Claassen and W. Westerveld: The photometric determination of cobalt with nitroso-R-salt (Rec. Trav. chim. Paysbas **67**, 720-724, 1948, No. 11).

As the result of a discussion on the choice of the most suitable wavelength to be used in the photometric determination of cobalt with nitroso-R-salt, a wavelength of 550 mμ is recommended, using absorption cells of 2 to 5 cm length.

The interference by copper and nickel is extensively dealt with. Some data are given on the interference by other elements.

1829*: J. H. van Santen and J. Th. J. Overbeek: Discrete energieniveaux in ionenroosters (Chem. Weekbl. 44 ,285-291, 1948, No. 21). (Discrete energy levels in ionic lattices; in Dutch.)

If ions of both signs unite to form a crystal lattice the discrete energy levels are generally broadened to energy bands. Transitions between such levels manifest themselves as rather broad emission or absorption bands. There are, however, some cases in which the interaction between ions of one kind and between these ions and neighbouring ions of another kind is so slight that the energy levels and the transitions remain sharp. This will occur especially if the radiation takes place in electron shells far from the periphery. A survey of several cases is given, showing how the energy levels of the free ions are influenced by the electric crystal fields. Especially rare earth ions and those of the iron group are dealt with. Finally the influence of extremely strong crystal fields is briefly discussed.

1830: J. A. Haringx: Het onderzoek van staalplaat met behulp van gekerfde proefstaven (De "Ingenieur" **60**, Mk 141-Mk 148, 1948, No. 51). (The examination of mild steel plate by means of notched test specimens; in Dutch.)

The examination of mild steel plates by means of notched test specimens aims at the determination of their disposition to brittle fractures. Although something is known about the factors giving rise to such brittle fractures, it is shown that for the time being we are not able to fix special specifications for the steels on account of theoretical considerations. One must therefore rely upon the results of more or less orientating tests. Some of these will mutually be compared in a research programme arranged by the Welding Society in the Netherlands in cooperation with the Belgian "Commission Mixte des Aciers", and are briefly described here. Further a review is given of the interesting investigations recently published by Mrs. Tipper and by Bagsar. The endeavour by Mc Gregor and Grossman to correlate a given stress distribution with the profile of the notch in the test specimen is also discussed.

1831: J. F. H. Custers: A new method for the determination of preferred orientations (Physica **14**, 453-460, 1948, No.7).

A new method is indicated for the determination of preferred orientations of flat specimens. It is shown that this method implies a somewhat easier construction of pole figures. Moreover it will simplify appreciably the calculation of the absorption in the specimen in its different positions during the taking of a series of X-ray pictures.

R 94: J. A. Haringx: On highly compressible helical springs and rubber rods, and their application for vibration-free mountings, I (Philips Res. Rep. 3, 401-449, 1948, No. 6).

A survey of the contents of a series of six papers, of which this is the first, is given in the introduction. The remainder of the first paper is devoted to a study of the elastic stability of helical springs under compression or under combined compression and twist. It is shown that, if the problem is simplified by replacing the helical spring by an elastic prismatic rod, for every point of the central line its rigidity with respect to shearing must be referred to the transverse force acting in that plane through the point considered, which in the unloaded state is normal to the central line of the rod but is no longer so in the deflected state. From this interpretation of the simplified model a new relation is deduced between the relative compression at which buckling occurs and the ratio of the length to the diameter of the spring. For this ratio there exists a limiting value below which no buckling occurs, not even under complete compression, and which is dependent only on the method of fixing the spring ends. A more accurate calculation, which takes the helical structure of the spring into account, confirms that the wire diameter and the number of coils are of secondary importance provided the pitch of the helix is less than, say, half its diameter.

R 95: W. Elenbaas: Dissipation of heat by free convection, II (Philips Res. Rep. 3, 450-465, 1948, No. 6).

Continuation of **R 90**. For the contents of this article see these abstracts No. **1773***.

R 96: C. Zwikker: Anticaustics - A cord construction and a general formula (Philips Res. Rep. 3, 466-473, 1948, No. 6).

The author describes a method for constructing anticaustics — if the caustic is given — by generalizing the well-known cord construction for the ellipse. An analytical method of calculation of the anticaustics is developed by means of the geometry of the complex plane.

VOL. 11 No. 3, pp. 65-100 SEPTEMBER 1949

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

BETATRONS WITH AND WITHOUT IRON YOKE

by A. BIERMAN and H. A. OELE.

621.384.62:621.3.042

Many developments in the last 20 years in nuclear physics have been directed towards the accelerating of elementary particles to ever greater energy. High-tension generators, which at first constituted the only means of attaining this object, have now been surpassed, in regard to the energy attainable, by such apparatus as the cyclotron, the betatron and many others capable of accelerating particles to energies corresponding to tens of millions and even hundreds of millions of volts. Electron accelerators, such as the betatron, are also beginning to find practical applications in the fields of medical therapy and the testing of materials. At Eindhoven two types of betratons have been built for relatively low energies, for 5 and 9 million electron volts. In the second type of betraton, in contrast to all others known to have been so far employed, there is no closed iron circuit for the magnetic field, and this has made it possible to build an apparatus weighing no more than 50 kg (excluding the supply unit).

The electrons travel around in the annular accelerating tube of a betatron at a velocity practically equal to that of light. In the small apparatus referred to they make about 60,000 loops and travel a distance of almost 30 km in each acceleration period of 1/10,000th of a second.

The acceleration of electrons

The acceleration of electrons, such as required for the generation of X-rays or for nuclear research, for instance, is usually effected with the aid of electrostatic fields. An electron passing through a potential difference V acquires a kinetic energy equal to eV (e = the electron charge). It is from this method that the measure has been derived which is commonly used for expressing the kinetic energy of particles: 1 eV or 1 electron volt is the kinetic energy of an electron (or of a particle having the same charge) that has passed through a potential difference of 1 V.

In practice one cannot well work with greater potential differences than about 3 to 5 million volts. Higher tensions can, in principle, be generated but owing to the requirements of insulation (apart from other complications) the dimensions of the apparatus become impracticable. Now, however, it is possible to accelerate electrons without using an electrostatic field. An apparatus with which this can be done is the betatron, so named after the old name of beta rays for the electrons emitted by radium. The idea underlying the betatron is about 25 years old, but it did not begin to materialize until 10 years ago. An account of the historical development of this idea has been given by D. W. Kerst [1]). In various countries betatrons are now in use or in course of construction which are capable of supplying electrons with energies of several millions to several tens of millions of electron volts. The largest, built by the General Electric Company at Schenectady [2]), produces an energy of 100 MeV.

Needless to say, such an enormous expansion in the range of energies attainable is of great importance for fundamental research in nuclear physics. There are, however, also interesting practical applications for these powerful electrons or for the ultra-short-wave X-rays derived from them. In the treatment of deep-seated tumours the medical practitioner can obtain with these rays a distribution

[1]) D. W. Kerst, Historical development of the betatron, Nature (London) 157, 90-95, 1946.

[2]) W. F. Westendorp and E. E. Charlton, J. Appl. Phys. 16, 581-593, 1945.

of the dosage entirely different from that obtainable with the rays from normal X-ray therapy apparatus. Further, these highly penetrating rays can be used for the macroscopic examination of workpieces of thicknesses which could not be penetrated by the X-rays hitherto available.

In the course of investigations into the possibilities presented here for the construction of X-ray apparatus, Philips have built in their Eindhoven Laboratories, in addition to a betatron for 5 MeV constructed in the normal way, a betatron of a new kind. This apparatus, which produced X-rays in July 1948 for the first time, has been calculated for an energy of 9 MeV. Having an intermittent action, it is not as yet suitable for the practical applications mentioned, but the maximum intensity of the radiation is extremely high. Both of these types will be described in this article.

Principle of the betatron

Application of the law of induction

Around a varying magnetic flux Φ, according to the law of induction, there is an electric field of which the field strength F is determined in magnitude by the equation

$$\oint F\mathrm{d}s = \frac{\mathrm{d}\Phi}{\mathrm{d}t}, \quad \ldots \ldots \quad (1)$$

whilst the direction is shown in *fig. 1*. The integral can be taken along any arbitrary closed path around the flux.

A free electron in this field is subjected to a force eF and is thereby accelerated. Suppose, now, that the electron can be made to travel along an orbit encircling the changing flux. After the electron has made a complete loop, the force eF acting upon the electron will have performed the work

$$\oint eF\mathrm{d}s = e\,\frac{\mathrm{d}\Phi}{\mathrm{d}t}, \quad \ldots \ldots \quad (2)$$

that is to say, the energy of the electron, expressed in electronvolts, is increased by $\mathrm{d}\Phi/\mathrm{d}t$. When making another loop the electron gains further in energy by the same amount, so that it can accumulate a vast amount of energy by making a large number of loops.

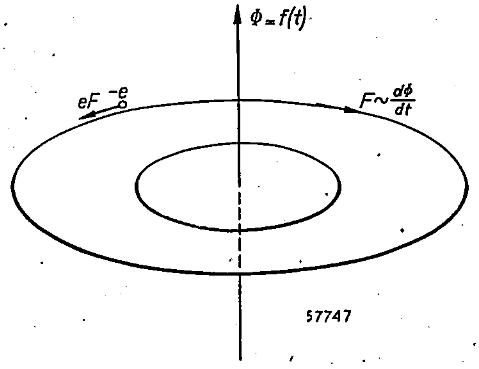


Fig. 1. Law of induction: round a changing magnetic flux Φ is an electric field of which the field strength F at any point is proportional to $\mathrm{d}\Phi/\mathrm{d}t$. The given direction F, applies for increasing Φ. (The force eF upon the electron acts in the opposite direction, because the charge e is negative.).

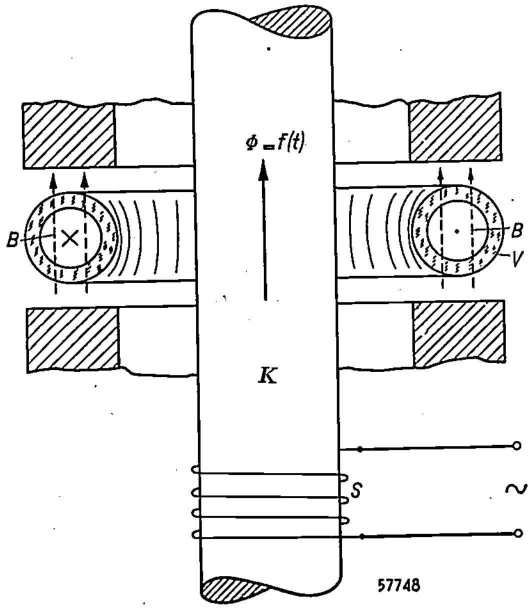


Fig. 2. Principle of the betatron. An alternating current flowing through the coil S produces a changing flux Φ in the iron core K. Around this core is a toroidal accelerating tube V. The electrons accelerated by the induction field are confined to a circular orbit in the toroid under the influence of a magnetic auxiliary field B which can for the time being be imagined as being excited by a separate magnetic circuit.

It is indeed possible to cause an electron to travel along a closed circular orbit around the flux Φ. This is done by applying along the whole of the path a magnetic auxiliary field B directed perpendicularly to the plane of the path (thus parallel to the flux Φ). This brings us to the fundamental arrangement of the betatron, as illustrated in *fig. 2*. The conditions which the magnetic auxiliary field has to satisfy and the manner in which it is generated will be dealt with farther on. First we shall discuss some other essential parts in the usual construction of the betatron, with reference to fig. 2.

To be able to circulate freely, the electrons must of course travel in a vacuum. The circular orbit along which the electrons are to travel is therefore enclosed in a toroidal and evacuated tube. Sealed into the tube is an electron gun supplying the necessary electrons. When the electrons are to be used for the purpose of producing X-rays a small metal target is fitted in the tube against which the accelerated electrons are caused to collide. Passing through the hole of the toroid is a core of soft iron in which the varying flux Φ is brought about by means of a coil energized with alternating current.

The electrons can only be accelerated by the electric induction field so long as the flux variation retains the same sign, i.e. at most during one half cycle of the alternating current producing the flux. As soon as the maximum flux has been reached and the sign of the flux variation is reversed the electrons have obtained their maximum energy; they would then be retarded owing to the reversal of direction of the electric field. It is so arranged, however, that at the moment the maximum flux is reached the electrons are deflected from their circular path and caused to collide with the target placed slightly to the side of the orbit. During the next half cycle, when the flux is changing to a maximum in the opposite direction, electrons could be accelerated in the reverse direction of travel, but this possibility is not as a rule utilized. It is not until the maximum flux is reached in the same direction that another group of accelerated electrons strikes against the target. Consequently the betatron has a pulsating action.

Velocity, mass, energy

The obvious question is what energy the electrons can ultimately obtain in the manner described.

For the case where the whole arrangement is rotationally symmetrical and thus the electrons travel in a circle around the alternating flux, it is obvious that the ultimate energy of the electrons depends only upon the total flux variation and not upon the manner in which the flux changes (thus not upon the form and frequency of the alternating current).

Let us consider the momentum mv of the electron (m = mass, v = velocity). According to the principal law of mechanics, force = the change of momentum per unit time. Hence

$$eF = \frac{\mathrm{d}(mv)}{\mathrm{d}t} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad (3)$$

Everywhere on the orbit along the circle (radius r) the magnitude of F is the same, and from (2) and (3) it follows that

$$2\pi r \, \frac{\mathrm{d}(mv)}{\mathrm{d}t} = e \frac{\mathrm{d}\Phi}{\mathrm{d}t}.$$

Hence the ultimate value of the momentum is

$$mv = \frac{e}{2\pi r} \, (\Phi - \Phi_0), \quad . \; . \; . \; . \; . \quad (4)$$

where Φ_0 is the flux at the moment that the electron begins to move ($mv = 0$). Given the momentum of a particle, its kinetic energy is also known, so that this energy likewise depends only upon the total flux variation $\Phi - \Phi_0$.

According to classical mechanics the kinetic energy $T = \frac{1}{2}\, mv^2$ and the relation between T and the impulse mv is $T = (mv)^2/2m$. For the betatron, however, we cannot apply these formulae because we have to do with velocities closely approaching the velocity of light (c). Account must therefore be taken of the relativistic increase of the mass, according to the well-known expression

$$m = \frac{m_0}{\sqrt{1-(v/c)^2}} \quad . \; . \; . \; . \; . \; . \quad (5)$$

(m_0 is the mass of the particle at rest). For the kinetic energy we must then write

$$T = mc^2 - m_0c^2\,, \quad . \; . \; . \; . \; . \quad (6a)$$

and the relation between T and the momentum mv becomes

$$T = \sqrt{(m_0c^2)^2 + c^2(mv)^2} - m_0c^2 . \, . \quad . \quad (6b)$$

For small velocities v these equations assume the form of the above-mentioned classical equations.

With the aid of equations (4) and (6b) the question as to what the ultimate energy of the electrons in the betatron will be can now be answered. Let us take for example a radius $r = 0.15$ m for the electron orbit and the reasonable value of 0.05 Vsec ($= 5 \cdot 10^6$ gauss cm^2) for the maximum flux increase [3]). Substituting the values

$$\begin{aligned} e &= 1.6 \times 10^{-19} \text{ coulomb,} \\ m &= 9.1 \times 10^{-31} \text{ kg,} \\ c &= 3 \times 10^8 \text{ m/sec} \end{aligned}$$

we get:

$$T \approx 15.5 \text{ million eV.}$$

The fact that with the betatron we come entirely within the sphere of the relativistic theory is most easily understood with reference to *fig. 3*. There the ratio v/c is plotted as a function of the kinetic energy T of the electron, whilst the quotient m/m_0 corresponding to the v/c ratio is indicated along the curve. Quite apart from all details of the mechanism of acceleration, we know that energies of several millions of eV are aimed at; from the graph we see that an energy of 1 MeV already corresponds to an electron velocity 0.94 times the velocity of light, owing to which the mass is increased by a factor 3.

[3]) Giorgi units are used in all formulae and calculations; a clear representation of the relation between this and other systems is to be found in Philips Technical Review **10**, 79-86, 1948 (No. 3).

As regards the formula (6*a*), readers not accustomed to working with the theory of relativity can best work this out for themselves by calculating the kinetic energy as the integral of force × path element, or

$$T = \int \frac{\mathrm{d}(mv)}{\mathrm{d}t}\,(v\mathrm{d}t) = \int v\mathrm{d}(mv).$$

Classically speaking, m is constant and one gets directly $T = \frac{1}{2}mv^2$. Relativistically, the expression (5) has to be substituted for m and after some calculation one arrives at (6*a*).

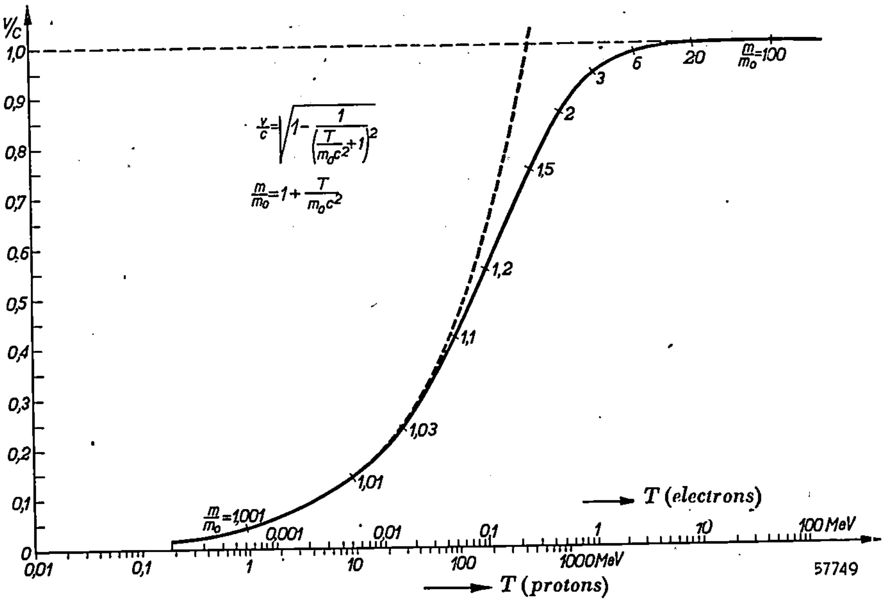


Fig. 3. The relation between the velocity v, the mass m and the kinetic energy T of an electron in motion. The quotient v/c (c = velocity of light) is plotted as a function of T (in millions of electron volts). Classically the dotted curve would apply, but at energies of the order of 1 MeV the velocity of the electron approaches the velocity of light, so so that the relativistic deviation from the classical relation is then very noticeable. The factor m/m_0, indicating the relativistic increase of the mass (m_0 = mass at rest) at high velocities, is shown along the curve. (For particles other than electrons the same graph holds, but with the logarithmic scale along the abscissa shifted. The scale for protons is also drawn in the diagram.)

To complete our picture of the motion of the electrons in the betatron we have to find the number of loops made by an electron during the short period of acceleration, and the total distance it thus travels. As we have already seen, the duration of the acceleration period depends upon the frequency of the alternating current producing the flux Φ. A usual value for the acceleration period is for instance 1/2000 sec. When dealing with large energies no great error is made when we take the velocity of an electron in the whole of the acceleration period as being equal to the velocity of light. Therefore the distance travelled by an electron in that period will be about 150 km. With a circular orbit having a radius of 15 cm this means that each electron makes well over 150,000 loops!

It is well to realize that in point of fact one can hardly speak of an "acceleration" of the electrons in this case. After the first two or three thousand loops there is practically no further increase in velocity, the energy $e\,\mathrm{d}\Phi/\mathrm{d}t$ imparted to the electron in each loop being manifested mainly in an increase of the mass m of the electron (cf. fig. 3).

The so-called "flux requirement"

When describing the principle of the betatron we mentioned the magnetic auxiliary field which forces the electrons to travel in a circular orbit. To be kept to a circular orbit with radius r an electron travelling at a velocity v must be subjected to a centripetal force mv^2/r. Presumably this force will have to be greater as the energy of the electrons increases. The magnetic auxiliary field mentioned, which may have a flux density (induction) B on the whole of the circular orbit, does indeed supply a centripetal force, the Lorentz force, having the value Bev. To keep the electrons to the circular orbit, B must vary with time such that at any moment $mv^2/r = Bev$, or

$$mv = Ber. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

From (4) and (7) it follows that

$$Ber = \frac{e}{2\pi r}\,(\Phi - \Phi_0)\,.$$

Let us assume that the initial flux $\Phi_0 = 0$. This means that the moment at which we let the electrons start (and at which, therefore, the field $B = 0$) is taken to be the zero point of the alternating current generating the flux Φ. Only one quarter cycle of the alternating current is then used for the acceleration. This is in fact the case in many constructions of betatrons [4]). The last equation given is then simplified and becomes

$$2\pi r^2 B = \Phi. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

This is the so-called "flux requirement", which plays an important part in the construction of the betatron. It shows that B must change proportionately with Φ and that therefore B and Φ can be generated by the same current.

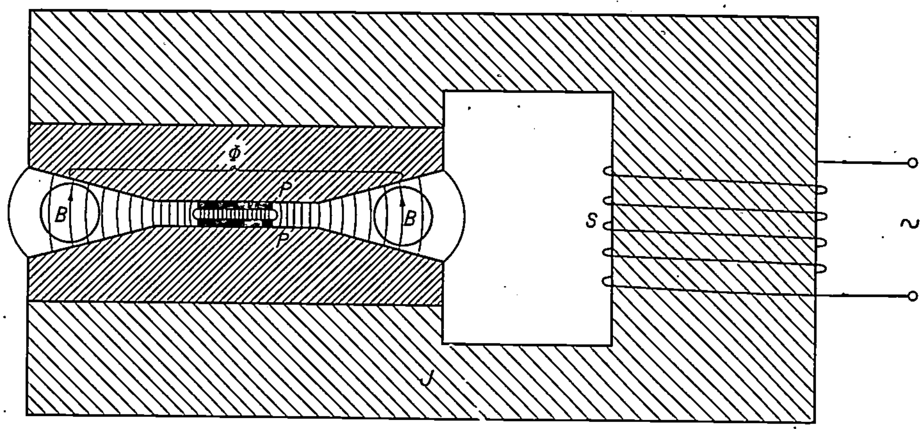


Fig. 4. According to the flux requirement (eq. (8)) the flux Φ and the auxiliary field B can be obtained with the same magnetic circuit. The iron core completed by the yoke J is then provided with an air gap in which the toroidal accelerating tube is placed. By a suitable choice of the shape and distance of the pole pieces P a particular radial slope of B is obtained, thereby satisfying not only the flux requirement but also the stability conditions.

We then come to an arrangement as sketched in *fig. 4*. Here there is only one magnetic field, in an air gap of the iron circuit. The inner part of this rotationally symmetrical field is enveloped by the toroidal acceleration tube placed in the air gap. The radial variation of the magnetic induction B must be such that where the desired electron orbit is to be situated, i.e. approximately on the centre circle of the toroid, the flux requirement (8) is satisfied. If the field within this circle were homogeneous and equal to B the enveloped flux would amount to $\pi r^2 B$. The flux requirement therefore implies that the induction inside the electron path must be greater than that on the path itself, on an average just twice as much. Consequently the iron core has to be provided with suitably shaped pole shoes.

Stability conditions

The electrons can only continue to travel round at that distance r from the centre that satisfies (8). There is, however, yet another requirement, namely that this circular orbit must be stable. This means to say that near to the orbit forces must exist to drive back the electron whenever it happens to leave the orbit through some cause or other.

Such stabilizing forces opposing both radial and axial deviations are obtained by causing the magnetic induction B in the vicinity of the orbit to decrease outwards in a certain manner. The relation between B and r, which within a small range can always be taken to be $B \sim 1/r^n$, must, near the orbit, be such that n lies between 0 and $+1$. Corresponding to this is a more or less "barrel-shaped" series of lines of force of the magnetic field as indicated in fig. 4. The desired variation in the magnetic field is obtained by giving the pole shoes a suitable shape. This need not conflict with the flux requirements, since there is an infinite number of different shapes of field to satisfy this condition.

The fact that the trenp variation described stabilizes the electron orbit in the centre plane of the air gap can be easily understood.

With a "barrel-shape" trend of the lines of force the induction B comprises, in addition to the axial component (B_z), also a radial component (B_r) which, proceeding in the direction of B, is directed outwards in front of the centre plane

[4]) It has certain advantages to take $\Phi_0 \neq 0$; we cannot go into this here and shall keep to eq. (8).

and inwards beyond that plane (that is to say $\partial B_r/\partial z < 0$, when the r-direction outwards and the z-direction in the direction of the field — in this case upwards — is taken to be positive). An electron travelling in a circle below the centre plane is subjected, through the radial component B_r, to an additional Lorentz force directed upwards, as can easily be deduced from the known rule about the direction of Lorentz forces. When the electron circles above the centre plane, a similar Lorentz force is directed downwards through the component B_r above the centre plane. Thus in both cases the electron is driven towards the centre plane, so that in that plane the orbit is stable with respect to axial deviations.

For a rotationally symmetrical magnetic field we have $\partial B_r/\partial z = \partial B_z/\partial r$. (This follows from the Maxwell equations.) In order to get the barrel-shaped field variation required for axial stability, for which $\partial B_r/\partial z < 0$, we must therefore ensure that $\partial B_z/\partial r < 0$, which means that in the centre plane the induction B must decrease outwards: n must be positive when B is taken to be proportional to r^{-n}.

Let us now consider the radial stability. On the desired orbit the Lorentz force Bev is equal to the necessary centripetal force mv^2/r, thus $Ber = mv$. If through some cause or other the electron travelling at a velocity v comes onto an orbit having a greater radius, $r_1 > r$, then it will be driven back to the path r if B_1ev on the greater circle $> mv^2/r_1$, or $B_1er_1 > mv$, thus $B_1r_1 > Br$. Similarily, on too small an orbit having a radius $r_2 < r$, B_2r_2 must be smaller than Br. The circular orbit with radius r is thus stable with respect to radial deviations if $B{\cdot}r$ increases with increasing r, or, if we introduce $B \sim r^{-n}$, if $1-n > 0$. Thus the two stability conditions $0 < n < +1$ mentioned above have been derived.

Some further details in the construction of a betatron, e.g. the injection of the electrons in the tube, will be mentioned farther on when describing the Eindhoven betatrons; for many other details the reader is referred to the literature on the subject [5]). Only two incidental remarks will be made in conclusion of this introduction:

1) In literature we find the betatron also described under the names of "induction-accelerator" or "ray transformer" (R. Wideröe). The last name is due to the fact that the arrangement (see fig. 4 and fig. 5 below) can be compared to a transformer; the secondary winding is replaced, so to speak, by a ray of electrons making a large number of loops around the core and thereby being given an energy (to be measured in volts) equal to the voltage between the extremities of the secondary winding having an equal number of turns.

2) Let us go back for a moment to the calculation of the ultimate energy of the electron. In the foregoing we derived the momentum from the total flux variation (equation (4)), but from eq. (7) we see that the ultimate value of the momentum of the electrons can also be denoted by the product Br, where B is the strength of the directing field at the end of the acceleration. Thus the ultimate energy can likewise be expressed by Br; by substituting (7) in (6) we get the formula.

$$T \text{ (in electron volts)} = \sqrt{\left(\frac{m_0c^2}{e}\right)^2 + c^2(Br)^2} - \frac{m_0c^2}{e}. \quad (9)$$

This formula can also be used without any relation to the betatron, and also for other particles, for instance protons, which have a mass m_0 1837 times greater than that of the electron (cf. fig. 10). B then has the general meaning of the magnetic induction required to force the moving particle of energy T into a circular orbit with radius r. The manner in which the particle gets its energy T is then immaterial. In nuclear physics it is in fact a common practice to measure the energy of particles by the product Br, by making their orbit visible in a Wilson cloud chamber and determining the radius of curvature of the orbit when the chamber is placed in a magnetic field of a known strength.

Design of a betatron with iron yoke

The magnetic circuit

Fig. 5 is an illustration of the betatron built by us in the usual way. The magnetic circuit has been completed by the addition of a yoke to the iron core with the pole pieces. In this circuit a flux is generated by two coils connected in series and placed around the core above and below the accelerating tube. The photograph in *fig. 6b* shows the peculiar shape given to the pole shoes in order to satisfy the flux requirement and the stability conditions.

To keep the apparatus within reasonable dimensions the diameter of the electron orbit has been made only 14 cm. The maximum value of the flux within this orbit is limited by the saturation of the iron. In our case this flux is 0.008 Vsec. The final energy of the electrons is thus 5 MeV.

The excitation current used has a rather high frequency, viz. 500 c/s. As we have seen, the choice of frequency does not affect the final energy with a given maximum flux, but a high frequency gives a rapid flux variation (large value of $\mathrm{d}\Phi/\mathrm{d}t$) and thus a large energy gain per loop of the electron. This has two advantages. The greater the energy gain per loop, the smaller is the number of loops and the shorter the distance that each electron has to travel until it has reached the ultimate energy; thus there is less chance of its colliding with residual gas molecules remaining in the evacuated tube. The second advantage is connected with the stabilization. When an electron does not follow precisely the equilibrium orbit and is driven back to it by the stabilizing forces it will as a rule follow a damped oscillatory movement around the equilibrium orbit. This oscillation, which may in various ways result in the electrons not participating in the acceleration process right through to the end, is damped more quickly the greater the energy gain per loop [6]).

[5]) See for instance D. W. Kerst and R. Serber, Electronic orbits in the induction accelerator, Phys. Rev. 60, 53-58, 1941.
W. Bosley, The betatron, J. sci. Instr. 23, 277-283, 1946.
R. Wideröe, Der Strahlentransformator, Schweizer Archiv 13, 225-232, 299-311, 1947.

[6]) See the article by Kerst and Serber quoted in footnote [5]).

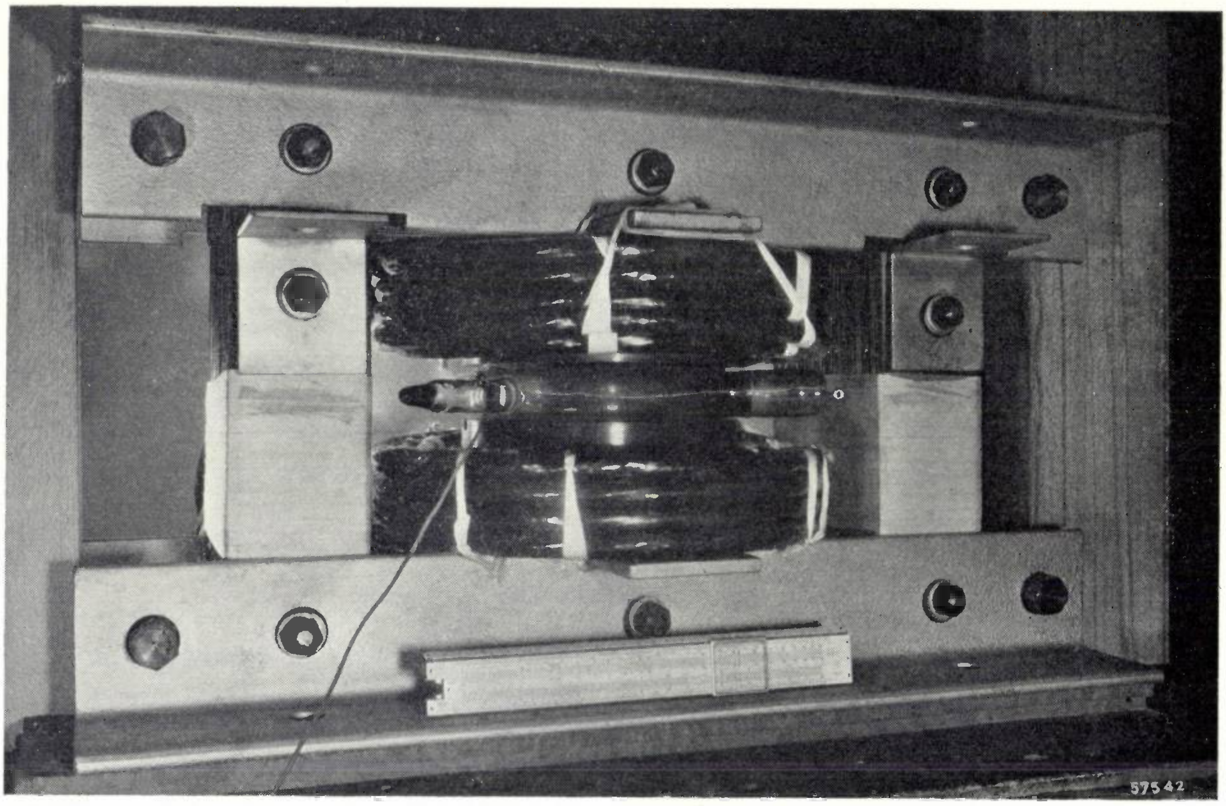


Fig. 5. Betatron with iron yoke built in the Philips laboratory at Eindhoven. The heavy excitation coils (reactive current 800 A) are seen above and below the flat, toroidal, sealed-off accelerating tube. The elements sealed into the tube, one of which is seen on the right and another on the left, contain, inter alia, the stems and wires for the electron gun and the target.

Fig. 6. *a*) The betatron illustrated in fig. 5 with the upper part of the yoke and the upper coil and pole piece removed. (For practical reasons each coil was composed of four parallel-connected parts; this explains the large number of cables seen at the back.)
b) Here the accelerating tube has also been removed, showing the peculiar shape and construction of the pole piece.

One cannot very well choose a frequency much higher than 500 c/s, because then difficulties arise in carrying off the heat developed in the iron circuit owing to the eddy current losses, which increase with the square of the frequency. In order to minimize these losses the core and the yoke are made of transformer sheet iron 0.35 mm thick. Moreover, eddy currents in the pole shoes adversely influence the magnetic field in the accelerating tube, and therefore the pole shoes are made of extra thin (0.12 mm) and mutually well-insulated laminations. To get the best possible axial symmetry, which is of importance for stabilization of the electron orbit, the pole shoes are built up in sectors made of stampings of different length; this can be seen in fig. 6*b*.

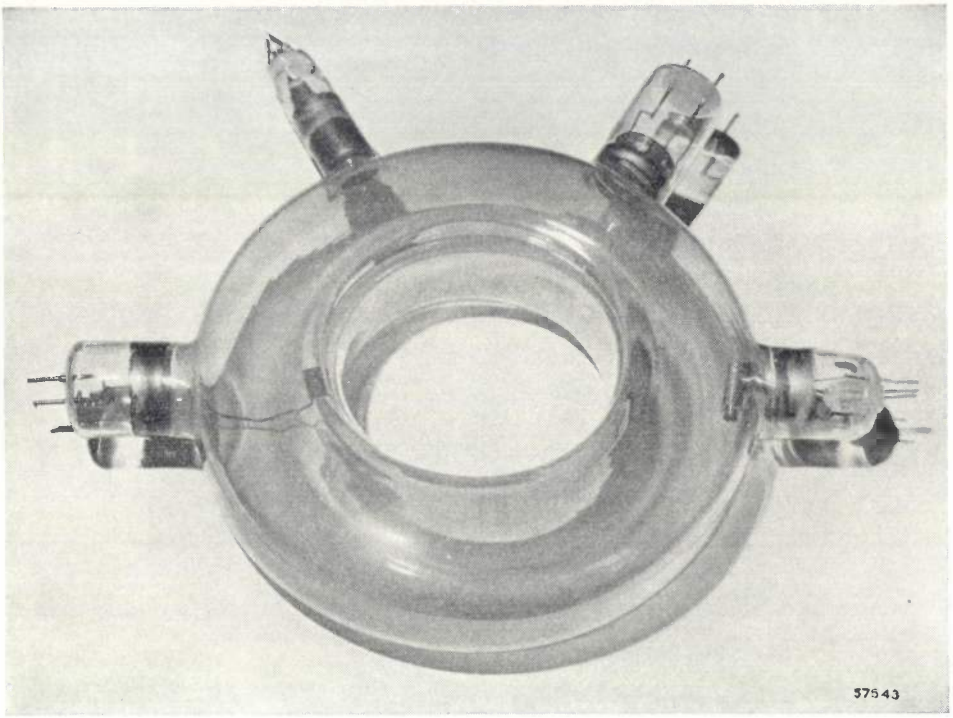


Fig. 7. Accelerating tube of the betatron. On the right, outside the centre circle of the toroid, is the electron gun injecting the electrons into the accelerating space in a direction approximately tangential to the circular orbit of the electrons. On the left, inside the centre circle, is the target against which the accelerated electrons are directed at the end of the acceleration period. The inner wall of the tube is covered with a transparent conductive layer

The accelerating tube

The toroidal glass accelerating tube is shown separately in *fig.* 7. The electron gun injecting the electrons into the accelerating field is seen on the right; it consists of a cathode, a Wehnelt cylinder for focusing the electron beam, and a cylindrical anode entirely enveloping the other electrodes, with the exception of the opening for the emergence of the beam. The anode voltage is a few kilovolts.

Injection has to take place in that part of the space where the magnetic induction answers the requirements for stability. Consequently the gun cannot be placed far outside the circular orbit of the electrons. To minimize the risk of electrons striking the back of the gun while oscillating around the stable orbit, the gun has been made as small as technically possible.

The inner wall of the tube is covered with a slightly conductive layer which is given the same potential as the anode of the injector. Thus the electrons move in a space free of electrostatic fields, which might possibly disturb the orbits. This conducting layer also carries off immediately any electrons which might reach the wall of the tube, so that no disturbing wall charges can arise. In our tubes this layer consists of a transparent semi-conductor, this making it possible to check the position of the component parts when they are being sealed into the tube. Incidentally a very useful property of the semi-conductor is the fact that it fluoresces when electrons strike it. In this way the behaviour of the electron beam can be studied, thereby facilitating adjustment of the apparatus.

The target against which the electrons are caused to collide after their acceleration is made of tungsten. As seen in fig. 7, it is mounted within the stable orbit. The electrons are directed towards the target by ensuring that at the end of the accelerating period the central part of the iron core (where the induction is greatest) is saturated. At that moment the directing field B on the electron orbit is still increasing proportionately with the magnetizing current, but the flux Φ inside the orbit with radius r rises less quickly. Consequently the flux requirement (8) is no longer satisfied and the centripetal Lorentz force predominates, the electrons then following a gradually constricting spiral course until they strike the target. This target is also connected to the wall of the tube to keep the space free of electric fields.

The electric supply

Current and voltage of the excitation coils of the betatron are shifted almost 90° in phase. The 220 V dynamo supplying thc excitation current with a frequency of 500 c/s is relieved of the extremely strong reactive current component (about 800 amperes!) by connecting in parallel to the coils a condenser of about 1000 μF, which together with the coil forms a resonance circuit. Thus the energy of the magnetic field is periodically stored and returned by the condenser, whilst the dynamo only has to supply the current required to compensate the losses. Thanks to the most careful lamination the losses are restricted to about 5 kW.

The total energy taken up by the electrons during their acceleration is only an extremely small fraction of this 5 kW, so that the "efficiency" — if this term were to be used here — is very small.

An unpleasant feature is the penetrating whistling sound of 1000 c/s produced by the iron of the betatron (mainly due to the phenomenon of magnetostriction). In the case of the small betatron described here this sound does not constitute an enormous problem, but with the large betatron for 100 MeV mentioned above it is an almost unbearable noise (120 db above the auditory threshold; see the article quoted in footnote [2])).

A betatron without iron yoke

A serious drawback about a betatron built according to the usual construction described above is the large amount of iron in the core and yoke of the magnetic circuit, which makes the apparatus almost unmanageable. When the betatron is to be used as an X-ray apparatus this is a great objection, since a source of radiation is required which is adjustable in all directions. Furthermore, the construction of the laminated iron circuit is very expensive, particularly owing to the complicated pole shoes, which have to be built up from thousands of thin laminations (about 8000 in the apparatus described) and thus cost a great deal of time and call for the utmost precision.

It is not, however, strictly necessary to use iron. The requirements for a betatron can also be met with properly dimensioned air coils. A much stronger current will of course be required to get the same magnetic induction as is obtained when iron is used, but this stronger current can easily be obtained by discharging via the air coils a condenser charged to a high voltage.

This idea, which in the meantime has also been suggested by several other investigators, has been worked out by us and has led to the development of a type of betatron which in many respects differs entirely from the usual constructions [7]).

The magnetic field

Fig. 8 shows the betatron materialized according to the above-mentioned idea. It has two coils connected in series. The field in the accelerating tube, which is placed between the coils, is given the desired variation by a suitable choice of the dimensions and mutual distance of the coils. In the centre an additional flux is needed to satisfy the flux variation. This flux could be obtained from a central coil in series with the other two coils, but, as a closer consideration shows, such a flux coil would take too much energy. The necessary central flux is therefore provided by means of a small iron core mounted in the axis of the coil system (not visible in the photograph). The core is interrupted midway so that the flux can be adjusted to the required value by varying the resultant air gap. The addition of the iron core, weighing only 5 kg, hardly affects the desired limitation of the weight of the whole apparatus. It has, moreover, the advantage that the orbit-contraction at the end of the accelerating period can be brought about in the same simple way as in the case of the betatron with closed iron circuit: the iron is saturated and

[7]) A. Bierman, A new type of betatron without an iron yoke, Nature (London), **163**, 649-650, 1949.
An earlier form of this kind of betatron, which proved unsuccessful at the time, has been described by E. T. S. Walton, The production of high-speed electrons by indirect means, Proc. Cambr. Phil. Soc. **25**, 469-481, 1929. Walton started from the so-called electrode-less annular discharge, which is applied in some spectroscopic investigations.

the increase of flux in the centre becomes relatively too small.

The contribution of the stray field of the air gap in the core is such that the stability of the electron orbit in the acceleration space is thereby improved.

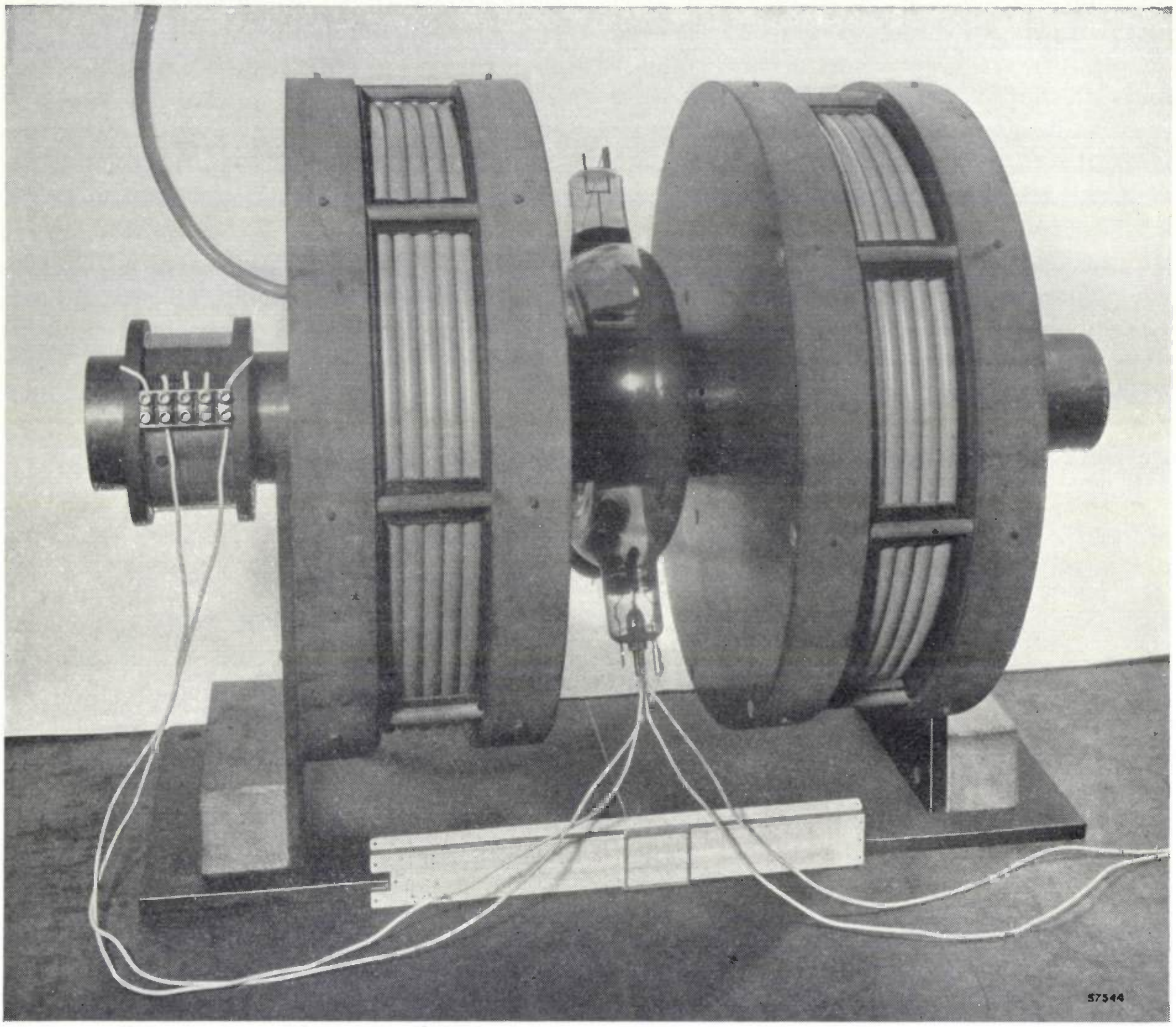


Fig. 8. The betatron with air coils as built in the Philips Laboratory at Eindhoven. The two coils are mounted in strong frames on a hard-paper cylinder. In the axis of this cylinder (thus not visible) is a small iron core. In the accelerating tube, fitting onto the hard-paper cylinder, the target can be seen at the top and the electron gun at the bottom. The anode voltage for the electron gun can be taken from the small coil on the extreme left, or from one of its tappings.

Each coil consists of 25 turns of a high-tension cable (with a copper area of over 20 mm^2). The peak value of the current passing through it is well over 5000 A, as we shall presently see. The flux thereby generated inside the electron orbit with radius 8 cm amounts to about 0.016 Vsec (maximum induction on the orbit B = 0.40 Vsec/m^2 or 4000 gauss; maximum induction in the centre about 2.0 Vsec/m^2 or 20,000 gauss). From these data the ultimate energy of the electrons is calculated to be about 9 million volts.

The radius chosen for the stable orbit is practically the same as that chosen for the conventional betatron first described. In other respects, too, the acceleration tubes are essentially identical for both betatrons. (The tube illustrated in fig. 7 is in fact that used in the betatron without yoke.) One point of difference will be dealt with below.

The electric supply

The coil system is connected to a condenser via a spark gap; see *fig. 9*. This condenser is charged from a high-tension supply unit. As soon as a certain tension is reached the spark-over takes place and the condenser discharges via the coils. By charging the condenser continuously this process can be repeated periodically at short intervals (see below); a single discharge can also be effected at a desired moment.

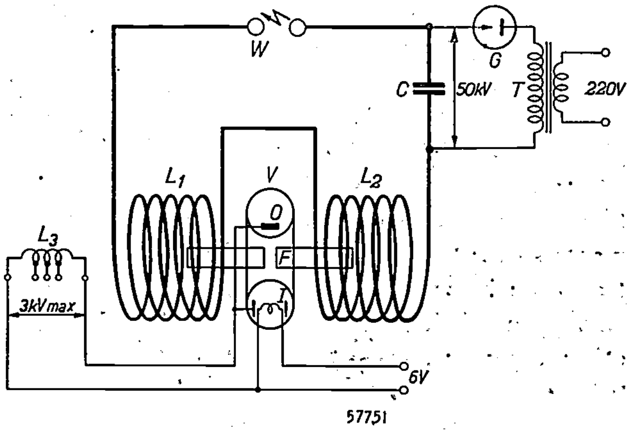


Fig. 9. Circuit diagram of the air-coil betatron. C = condenser battery charged up to 50 kV by the high-tension transformer T and the valve G. When a spark-over takes place in the spark gap W, C discharges via the two coils L_1 and L_2 of the betatron. F = iron core with an air gap; V = accelerating tube with electron gun I and target O; L_3 = coil supplying the anode voltage for the electron gun.

The discharge current generating the magnetic field has the form of a strongly damped oscillation. Since this is a free oscillation, its frequency is the natural frequency of the circuit formed by the condenser and the coils. The self-inductance of the coils is $L = 625$ μH, the capacitance of the condenser $C = 6.5$ μF and thus the frequency $f = 2500$ c/s. The betatron "works" only during the first few cycles of the damped oscillation. Electrons are of course accelerated to a fairly high energy also in the following 10 or 20 cycles, but after the first few cycles the magnetic field strength in the iron core remains below the minimum required for contracting the orbit of the electrons to the target by the saturation effect.

The maximum current I in the coils can be calculated by taking the maximum energy $\frac{1}{2}LI^2$ of the magnetic field to be equal to the energy $\frac{1}{2}CV^2$ of the charged condenser (disregarding losses); thus $I = V\sqrt{C/L}$. Taking $V = 50{,}000$ V, with the above-mentioned values of C and L, we get $I = 5100$ A.

At the beginning of the damped oscillation the condenser therefore yields a reactive power of about 250,000 kVA. The losses in the dielectric of the condenser and the copper losses in the coils amount to some thousands of kilowatts. Compared with these, the eddy current losses in the iron core are negligible, and the core is therefore made of normal laminations 0.35 mm thick. Because of the enormous power required it would not be practicable to work the betatron continuously by connecting the *L-C* circuit to an A.C. generator supplying the exact frequency. Furthermore, the small apparatus of the present design needs at least one second after each discharge to dissipate the heat generated.

The anode voltage for the electron gun can easily be obtained by placing a small coil in the field of the large ones (see figs 8 and 9). The voltage induced in the auxiliary coil reaches its maximum when the magnetic field passes through zero, and just about at that moment the injection has to take place, namely when the induction is at the low value which according to equation (7) corresponds to the radius of the orbit and the relatively small injection energy (a few keV). Compared with a constant anode voltage supply, this method has, moreover, the advantage that the gun operates only during the short intervals at which the current is flowing through the main coils. Consequently there is no unnecessary heating of the glass wall of the accelerating tube at the spot, opposite the mouth of the gun, against which the electron beam strikes if no magnetic field is present.

It is even more advantageous to feed the gun with very short pulses at the required moments. All electrons injected then have a reasonable chance of being "captured" by the accelerating field and brought into the stable orbit; this reduces disturbing space charge and local magnetic fields of electrons that have not been captured. With this injection method, with which we have already experimented, the apparatus becomes, of course, somewhat more complicated.

Comparison of the two betatrons

The air-coil betatron [8]) has a total weight of about 50 kg. In addition to the coils and the core, it is particularly the coil frame that counts in this weight. This coil frame had to be made rather heavy because at the maximum current an attractive force of more than 10,000 newton (about 1000 kg_{force}, since 1 newton = 1/9.81 kg_{force}) occurs between the two coils. Also the separate windings are subjected to strong forces, about 5000 newton acting upon the outermost turns.

With this weight of 50 kg the air-coil betatron

[8]) We can still speak of air coils here, although the use of iron is not entirely avoided.

compares very favourably with the conventional construction. For the same ultimate energy of the electrons a betatron with iron circuit would weigh at least ten times as much. Even the conventional betatron described above, which is calculated for an ultimate energy of 5 MeV, weighs much more than our air-coil betatron, namely 270 kg.

In addition to the ultimate energy, various other data are of importance, some of which are given in the table below.

Particularly striking is the high gain in energy per loop in the case of the air-coil betatron owing to the exceptionally high frequency. This accounts for the very much shorter distance the electrons have to travel. It has already been pointed out that owing to this effect and to the stronger damping of the oscillatory movement of the electrons, the high frequency favourably influences the number of electrons accelerated to the end of the acceleration period.

There are two other items making the air-coil betatron better in this respect than the betatron with iron yoke. In the first place, with the air-coil construction a perfect axial symmetry can be obtained, this being desirable for proper stabilization of the electron orbits. It is only necessary, for this purpose, to see that the two coils are regularly wound and properly centred. In the case of the iron betatron axial symmetry is limited by the lamination of the pole pieces.

Secondly, in the iron betatron the toroidal accelerating tube has to be rather flat, because otherwise the air gap in the magnetic circuit would be too large. This is not the case with the air-coil betatron, the tube being in fact twice as thick as that in the iron betatron. As a consequence the space within which the electron orbit can make an oscillatory movement without electrons being lost through collision with the wall is twice as large. To take full advantage of this it has to be ensured that the ratio of the stabilizing forces in the radial direction to those in the axial direction is adapted to the freedom of the electrons to deviate in these two directions. The power n for the field variation ($\sim r^{-n}$, see above), upon which this ratio depends, has been chosen $^1/_2$ for the air-coil betatron and $^3/_4$ for the iron betatron.

Because of the large number of electrons fully participating in the acceleration in the air-coil betatron, an X-ray radiation of high peak intensity could be expected. No definitive figures can yet be given for the intensities obtainable with our two betatrons, because experiments in this direction have not yet been completed. However, measurements so far taken indicate that the air-coil betatron does indeed generate (for the same ultimate energy) a much greater peak intensity than the other apparatus.

This great intensity is, it is true, only obtained during very short periods, so that in its present form the apparatus is not suitable for the applications mentioned above in the fields of therapy and the testing of materials, as already observed in the introduction. There are cases, however, where such a limitation to short radiation pulses does not constitute an objection. We have in mind for instance the use of this apparatus as a source of radiation for experiments in nuclear physics with the Wilson cloud chamber. For such work the air-coil betatron described should certainly merit preference over the expensive and cumbersome type of betatron with iron yoke.

Finally, it is to be pointed out that with the air-coil betatron it is a very simple matter to vary the ultimate energy of the electrons. If a smaller energy is required it is only necessary to replace

Type of betatron	Ultimate energy of the electrons *) MeV	Frequency of excitation current c/s	Duration of operation	Energy gained per ($d\Phi/dt$) loop **) eV	Number of loops to attain ultimate energy	Total distance travelled by an electron ***) km	Weight kg
With iron yoke	5	500	each cycle	25	330,000	150	270
With air coils	9	2500	a few cycles at intervals of abt. 1 sec	240	60,000	30	50

*) Value derived from B and r.

**) At the beginning of the acceleration.

***) This follows approximately from $c/(4 \times \text{frequency})$, since during the greater part of the acceleration the electrons have practically the velocity of light c.

the iron core by a thinner one, which is sooner saturated. Contraction of the orbit then takes place earlier, that is to say the acceleration process ends at an earlier point of the cycle and thus the ultimate energy is smaller. Although the excitation current has the form of a damped oscillation, the ultimate energy is the same for all active periods of this oscillation, since the saturation required always takes place at the same field strength. In the second cycle this saturation and, with it, the end of the acceleration process occurs somewhat later than in the first cycle, in the third one still later, until in the last active cycle the current amplitude has dropped so far that the contraction takes place roughly at the peak of the current curve. From this it follows that the smaller the ultimate energy is chosen (thinner core saturated with smaller current) the larger will be the number of active cycles.

Appendix: Betatron and cyclotron

The question arises whether the betatron can also be used for accelerating particles other than electrons, for instance for the 1837 times heavier, positively charged protons. In principle this is indeed so, but this would be putting the cart before the horse. For the acceleration of protons (and other heavy particles) to energies greater than can be attained with high-tension generators there is a much older instrument than the betatron, namely the cyclotron, which is much more suitable for this purpose. Historically speaking, the development of the betatron is in fact due mainly to the cyclotron not being suitable for the acceleration of electrons. The reason for this, and why on the other hand the betatron is less suitable for heavy particles, will be briefly explained.

In the cyclotron the particles are accelerated by a relatively small, alternating potential difference (20 to 200 kV), the particles being subjected to this accelerating action several times in succession, because a constant magnetic field B on a circular orbit forces them back to the accelerating field at the right moment.

With each gain in energy the radius r of the path becomes somewhat larger. A condition is that the angular velocity v/r of the particles on their circular orbits must be practically constant. According to the general formula (7) the angular velocity is determined by

$$\frac{v}{r} = \frac{Be}{m}.$$

B and e are constant, but the mass m increases with the velocity of the particle according to eq. (5). This relativistic change of the mass begins to upset things, in the case of the cyclotron, as soon as it reaches a value of about 1%, that is to say when the velocity v becomes equal to 0.15 times the velocity of light.

The energy corresponding to this velocity is ouly a few thousand eV for an electron (see fig. 3). Since such energies can much more easily be obtained with a normal high-tension generator, there is no purpose in using a cyclotron for accelerating electrons. For a proton on the other hand the energy corresponding to the said velocity is 10 MeV (fig. 3). Therefore the cyclotron is quite suitable for accelerating protons, and still better for deuterons and other heavier particles where greater energies can be reached before the relativistic change in mass interferes [9]).

In the betatron the particles likewise travel along circular orbits, as we have already seen, but here r is constant and the induction increases (maximum value B). With electrons reasonably high energies are reached with moderate values of the product Br, but with protons much greater values of Br would be needed. This can be seen at once from *fig. 10*, where the kinetic energy T according to equation (9) is plotted as a function of Br for electrons and for protons. With a field B having the peak value 0.4 Vsec/m^2 a proton energy of 10 MeV requires an orbit radius r of at least 1.1 m. When magnets of such dimensions have to be employed the cyclotron is much more economical, 1) because one can work with higher values of B on the orbit (one is not restricted by the flux condition); 2) because B is constant, so that no eddy currents arise in the iron and lamination is not necessary;

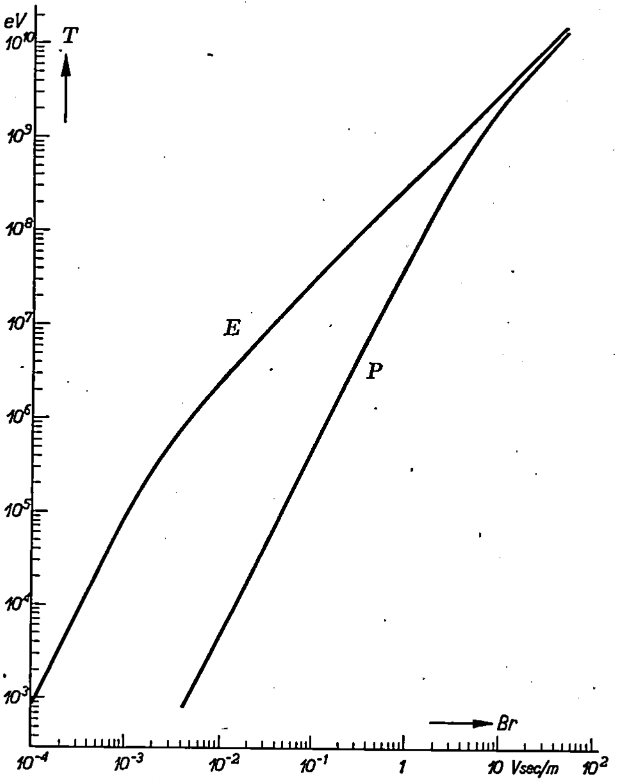


Fig. 10. A moving particle with energy T is kept to a circular orbit with radius r by a magnetic field with the inductance B. The product $B{\cdot}r$ is a measure for the energy when the mass and the charge of the particle are known. T is plotted (in electron volts) as a function of Br (in Vsec/m) for electrons (E) and for protons (P), according to eq. (9). Since for electrons the "rest energy" m_0c^2/e in MeV amounts to 0.51, the equation for electrons is

$$T_{\mathrm{MeV}} = \sqrt{(0.51)^2 + 9{\cdot}10^4\,(Br)^2} - 0.51.$$

For the proton the rest energy is 940 MeV. The two curves approach the asymptote with the equation $T = 300\ Br$ (on a non-logarithmic scale, to the two asymptotes $T = 300\ Br - 0.51$ and $T = 300\ Br - 940$ respectively).

[9]) A modern variation of the cyclotron which is suitable both for protons and for electrons and is capable of generating much higher energies than 10 MeV, namely the synchrotron, must for the sake of brevity be passed over here.

3) because a continuous current of accelerated particles is obtained, whereas the betatron, since it is based upon a changing flux, must always have a pulsating action.

The fact that the relativistic increase of mass, which makes the cyclotron unsuitable for electrons, does not interfere with the working of the betatron is due to this change in mass having the same influence upon the tangential acceleration as upon the radial acceleration. In the combination of the equations (4) and (7) to the ultimate formula (8) the mass is thereby eliminated. The maximum energy that can be given to electrons in a betatron is in fact determined by an entirely different "classical" effect: the electromagnetic radiation emitted by the electron while it is travelling round [10]). This limit, however, lies so high that even greater electron energies can be reached than the 100 meV hitherto realized.

[10]) D. Iwanenko and I. Pomeranchuk, On the maximum energy attainable in a betatron, Phys. Rev. **65**, 343, 1944.

Summary. After an explanation of the fundamental principles of the acceleration in the betatron and the behaviour of the electrons in that apparatus, this article gives a description of two betatrons built in the Philips Laboratory at Eindhoven. The first has been constructed on more or less conventional lines; the required magnetic flux and the magnetic directing field which must exist along the circular orbit of the electrons, are generated in an air-gap of an iron circuit excited with an alternating current of 500 c/s and provided with carefully laminated pole pieces of a special shape. This apparatus weighs 270 kg and generates in 500 short pulses per second electrons which have an energy of 5 million electron volts. In the second apparatus the magnetic field is obtained by means of air coils through which the discharge current from a condenser battery flows. Only a small iron core is introduced in the coils. Thanks to this core it has been easy to satisfy the so-called flux requirement and, owing to the saturation of the iron, it has also been possible to obtain in a simple manner the desired contraction of the electron orbit at the end of each acceleration period. The discharge has the form of a strongly damped oscillation with a frequency of 2500 c/s, whilst the reactive power of the first oscillations amounts to 250,000 kVA. This apparatus has been constructed for supplying electrons with an energy of 9 million electron volts in a few very short pulses at the beginning of the discharge, which can be repeated at intervals of about 1 second. Provisional measurements indicate that the number of electrons fully participating in the acceleration process is much greater in this apparatus than in the betatron with iron yoke. This is due, among other things, to the high energy gain per loop of the electron (240 eV), limiting the total distance to be travelled by the electrons to 30 km (60,000 loops in the toroidal accelerating tube). The air-coil betatron weighs only 50 kg.

TWO TRIODES FOR RECEPTION OF DECIMETRIC WAVES

by K. RODENHUIS.

621.396.694.029.63:621.385.33

Decimetric waves are very suitable for many forms of radio communication over limited distances, such as private radio links (for which a frequency band has already been made available in Great Britain and the U.S.A.), telephonic communication with motorcars or with tugs in a port, sound and/or picture transmission between a studio and a transmitter, telemeter installations and finally a number of military uses. Moreover, the metric waves are already showing signs of overcrowding with the ever-increasing number of broadcasting stations working on a wide frequency band (television, broadcasting with frequency modulation), so that in all probability shorter waves will have to be used for these services. That is why attention is being centered upon decimetric waves.

Normal radio valves are not suitable for the reception of these decimetric waves (frequencies higher than 300 Mc/s). In this article two valves specially designed for these ultra-high frequencies are described, namely a triode for high-frequency amplifying and mixing and an oscillator valve. In contrast to the present type of valves for this frequency range they have the appearance of conventional valves. Thus with these new valves a receiver for decimetric waves can be built which is simple in construction and easy to operate.

The shortcomings of normal radio valves at ultra-high frequencies

Most decimetric-wave receivers are nowadays built on the superheterodyne principle: the signal received, possibly after being amplified, is fed into a mixer valve together with the voltage generated by a local oscillator. Owing to its non-linear characteristic, the mixer valve produces a large number of combinations of frequencies, one of which (generally the difference between the frequency received and the locally generated frequency) is further amplified in a so-called intermediate-frequency amplifier. As a rule the intermediate frequency does not exceed 60 Mc/s. No special valves are required for amplifying such a frequency, but for the reception of decimetric waves special valves are indeed required for the following functions:

a) amplification of the aerial signal,
b) local generation of oscillations,
c) mixing of the two voltages.

At frequencies higher than 300 Mc/s (1 m wavelength) normal receiving valves no longer give any amplification. Neither can they be made to oscillate at these frequencies, at least not in a reliable way. Mixing is still possible, but the conversion "amplification" is less than 1 (that is to say, the intermediate frequency signal obtained is smaller than the high-frequency signal). Moreover, at ultra-high frequencies an important advantage of the normal mixer valve (for which a hexode is usually chosen at lower frequencies) is lost, namely, prevention of the energy from the local oscillator reaching the aerial and being radiated, thereby interfering with neighbouring receivers.

These shortcomings of normal radio valves at ultra-high frequencies can be ascribed to a number of phenomena which can be grouped under three headings.

In the first place there are the transit-time effects, which become of importance at such high frequencies that the oscillation period no longer extends over a sufficient interval of time compared with that required by the electrons to traverse the path from cathode to anode. These effects manifest themselves in a certain form of damping and a reduction of the mutual conductance [1]).

A second group of phenomena comprises the undesired couplings, the cause of which is to be found in the self-inductance and the mutual inductance of the electrode connections.

The third group consists of losses which increase with the frequency f: firstly the dielectric losses (in the glass and possibly in a plastic base), which increase roughly in proportion to f, and secondly the dissipative losses in the contact and lead pins and in the electrodes themselves. The last-mentioned losses are roughly proportional to $f^{5/2}$ and as f rises they soon become much more important than the dielectric losses.

[1]) See for instance C. J. Bakker, Some characteristics of receiving valves in short-wave reception, Philips Techn. Rev. 1, 171-177, 1936, and M. J. O. Strutt and A. van der Ziel, The behaviour of amplifier valves at very high frequencies, Philips Techn Rev. 3, 103-111, 1938.

Nearly all these phenomena have the tendency to reduce the amplification as the frequency becomes higher. Ultimately a frequency is reached where the amplification is so small that the increase in the signal strength is less than the increase of the noise contributed by the amplifying stage. Thus the highest frequency at which an amplifying valve can be used to advantage also depends upon the noise properties.

There has been no lack of attempts to make valves suitable for higher frequencies. Important advances in the development that has resulted therefrom have been mentioned in this journal. Mention may be made of the change-over from the glass "pinch" to a flat glass base with lead pins acting also as contact pins [2]). With this construction there is no longer any need of a base of insulating material, so that one source of dielectric losses has been eliminated. Moreover with this construction the connections between the electrodes and the contact pins are much shorter, thereby reducing troublesome self-inductances, capacitances and resistances. This is all the more marked in the case of the valves made in the so-called A-technique [3]), such as the "Rimlock" valves.

Finally mention may also be made of the double (push-pull) pentode, in which the effect of the self-inductance of the cathode lead has been eliminated [4]).

Notwithstanding all these improvements, the valve problem for decimetric-wave reception could not by any means be said to have been solved. Such is not even the case when considering valves in the construction of which the problem has been approached from an entirely different angle, such as the magnetrons [5]), velocity modulation valves [6]) and suchlike, where the transit-time effect is turned to advantage.

During the war, triodes working on the normal principle were developed in Great Britain and the U.S.A. and important results have been achieved with these. These triodes are of an uncommon construction and are referred to as "disc-seal" valves. The anode and the grid each have a flange (disc fused into and protruding beyond the wall of the glass bulb). The protruding rim forms the connection, so that the self-inductance and resistance are extremely small. The capacitances, too, are very small. The "disc-seal" valves can be used as amplifiers up to a frequency of about 1000 Mc/s (wavelength $\approx$ 30 cm) and as oscillator even up to about 3000 Mc/s (wavelength $\approx$ 10 cm).

The abnormal construction of the "disc-seal" valve however has also its disadvantages: these valves take up a great deal of space in the receiver, and they do not lend themselves for mass production with the machines designed for conventional valves. Therefore, at least for the present, this kind of valve will continue to be fairly expensive. Philips have nevertheless taken in hand the manufacture of "disc-seal" valves (type EC 55), but at the same time a study has been made of the problem in how far good high-frequency properties can be obtained with valves which can indeed be manufactured on a large scale with the present factory equipment.

As a result of the study of this problem two valves have been designed, the EC 80 for amplifying and mixing (to be used as amplifier up to frequencies of about 600 Mc/s, wavelength 50 cm), and the oscillator valve EC 81 (up to about 1500 Mc/s, wavelength 20 cm).

Features common to both valves EC 80 and EC 81

Both these valves (*fig. 1*) are made according to the A-technique already mentioned, thereby obtaining short connections between the electrodes and the valve holder.

This measure, which had already been applied in the "Rimlock" valves, is not, however, by any means sufficient for the frequency range of decimetric waves. As regards the resistance of the connections between valve holder and electrode this may be explained by an example.

In a normal radio valve this resistance R at a frequency of 300 Mc/s ($\lambda = 1$ m) amounts to about 3 Ω. More striking is the value of the damping, i.e. the conductance g (across the input or output circuit), in which just as great a loss occurs as in the series resistance R. A simple calculation shows that

$$g = \omega^2 C^2 R, \quad \dots\dots \quad (1)$$

in which ω = angular frequency and C = capacitance of the electrode to which the series resistance R applies. With C=10 pF; $\omega=2\pi\times300\times10^6$ sec^{-1}

[2]) A new principle of construction for radio valves, Philips Techn. Rev. 4, 162-166, 1939; Th. P. Tromp, Technical problems in the construction of radio valves, Philips Techn. Rev. 6, 317-323, 1941.

[3]) In the A-technique the base and the bulb are fused together with the aid of a kind of glass having such a low melting point that the electrode system can be mounted close to the base without risk of overheating the cathode while fusing. See G. Alma and F. Prakke, A new series of small radio valves, Philips Techn. Rev. 8, 289-295, 1946.

[4]) M. J. O. Strutt and A. van der Ziel, A new push-pull amplifier valve for decimetre waves, Philips Techn. Rev. 5, 172-181, 1940.

[5]) See, for instance, The magnetron as a generator of ultra short waves, Philips Techn. Rev. 4, 189-197, 1939.

[6]) F. M. Penning, Velocity-modulation valves, Philips Techn. Rev. 8, 214-224, 1946; F. Coeterier, The multireflection tube, a new oscillator for very short waves, Philips Techn. Rev. 8, 257-266, 1946.

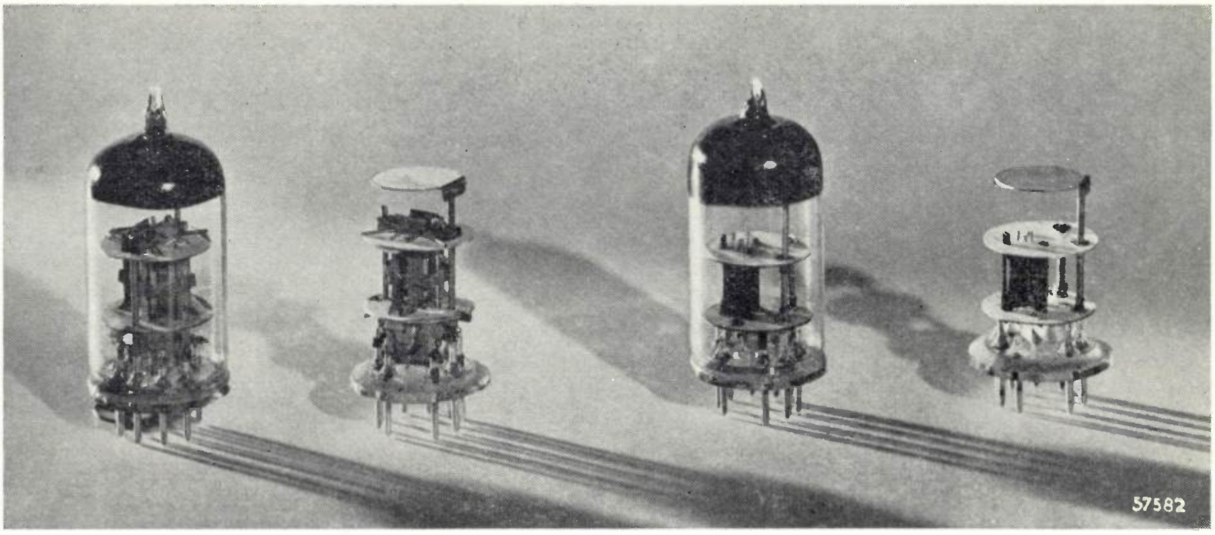


Fig. 1. The amplifying and mixing triode EC 80 (left) and the oscillator triode EC 81 (right). A specimen of each is shown without bulb.

and $R = 3\ \Omega$ one finds $g \approx 1000\ \mu\text{A/V} = (1000\Omega)^{-1}$. Thus the series resistance of 3 Ω has the same effect as a resistance of only 1000 Ω parallel to one of the circuits.

The rather high value of R is due to the skin effect which occurs at high frequencies. For the resistance of a round conductor we then have:

$$R = 2\,\frac{l}{D}\,\sqrt{\varrho\,\mu_r f \cdot 10^{-7}}\ \text{ohms}, \quad (2)$$

in which l and D represent respectively the length and diameter of the conductor (expressed in the same length unit), ϱ the specific resistance (in Ω·m), μ_r the relative permeability, and f the frequency (in c/s).

The material commonly used for the contact (and lead) pins is chrome iron and that for the connecting rods (connectors) inside the valve nickel. Both these materials are most unfavourable from the point of view of high-frequency resistance, not only because of their rather high specific resistance but particularly on account of their ferromagnetic properties, which determine the value of μ_r in equation (2).

A considerable improvement may be obtained by covering the chrome-iron pins and the nickel connectors with a layer of non-ferromagnetic and good conducting metal (silver or copper), which, owing to the skin effect, takes over the conduction of the whole of the current. In the core the magnetic field strength is therefore nil, so that the magnetic properties of the material of the core are of no consequence.

As regards the application of this principle, it was not difficult to give the lead pins, in so far as they protrude either side of the glass base, and the connectors a plating of copper or silver. In the case of the lead pins the resistance (at 300 Mc/s) was thereby reduced from a few ohms to about 0.5 Ω. The greater part (0.42 Ω) of this resistance is due to the non-plated part of the pin contained in the glass. Therefore it was necessary to investigate how this part too could be covered with a good conducting layer. This layer, however, has to be able to withstand the temperature reached in the fusing-in process, and it must not endanger the vacuum seal. Only after extensive investigations has it been possible to find a way of plating these pins with copper which ensures a good seal. The resistance of the fully copper-plated pins used in the valves EC 80 and EC 81 is now only 0.03 Ω (at 300 Mc/s).

The valves EC 80 and EC 81 have nine contact pins ("Noval" base) and fit into an internationally standardized valve holder. For the construction of receiving sets this means a considerable saving of space in comparison with the "disc-seal" valves.

The EC 80 amplifier and mixer

Triode or pentode?

In the designing of an amplifying valve for decimetric waves the first question to be decided is whether the valve is to be a triode or a pentode.

For the normal broadcasting wavelength it is a matter of obtaining the greatest possible amplification with the least possible feed-back from the anode circuit to the grid circuit, less attention having to be paid to the noise because this is mainly determined by the circuits. In this range of wavelengths the pentode (or hexode, or heptode) is the more suitable and is therefore generally the type of valve used.

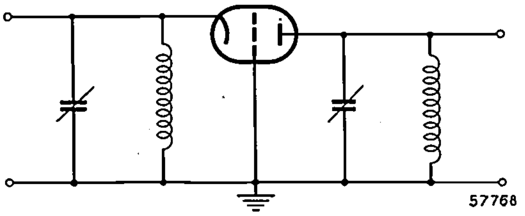


Fig. 2. Triode in grounded-grid connection. The direct-voltage sources have been omitted.

For decimetric waves it is a different matter and here the triode is preferred because of the following considerations.

Not only is a triode simpler than a pentode but it is also free of the so-called partition noise inherent in a pentode: in a pentode the electrons coming from the cathode pass partly to the screen-grid and partly to the anode, and the fraction of the cathode current formed by the anode current is subject to statistical fluctuations, the noise emanating therefrom being called partition noise. A triode is obviously free of at least this source of noise, a factor of great importance when the noise originates mainly in the valve.

An argument apparently in favour of the pentode is the feed-back of the anode upon the grid via the capacitance between these two electrodes. In a pentode this capacitance is much smaller than in a triode. This argument, however, carries no weight when the pentode is compared with a (specially constructed) triode in grounded-grid connection. In this form of connection it is the grid and not the cathode that forms the common electrode for the input and output circuits (*fig. 2*)[7]. Properly constructed, the grid acts as a screen between anode and cathode and there is no necessity for a separate screen-grid. It is true that with the triode in grounded-grid connection the self-inductance of the grid lead may cause instability, but by connecting a number of pins in parallel this self-inductance can be kept very low.

An advantage of the grounded-grid connection is that the self-inductance of the cathodc lead is harmless. In a grounded-cathode connection —which is used with a pentode—this self-inductance, in combination with the capacitance between cathode and grid, causes damping of the input circuit, and this damping is all the greater as the frequency rises. This difficulty does not occur in the grounded-grid connection.

Details of construction

The electrodes of the EC 80 triode differ considerably in shape from those found in conventional radio valves; see *fig. 3*. The cathode has two wide, flat, emitting surfaces. The grid proper is extended upwards and downwards by a metal plate acting as screen between anode and cathode (including the leads). As a consequence the capacitance C_{ak} between anode and cathode is less than 0.06 pF. The "fins" of the anode, which consist of two halves, are bent away from the grid; they serve for cooling and for fixing the anode in the mica discs between which the electrode system is located. Owing to this shape of the anode its output capacitance C_a is only 4.3 pF.

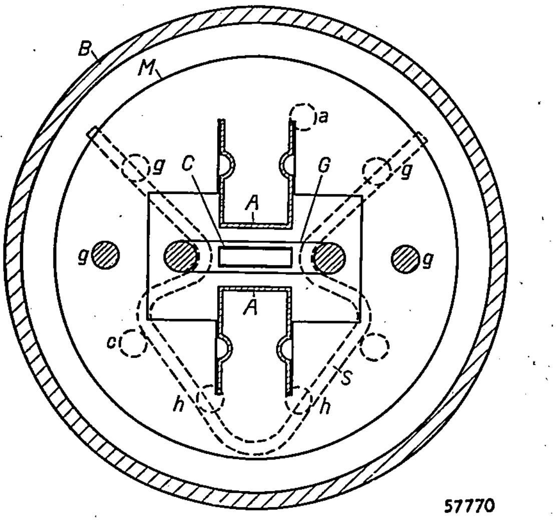


Fig. 3. Top view of a horizontal cross-section of the electrode system of the EC 80 triode. C = cathode (connected to the pin c), G = grid (connected to the four pins g), A = anode made in two halves (connected to the pin a), h = pins connected to the filament, M = bottom mica disc, S = tungsten spring keeping the grid windings taut, B = glass bulb.

The lower screen forming an extension to the grid is connected to four of the nine contact pins by means of short, wide strips. The corresponding contacts of the valve holder must all be connected to the chassis. This quadruple connection between grid and chassis helps to keep the series resistance and the series self-inductance of the grid at a low value. Furthermore, the series resistance of the electrodes is kept low by plating the respective parts with silver or copper. The grid itself consists

[7] Such circuits have been mentioned in connection with transmitting valves in an article by E. G. Dorgelo, Glass transmitting valves with high efficiency in the 100 Mc/s range, Philips Techn. Rev. **10**, 273-281, March, 1949 (No. 9).

of wire wound round two support rods. Two springs of tungsten force the rods apart, so that the windings are kept taut. Thus the distance between grid and cathode can be made very small without fear of short-circuiting. This is favourable for ensuring a short transit time between cathode and grid and a high mutual conductance.

As the distance d between cathode and grid is reduced, so the capacitance C between the active parts of these electrodes increases:

$$C = k_1 \frac{O}{d},$$

in which k_1 is a constant and O represents the surface area of the cathode.

In itself any increase of C is unfavourable, but the essential quantity, which should be as large as possible, is S/C (S = mutual conductance) and this increases as d diminishes, as will be understood from what follows.

According to Langmuir and Child, for an anode current I_a we have

$$I_a = k_2 \frac{O}{d^2} V_c^{3/2}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

where k_2 is a constant and V_c is the control voltage, i.e. the active voltage in the grid plane. From (3) it follows that:

$$S = \frac{\partial I_a}{\partial V_c} = \frac{3}{2} k_2 \frac{O}{d^2} V_c^{1/2}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

so that we find that

$$S/C = \frac{3}{2} \frac{k_2}{k_1} \frac{V_c^{1/2}}{d} = k_3 \frac{V_c^{1/2}}{d}$$

indeed increases when d is reduced.

C forms a part of the cathode input capacitance C_k, which in the case of the EC 80 valve amounts to 6.2 pF, whilst S = 12 mA/V.

S depends not only upon the geometry of the valve but also upon V_c and I_a. Eliminating O/d^2 from (3) and (4) we get

$$S = \frac{3}{2} I_a V_c^{-1}.$$

In order to get a high mutual conductance it is therefore necessary to work with a high anode current and a low control voltage. These two measures can only be applied, however, to a limited extent, because high anode current involves, inter alia, an expensive supply apparatus, and reduction of the control voltage increases the transit time τ between cathode and grid, since

$$\tau = k_4 \frac{d}{\sqrt{V_c}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

(where k_4 is a constant) and thus the damping due to the transit time, to which we shall revert later, is increased.

From (5) it appears that τ is proportional to d, so that the compromise that has to be made between S, I_a and τ is all the more favourable the smaller the value of d. That is why means are being continuously sought to reduce this clearance to the utmost.

Amplification

In the frequency range of decimetric waves the relations existing between the admittance of tuned circuits (or of whatever takes their place, e.g. cavity resonators) on the one hand and the dampings connected in parallel thereto on the other hand are quite different from those existing in the case of lower frequencies. With the aid of a simple example it will be shown how this affects amplification.

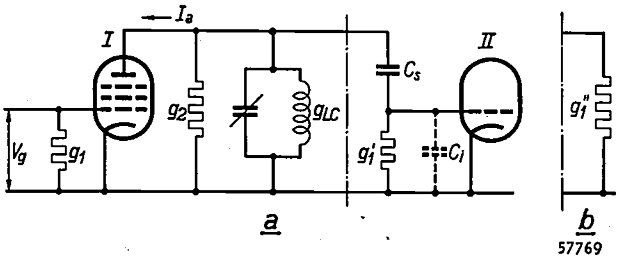


Fig. 4. *a*) Circuit for high-frequency amplification with a pentode (*I*). g_1 = input damping, g_2 = output damping, g_{LC} = conductance of the tuned circuit, g_1' = input damping of the next stage (*II*). V_g = grid alternating voltage, I_a = anode alternating current (both R. M. S. values). The capacitances C_s and C_i form a voltage divider used for matched coupling between g_2 and g_1'. The two capacitances and g_1' can be imagined as being replaced by the transformed input damping g_1'' (see (*b*)).

Let us first consider an amplifying stage (*fig. 4a*) consisting of a pentode *I* (mutual conductance *S*). In the anode circuit is a tuned circuit to which the next valve (*II*) is connected. The voltage amplification of this stage is S/g_0, where g_0 represents the conductance arising from the parallel connecting of the conductance g_{LC} of the tuned circuit, of the output damping g_2 of the valve *I* and of the input damping g_1' of the next stage: $g_0 = g_{LC} + g_2 + g_1'$.

At not very high frequencies $g_2 = 1/R_i$ and $g_1' = 1/R_g$, where R_i = the internal resistance of the valve *I* and R_g = the resistance via which the control grid of the valve *II* receives negative grid bias. g_2 and g_1' are small compared with g_{LC}, which depends upon the quality of the circuit. Thus this mainly determines the amplification.

At the frequencies of decimetric waves the situation is quite different. Transit-time effects and the resistance and self-inductance of the electrode leads increase the dampings g_2 and g_1' to values greater than g_{LC}. The circuit quality is then only of minor importance in determining the voltage amplification, which is much less than that obtained at lower frequencies.

In order to get the highest possible voltage on the grid of the next valve a sort of transformer has to be used to provide such a matching that as much power as possible is fed into the damping g_1'. This is the case when g_1'' (fig. 4*b*), the transformed value of g_1', is equal to g_2. As "transformer" one can use the capacitive voltage divider formed by the coupling capacitance C_s and the hitherto disregarded input capacitance C_i of the valve *II*, C_s being given a suitable value in that case.

If I_a is the R. M. S. value of the anode current, then, assuming that matching is correct, there flows through g_1'' a current $\frac{1}{2}I_a$ producing the power $P_2 = \frac{1}{4}I_a{}^2/g_1'' = \frac{1}{4}I_a{}^2/g_2$. Now $I_a = SV_g$ when V_g is the alternating voltage on the grid of the valve *I*. If the input power is P_1 and the input damping of the valve *I* is g_1 then $V_g = \sqrt{P_1/g_1}$. The power amplification G is then

$$G = \frac{P_2}{P_1} = \frac{1}{4}\,\frac{S^2}{g_1 g_2}.$$

The foregoing considerations have been based upon a pentode in grounded-cathode connection (fig. 4*a*). Let us now consider the amplification in the case of a triode in grounded-grid connection, for which the EC 80 has been designed. With this connection (fig. 2) the input damping g_1, as already stated, is formed by the mutual conductance S and the output damping g_2 by $1/R_i = S/\mu$ (where μ = amplification factor), at least for frequencies at which other causes of damping may be disregarded. It would lead us too far to go into an analysis of the circuit here, but it must be borne in mind that there is a feedback via R_i. The maximum power amplification with the most favourable matching is then found to be $G = \mu + 1$. However, owing to circuit losses and the extra damping at higher frequencies G is usually less than $\mu + 1$.

As a general rule it may be concluded that to get a high power amplification the valve should have a high value of μ.

For the EC 80 $\mu = 80$. At a wavelength of 1 m and with a bandwidth of 4 Mc/s this valve can yield a power amplification of 20 and at a wavelength of 75 cm an amplification of 13. (The bandwidth must be mentioned, since the damping is roughly proportional to the bandwidth.)

Noise

Nowadays the noise properties of a receiver (or of a part of it) are often denoted by the noise figure F, representing the signal-to-noise ratio at the output of the respective part of the receiver divided by the signal-to-noise ratio in the aerial. (Here both the noise and the signal are expressed as power and the noise has to be taken over the bandwith of the receiver or of the succeeding amplifying stages.) So the smaller the noise figure, the better it is.

Instead of the quotient of the two signal-to-noise ratios being given direct, it is also a common practice to express it in decibels.

In the case of an amplifying stage containing the EC 80 valve the lowest possible noise figure at 300 Mc/s is about 5 (or about 7 decibels). If the amplifier is adjusted for maximum gain this noise figure is about 6 (or about 8 decibels). By way of comparison, the noise figure of the best high-frequency pentode so far known is about 20 (13 db). The difference is due for the greater part to the absence of partition noise in the EC 80.

The EC 80 as mixer valve

For mixing the EC 80 can be used either as a diode (grid connected to anode) or as a triode. In the latter case the grid may be earthed and the amplified aerial signal and the locally generated voltage may both be applied to the cathode. The anode circuit is coupled to the intermediate-frequency amplifier. We shall presently give an example of the use of the EC 80 as a mixing triode.

Other applications of the EC 80

Thanks to its high mutual conductance (12 mA/V) and low noise figure the EC 80 can often be used to advantage for other purposes than amplifying and mixing in the range of decimetric waves. There should be mentioned, for instance, amplifiers with a wide frequency band and intermediate-frequency amplifiers following a crystal mixing stage (radar, communications with beamed transmitters, and suchlike).

The oscillator valve EC 81

Frequency limit

For a triode to oscillate it is necessary that certain properties of the valve satisfy the equation [8]):

$$(s-2g_{ag})^2 + z^2 - 4(g_{gk} + g_{ag})(g_{ak} + g_{ag}) > 0, \quad (6)$$

where s and z represent respectively the real and the imaginary components of the complex mutual conductance S, and g_{ag}, g_{gk} and g_{ak} represent the damping between anode (a), grid (g) and cathode (k). All these factors are a function of the frequency.

The fact that there is some purpose in regarding the mutual conductance as a complex quantity may be explained as follows.

Since the electrons have only a finite velocity certain phase differences arise between the alternating voltage v_g on the control grid of the valve and the alternating currents flowing to the various electrodes. For instance, the cathode current

[8]) See an article by the present author shortly to appear in Philips Research Reports.

i_k lags in phase behind v_g (see the vector diagram in *fig. 5a*, applying for such a high frequency that the phase angles in question assume considerable values). There is a still greater phase shift between v_g and the anode alternating current i_a, because the electrons reach the anode later than the grid plane.

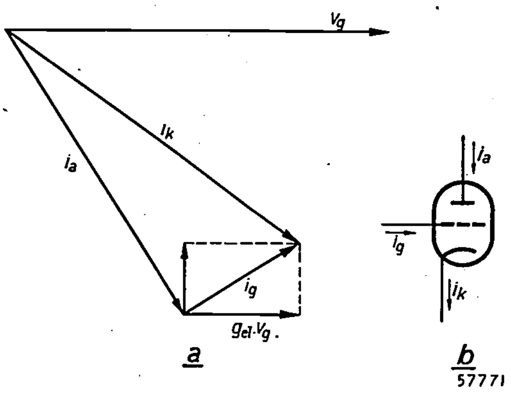


Fig. 5. *a*) Vector diagram showing the consequences of the finite transit time of the electrons in a triode. $v_g =$ grid voltage, $i_k =$ cathode current, $i_a =$ anode current, $i_g =$ grid current, $g_{el} \cdot v_g =$ component of i_g in phase with v_g.
b) Triode with the alternating currents i_k, i_a and i_g..

Owing to these phase angles the relation between v_g and the current i_a can be expressed by a complex mutual conductance $\mathbf{S} = i_a/v_g$, which is characterized by a negative phase angle and the modulus of which is less than the "ordinary" mutual conductance S measured at a low frequency. So long as the absolute value of the phase angles is not too great, one may to a first approximation disregard the difference between S and the modulus of $\mathbf{S}$.

The transit-time effects find expression not only in the fact of the mutual conductance becoming complex but also, as is apparent from what follows, in a damping across the input circuit.

The difference between the currents i_k and i_a is the alternating current i_g (fig. 5*b*) flowing to the grid and also represented in the vector diagram of fig. 5*a*. It is seen that i_g has a component in the direction of v_g. This component may be represented by $g_{el} \cdot v_g$, in which g_{el} is a real factor having the dimension of a conductance. The finiteness of the velocity of the electrons manifests itself in an apparent conductance g_{el} — i.e. a damping — between grid and cathode. This is what we already had in mind in the first paragraph of this article.

If the condition (6) is not satisfied the valve cannot by any means be made to oscillate, at least not at the frequency at which the factors occurring in (6) have the substituted value.

Owing to the transit-time effects, as the frequency rises so the modulus of the mutual conductance and thus also its square $(s^2 + z^2)$ decrease; at the same time, due to transit-time and other effects, the dampings increase. As a consequence the left-hand member of (6) becomes smaller and at a certain frequency reaches the value zero. This is the highest frequency at which the value can oscillate.

Measures for raising the frequency limit

In order to get the highest possible frequency limit it is necessary that the damping should be as small as possible in relation to the mutual conductance. As regards the part of the dampings resulting from the finite transit time, there are two measures to be taken, both of which aim at shortening the transit time: reducing the clearances between the electrodes in so far as mass production allows it, and applying high voltages.

These measures are of most importance for the space between the cathode and the grid, where the voltage is much lower, so that — notwithstanding the shorter distance — the transit time is longer than in the space between grid and anode. Therefore a high control voltage (potential in the plane of the grid) is desired.

As a result of a high control voltage (V_c) and a short distance (d) between cathode and grid there is a great current density at the cathode, since

$$I_a = k \frac{O}{d^2} V_c^{3/2}, \quad . \; . \; . \; . \; . \; . \; . \quad (3)$$

(where k is a constant and O the surface area of the cathode) and thus the current density is:

$$\frac{I_a}{O} = k \frac{V_c^{3/2}}{d^2}.$$

With a certain anode current O should therefore be small, this being favourable for keeping the inter-electrode capacitances low and thus at the same time reducing that part of the dampings that is due to the losses caused by capacitive currents in the series resistance of the leads. These dampings are proportional to the square of the inter-electrode capacitances (see eq. (1)) and are therefore strongly reduced by the reduction of the cathode surface area mentioned.

These principles have been embodied in the construction of the EC 81 triode, an idea of which is given in *fig. 6*. It may be noted that here there is no need of a spring construction of the grid as employed in the EC 80, because the grid of the EC 81 is much smaller and oval in shape, so that it is of itself already sufficiently rigid.

Further details of the EC 81 valve

The anode has been blackened for better radiation. It is welded to a copper-plated chrome-iron lead pin. The grid and the cathode are both connected to a similar pin with a wide silver-plated strip.

Like the EC 80, the electrode system is mounted as close as possible to the base so as to minimize

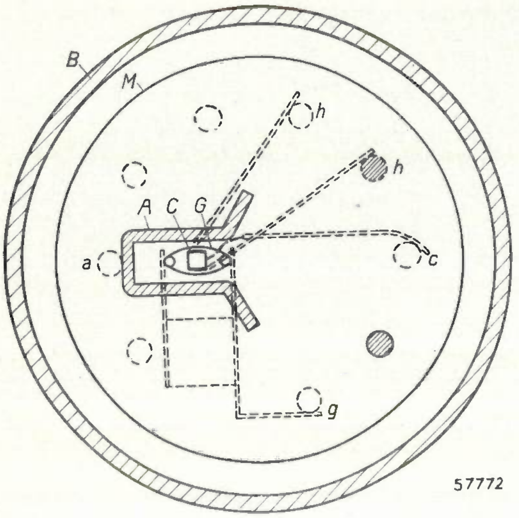


Fig. 6. Top view of a horizontal cross-section of the electrode system of the EC 81 triode. C = cathode (connected to the pin c), G = grid (connected to the pin g), A = anode (welded to the pin a), h = pins to which the filament is connected, M = bottom mica disc, B = bulb.

the self-inductance of the connections. The reason why this self-inductance must be small in an oscillator valve is that with a low self-inductance the wavelength at which the electrode connections and the capacitances connected thereto come into resonance is small. Near this resonance frequency the valve cannot be caused to oscillate at any arbitrary frequency without more or less complicated measures being taken. The resonance frequency must therefore be raised to a value as close as possible to the frequency limit.

Measurements taken on the EC 81 valve

Fig. 7 shows the output power and the efficiency of the EC 81 as functions of the wavelength and frequency. It is seen that the frequency limit is about 1500 Mc/s (wavelength = 20 cm) [9]). To reach this frequency circuits of good quality must be used.

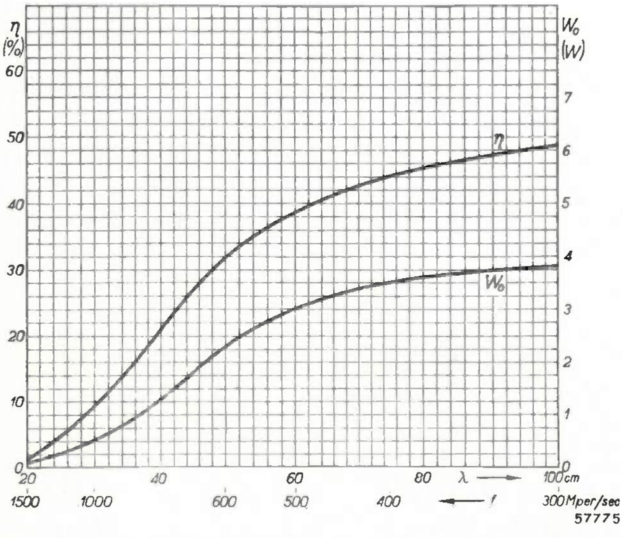


Fig. 7. Maximum high-frequency output W_0 and efficiency η as functions of the wavelength λ and the frequency f for the EC 81 triode.

At an anode current of 30 mA the mutual conductance amounts to 5.5 mA/V. The anode dissipation must not exceed 5 W. The capacitances with respect to the grid are C_{gk} = 1.7 pF and C_{ag} = 1.5 pF.

With the aid of an oscillator (*fig. 8*) the EC 81 valves are tested for their output at a wavelength of 40 cm. *Fig. 9* is a diagrammatic representation of this oscillator and its circuit.

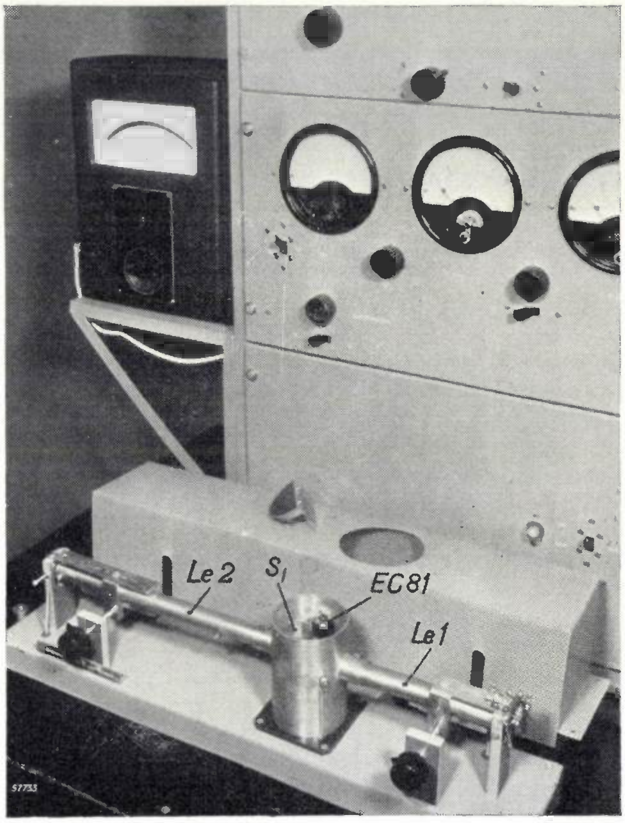


Fig. 8. Photo of the oscillator in which the EC 81 valves are tested at a wavelength of 40 cm. In the opened container the EC 81 valve is on the right and the oscillatory circuit (S_1) on the left. Le_1 and Le_2 are transmission lines.

Applications of the EC 81

In addition to the object already mentioned (oscillator valve in receivers for decimetric waves, an example of which will be given presently) the EC 81 can be used for small transmitters, for instance for "business-" or "citizen's radio", i.e.

[9]) For a description of the apparatus used for measuring the curves of fig. 7 see the article announced in footnote [8]).

private radio communications for which a frequency band near 470 Mc/s has been allotted. At this frequency the EC 81 can give an output of 3 W.

Due to the small values of the capacitances C_{ag} and C_{gk} previously mentioned, the capacitance variations (for instance those due to temperature changes) are also small. This makes the EC 81 eminently suitable as an oscillator valve in all sorts of radio-measuring apparatus, such as a standard-signal oscillator, where the frequency has to be very constant. Other favourable factors for this application are the low filament consumption (1.26 W) and the small dimensions.

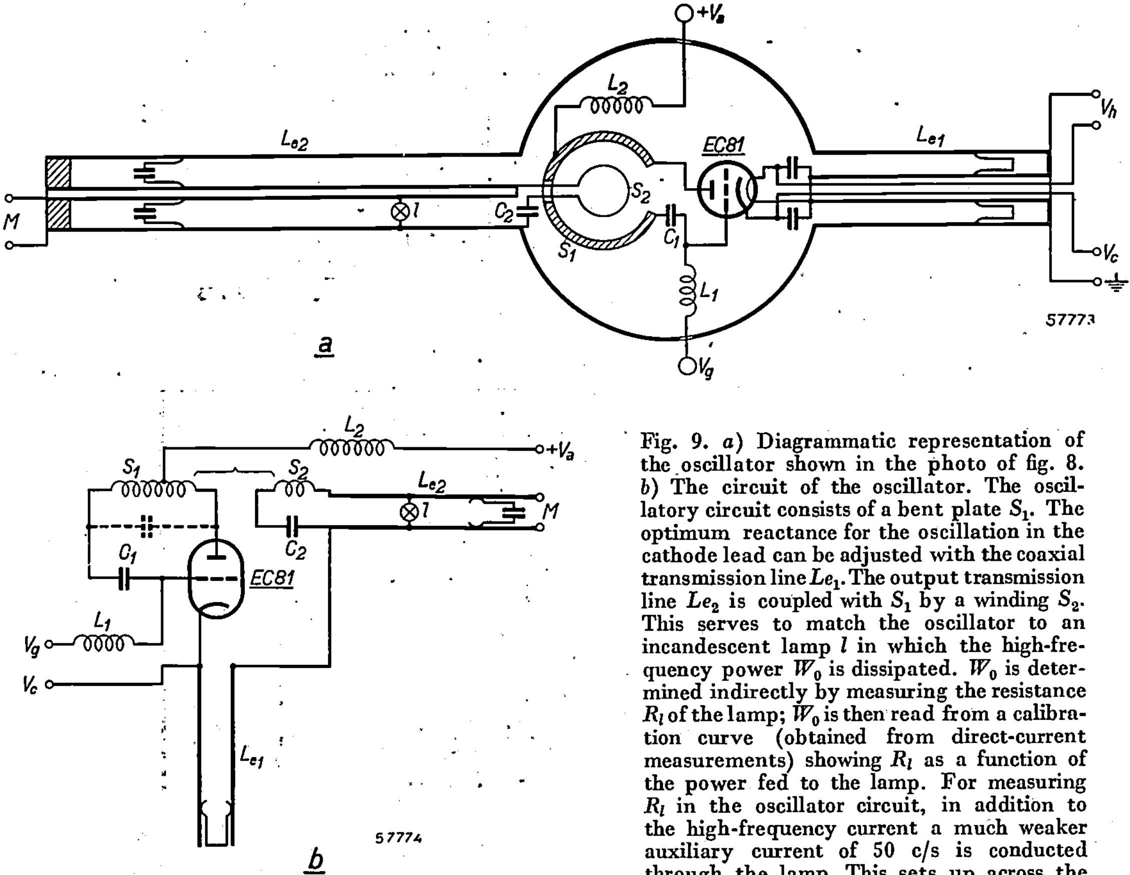


Fig. 9. *a*) Diagrammatic representation of the oscillator shown in the photo of fig. 8. *b*) The circuit of the oscillator. The oscillatory circuit consists of a bent plate S_1. The optimum reactance for the oscillation in the cathode lead can be adjusted with the coaxial transmission line Le_1. The output transmission line Le_2 is coupled with S_1 by a winding S_2. This serves to match the oscillator to an incandescent lamp l in which the high-frequency power W_0 is dissipated. W_0 is determined indirectly by measuring the resistance R_l of the lamp; W_0 is then read from a calibration curve (obtained from direct-current measurements) showing R_l as a function of the power fed to the lamp. For measuring R_l in the oscillator circuit, in addition to the high-frequency current a much weaker auxiliary current of 50 c/s is conducted through the lamp. This sets up across the lamp a low-frequency voltage which is measured with a valve voltmeter (type GM 4132, seen in the top left-hand corner of fig. 8) connected to the terminals M. From the value of this voltage and that of the auxiliary current the value of R_l is derived. The diagram also shows how the supply voltages are fed into the valve (V_h = filament voltage, $+V_a$ = + pole anode voltage; between V_c and V_g a direct voltage source or a resistor can be connected). L_1 and L_2 are high-frequency chokes.

Receiver for frequencies of 300 to 400 Mc/s

In conclusion some details are given of a receiving set [10]) for decimetric waves in which the EC 80 and EC 81 valves are employed.

The high-frequency part contains two amplifying stages, a mixing stage and an oscillator. The circuit diagram is given in a simplified form in *fig. 10a*. The three EC 80 valves are used with earthed grid. The oscillator circuit is in essence a Colpitts circuit, as will be understood, considering the capacitances C_{ak} and C_{gk}.

Actually the high-frequency circuits A, B and C are not ordinary L-C circuits as shown for the sake of simplicity in fig. 10*a*, but coaxial transmission lines (Lecher systems) (figs. *10b* and *11*). These have a short-circuiting plug which is adjustable for varying the tuning. When the tuning knob is turned this action is transmitted via a pinion and rack to the piston, to which a pointer is connected. This pointer moves along a

[10]) This receiver has been designed for a radio link in a carrier telephone system with 48 channels. The designer is J. M. van Hofweegen.

calibrated scale (*fig. 12*) covering a range from 300 to 400 Mc/s (100 to 75 cm). This makes it very easy to operate this receiver. It is quite possible that after further development of this construction tuning in this ultra-short-wave range can be done with a single knob.

Such a simple construction is not possible with "disc-seal" valves. Another advantage of the EC 80

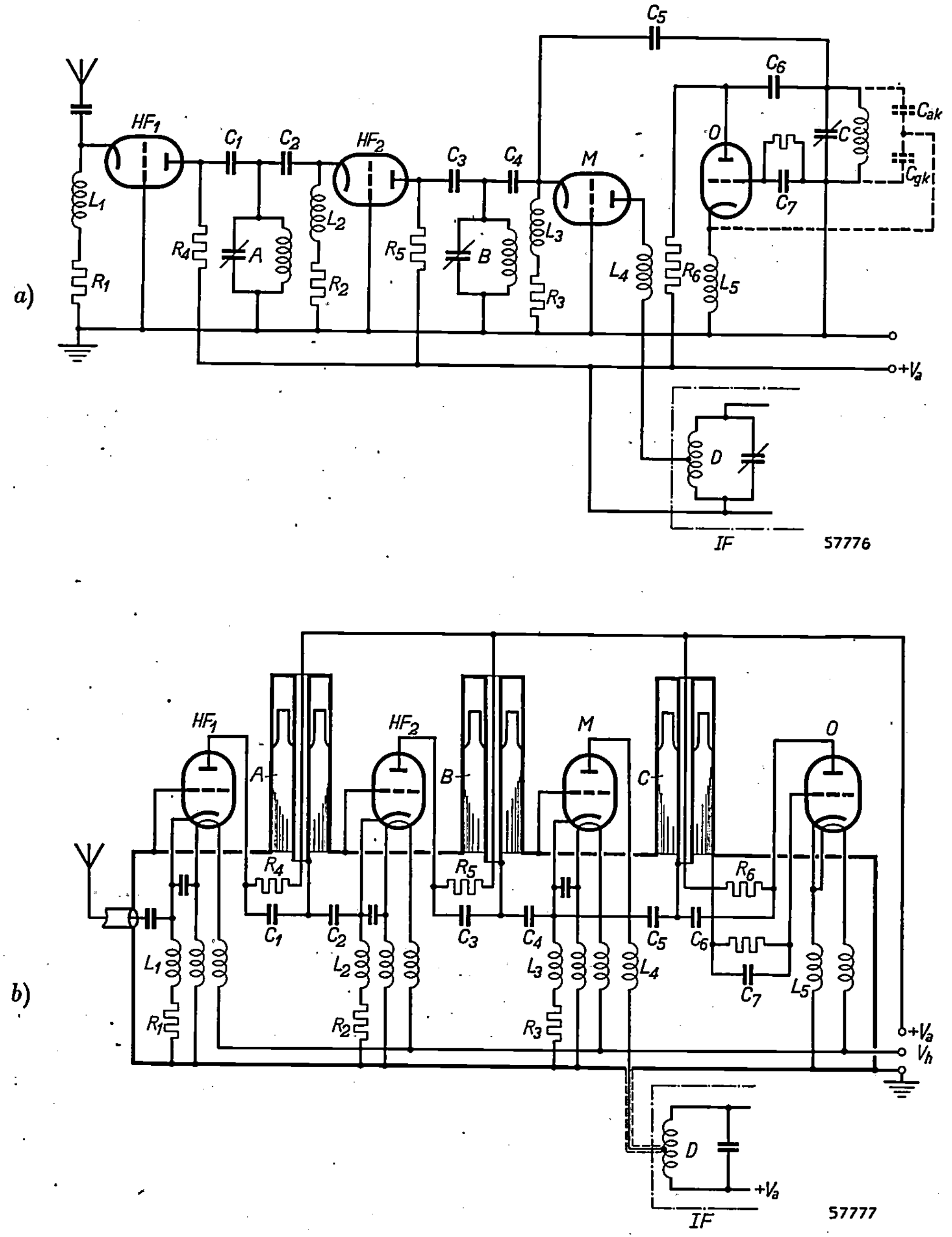


Fig. 10. High-frequency part of a receiver for 100 to 75 cm waves (see footnote [10])), (*a*) in simplified form, (*b*) complete. *HF1, HF2* = first and second high-frequency amplifying valves (EC 80); *M* = mixing valve (EC 80); *O* = oscillator valve (EC 81); *A, B, C* = oscillatory circuits in the form of coaxial transmission lines (in (*a*) represented for the sake of simplicity as normal *L-C* circuits); *IF* = intermediate frequency amplifier with intermediate frequency circuit *D*; $C_1 \ldots C_6$ = separating capacitors, C_7 = grid capacitor; C_{ak}, C_{gk} = valve capacitances making the oscillator a Colpitts oscillator; $L_1 \ldots L_5$ = high-frequency chokes; R_1, R_2, R_3 = resisistors for automatic positive cathode voltage (with respect to the earthed grid). R_4, R_5, R_6 = resistors via which anode current is fed. The capacitors C_2 and C_4 are not only separating capacitors but form at the same time, together with the input capacitance of the next stage, voltage dividers matching the low input resistance of the following stage to the higher output resistance of the preceding stage. In the complete diagram (*b*) a number of chokes are shown in the filament current leads which have been omitted in (*a*).

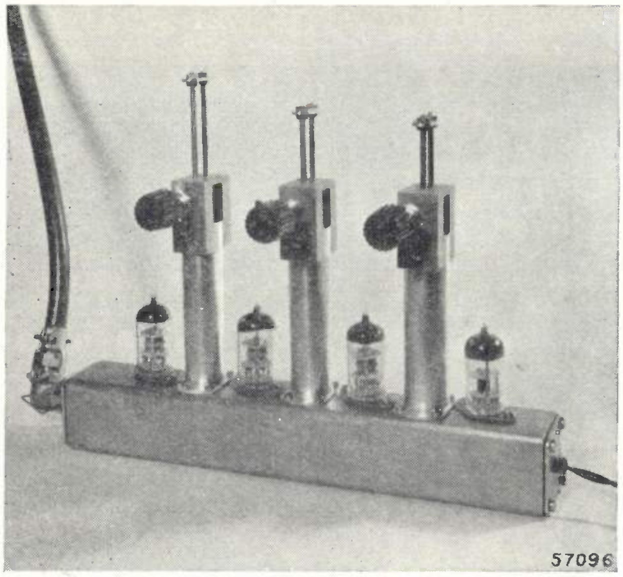


Fig. 11. High-frequency part of the receiver for waves of 100 to 75 cm the circuit of which is represented in fig. 10. Mounted on the chassis are three EC 80 valves, one EC 81 valve and three variable coaxial transmission lines. On the extreme left the aerial cable, on the right the supply lead for filament and anode current. Dimensions 24 cm × 4 cm × 20 cm ($9^1/_2'' \times 1^5/_8'' \times 8''$).

and EC 81 valves with their normal base is that they are easily interchangeable.

The bandwidth for which the receiver illustrated in fig. 11 has been designed is 5 Mc/s, at which a power gain of 250 is obtained in the two high-frequency stages together. The noise figure F is about 6.

The most important data of the EC 80 and EC 81 valves are tabulated below.

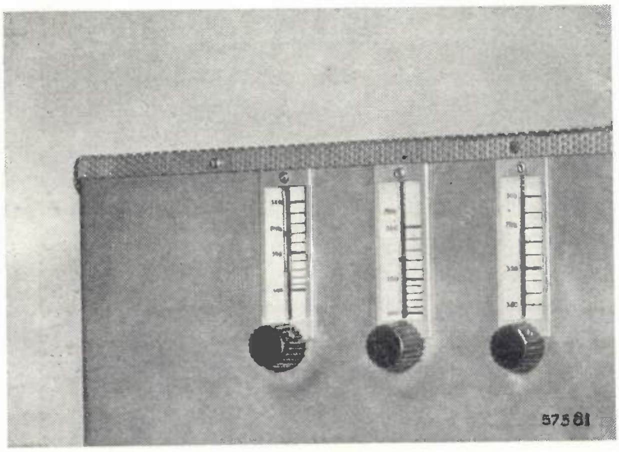


Fig. 12. Part of the front panel of the receiver the high-frequency part of which is shown in fig. 11. Here the three tuning knobs and calibrated scales (300-400 Mc/s) can be seen.

Table. Some data of the amplifying and mixing triode EC 80 and the oscillator triode EC 81.

	EC 80	EC 81
Filament voltage	6.3 V	6.3 V
Filament current	0.45 A	0.2 A
Anode current	15 mA	30 mA
Mutual conductance	12 mA/V	5.5 mA/V
Anode dissipation, max.	4 W	5 W
Amplification factor μ	80	16
Capacitance C_{k+f}	6.2 pF	—
Capacitance C_a	4.3 pF	—
Capacitance $C_{g(k+f)}$	5.4 pF	1.7 pF
Capacitance C_{ag}	3.4 pF	1.5 pF
Capacitance $C_{a(k+f)}$	0.060 pF	0.5 pF
Power gain G (at 300 Mc/s and over a bandwidth of 4 Mc/s)	20	—
Noise figure F (at 300 Mc/s)	5-6	—

Summary. A description is given of two receiving valves for decimetric waves: a triode (EC 80) for high-frequency amplifying and mixing, and an oscillator triode (EC 81). In appearance they are similar to conventional radio valves. They are made according to the A-technique (largest diameter 22 mm $\approx$ $^7/_8''$) and have an internationally standardized base with 9 pins. The damping due to the resistance in series with the electrodes is greatly reduced by plating the lead pins, the connecting rods and partly also the electrodes themselves with copper or silver. — The EC 80 is designed for grounded-grid connections. Its most important electrical data are: mutual conductance 12 mA/V at an anode current of 15 mA, optimum gain $G = 20$ (at 300 Mc/s and over a bandwidth of 4 Mc/s), noise figure $F=5$ to 6 (at 300 Mc/s). — The oscillator triode EC 81 has a frequency limit of about 1500 Mc/s. At 750 Mc/s it has an output of about 1.3 W at an efficiency of 20 %. Mutual conductance is 5.5 mA/V at an anode current of 30 mA. The anode can dissipate 5 W. — In addition to the purposes mentioned, the EC 80 is very suitable for amplifiers with a wide frequency band and for the intermediate frequency amplifiers of radar and beam receivers. The EC 81 is also very useful in small transmitters and standard-signal oscillators. Finally mention is made of a receiving set having two stages of high-frequency amplification (EC 80), a mixing stage (EC 80) and an oscillator (EC 81). As oscillatory circuits coaxial transmission lines are used, the tuning of which is continuously variable between 300 and 400 Mc/s. The tuning of each circuit can be read from a scale.

PARTS OF THE PHILIPS WORKS AT EINDHOVEN

The upper photograph shows on the extreme left the old glass factory and the new one. On the right in the background is the "Philite" factory. The row of factories in the middle is where radio sets, components and all sorts of electrotechnical products are made. In the lower photograph in the foreground is the paper and cardboard factory, and to the rear and right of that the glass works, the background being formed by the above-mentioned row of radio and other factories with on the extreme left the warehouse for this group of factories (Strijp Works). The lamp and valve factories are situated in another part of the town (Emmasingel Works).

A METHOD FOR DETERMINING THE MERCURY CONTENT OF AIR

by H. van SUCHTELEN, N. WARMOLTZ and G. L. WIGGERINK.

544.4.082.5:546.49:613.632

Periodical medical examination of persons coming into contact with mercury vapour is essential in many industries but is cumbersome and expensive. It means a great improvement if one is able to determine directly the amount of mercury vapour in the atmosphere. In the Philips factories at Eindhoven an apparatus is being used for this purpose by means of which it is possible to detect the presence of one microgram of mercury per cubic metre of air (that is about 1% of the concentration dangerous for human beings). This is done by an electronic measuring method.

Introduction

Scientists and workers who have to carry out their work in an atmosphere containing mercury vapour are exposed to the danger of mercury poisoning. The mercury inhaled with the air is for the greater part retained in the body. Tests with animals have shown that the mercury accumulates mainly in the brain and in the kidneys, up to certain quantities [1]).

Acute poisoning can usually be quickly diagnosed as such and steps can then be taken by the doctor to check it. It is more difficult however to diagnose chronic mercury poisoning, so that both the doctor and the patient may be left in doubt for quite a time. The symptoms — forgetfulness, aversion to work, irritability and in more serious cases headache, alimentary disorders, decay of the teeth — are not so specific, for they may also be found among persons who have not inhaled any mercury vapour before.

It is therefore highly reassuring for the personnel concerned if a check is carried out to determine with certainty that no mercury vapour or at most a harmless quantity is present in the workshops. Periodical testing of the atmosphere in the workshops is also much less costly than a periodical medical examination of the people working in such an atmosphere.

It is first of all necessary to know what quantity of mercury can be tolerated. Among those who have investigated the matter there is some difference of opinion, which is understandable considering the great variation in individual susceptibilities for mercury vapour, as also exists for other poisons. In particular it appears that once a person has had this poisoning he is apt to be more susceptible to it for some time. From extensive investigations carried out among people working with mercury in a felt factory and in a scientific laboratory Flinn, Hough and Neal [2]) came to the conclusion that it may be injurious to health to stay for a long time in an atmosphere containing more than 100 μg mercury per m^3. This we shall call the dangerous concentration [3]).

Some 20 years ago in the United States considerable attention was paid to the hazard of mercury poisoning as a result of experience in some mercury distilleries. Various methods were then worked out for determining the mercury content of air.

Among these there are to be distinguished chemical and optical methods. Of the chemical methods the one most known is that whereby selenium sulphide is used as indicator. Activated SeS_2, a yellow powder, is precipitated in a thin layer on paper. If Hg is present then the black HgS that is formed gives this paper a darker colour. Where it is a question of quantitative determinations this method is cumbersome and not very accurate, as is in fact the case with the other chemical methods.

A method will now be described whereby the mercury content of air can be determined by optical-electrical means.

Principle of the method

The method described here is based upon the fact that the mercury atom absorbs radiation of the wavelength 2537 Å (one of the resonance lines of this atom). Radiation of this wavelength is ob-

[1]) A. Stock, Die chronische Quecksilber- und Amalgamvergiftigung, Archiv Gewerbepathologie und Gewerbehygiene 7, 388, 1936. See also A. Kreyer, thesis Karlsruhe, 1936.

[2]) J. Res. Nat. Bur. of Stand. 26, 357-375, 1941. The scientists mentioned used an apparatus designed by Woodson which is based upon the same principle as that of the apparatus described in this article but in which only one photocell is employed; see T. T. Woodson, A new mercury vapor detector, Rev. sci. Instr. 10, 308-311, 1939.

[3]) The limit of 2 to 20 μg mercury per m^3 given by Stock applies presumably for highly sensitive people and for those who have already suffered from mercury poisoning.

tained by employing a low-pressure mercury lamp with a filter. This radiation is directed through a test tube through which the air to be tested is passed and then "picked up" by a photocell. If the air contains mercury vapour then, owing to the absorption referred to, the photocell will produce a weaker current than is obtained with pure air.

The absorption of radiation of the wavelength 2537 Å by mercury vapour can be demonstrated visually. The radiation from a low-pressure mercury lamp provided with a filter is directed through the atmosphere over a small open vessel containing mercury onto a fluorescent screen. Owing to the absorption of the radiation by the mercury vapour dark clouds are seen on the screen rising above the surface of the mercury, even when the mercury is at room temperature (see *fig. 1*).

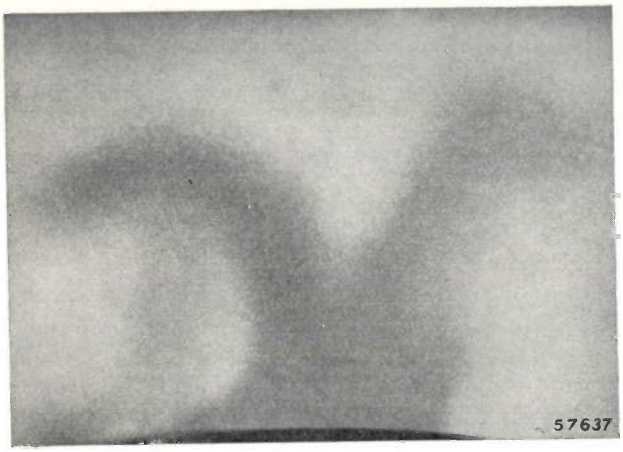


Fig. 1. The absorption of radiation of the wavelength 2537 Å by mercury vapour. The shadow of the mercury clouds has been made visible by directing the radiation of the said wavelength onto a fluorescent screen. The surface of the mercury is at the bottom of the picture (the rim of the vessel can just be seen). The temperature of the mercury was about 25 °C.

To get an apparatus which registers with sufficient accuracy the small differences in current intensity that have to be measured here, and which is not affected by fluctuations in the voltage of the electric mains, two photocells are employed in the mercury vapour detector used in the Philips works at Eindhoven. In front of the lamp is a semi-transparent mirror set at an angle of 45° to the optical axis, so that the transmitted radiation falls upon one photocell via the test tube already mentioned and the reflected part of the radiation falls directly onto the second photocell. The two photocells are taken up in compensating circuit, and as zero indicator use is made of a so-called magic eye, a tuning indicator commonly used in radio receivers.

The signal from the measuring branch of the bridge circuit (the "zero signal") is amplified with a simple amplifier and observed on the tuning indicator. Advantage is thereby taken of the fact that the exposure is intermittent (the mercury lamp is fed with alternating current), so that also the zero signal is an alternating voltage. This makes it possible to use resistors and capacitors as coupling elements for the amplifier. The frequency of the alternating voltage is 100 c/s.

Before a series of measurements can be taken the zero position has to be checked. This is done by passing air free of mercury through the measuring tube and adjusting the tuning indicator to the minimum deflection. If the air subsequently passed through the measuring tube contains mercury then the balance is disturbed, and this can be restored with the aid of a potentiometer. From the position of the potentiometer, which can be read on a scale, the mercury content of the air can be found by means of a calibrating curve belonging to the apparatus.

Description of the apparatus

A diagrammatic representation of the apparatus, without the electrical circuit, is given in *fig. 2*, which clearly shows how it is arranged. We shall first describe some of the component parts and after that discuss the circuit.

The light source

The low-pressure mercury-vapour lamp (a quartz-glass tube with a diameter of about 15 mm) is coiled in the shape of a flat helix, with the coils lying close up against each other. In this way the luminous surface is made as uniform as possible.

When the mercury lamp is burning it produces a rather large amount of ozone. To prevent the deleterious action of this gas upon parts of the apparatus (aluminium mirror, flex, rubber tube) the lamp is placed in a housing shut off by neoprene [4]) and having some holes in the bottom and one in the lid to provide for ventilation.

Immediately in front of the lamp is a so-called black filter (Bläckström filter) consisting of a quartz tube (2 cm thick) containing a solution of nickel-cobalt-sulphate (3 vol. 20.7% $NiSO_4$ + 2 vol. 24.1% $CoSO_4$). This filter allows only ultra-violet rays (of 2100-3300 Å) to pass through, namely 80% radiation of the wavelength 2537 Å, 10% of 2200 Å and 40% of 3200 Å.

Since the photocells are taken up in a compensating circuit any fluctuation in the total brightness of the lamp has no effect. A local variation in the brightness of the luminous surface may, however, be disturbing because such a fluctuation might

[4]) A plastic material which is not affected by ozone.

not have the same effect upon both the photocells. In order to "smooth out" the fluctuation in such a case a frosted quartz plate is placed in front of the liquid filter. In front of this plate is a diaphragm of black-burnt brass to avoid light from the lamp falling directly upon the photocell *A*.

The semi-transparent mirror already referred to is a quartz plate on which a thin layer of aluminium has been precipitated.

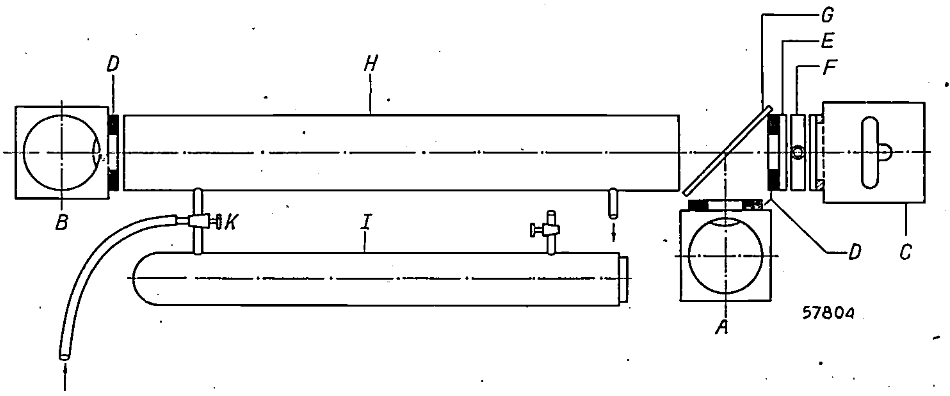


Fig. 2. Diagram showing the arrangement of the apparatus for determining the mercury content of air, without the electrical circuits. *A* and *B* are photocells, *C* is the mercury-vapour lamp. *D* diaphragms, *E* frosted quartz plate, *F* liquid filter, *G* semitransparent mirror, *H* measuring tube, *I* hopcalite filter, *K* three-way cock.

The measuring tube

The measuring tube is a glass tube (length 30 cm, diameter 7 cm) with quartz windows cemented onto it. The inner wall of the cylinder is frosted in order to avoid interference from specular reflection.

Leading off from the side of this tube are two small pipes for passing the air through the tube. One has to be connected to a pump or a vacuum pipe, the other serving to connect the tube via a three-way cock either direct to the locality from which the air is to be tested or to a hopcalite filter, through which the air is sucked in when determining the zero position. The hopcalite, which is a mixture of iron, nickel, copper and manganese oxides, has the property of completely absorbing mercury vapour. This hopcalite filter can be disconnected by means of a second cock when not in use.

The photocells

The two cells are caesium-antimony vacuum cells with a quartz-glass bulb. These are electrically screened off and in front of them are mounted adjustable diaphragms, with the aid of which the tuning indicator is adjusted to the minimum deflection for determining the zero position.

The two photocells are placed in such a position that the paths travelled by the ultra-violet rays outside the measuring tube are the same for both. There can then be no interference from fluctuations in the concentration of the mercury in the air outside the tube.

The circuit

The circuit is so arranged that the two photo-electric currents are compared one with the other immediately behind the photocells. This takes place in a bridge circuit, the principle of which is indicated in *fig. 3*. *Fig. 4* shows the complete circuit in a simplified form.

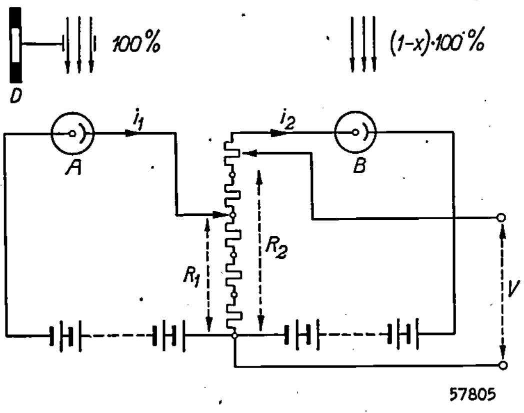


Fig. 3. Principle of the bridge circuit. *A* and *B* two photocells, *D* a diaphragm, *V* the amplifier.

The alternating voltage supplied by the amplifier when the balance is disturbed is applied direct, without rectification, to the grid of the tuning indicator. Thus the light vanes are thrown out of their state of rest 100 times per second. This gives the impression of a continuous deflection, but at the same time the light becomes hazier over the range of the deflection. The disadvantage that the edges of the picture become somewhat blurred is outweighed by the advantage that a greater sensitivity is obtained than would be the case if the output voltage were first rectified.

The gain in the first and second stages is respectively 80 and 50 times. The filter circuit, tuned to 100 c/s, in the second stage has a quality factor $\omega L/r = 90$.

It was found necessary to screen off the potentiometer, which has a coarse and a fine adjustment, in order to avoid stray voltages of 100 c/s being picked up from the mains or other parts of the circuit.

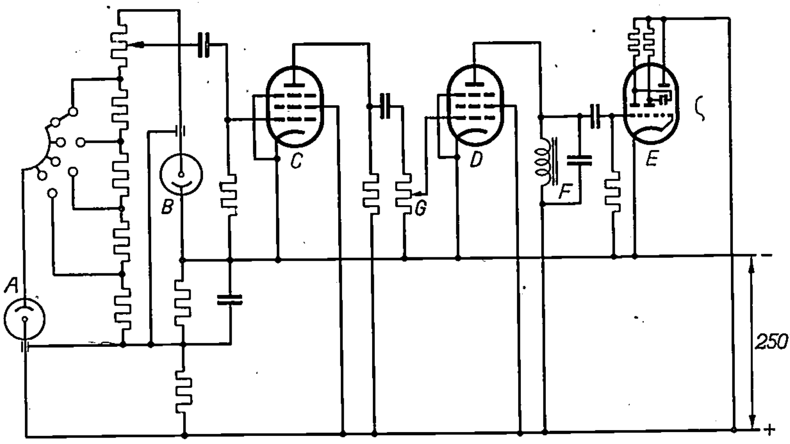


Fig. 4. The full circuiting in simplified form. *A* and *B* the two photocells, *C* and *D* the amplifying valves, *E* the tunig indicator, *F* the 100 c/s filter, *G* a sensitivity regulator to be used for adjusting the diaphragms.

From the bridge circuit in fig. 3 it can be seen how the zero adjustment of the potentiometer is related to the absorption by the mercury vapour. Suppose that of the beam of ultraviolet rays passing through the measuring tube a fraction x is absorbed and that the two beams of 100% and $(1-x)\cdot 100\%$ respectively produce proportional currents i_1 and i_2, then $i_2 = (1-x)i_1$. The condition $i_2 = i_1$ for $x = 0$ is satisfied by adjusting a variable diaphragm (*D* in fig. 3). The current i_2 traverses the resistance branch R_1 in a direction opposite to i_1. When the sliders of the potentiometer are so adjusted that $i_2R_2 = i_1R_1$ no voltage occurs at the input of the amplifier. Thus with this zero adjustment

$$x = \frac{R_2 - R_1}{R_2} = 1 - \frac{R_1}{R_2}.$$

In the first instance therefore x is given by the resistance between the two branches. The attractive feature about this circuit is the fact that the potentiometers can, if desired, be provided with a percentage scale for x and thus one can easily combine the readings on the coarse and the fine scales.

No true zero adjustment of the bridge can be obtained because the curves of i_1 and i_2 as functions of time are never absolutely identical. This is due to various causes, the most important of which are 1) that the light used is not entirely monochromatic, and 2) that the photocells are not identical in the spectral distribution of the sensitivity. In order to get the best possible zero adjustment under these conditions the Bäckström filter is employed and the amplifier is tuned to the flicker frequency of 100 c/s by means of a tuned circuit.

Fig. 5 shows the apparatus described with the metal cover removed.

Calibration of the apparatus

When one can work with a monochromatic light beam which is absorbed by the vapour being tested then a simple relation exists between the absorption x and the concentration of the vapour. This relation is given in the well-known Beer's law.

Any other component which is contained in the beam and is not absorbed disturbs this relation and obviously makes the method of measuring less sensitive. With absorption meters one therefore generally tries to approximate the monochromatic light. The Bäckström filter employed with this apparatus is likewise an attempt in that direction. Since, as the figures given show, this filter cannot be said to be ideal, the relation between absorption and concentration has to be determined by calibration. For routine work, however, the value of x is not of primary importance and one is more interested in a calibration which gives the mercury concentration in μg/m^3 as a function of the potentiometer reading.

The detector has been designed so as to be able to measure mercury concentrations varying between 1% and 200% of the danger limit, that is to say between 1 μg and 200 μg mercury per m^3 air [5]). For the calibration, therefore, air has to be passed

[5]) The sensitivity is per scale division 1 μg/m^3, corresponding to about 10^{-10} vol. parts.

through the apparatus with a variable mercury-vapour content lying within this range. This has been achieved in two ways. By the first method air was passed over mercury kept at 0 °C in a thermostat with ice water. At a rate of flow of 0.5 litre per minute the air was found to be saturated with mercury vapour. In order to get the concentration desired for the calibration this air was mixed in the required proportions with air that was free of mercury vapour. The advantage of this method is that the saturation pressure from which one starts can be determined in a simple way with the aid of ice water.

By the second method air was conducted over mercury contained in a thermostat with variable temperature. The temperature was chosen low enough to be able to reach directly the mercury concentration required for the calibration [6]). These temperatures, from —22 °C to —45 °C, were obtained by cooling a bath of ethylene trichloride by means of the addition of liquid nitrogen. A stirring device and heat-insulating walls provided for a uniform distribution of temperature in the liquid. The melting point of mercury, —38.9 °C, served as a check for the temperature.

As the results of the two calibrations did not agree over the whole range the mean value of the two measurements was taken. The difference may be attributed in part to the lack of accuracy in the known vapour-pressure curve of mercury at these low temperatures. The absolute accuracy reached in the measurements with this apparatus is therefore 10%. Except for the very small concentrations the relative accuracy is greater, amounting to about 2% for concentrations between 25% and 200% oı the dangerous limit.

Use of the apparatus

A rubber tube can be attached to the apparatus for drawing in air from different points in a locality so as to trace any sources of mercury vapour. The rate of suction should not be chosen higher than is necessary, so as to draw in only as much air as possible from the suspected part of the locality and not from the surroundings. On the other hand the rate of suction must not be too low either, because during the measuring process a "bleaching" takes place, the concentration of mercury vapour in the measuring tube being reduced through oxidation of excited mercury atoms. A suitable rate of suction is from 1 to 2 litres per minute.

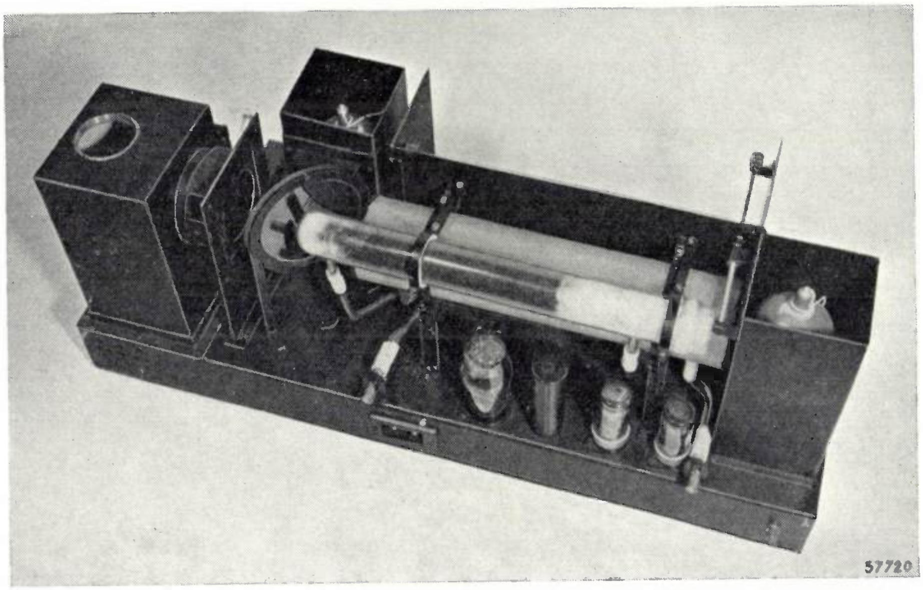


Fig. 5. Photograph of the mercury-vapour detector used in the Philips works at Eindhoven, with the metal cover removed.

When working with the apparatus described one should wait about 1 minute before taking a new measurement, so as to allow the air in the measuring tube to be entirely replaced by the air to be tested.

Applications

The apparatus described can be used primarily for measuring concentrations of mercury in the air

[6]) The vapour pressure of Hg has been taken from R. W. Ditchburn and J. C. Gilmour, The vapor pressures of monatomic vapors, Rev. mod. Phys. 13, 310-327, 1946.

in factories and laboratories and all other places where metallic mercury and mercury compounds are used. One can then ascertain in a simple and direct way whether the concentration is below the limit deemed to be dangerous.

We have in mind here not only those parts of a factory where a product is turned out which contains mercury, such as mercury-vapour rectifiers, but also workshops where mercury is used in some part or other of the manufacturing process, as is the case for instance in the electrolytic separation of alkaline metals by means of an amalgam electrode. Furthermore, mercury is frequently used in measuring instruments and accessories like mercury switches for regulating the temperature of ovens, vacuum valves, mercury-vapour-diffusion pumps and manometers. In all these cases there is a possibility, depending upon the care taken and the efficiency of the ventilation, that in course of time there may be a not inconsiderable concentration of mercury vapour in the air, either through direct evaporation of this metal from the apparatus concerned, or owing to a gradual or sudden contamination of the walls and floor of the workshop (breaking of an apparatus). Particularly interstices in a floor that is not absolutely smooth and the wainscoting around the room may constitute a source of mercury vapour.

In order to give an idea of what concentrations of mercury vapour can sometimes be found in laboratories and workshops under various circumstances, some results are given of measurements taken with the detector described above.

In a workshop where mercury-vapour rectifiers are evacuated the mercury concentration in the air after the workshop had been kept closed all night rose to about the danger limit (100 μg/m^3). This was also the case on warm days when the doors and windows were shut for some hours. When, however, the workshop was ventilated by causing a good draught to blow through it the concentration dropped to 10% of the limit and even less. In other cases too, a draught of pure air was found to be the best means of reducing the mercury concentration.

In laboratory rooms where only mercury diffusion pumps were used the mercury vapour concentration was small. In other rooms in the same laboratory where no mercury was used no trace of mercury could be detected in the air (that is to say the concentration was less than 1 μg/m^3).

Moreover, with this apparatus one can collect data in a short time regarding the variation of the mercury vapour concentration in a certain locality, as may be illustrated by the following example.

In a certain workshop where mercury was used the concentration was found to average 20% of the danger limit, but in one corner where there was little ventilation it rose to the danger limit. In the passage running past this department, where there was a perceptible draught, the measured concentration in front of the workshop was 8% and just past it 10% of the danger limit.

In large spaces and where there is proper ventilation the mercury concentration diminishes rapidly with the distance from the open mercury surface. At about 2 m away from a pool of mercury only a small percentage of the danger limit is measured, so that only the person actually working with the mercury is exposed to danger.

The use of the mercury-vapour detector is not confined to the protection of people against poisoning. In a mercury distillery for instance one can easily trace a leak in the installation by drawing in air from various places by means of the tube attached to the apparatus.

An often forgotten source of mercury vapour is the vapour from mercurous salt solutions. Sometimes the evaporation from such solutions and from paper and foil saturated with them is stronger than that from an open mercury surface. From mercuric salt solutions, on the other hand, there is no evaporation of mercury.

Not only the mercury atom absorbs rays of the wavelength 2537 Å. Various organic vapours also do so to a considerable extent. Therefore in some cases the mercury vapour detector can also be used for determining the concentrations of such vapours as these in the air.

Among the organic vapours to which the detector reacts there are in the first place the organic solvents nowadays being used more and more in industries for lacquers and degreasing. *Table I* gives some of these substances whose presence in air can be recorded with the apparatus. This table also indicates the sensitivity of the apparatus for each substance and what, according to the figures collected from literature by Jacobs[7]), is to be considered the dangerous concentration. A comparison of the figures given in the second and third columns shows that although this apparatus is less sensitive to these vapours than to mercury vapour it is nevertheless useful for detecting the presence of these substances, since the danger limit lies at very much higher concentrations.

In connection with the foregoing one might ask in how far the presence of organic vapours may interfere with the determination of the concentration of mercury vapour. Owing to the fact that

[7]) M. B. Jacobs, Analytical chemistry of industrial poisons, hazards and solvents, New York 1944.

Table I. Organic solvents whose presence can be detected with the mercury-vapour detector.

Substance	Sensitivity per scale division (in vol. parts)	Dangerous concentration (in vol. parts)
Xylol	2×10^{-7}	1 to 2×10^{-4}
Monochloro benzene	3×10^{-7}	7.5×10^{-5}
Aniline	3×10^{-7}	2.5 to 7×10^{-6}
Toluene	10^{-6}	5.3×10^{-5} to 2×10^{-4}
Benzene	1.2×10^{-6}	1.5×10^{-5} to 10^{-4}
Acetone	5×10^{-6}	2 to 4×10^{-4}
Triochloro ethylene	10^{-5}	1 to 2×10^{-4}
Benzine	5×10^{-5}	1 to 1.5×10^{-3}

this apparatus is less sensitive to these substances than to mercury vapour there need be no fear of a mistake being made and passing unnoticed. The concentration of these vapours that can be perceived with our sense of smell is less than the smallest concentration recorded by the apparatus. Consequently our sense of smell warns us in time of the presence of organic vapours, which have to be removed first; if it is impracticable to remove them then both concentrations can be determined separately by employing specific absorbents. Many organic substances are absorbed by active carbon, which is permeable for mercury vapour. On the other hand the hopcalite filter, which completely absorbs mercury vapour, in some cases lets organic vapours pass through.

Substances of another kind to which the apparatus reacts are fumes (such as tobacco smoke), since there is naturally absorption and scattering by the solid particles.

Also the presence of ozone causes a reaction of the mercury vapour detector. It is therefore necessary to take care that the ozone produced by the mercury lamp does not get into the measuring tube.

It is perhaps useful to mention some frequently occurring substances which do not affect the determination of the mercury vapour concentration. Among others, there are aqueous vapour, carbon tetrachloride, methyl-, ethyl- and amyl alcohols, chloroform, ethyl acetate, dichloro ethane, methyl chloride.

Finally it should be mentioned that in principle the mercury-vapour detector could also be coupled to an alarm device coming into action when a certain concentration is reached, or to a recording apparatus which records the concentration after certain intervals of time.

Summary. Exposure for some length of time to an atmosphere contaminated with mercury vapour may be dangerous for human beings when the air contains more than 100 μg mercury per m^3, as established by American scientists. The method described in this article for determining the mercury content of air is based upon the property of the mercury atom to absorb radiation of the wavelength 2537 Å. Part of this radiation, supplied by a low-pressure mercury-vapour lamp, passes through a measuring tube through which the air to be tested is conducted and is then directed upon a photocell, whilst another part is thrown directly upon a second photocell. The two photo-electric currents are compared with the aid of a compensation circuit, with a tuning indicator used as zero device. The two methods by which the apparatus described is calibrated are briefly discussed. Mercury concentrations of 1 to 200% of the limit dangerous for human beings can be measured. Various possibilities of application are mentioned. This mercury-vapour detector appears to be sensitive also to all kinds of organic vapours, but to a much less degree compared to mercury, so that this does not cause serious interference. Advantage can be taken of this property of the detector for measuring also the concentration of these organic vapours.

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

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

1832: J. F. H. Custers: The intensity distribution along the Debye halo of a flat specimen in connection with a new method for the determination of preferred orientations (Physica 14, 461-474, 1948, No. 7).

In a previous article (see No. **1831**) a new method for the determination of preferred orientations has been introduced. In this paper the formulae for the intensity distribution along the Debye halo are given when assuming a specimen with random orientation of crystallites. The same formulae may be used for making corrections for the strongly varying absorption along the Debye halo when preferred orientations have to be determined quantitatively. Moreover an expression is given for the breadth of the Debye circle, which may vary strongly with the azimuth angle.

1833: J. H. van Santen and W. Opechowski: On a generalization of the Lorentz-Lorenz formula (Physica **14**, 545-552, 1948, No. 8).

The well-known Lorentz-Lorenz formula holds only for a crystal in which all atoms have environments with cubic symmetry. In this paper a derivation is given of an analogous formula valid for the more general case that the environments of atoms are not necessarily cubic, the macroscopic symmetry of the crystal being still cubic. The formula is applied to the perovskite lattice.

1834: Th. P. J. Botden and F. A. Kröger; Energy transfer in sensitised $Ca_3\ (PO_4)_2$-Ce-Mn and $CaSiO_3$-Pb-Mn (Physica 14, 553-566, 1948, No. 8).

Calcium phosphate activated with cerium shows an ultra-violet emission consisting of two overlapping bands, with maxima at approximately 3470 Å and 3660 Å, when excited by $\lambda<3500$ Å in the absorption bands due to the Ce^{3+}ion. Calcium silicate activated with lead shows an ultra-violet emission with a maximum at 3300 Å when excited by $\lambda <$ 2900 Å in the absorption bands due to the Pb^{2+}ion. With manganese as a second activator, in both phosphors a large part of the absorbed energy is transferred from the cerium (respectively lead) centres to the manganese centres. The fluorescence spectrum consists of a faint ultra-violet cerium (or lead) emission and a strong orange-red manganese emission. The quantum efficiency of the red emission is about 55% for calcium phosphate activated with lead and manganese. For both phosphors the quantum efficiency is almost constant between the temperatures —150 °C and +250 °C. The mechanism of excitation by ultra-violet is discussed.

1835: A. van der Ziel: On the mixing properties of non-linear condensers (J. Appl. Phys. **19**, 999-1006, 1948, No. 11).

The theory of a mixer circuit containing a non-linear condenser as a mixing element is developed. It is shown that such a condenser has an imaginary conversion transconductance and a capacitive input and output impedance. It is found that the circuit has widely different properties, depending upon the choice of the frequency, f_m, of the local oscillator, the intermediate frequency f_0, and the input frequency f_i. The following three cases are considered: (*a*) $f_m = f_i + f_o$, (*b*) $f_m = f_i — f_o$ and (*c*) $f_m = f_o — f_i$. A non-linear condenser in a mixing circuit acts as a transformer; in the cases (*b*) and (*c*) the mixing condenser transforms an i.f. impedance into a positive input load; in case (*a*) an i.f. impendance is transformed into a negative input load, so that even oscillations may occur. In case (*c*) the power gain is larger than unity, the power delivered to the circuit by the antenna is smaller than the power dissipated by the output load, and the difference is due to the power delivered by the local oscillator. In case (*b*) the power gain is smaller than unity, because in this case power is dissipated by the local oscillator. In case (*a*) instability may occur; the circuit is then capable of splitting the local-oscillator signal into two signals of frequencies f_i and f_o respectively. The band width and the noise factor of the circuit are also discussed; it is shown that the circuit might have a very low noise factor. A few experiments are given which show qualitative agreement with theory.

1836: K. F. Niessen: The earth's constants from combined electric and magnetic measurements, partly in the vicinity of the emitter (Z. Naturforschung **3a**, 552-558, 1948, No. 8/11).

In this article a method is indicated for the determination of the earth's constants based upon measurements of both the electric and the magnetic field (remote from but also in the vicinity of the emitter). The method is based especially upon the different behaviour of the electric and magnetic fields as a function of the distance in the vicinity of the emitter. Owing to this difference and its mathematical form it will be possible to derive the required constants of the earth by means of one simple system of curves, which may be used again in every other case, as the curves do not depend upon the distance of the points of observation from the emitter. This greatly simplifies the solution of the problem.

1837: W. J. Oosterkamp: Calculation of the temperature development in a contact heated in the contact surface, and application to the problem of the temperature in a sliding contact (J. Appl. Physics **19**, 1180-1181, 1948, No. 12).

The similarity between the problem of heat dissipation in a sliding contact, as treated by Holm and that of the heat dissipation in an X-ray tube anode is pointed out. The anode focal spot corresponds to the contact area, a stationary anode to a stationary contact, and a rotating anode to a sliding contact. The formulae previously obtained are applied to the contact problem and the results, shown graphically, are compared with Holm's (see these Abstracts, Nos. **R 71**, **R 78** and **R 88**).

1838: F. L. H. M. Stumpers: Theory of frequency-modulation noise (Proc. Inst. Radio Engrs. **36**, 1081-1092, 1949, No. 9).

The output energy spectrum of frequency-modulation noise is computed for different ratios of input signal to noise. Numerical values are calculated for some simple filter amplitude characteristics. The theory is based on the Fourier concept of noise and treated in three steps: no signal, signal without modulation, and modulation signal. The result is given in the form of a series, and it is shown that this development is convergent. The suppression of the modulation by noise is also discussed.

1839: W. Elenbaas: The dissipation of heat by free convection from vertical and horizontal cylinders (J. Appl. Physics **19**, 1148-1154, 1948, No. 12).

In this paper a simple deduction is given of the formulae for the dissipation of heat of horizontal and vertical cylinders with cooling by thermal convection. (See these abstracts Nos. **1773***, **R 90** and **R 95**).

1840: P. C. van der Willigen and G. Zoethout: Contact electrodes and applications of contact arc welding (Welding J. **27**, 615-620, Aug. 1948).

This paper describes "Contact" electrodes (C-18 and C-20), which have been derived from standard free arc electrodes (of the E60 12-13 and E60 20-30 classification, respectively). The special characteristics of "Contact" arc welding, described in a previous paper in Welding J. are now demonstrated in applications of vertical down and horizontal-vertical welding with C-18, and groove welding with C-20. The special welding technique, resulting in greater speed of welding, is discussed and the principles on which this technique is based are outlined. A new welding method called contact arc spot welding is also discussed. See these abstracts, No. **1708**.

1841*: F. de Boer: Structure and conductivity of the VI B group of the periodic system (J. Chem. Phys. **16**, 1173-1174, 1948, No. 12).

Referring to an article by von Hippel (J. Chem. Phys. **16**, 372, 1948) the author describes some view points regarding the crystal lattice of trigonal Se and Te. In this lattice the Se and Te atoms are arranged in parallel spiral chains. The structure found cannot be explained by covalent forces in the chains and Van der Waals-London forces between the chains. Additional forces of the metallic type must be present.

1842*: P. J. Bouma: Les couleurs et leur perception visuelle. Introduction à l'étude scientifique des excitations ct sensations de couleur (348 pages, 113 fig., 15 tables; edited by Philips technical Library Dept. 1949).

French translation of "Physical aspects of colour" by the same authoιs. See these abstracts, No. **1768*** and Philips Techn. Rev. **9**, 158, 1947.

1843: A. Claassen: A continuous reading vacuum tube voltmeter for electrometric titrations (Analytica chimica Acta **2**, 602-605, Dec. 1948).

A simple, continuous-reading, mains-operated, vacuum-tube voltmeter for electrometric titrations is described, using two tubes EBC 3 in a Wheatstone-bridge circuit.

1844: W. Elenbaas: Intensity measurements on water-cooled high-pressure mercury lamps with additions of Cd and Zn (Rev. Optique 27, 683-692, 1948, No. 11).

Description of water-cooled high-pressure mercury lamps (500 V, 1.6-2 A, d.c.) with addition of zinc and cadmium. The spectral energy distribution has been measured relative to tungsten (colour temperature 2370 °K and 3700 °K). Curves are given for Hg, 80 Hg + 20 Cd, 75 Hg + 25 Zu, 70 Hg + 13 Cd + 17 Zn (atom percent). By means of a division of the spectrum in blocks (Bouma's method) it is found that the lamp with 75 Hg + 25 Zn approaches most nearly the high intensity carbon arc.

R 97: H. Bremmer: Some remarks on the ionospheric double refraction, I (Philips Res. Rep. 4, 1-19, 1949, No. 1).

After a survey of the general theory some details concerning short waves are worked out. An explicit form is given for Snellius's law determining the direction of propagation. The connection between this direction (called normal) and that of the corresponding ray is derived (*a*) from the mathematical theory of the characteristic surfaces of a partial differential equation, (*b*) from a consideration of Fresnel's indicial surface, (*c*) from Fermat's principle, and (*d*) from the Poynting vector. For a given primary ray 1) the splitting into an ordinary and an extraordinary ray at the entrance into the ionosphere, and 2) the state of polarization, when leaving the ionosphere, are investigated. Finally the corresponding theory concerning the reflection of long waves is summarized and illustrated by a numerical example.

R 98: B. B. van Iperen: On the generation of electromagnetic oscillations in a spiral by an axial electron current (Philips Res. Rep. 4, 20-30, 1949, No. 1).

When an electron current is sent along the axis of a spiral, electromagnetic oscillations of very high frequency may occur in that spiral. A simple small-signal theory of such an oscillator is given, from which the accelerating voltages for maximum oscillation strength are calculated. The results are in agreement with some measurements taken.

R 99: B. D. H. Tellegen: The synthesis of passive two-poles by means of networks containing gyrators (Philips Res. Rep. 4, 31-37, 1949, No. 1).

Any passive two-pole of a certain order may be realized by connecting a resistance between a terminal pair of a passive, resistanceless four-pole of the same order that may contain gyrators. The most general passive two-pole of order *n* can be realized by one network containing the minimum number, $2n+1$, of elements, including one resistance, *n* capacitances and inductances, and *n* ideal transformers and gyrators.

R 100: K. F. Niessen: Relaxation in the anomalous skin effect (Philips Res. Rep. 4, 38-48, 1949, No. 1).

If for infinitely long free paths of the electrons in Pippard's one-dimensional theory of the anomalous skin effect the relaxation is taken into account, a skin impedance is found that is dependent on the conductivity. The influence of relaxation on Pippard's concept of ineffectiveness is considered.

R 101: J. A. Haringx: On highly compressible helical springs and rubber rods, and their application for vibration-free mountings, II (Philips Res. Rep. 4, 49-80, 1949, No. 1).

This paper (a continuation of **R 94**) comprises the calculation of the lateral rigidity of helical springs under compression and their natural frequencies for transverse vibration, starting once more from the simplification, mentioned in the first paper, that the helical spring may be treated as an elastic prismatic rod, provided due allowance is made for the rigidity against shearing. Although here, too, this simplification leads to reliable results, the fact that the spring is actually a helically wound wire structure requires special attention in respect of the spring ends, which are either rigorously fixed or elastically constrained by means of flat-wound end coils. Special precautions, without which the central line of the spring when compressed takes a curved and oblique position, are indicated for the two cases separately. Further, particular attention is drawn to the negative lateral rigidity of helical springs occurring at certain compressions, in connection with its importance for compensating another positive rigidity.

VOL. 11 No. 4, pp. 101-132. OCTOBER 1949

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

CAUSES OF POROSITY IN WELDS

by J. D. FAST.

621.791.056:620.192.34

When consideration is given to the very great importance of electric arc welding in engineering it is surprising that this process should be based upon such scanty scientific grounds. This is manifest in the number of phenomena still unexplained and the surprises, sometimes unpleasant, encountered in practice. With the aid of thermodynamic and metallurgical theories the writer endeavours to throw some light upon one of the most common causes of disappointments in the making of welds.

In electric arc welding under certain conditions the welds are likely to be porous. It may be taken as an established fact that this porosity must be related to the formation of gas bubbles in the metal during the process of welding, but no deeper insight has yet been obtained into the underlying phenomena. For instance, no explanation at all has yet been given for the fact that certain types of welding electrodes are particularly sensitive, as regards porosity, to the presence of relatively small quantities of sulphur in the metal to be welded. For instance from investigations carried out by J. ter Berg [1]) it appears that with type Ph 50 electrodes highly porous welds are obtained if the basic metal contains 0.1% sulphur, whereas with type Ph 56 electrodes even free-cutting steels containing about 0.25% sulphur can be welded without the welds being porous [2]).

In this article we shall outline a mechanism whereby porosity may arise in arc welds. The rate at which gas bubbles are formed and any factors which hinder their formation will play an essential part in this mechanism. Based upon this picture, an attempt will be made to explain the mysterious action of sulphur and some other phenomena hitherto not understood.

The formation of cavities in solidifying metal

The solubility of most gases in metals diminishes as the temperature drops, and when the freezing point of the metal is reached it may even show a sudden drop. This may be seen, for instance, from the curves in *fig. 1* showing the solubility of hydrogen and nitrogen in iron at a pressure of 1 atm.

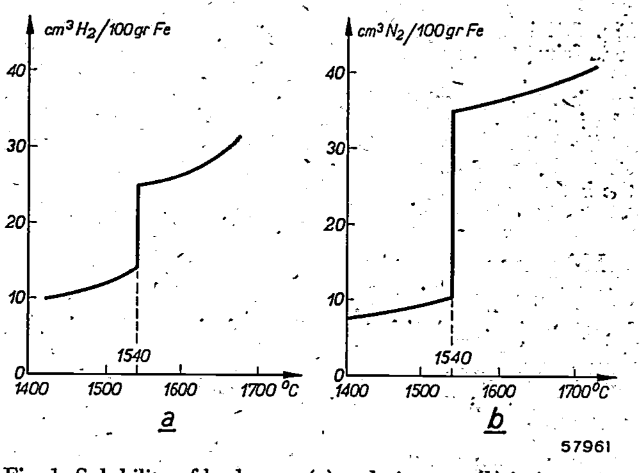


Fig. 1. Solubility of hydrogen (*a*) and nitrogen (*b*) in iron as a function of the temperature, showing the quantities of gas (expressed in volume at 1 atm. and 0 °C) absorbed by 100 grams of iron in atmospheres of hydrogen and nitrogen of 1 atm.

[1]) J. ter Berg, Philips Techn. Rev. 7, 91-93, 1942.

[2]) For a description of the composition of the various types of welding electrodes see J. D. Fast, Philips Techn. Rev. **10**, 114-122, 1948 (No. 4). As appears from this article, the Ph 56 electrode is identical in composition to the older type Ph 55. The only difference is that the coating materials of the Ph 56 are applied around the core in two different layers.

A metal which in the molten state has absorbed gases will therefore usually have to give off some of this gas again when cooling down and especially when freezing. The excess gas present in the metal in the atomic state can leave the metal in two ways: 1) the atoms may diffuse through the liquid or through the metal meanwhile solidified to the outer surface and there recombine into gas molecules; 2) gas bubbles may form in the liquid metal, these growing in size owing to the diffusion of atoms from the surrounding metal, and rising to the surface under the influence of the hydrostatic pressure.

It will depend upon the rate at which these two processes take place and upon the rate of freezing whether cavities occur in the solidifying metal. Some experiments will explain this.

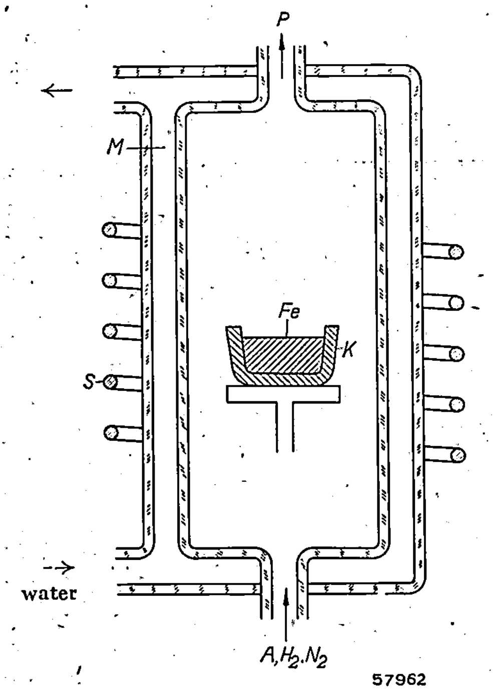


Fig. 2. Diagram showing the arrangement for melting and solidifying small quantities of iron in an atmosphere of certain pure gases. The iron (*Fe*) is contained in a crucible *K* of aluminium oxide placed in a glass vessel with a double wall *M*. The iron is heated by the high frequency method by means of the coil *S*; rapid cooling is brought about by passing a stream of water through the jacket *M*; the gases are led in at the bottom and pumped out at *P*.

100 gram portions of iron were melted in crucibles of refractory oxide by means of high-frequency heating (*fig. 2*), the crucibles being held in an atmosphere of carefully purified hydrogen or nitrogen at 1 atm pressure. The molten iron becomes saturated with gas and is then cooled down. From fig. 1 it can be seen that in the solidification of 100 grams of iron, in equilibrium with hydrogen at 1 atm, 11 cm³ H_2 (volume converted to 0 °C) had to escape from the metallic solution, i.e. 72 cm³ H_2 at the prevailing temperature of 1540 °C. With nitrogen this volume is still greater. Thus in our experiments the volume of the gas liberated at the freezing point is much greater than the volume of the iron itself!

The experiment was repeated in vacuum and in an atmosphere of argon, which is quite insoluble both in liquid and in solid iron.

Four small lumps of iron obtained in this way by melting and rapid cooling are shown in *fig. 3a* and, sawn through, in fig. 3*b*. As was to be expected, the lumps produced in vacuum and in argon showed no cavities or pores, whereas those produced in atmospheres of hydrogen and nitrogen show large cavities. The formation of these cavities could be observed during the experiments through the glass wall of the vessel. Since the molten iron cooled down by radiation and convection, it solidified first on the surface. Owing to the sudden drop in solubility at the freezing point the solidifying iron retains little gas in solution, so that the liquid iron underneath becomes temporarily supersaturated with gas. As a result a vigorous formation of gas bubbles takes place in the liquid iron and it can be observed how at a certain moment the solidified surface is in places pushed upwards. In some of our experiments, repeated several times, the liquid then broke through the protruding crust and flowed over the solid surface with "boiling" phenomena. Thus there may ultimately be not only cavities underneath the surface but also holes in the surface.

In the case of such an "eruption" it may happen that the protrusion first formed melts away again; see *fig. 4a* and *b*. In some experiments this phenomenon of the formation and collapse of the protruding surface even repeated itself several times. This shows a striking resemblence, in miniature, to volcanic eruptions where a liquid lava supersaturated with gas breaks through a solid crust and flows outwards.

In a further series of experiments the molten iron was cooled more slowly by somewhat insulating the crucible (placing it in a second crucible with the surrounding space filled with powdered aluminium oxide). As shown in *figs 5a* and *b*, a lump of iron formed in this way in an atmosphere of hydrogen has but relatively few holes, whereas one formed in nitrogen shows the same phenomena as found in the case of rapid cooling. The explanation of this is that when the metal solidifies more slowly the first process mentioned above — diffusing of the dissolved atoms through the liquid and solid metal followed by recombination at the surface — has a better chance of assisting the removal of the surplus gas. In the case of hydrogen this process takes place

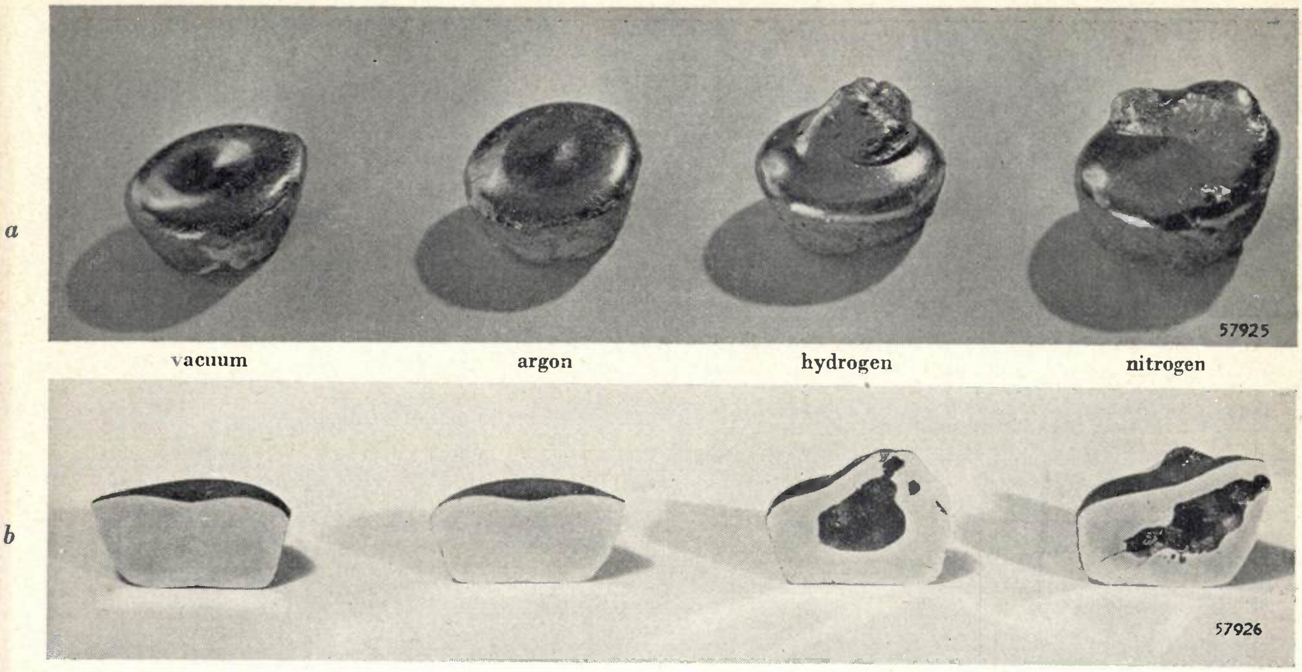


Fig. 3. *a*) Four lumps of iron melted and rapidly cooled respectively in vacuum, in argon, in hydrogen and in nitrogen. *b*) The same lumps sawn through. The lumps obtained by melting in vacuum and in argon show no cavities, whereas those obtained by melting in hydrogen and nitrogen show the traces of strong gas formation in the course of solidification.

relatively quickly [3]), so that this gas can escape from the supersaturated solution without large bubbles being formed. In the case of nitrogen the diffusion is much slower and, moreover, according to fig. 1 the supersaturation is much greater, so that the situation is not appreciably improved by the slower rate of solidification.

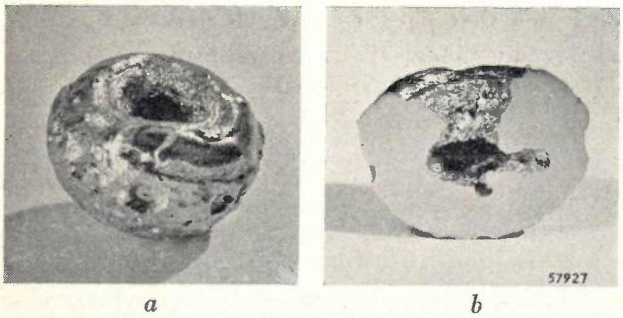


Fig. 4. A lump of iron melted and solidified in hydrogen analogous to that in fig. 3, which behaved as a miniature volcano in the process of solidification and showed the formation of a "crater" and "stream of lava".

The experiments described throw some light upon some of the factors taking part in the formation of pores when welding. But the situation in welding is in several respects different and more complicated. There is one difference which we must discuss at this stage: in welding, the iron deposited does not solidify inwards but outwards, beginning at the interface between the workpiece and the shallow pool of molten metal.

If in this case bubbles are still being formed while the outer layer of metal (adjacent to the floating slag) is about to solidify, there will obviously be holes in the final weld surface or, in a less serious case, small depressions. But even if no more gas bubbles are formed at that stage and the solidified surface therefore shows no defects, there may nevertheless still be internal cavities. To understand this we must look more closely into the actual process of the formation of the gas bubbles.

This takes place at the boundary between the solid and the liquid metal, as we shall see in the following sections. So long as they are small the bubbles adhere to this interface and it is not until they have grown beyond a certain size that they can detach themselves from the interface and rise. It may, however, happen that the bubbles grow more slowly than or at most not much quicker than corresponds to the displacement of the said interface as the process of solidification proceeds. The bubbles are then held in the metal and take the shape of elongated gas inclusions with their longitudinal axis roughly perpendicular to the solidifying surface. This is illustrated in *fig. 6*.

[3] See for instance J. D. Fast, Experiments on the permeation of gases through metal walls, Philips Techn. Rev. 7, 74-82, 1942.

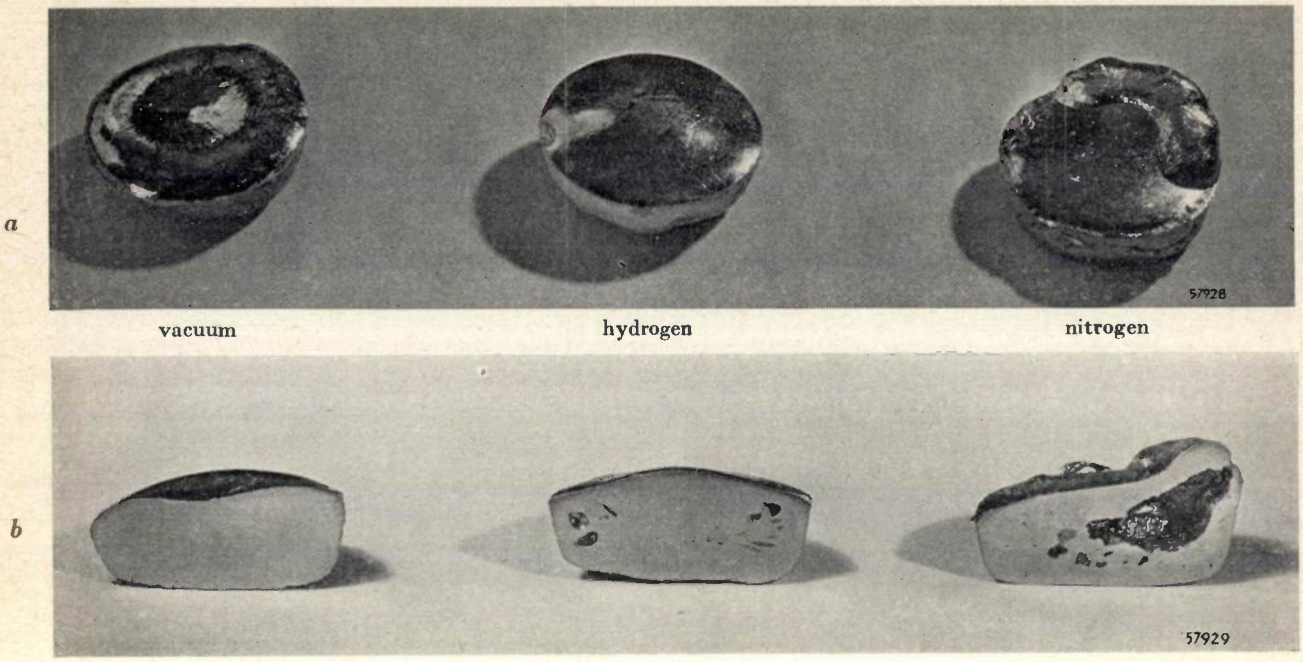


Fig. 5. Three lumps of iron melted and solidified respectively in vacuum, hydrogen and nitrogen. These lumps were cooled down much more slowly than those in fig. 3. In the case of the lump melted in hydrogen most of the dissolved gas was able to escape through diffusion, but nitrogen diffuses too slowly to show any difference compared with the corresponding lump which solidified rapidly.

Experiments where this phenomenon occurred, due to the oxygen content of the deposited metal being reduced by the addition of Al, Ti or Zr, have been described in the article quoted in footnote [2]).

As a particular case, in the race between a growing gas bubble and the advancing interface it may happen that, while the bottom part of the bubble remains held in the metal, at the top where the bubble has already grown considerably in size at a certain moment a part of it breaks away. The liquid metal flowing inwards then causes at that place a constriction in the gas cavity which is meanwhile growing further again. This phenomenon may repeat itself several times, causing the formation of cavities with alternating constrictions and expansions, which may occur not only in welds but also in steel ingots [4]).

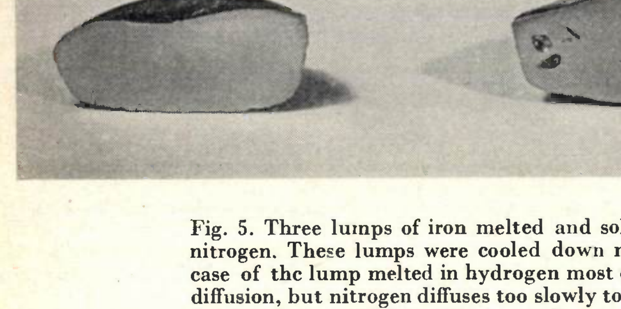


Fig. 6. In welding a pool of molten metal is deposited on the solid metal of the workpiece. During the process of solidification the interface between the solid and the liquid metal moves gradually towards the surface. Gas bubbles are formed preferentially on the interface. If they grow quickly enough then they soon break away and rise to the surface (*a*). If they grow too slowly they are apt to become enclosed by the solidifying metal and then give rise to the formation of cavities (*b*). This process may be promoted by the interface between the solid and the liquid not advancing in a smooth plane, as drawn in the diagram for the sake of simplicity, but more or less irregularly due to the crystals (dendrites) growing in a criss-cross fashion and with many spurs.

These considerations lead to the rather surprising conclusion that porosity in welds may be due not only to an abnormally strong evolution of gas but possibly also, in the case of normal or even subnormal quantities of gas in the metal, to a retardation of the evolution of gas. Such a retardation may on the one hand cause the evolution of gas to be postponed too long, so that it is not completed by the time the outermost layer of metal solidifies, while on the other hand it may delay the growth of the gas bubbles, causing them to remain held in the solidifying metal.

We shall now consider firstly the conditions under which evolution of gas can take place in the deposited metal when welding, and secondly under what conditions there may be a hindrance to the evolution of gas as referred to above.

Conditions for the evolution of gas in the deposited liquid metal

Just as was the case in the experiments already described, in the process of welding the molten

[4]) See also A. Hultgren and G. Phragmen, Trans. A.I.M.E. **135**, 133-244, 1939.

metal tends to become saturated with gas from the surrounding atmosphere, in particular with oxygen and nitrogen when welding in air. Apart from the likelihood of porosity this is undesirable because it spoils the mechanical properties of the weld [5]). The molten metal is therefore protected as far as possible against the attack of the air by means of a coating round the electrodes. This provides a protective layer of slag on the molten metal (and also around the droplets as they are being transferred). This coating, however, must also produce a powerful stream of gas which gives a steady deposition of droplets and a good penetration. For this purpose the coating contains chemically bound gases which are released when welding: bound water, H_2O, and carbon dioxide, CO_2, in the case of type Ph 50 electrodes, and almost exclusively bound CO_2 in the type Ph 56 electrodes. Due to reaction with reducing metal powders in the coating the Ph 50 releases a mixture of CO_2, H_2O, CO and H_2, whilst the Ph 56 produces a mixture of CO_2 and CO. In the case of the Ph 56 the gases thus evolved have little influence upon the deposited metal: a small quantity of oxygen, 0.03%, is thereby added to the 0.08% — 0.10% of carbon already present. In the case of the Ph 50 the situation is much less favourable: a large amount of the hydrogen contained in the gases is absorbed by the liquid metal (more than 0.0014% by weight) and since this type of electrode contains, among other compounds, iron oxide in the coating, also more oxygen is taken up in the deposited metal (more than 0.1%). Furthermore, there is a fairly considerable amount of nitrogen taken up when welding with type Ph 50 electrodes (about 0.03%).

Therefore, when welding with the Ph 50, one obtains liquid iron containing dissolved carbon, oxygen, hydrogen and nitrogen together. Such a solution is only stable if there is an atmosphere over the liquid metal containing, among others, CO_2, H_2O, CO, H_2 and N_2 with certain partial pressures. The values of these partial pressures necessary for stability depend both upon the temperature and upon the quantities of dissolved C, O, H and N. If these equilibrium pressures (which we shall term the "latent" partial pressures of the gases in the metal) are greater than the partial pressures actually present in the gas phase, then gas will be released from the metal. We have already seen that this can happen either through diffusion of the gas atoms to the surface of the metal and the formation of gas molecules at that surface, or through the formation of gas bubbles in the liquid. For the latter case — which is of most interest to us here — there is, however, a second requirement: a gas bubble can only exist in the liquid if its internal pressure (which can never exceed the total "latent pressure" of the dissolved gases) is greater than the total pressure of the gas atmosphere over the metal. If, for instance, an extra pressure of argon of say 1000 atm were applied over the metal, gas bubbles could only be formed with a pressure greater than 1000 atm, which means to say that it would be practically impossible for gases to escape via the bubbles, although there had been no change in the position of the equilibrium between the solution and the gas phase. The equilibrium could only be established via the slow process of diffusion and recombination.

When welding in air the total pressure over the metal amounts to 1 atm. The obvious question is how great is the pressurc available for the formation of bubbles, i.e. the total latent pressure of the dissolved elements in the liquid metal of the weld.

Latent pressure of the dissolved gases in the deposited metal

The concentrations of C, O, H and N in the iron after welding with the Ph 50 and Ph 56 electrodes have already been indicated roughly in the foregoing section. With these relatively small concentrations it may be assumed with a sufficient approximation that as far as the latent pressures are concerned the dissolved elements do not influence each other. In that case, for a temperature immediately above the melting point of iron (1813 °K) these pressures can be calculated from the concentrations with the aid of the formulae [6]):

$$p_{CO}^{1813} = 500\,[\%\,O]\,[\%\,C]\ \text{atm},$$
$$p_{CO_2}^{1813} = 580\,[\%\,O]^2\,[\%\,C]\ \text{atm},$$
$$p_{H_2}^{1813} = 1.92 \cdot 10^5\ [\%\,H]^2\ \text{atm},$$
$$p_{H_2O}^{1813} = 8.8 \cdot 10^5\,[\%\,O]\,[\%\,H]^2\ \text{atm},$$
$$p_{N_2}^{1813} = 520\,[\%\,N]^2\ \text{atm}.$$

For the concentrations in the solidified metal deposited from the Ph 50 let us take the following values which have been found experimentally: 0.06% C; 0.12% O; 0.03% N; 0.0014% H. The

[5]) See for instance P. C. van der Willigen, The mechanical properties of welded joints, Philips Techn. Rev. 6, 97-104, 1941.

[6]) These formulae are derived in an article by the present author shortly to be published in Philips Research Reports. Strictly speaking, also oxygen (O_2), methane (CH_4) and many other reaction products in the gas phase should be included, but their equilibrium pressures are all so small that the presence of these products may be ignored.

Table I. Latent partial pressures of the gases which may be formed from the dissolved elements in the weld metal of the Ph 50 and Ph 56 electrodes.

gas	Ph 50 0.06% C; 0.03% N; 0.0014% H		Ph 56 0.08% C; 0.01% N; 0% H	
	upper limit (0.12% O) atm	lower limit (0.043% O) atm	upper limit (0.03% O) atm	lower limit (0.001% O) atm
1	2	3	4	5
CO	3.60	1.29	1.20	0.04
CO_2	0.50	0.06	0.04	0.00
H_2	0.38	0.38	—	—
H_2O	0.21	0.07	—	—
N_2	0.47	0.47	0.05	0.05
Total latent pressure	5.16	2.27	1.29	0.09

latent partial pressures calculated from these values are given in *table I* column 2; column 4 gives the latent pressures similarly calculated from the concentrations of C, O and N found experimentally in the metal deposited from the Ph 56.

In the case of the gases containing oxygen as a component the pressures indicated are to be taken as the upper limit, for in the deposited metal there are also silicon and manganese, which have a great affinity for the oxygen dissolved in the metal. The influence of these components upon the latent gas pressures (available for the formation of bubbles) depends not only upon the concentrations of Si, Mn, etc. in the metal but also upon the composition and amount of the slag, which likewise contains these elements. Since in the process of welding no equilibrium is established between the metal and the slag, it is difficult to determine this influence quantitatively. We shall, however, certainly get a lower limit for the latent pressures if we take into account only that percentage of oxygen which is found in the form of FeO by chemical analysis of the deposited metal. According to investigations carried out by Andrews and by Sloman, Rooney and Schofield [7]) this figure is 0.043% O (0.192% FeO) in the case of electrodes of the type Ph 50, and only 0.001% O in the case of the type Ph 56. The latent pressures of CO, CO_2 and H_2O calculated from these figures are given in columns 3 and 5 in table I.

Adding up the pressures given in the table for the various gases, we get in the case of the Ph 56 a total latent pressure which is certainly not appreciably more than 1 atm. In the case of the Ph 50, however, we find that the latent pressure available for the formation of bubbles at the moment of solidification will lie between about 2 and 5 atm, thus in any case much greater than 1 atm.

From this it might be concluded that when welding with the Ph 50 there is a vigorous evolution of gas in process at the moment of solidification. So, while we began this article with the problem of accounting for the porosity sometimes found in welds, we now find ourselves faced rather with the question why it is that under normal conditions *no* porosity occurs when welding with the Ph 50 electrode, a fact well known from experience.

Formation of gas nuclei

The explanation is fairly simple. A necessary condition for the evolution of gas bubbles has been mentioned above, namely that the pressure in the bubble must be greater than the total gas pressure (p) above the liquid. But this condition is not sufficient. We have entirely disregarded the fact that the formation of a gas bubble entails a disruption of the cohesion in the liquid metal. The forming of the new surface corresponds to an extra "capillary pressure" which is given by $2\sigma/r$, where σ represents the surface tension of the liquid metal and r the radius of the gas bubble. The total pressure inside a gas bubble must therefore exceed a value of

$$P = p + \frac{2\sigma}{r} \quad . \; . \; . \; . \; . \; . \quad (1)$$

before the bubble can grow. (Strictly speaking, there should also be added the hydrostatic pressure of the column of liquid over the gas bubble, but in the shallow pool of the deposited metal we can ignore this pressure.)

A growing bubble must of necessity pass through an initial stage in which it is extremely small. In this stage, according to the formula (1), the required pressure in the bubble is therefore very great. In principle this already provides the explanation for the fact that even at a latent pressure exceeding 2 atm no bubbles are formed in the metal deposited from the Ph 50.

The pressure that would be sufficient to bring this about cannot be derived directly from formula (1), since it is impossible to apply this formula for atomic dimensions. The conception of surface tension is in fact essentially macroscopic and loses its significance when it becomes a question of atomic dimensions. The formation of the nucleus of a gas bubble consisting only of relatively

[7]) W. Andrews, Trans. Inst. Welding 8, 119-132, 1945. H. A. Sloman, T. E. Rooney and T. H. Schofield, J. Iron & Steel Inst. 152, 127P-153P. 1945.

few molecules has to be treated as a statistical problem. Although the required latent pressure cannot be indicated, something may be said about the factors playing a part in the statistical problem.

One can imagine that now and again gas molecules (CO, H_2, etc.) may be formed in the liquid metal when the atoms migrating in the metal meet under favourable conditions. These molecules will have a very short life, dissociating again into atoms after a short time. For the formation of a very small gas bubble it will be necessary for a gas molecule in the liquid during its short existence to meet a number of other molecules (likewise having a very short life). The chance of the required rapid succession of favourable encounters increases with the concentrations of the atoms and is further proportional to the time available. The question whether or not gas bubbles will be formed depends therefore also upon the rate at which the metal solidifies.

In order nevertheless to get some indication of the latent pressure required for the formation of bubbles, we shall assume, peihaps not unjustifiably, that formula (1) may be applied for a bubble with a radius of 10^{-5} cm. According to the data available the surface tensions of liquid metals are of the order of magnitude of 0.5 newton/m (500 dynes/cm). According to formula (1) the bubble can then only grow if its internal pressure is of the order of 10^7 newton/m^2 (10^8 dyne/cm^2) ≈ 100 atm. Considering the manner in which it is calculated, this value of 100 atm must be regarded as the lower limit for the latent pressure required. It may have to be placed a factor 10 or 100 higher.

Thus the formation of gas bubbles in the interior of the liquid metal is virtually precluded.

This difficulty of the formation of gas nuclei gives rise to the occurrence of many phenomena which have been hardly investigated at all in connection with liquid metals but which are only too well known in the case of water and aqueous solutions. In this connection we would refer for instance to the following experimental evidence. Pure water, from which all "gas nuclei" (minute gas bubbles) previously present have been removed by special treatment, can be heated in a glass vessel with flawless walls to over 200 °C without boiling. A temperature of 200 °C corresponds to a water-vapour pressure of 15 atm, which is apparently still insufficient to disrupt the cohesion of the liquid. Water treated in the same way and placed in a glass vessel under a piston can withstand a tensile stress of even more than 50 atm before a break takes place in the liquid. Further, in flawless glass vessels water can be strongly supersaturated with CO_2 without carbon dioxide bubbles being formed.

With all these phenomena it is essential that the vessel should be of glass and that its walls should be perfectly clean. If the water is contained in a vessel the walls of which exercise a smaller adhesive force upon water than that exercised by glass, then the phenomena occur to a much lesser extent because the cohesion between the liquid and the walls can be broken comparatively easily. In a vessel made for instance of a solid hydrocarbon it is not possible to maintain any marked supersaturation of CO_2 in the water. The bubbles are formed at the "weakest" spot of the system, i.e. at the wall. Similarly, if the formation of bubbles of dissolved gas occurs in liquid iron it will usually take place at the interface between the liquid and the solid. This has been well demonstrated in experiments carried out by Oelsen [8]), where liquid iron was surrounded on all sides by liquid slag. In this case even with a C content of 2% and an O content of 0.035% there was no boiling, that is to say no CO bubbles were formed. The given contents correspond at 1813 °K to a latent pressure of CO of about 13 atm [9])! When an iron rod was immersed in the melt supersaturated with C and O there was a strong evolution of gas, which immediately ceased when the rod was withdrawn.

In the case of welding we shall therefore have to bear in mind that the latent pressure required for the formation of bubbles may be very much reduced at the advancing interface between liquid and solid; this applies particularly for sharp corners and edges of crystals which lie in this interface and which grow as the metal solidifies.

Summary of considerations regarding normal welding with electrodes of the Ph 50 type

In the foregiong it has been shown that when welding with electrodes of the Ph 50 type the dissolved gases have a latent pressure of more than two atmospheres at the moment the weld metal solidifies. The experience that under normal conditions non-porous welds are nevertheless obtained with these electrodes has been explained by the fact that for the formation of gas bubbles an extra pressure is needed which in the case in question exceeds the difference between the latent and the atmospheric pressure, notwithstanding the reduction of the

[8]) W. Oelsen, Stahl und Eisen 56, 182-188, 1936.

[9]) The formula given above for the CO pressure would yield a pressure of 35 atm. With a C content of 2%, however, this formula can no longer be applied because there is then already a strong interaction between C and O in the melt. This interaction reduces the CO equilibrium pressure, which according to the findings of S. Marshall and J. Chipman (Trans. Amer. Soc. Metals 30, 695-746, 1942) in the case of 2% C is given by $p_{CO}^{1813} = 190$ [%O] [%C].

required latent pressure at the interface between the liquid and the solid metal.

This applies, however, only for the end of the solidification, that is to say for the moment at which the outermost layer of deposited metal solidifies. It is almost certain that in the first stage of the solidifying process bubbles are indeed formed, as appears, for instance, from the fact that the carbon content of the weld metal drops from 0.08-0.10% before welding to 0.06% after welding. The difference in carbon (in so far as it is not already oxidised during the melting of the electrode) must have escaped from the pool in the form of gaseous CO and CO_2. The initial formation of bubbles is also plausible for various other reasons. In the first place we have to bear in mind that C, O, H and N are much less soluble in solid iron than in liquid iron (cf. fig. 1), so that the growing crystals are continuously driving ahead C, O, H and N. In the layer of liquid immediately adjacent to the already solidified metal the concentrations of C, O, H and N will consequently not only be greater than those corresponding to the overall composition but even greater than those corresponding to the composition of the liquid outside this layer. In the second place it may be assumed that the contents not only of C but also of O, N and H in the liquid metal — which unfortunately are not known — are greater than those found in the solid iron after welding and which have been mentioned above. As far as hydrogen is concerned this can be derived from measurements taken by Mallett and Rieppel[10]), who found 42 vol. % hydrogen in the gases in the immediate vicinity of the arc when using an electrode corresponding to the Ph 50 type. To the partial pressure of 0.42 atm H_2, a dissolved quantity of 16 cm^3 per 100 gm of iron corresponds, but since the arc produces large quantities of atomic hydrogen, much more than 16 cm^3 H_2 per 100 gm will be absorbed by the liquid iron. Immediately after welding Mallett and Rieppel found no more than 15.5 cm^3 H_2 per 100 gm in the solidified metal. The difference must, at least for the greater part, have escaped from the metal in the form of gas bubbles while the metal was solidifying.

Going back to the picture we formed in the beginning of this article about the formation of holes duc to gas bubbles, we may describe the situation when welding with electrodes of the Ph 50 type as follows: while the metal is solidifying there is some development of gas bubbles but under normal conditions this development ceases and the gas bubbles disappear before the outermost layer of the weld metal solidifies. By the time this takes place the concentrations of the dissolved gases have diminished to such an extent that the latent pressure no longer reaches the value required for the formation of gas bubbles.

The influence of sulphur

As an "abnormal condition" under which welds made with electrodes of the Ph 50 type *do* show porosity we mentioned in the beginning the presence of a relatively small quantity of sulphur in the iron to be welded. Can we now understand why there should be porosity in such a case?

Sulphur is practically insoluble in solid iron, so that crystals growing in liquid metal containing sulphur are continually driving the sulphur ahead. In other words, during solidification the layer of liquid immediately adjacent to the solid metal will contain an abnormally high concentration not only of C, O, H and N but also of sulphur. In the picture we formed above concerning the evolution of gas the gas nuclei are formed at the advancing solid-liquid interface. The bubbles probably develop from a small layer adsorbed on the wall, mainly on the sharp corners and edges of the growing crystals. We must now imagine that these favourable ("active") places for the formation of gas nuclei are blocked up by the sulphur, since this element is preferably adsorbed there. This "poisoning" of the interface causes the evolution of gas to be delayed: the bubbles are not formed until in a later stage of the solidification, when the supersaturation in the still liquid material has become so great that the formation of bubbles can no longer be prevented by the increase of the required latent pressure due to the lack of full cooperation of the poisoned interface. Thus this may lead to the formation of bubbles in the last stage of solidification, causing porosity. This would explain the sensitivity of the Ph 50 towards sulphur.

With type Ph 56 electrodes the situation is more favourable because, as appears from table I, the latent pressure of the dissolved gases is much lower. The supersaturation in the liquid metal remains so small that there is probably no formation of bubbles at all either at the beginning or at the end of the solidification, so that the presence of sulphur is not expected to have any effect.

It is probably due to the entire absence of bubbles that the Ph 56 electrodes lend themselves so well for overhead welding. The Ph 50 type of electrodes are not at all suitable for

[10]) M. W. Mallett, Welding J. 25, 396S-399S, 1946; M. W. Mallett and P. J. Rieppel, Welding J. 25, 748S-759S, 1946.

welding in this position, owing partly to the too low viscosity and surface tension of the metal deposited with that type of electrode, but probably also in part on account of the formation of bubbles at the beginning of freezing; the fact is that in overhead welding the bubbles are not able to rise and escape (fig. 6).

The explanation given for the sensitivity of electrodes of the Ph 50 type for sulphur is supported by a number of experiments and many known facts, some of which will be mentioned.

Adsorption at "active" places on a surface of metal is also found in the phenomenon of heterogeneous catalysis. Now the action of sulphur in welding does not appear so strange when remembering that sulphur is known to be a catalyst poison. As far back as the first half of the last century it was known that the catalytic action of finely divided platinum upon the reaction between oxygen and hydrogen is eliminated by the addition of small quantities of sulphur compounds to the gas mixture. A more recent example is the dissociation of gaseous molybdenum carbonyl Mo $(CO)_6$ on hot surfaces [11]). The layers of molybdenum obtained generally contain carbon owing to the dissociation of a part of the CO released. The precipitation of C can, however, for the greater part be avoided by adding a little H_2S to the carbonyl. This is probably caused by poisoning through the adsorption of sulphur on the active spots of the surface at which normally the dissociation of CO takes place. In welding, it is true, we are concerned with the formation of CO and not with its dissociation, but since a catalyst cannot change the position of an equilibrium the poisoning will delay both the formation and the dissociation of CO.

This phenomenon is likewise encountered in the manufacture of steel. Often silicon or aluminium is added to the liquid steel to "kill" it: the oxygen present is chemically bound and as a consequence the evolution of CO is prevented and the steel solidifies without gas bubbles being formed. Now it is known that sulphur also has a killing action: a quantity of 0.1% S is sufficient to suppress almost entirely the evolution of CO during the solidification [12]). There is a very striking analogy with the effect of sulphur when welding.

We shall not go into various other cases here, some of which have been known for a long time, in which the formation of gas molecules (particularly of hydrogen) is prevented by catalyst poisons.

The fact that the dissolved gases are responsible for the sulphur-sensitivity of the Ph 50 electrode is demonstrated in *fig. 7*, showing the result of welding tests on an iron plate free of carbon but with a high sulphur content (0.25%). Welding with a normal Ph 50 electrode, the iron core of which contains 0.1% C, gave the highly porous bead at the top left-hand side of the picture. With a carbon-free core the less porous bead at the top right-hand side was obtained. The fact that the latter bead still shows porosity is due to the presence of hydrogen. The two lowermost beads are much less porous and were obtained after heating two of the electrodes (with and without C in the core) for two hours in a stream of argon at 850 °C, the bound hydrogen being thereby driven out almost completely.

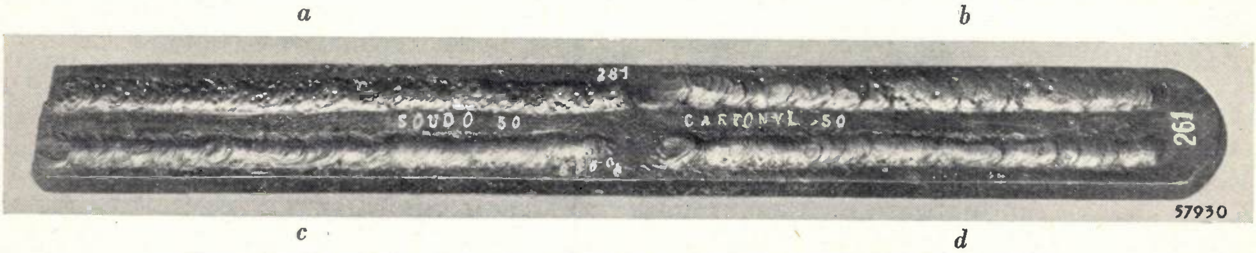


Fig. 7. Beads welded onto an iron plate free of carbon but containing 0.25% sulphur, *a*) with a mornal Ph 50 electrode; *b*) with a Ph 50 electrode the core of which contained no carbon; *c*) with a Ph 50 electrode from the coating of which the hydrogen had been expelled; *d*) with a Ph 50 electrode in which the conditions (*b*) and (*c*) were combined

Welding under elevated pressure

In all the foregoing considerations the external pressure while welding has been assumed to be 1 atm. It is worth while considering what would happen if this pressure were greater than 1 atm.

The elevated pressure will have relatively little effect upon the amounts of gas taken up by the liquid metal: the nitrogen content of the metal deposited

[11]) J. J. Lander and L. H. Germer, Metals Technology 14, T.P. 2259, September 1947.

[12]) Second report on the heterogeneity of steel ingots, Iron and Steel Institute, London, 1928.

will certainly increase (proportional to the square root of the external pressure) but the oxygen content, which according to table I forms by far the most important source for any subsequent evolution of gas, depends mainly upon the composition of the electrode coating, or of the slag.

The latter is borne out i.a. by experience when welding with "Contact" electrodes. These form a longer cup and thus give a better protection to the depositing material against the influence of the atmosphere; the nitrogen content of the weld metal appears in fact to be less than that in welds made with the corresponding normal electrodes. However, one still finds the same oxygen content.

Assuming that the ideas developed in this article regarding the cause of porosity are correct, it may be expected that an increase of the external pressure might lead to the same result as the presence of sulphur, since it would make the formation of gas nuclei more difficult, whilst, as we have just seen, it does not cause any appreciable increase in the latent pressure available in the liquid metal. If the external pressure is very high it may even entirely prevent the formation of nuclei, as already stated. But if it is not too high the arresting of the formation of nuclei may result in the escape of gases being only delayed, possibly until the critical last stage is reached in the solidification, and in that case porosity is to be expected.

This effect has indeed been known for some ten years already as an experimental and hitherto unexplained fact: with electrodes of the Ph 50 type porous welds are obtained if the total external pressure amounts to say 2 or 3 atmospheres, as was experienced in the construction of the tunnel under the river Meuse at Rotterdam, where welding had to be done in the caissons under such a high pressure [13]).

In conclusion it seems advisible to recall what was said in the introduction to this article: an attempt was to be made to indicate a mechanism for the phenomenon of porosity in welds. The critical reader will have noticed that the picture developed is not in all details based upon exact data but rather upon plausible theories. The only justification for our considerations lies therefore in the fact that as yet nothing better seems to be available on the subject. We have formulated a working hypothesis; the quantitative development of the theory of the phenomenon may be a long and difficult task.

[13]) W. Gerritsen and F. G. van Riet, Smit-Lastijdschr. 3, 2-12, 1939. According to a verbal statement made by the firm of Smit, no porosity occurs when welding under elevated pressure with electrodes of the Ph 55 and Ph 56 types!

Summary. When welding with electrodes of the type Ph 50 the liquid metal unavoidably absorbs certain quantities of oxygen, hydrogen and also nitrogen. In the solidification of the metal (already containing carbon in solution) CO, CO_2, H_2, H_2O and N_2 are released in the form of gas bubbles. This liberation of gases from the metal, however, takes place, under normal conditions, before the outer layer solidifies. The "latent pressure" of the said gases in the metal is then, it is true, still greater than 1 atm (at least 2.27 atm, as found from a close consideration) but the formation of gas nuclei is prevented because, notwithstanding the cooperation of the "active places" of the interface between the solid and the liquid metal, this requires an additional pressure (capillary pressure); as a rule such a great latent pressure is no longer present at the end of the solidification, when most of the gas has already escaped. If there is any hindrance to the formation of gas nuclei, either through sulphur blocking the active places (poisoning) or in consequence of increased external pressure or through any other causes, then the formation of gas bubbles at the beginning of the solidification can be suppressed; it then takes place at the end of the solidification owing to the very much greater supersaturation. Such a "delayed" evolution of gas may account for porosity in a weld and in particular for the known sensitivity of electrodes of the Ph 50 type for sulphur. Also other facts, such as the insensitivity of the Ph 56 electrode to sulphur, fit well into the picture, but many other problems remain to be investigated further.

AN EASILY PORTABLE CATHODE-RAY OSCILLOGRAPH

by E. E. CARPENTIER.

621.317.755

For a number of years the cathode-ray oscillograph has been a widely known and indispensable instrument for the investigation of electrical phenomena (wave form, amplitude, frequency, phase shift, modulation, etc.). Much more interest has been taken in the oscillograph, however, since it came to be realized that mechanical and other non-electrical quantities can easily be converted into electric voltages. But the weight and dimensions of an oscillograph have often formed an obstacle to a wider use of this instrument. In this article the author explains how he designed an exceptionally small and light (and thus cheaper) apparatus which is eminently suitable for all sorts of routine and service jobs.

A few years ago there appeared in this journal a description of a cathode-ray oscillograph (type GM 3159) which was distinguished from older types by the fact that it contained two amplifiers, one for vertical and one for horizontal deflection [1]), thus making it more generally useful.

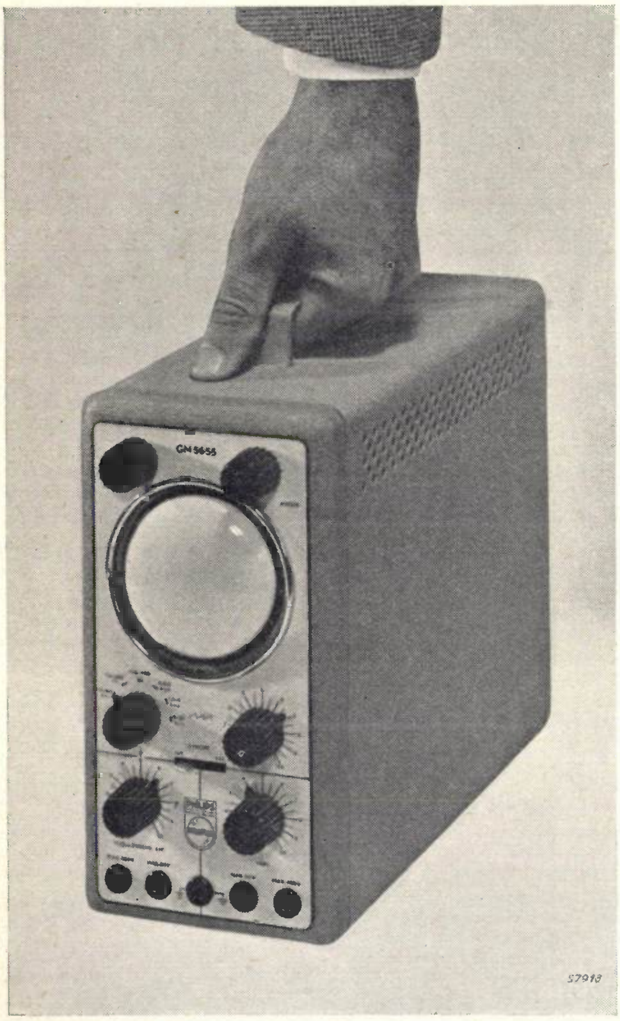


Fig. 1. Front panel of the small cathode-ray oscillograph GM 5655 (dimensions 11.5 cm × 29.5 cm × 24 cm or $4^{1}/_{2}''$ × $11^{1}/_{2}''$ × $9^{1}/_{2}''$, weight 6.4 kg or 14 lbs). At the top are the knobs for adjusting the intensity of the beam and for focusing. Below the screen of the cathode-ray tube: knobs for controlling the time-base frequency step by step and continuously, underneath these the input potentiometers for the two directions of deflection, and right at the bottom the terminal sockets. Qver the Philips emblem is the "Intern."/"Extern." switch (see text).

Although this oscillograph was much smaller and lighter than its predecessors, the need was still felt for a type that could be more easily carried about. The fact is that an oscillograph has gradually become an indispensable instrument for maintenance work and for the tracing of defects in telephone communications, amplifying or radio installations, etc., and for such purposes easy portability is of great value.

To meet this need a new type of oscillograph (GM 5655) has been developed which will be described in this paper and compared in various respects with the type GM 3159 mentioned above. This small oscillograph, illustrated in *fig. 1*, has the following dimensions: width 11.5 cm, depth 29.5 cm, height 24 cm ($4^{1}/_{2}''$ × $11^{1}/_{2}''$ × $9^{1}/_{2}''$). It weighs 6.4 kg (14 lbs). Its volume (8.2 dm^3) is only 40 % and its weight 50 % of that of the GM 3159. *Fig. 2* gives a view of the inside of this apparatus.

Naturally it has not been possible to make such a drastic reduction in size without some sacrifice in performance compared with the type GM 3159, but it was not found necessary to make very great concessions — as will be shown farther on — and this small oscillograph will therefore also find a wide field of application.

General construction

Just like the GM 3159, this small oscillograph is fitted with a cathode-ray tube with a screen of 7 cm diameter, so that nothing has been sacrificed in picture size, and it has two amplifiers, the usual one for vertical deflection and another for horizontal

[1]) E. E. Carpentier, A cathode-ray oscillograph with two push-pull amplifiers, Philips Techn. Rev. 9, 202-210, 1947.

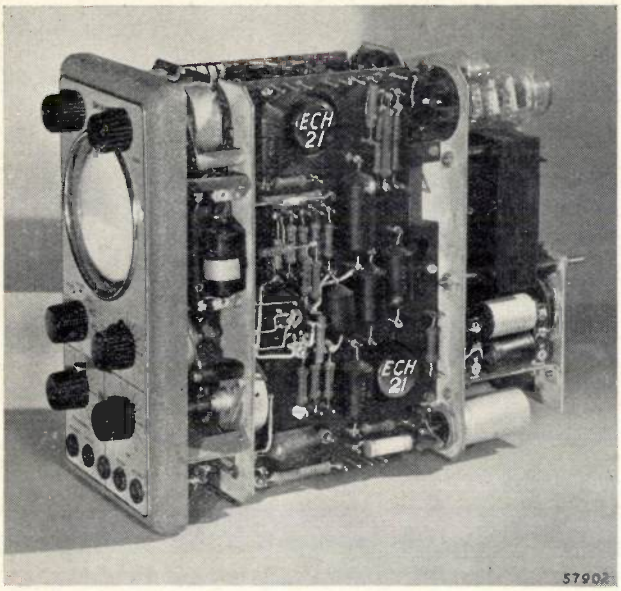


Fig. 2. The inside of the oscillograph GM 5655, showing two of the three ECH 21 valves and (on the right at the top) the two rectifying valves (EZ 2).

deflection. This second amplifier is often useful when studying a voltage as a function of some other arbitrary voltage (Lissajous figures); thanks to this second amplifier the latter voltage need have only a small amplitude to give quite a wide picture.

But this second amplifier also has its advantages for studying a voltage as a function of time, for the following reasons. If a reasonably linear saw-tooth voltage, required for the time base, is to be generated directly and with sufficient amplitude to give a picture of the full width allowed by the screen, this involves a rather complicated circuit. A saw-tooth voltage with sufficient linearity can, however, be generated in a very simple manner if its amplitude is small, but then an amplifier is needed to give it the amplitude desired. When such an amplifier is already to hand, as it is here, this method is by far preferable to the other.

As shown by the block diagram (*fig. 3*) of the GM 5655, an amplifier (A_v, A_h) is permanently connected to each pair of plates of the cathode-ray tube. When it is desired to study a particular voltage as a function of some other arbitrary voltage these voltages are applied to terminals *I* and *II* respectively and the switch S_1 is put in the position *1*. If, on the other hand, a voltage is to be studied as a function of time it is likewise applied to terminal *I* but S_1 must then be in the position *2*. Via the amplifier A_h the generator *TB* then supplies a saw-tooth voltage required to get a horizontal deflection proportional to the time. This saw-tooth voltage, the frequency of which is adjustable between wide limits, can be synchronized either with the voltage bringing about the vertical deflection or with some other arbitrary voltage. In the former case the switch S_2 has to be in the position "Intern.", synchronization then being brought about by the output voltage from A_v, whereas in the latter case S_2 is put in the position "Extern." and the synchronizing voltage is applied to the terminal *II*. Up to this point the arrangement of the oscillograph GM 5655 corresponds on broad lines to that of the larger type GM 3159.

It might be asked how it has been possible to reduce the dimensions and weight of this apparatus. Some concessions have had to be made as regards the quality of the picture and the frequency range for which the oscillograph is suitable, and from what follows it will be seen how by making such concessions it has been possible to reduce the volume and weight of this oscillograph.

Let us first consider the quality of the picture.

Symmetrical control of the cathode-ray tube has been dispensed with in this small oscillograph. As a result some distortion takes place through focusing errors [2]) and the light spot is not equally sharp all over the screen. In the case of the type GM 3159 symmetrical control has been chosen because good picture quality is very valuable for certain measurements, but for a great many investigations focusing errors are permissible provided they do not occur to any great extent.

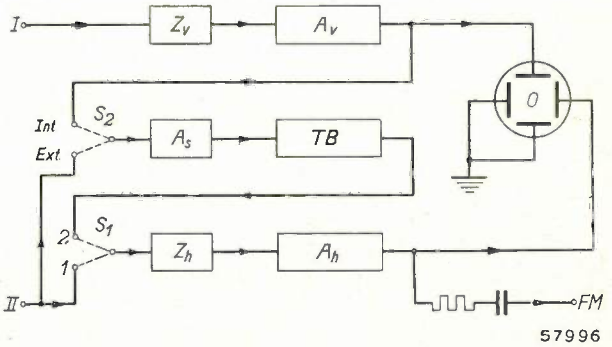


Fig. 3. Block diagram of the small oscillograph. *I*, *II* = input terminals, Z_v and A_v = attenuator and amplifier for vertical deflection, Z_h and A_h = ditto for horizontal deflection, *O* = cathode-ray tube, *TB* = time-base apparatus (saw-tooth generator), A_s = amplifier for the synchronization voltage, S_1, S_2 = switches, *FM* = connection for the frequency modulator GM 2881 (see the last paragraph of the text).

The advantage of asymmetrical control lies in the fact that it does not need a push-pull circuit, so that one valve can be saved in each amplifier. We shall now consider the amplifiers in more detail.

[2]) See Philips Techn. Rev. 4, 198-204, 1939, particularly figs. 3 and 4, page 199.

Amplifiers

With reference to the article quoted in footnote [1]), we would recall that the amplifiers of the oscillograph type GM 3159 each consist of a push-pull stage (without pre-amplification) which has to provide an amplification of 560 times to give the oscillograph the sensitivity required (corresponding to an input voltage of 25 mV R.M.S. per cm picture height). To maintain this high amplification factor in a wide frequency range it is necessary to have — as explained in the article referred to — a high supply voltage and a considerable supply current (675 V × 35 mA ≈ 24 W for the anodes and screen-grids of the four EF 50 valves together).

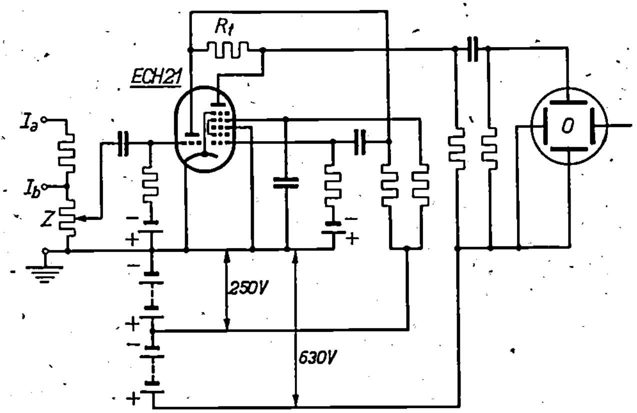


Fig. 4. Circuit diagram of the amplifiers of the oscillograph GM 5655. Each amplifier contains one ECH 21 valve. I_a = input for voltages up to 300 V, I_b = input for voltages up to 50 V, Z = input potentiometer, R_t = negative feedback resistor, O = cathode-ray tube (DG 7-2). The H.T. (250 V + 380 V) is supplied by two rectifiers in series, the negative grid biases are derived from a small selenium rectifier.

As regards the power a great saving has been reached in the amplifiers of the oscillograph GM 5655. Each of these consists of two stages (*fig. 4*) employing respectively a triode and a heptode which are incorporated in one bulb (ECH 21). The frequency range being limited to about 100 kc/s (in the case of the GM 3159 it extends to 500-1000 kc/s), only 4 mA is required in the output stage, with a supply voltage of 630 V. Owing to the smaller voltage amplitudes in the driver stage this stage needs no more than 250 V supply, again with a current of 4 mA. Thus the total D.C. power of all the amplifiers together amounts to 2 × (250 + 630) V × 4 mA ≈ 7 W, which is only 30 % of the corresponding power in the case of the GM 3159. This means a considerable saving in weight, not only of the mains transformer (the heaviest component of the oscillograph) but also of the chassis on which the transformer is mounted. The smaller size of the transformer and the saving of two valves and minor accessory parts result in a saving in volume, which in turn allows of a lighter cabinet. Obviously all this results in a lower price too.

If no negative feedback were applied the amplification of the first stage (triode) would be 12 times and that of the output stage (heptode) 400 times, thus making in all 4800 times. In order to get a flat frequency characteristic and a constant amplification whilst at the same time minimizing non-linear distortion, negative feedback has been applied which reduces the amplification to 500 times. With this amplification the sensitivity (corresponding to 30 mV per cm picture height) is hardly less than that of the oscillograph GM 3159. The negative feedback is brought about by means of a network represented in fig. 4 for the sake of simplicity as a resistor (R_t) but in reality having such a frequency-dependence as to improve the frequency characteristic of the amplifiers (*fig. 5*). Between 6 c/s and 100,000 c/s the amplification is within +5 % (0.4 db) and —25 % (—2.5 db) of the nominal value.

On the input side of each of the amplifiers there are two attenuators, one of which (Z, fig. 4) is continuously variable in the ratio 1 : 10,000 and can withstand 50 V, whilst the other, allowing the input voltage to be raised to 300 V, consists of a fixed resistor. For measuring phase relations the amplifiers can be used up to a frequency of 10,000 c/s when only the continuous attenuators are used, or to a frequency of 2500 c/s when the fixed attenuators are also in use; up to these limits the phase-shifts in the two amplifiers and the corresponding attenuators are equal and negligible.

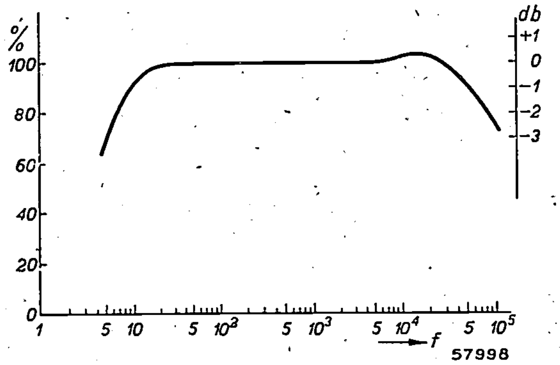


Fig. 5. Frequency characteristic of the amplifiers of the oscillograph GM 5655. The nominal amplification is taken as 100 %. On the right is a scale (in db) showing the deviation from this rating.

Time-base

The saw-tooth voltage for the time-base is generated in the same way as in the GM 3159, i.e. by a squegging oscillator. The valve used in this oscillator is of the same type as that employed in the amplifiers, an ECH 21, the heptode part of which is used for this purpose (*fig. 6*).

In connection with the afore-mentioned frequency range of the amplifiers the limits between which the time-base frequency is continuously adjustable have been chosen at 15 c/s and 20,000 c/s.

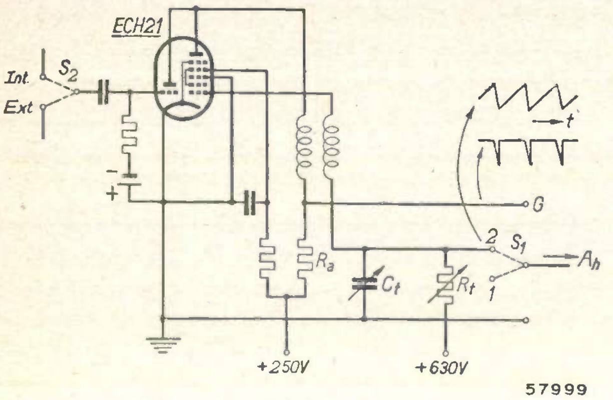


Fig. 6. Circuit diagram of the saw-tooth generator and the amplifier for the synchronizing voltage. The anode circuit of the heptode part of the ECH 21 valve is so tightly coupled to the grid circuit that a state of squegging arises. Periodically the grid capacitor C_t is thereby gradually charged through the resistor R_t and rapidly discharged across the grid. The saw-tooth voltage across C_t can be applied to the amplifier A_h. During the rapid discharge of C_t the valve oscillates for a moment, thereby taking up anode current and causing a sudden voltage drop in the resistor R_a. The voltage pulses thus obtained at the terminal G are used for blocking the electron beam in the cathode-ray tube during the flyback. The triode part of the ECH 21 valve serves as amplifier for the synchronizing voltage applied via the switch S_2 (cf. fig. 3).

A purpose has also been found for the triode part of the ECH 21 valve used in the time-base circuit: it serves as amplifier for the voltage with which the time-base can be synchronized [3]). As a result a smaller voltage suffices for synchronization than is required in the GM 3159, which lacks this amplifying stage.

During the flyback the electron beam in the cathode-ray tube is blocked, thus making the oscillogram clearer. As is the case with the GM 3159, this is brought about supplying to the control electrode of the cathode-ray tube during each flyback a negative voltage pulse derived from a resistor (R_a, fig. 6) in the anode circuit of the saw-tooth generator.

Use of the oscillograph at frequencies higher than 100 kc/s

Voltages with frequencies higher than 100 kc/s can also be investigated with this oscillograph provided they are amplitude modulated. In that case an auxiliary apparatus, the so-called measuring head GM 4575 (*fig. 7*), is needed for rectifying the voltage applied. This is useful for tracing defects in and testing radio sets, carrier telephony installations, etc.

So as not to be dependent upon some particular transmitter when taking such measurements of radio receivers, it is advisable to connect the aerial terminal of the receiver being tested to a service oscillator supplying a high-frequency voltage of variable amplitude and frequency. The modulation depth being known, the low-frequency voltage obtained after detection gives an indication of the value of the high-frequency voltage (without modulation).

The auxiliary apparatus referred to consists mainly of a small "Rimlock" pentode EF 41 acting as grid detector and contained in a case of "Philite". At one end this has a metal probe and at the other end a cable which has to be plugged into a socket at the back of the oscillograph. The neces ary supply voltages are fed to the valve EF 41 via this cable, whilst its output voltage is connected to the vertical deflection amplifier. The circuit diagram is reprcsented in *fig. 8*.

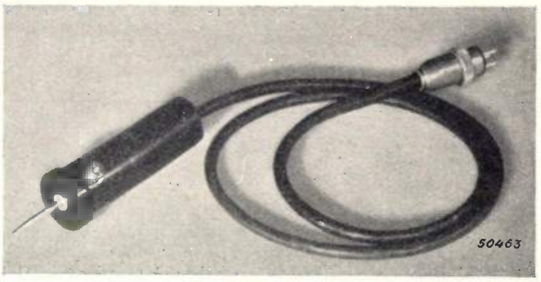


Fig. 7. Auxiliary apparatus (GM 4575) with which an oscillogram can be made of the audio-frequency voltage modulated upon a high-frequency voltage.

When the probe is connected in succession to, say, the input and the output of a certain h.f. or i.f. stage, with a suitably chosen sensitivity one can determine the amount of the amplification from the ratio of the amplitudes of the two oscillograms. A rough comparison of the amplitudes is often sufficient to localize a defect very quickly.

A switch on the measuring head allows a choice of two sensitivities corresponding to 65 and 650 mV (R.M.S. value) per cm picture height with a modulation depth of 30 %. Furthermore, the input

[3]) A description was recently given in this journal of a circuit in which the two systems of an ECH 21 valve are similarly used for a saw-tooth generator and for the synchronization of a saw-tooth voltage (see J. Haantjes and F. Kerkhof, Projection-television receiver, V. The synchronization, Philips Techn. Rev. **10**, 364-370, 1949, No. 12) but in which the functions of the triode and heptode parts were just the reverse of those in the present case. This is related to the fact that for separating television synchronization signals a multiple-grid valve is needed (as fully described in the article referred to). In the case of the oscillograph a reversal of the functions is more satisfactory.

potentiometer of the amplifier incorporated in the oscillograph is available for continuous adjustment of the sensitivity.

The frequency range of the measuring head extends to about 30 Mc/s.

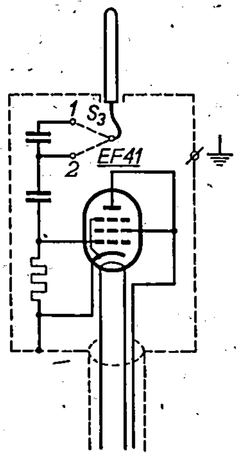


Fig. 8. Circuit diagram of the auxiliary apparatus GM 4575. The pentode EF 41 (connected as triode) serves as detector. S_3 = switch for low or high sensitivity (positions *1* or *2*).

Another purpose for which this small oscillograph can quite well be used in taking measurements on high-frequency apparatus is the recording of frequency characteristics (e.g. resonance curves). The manner in which this is done has already been explained at length in this journal [4]). All we need say here is that a high-frequency voltage modulated in frequency with the low-frequency saw-tooth voltage of the time-base of the oscillograph is applied to the receiver. This time-base is synchronized for instance with the mains frequency. The frequency modulator required (e.g. type GM 2881) has to be controlled by the saw-tooth voltage of the time-base. For this reason the oscillograph GM 5655 is provided with a connection (*FM*, fig. 3) from which this voltage can be taken off.

Summary. A description is given of an exceptionally small and light cathode-ray oscillograph (type GM 5655; 11.5 cm × 29.5 cm × 24 cm or $4^1/_2''$ × $11^1/_2''$ × $9^1/_2''$; 6.4 kg or 14 lbs). The small dimensions and light weight have been obtained by means of an asymmetrical control of the two pairs of deflecting plates of the cathode-ray tube and by limiting the frequency range for which the oscillograph is suitable to 100 kc/s. This has reduced the D.C. power required to only 3.5 W per amplifier. — The oscillograph contains two amplifiers, one for each direction of deflection. The horizontal deflection amplifier serves for amplifying either a saw-tooth voltage (as time-base) or some other arbitrary voltage (Lissajous figures). The two amplifiers are identical and each contains a triode-heptode ECH 21, with the two systems connected in cascade. For vertical deflection 30 mV (R.M.S. value) per cm picture height is required at the input. — In the time-base circuit (frequency 15 c/s to 20,000 c/s) an ECH 21 valve is likewise used, the heptode part of which is used for the saw-tooth generator whilst the triode part amplifies the synchronizing voltage. With the aid of an auxiliary apparatus (GM 4575) the amplitude modulation of a high-frequency voltage can be shown (for the testing of carrier telephony apparatus, radio sets, etc.). The oscillograph described is also suitable for recording frequency characteristics.

[4]) H. van Suchtelen, Application of cathode-ray tubes in mass production, Philips Techn. Rev. 4, 85-89, 1939.

CONSTRUCTION AND APPLICATIONS OF A NEW DESIGN OF THE PHILIPS VACUUM GAUGE

by F. M. PENNING and K. NIENHUIS

531.788.732

The Philips vacuum gauge, with which the pressure of a gas is measured by means of the intensity of a gas discharge in a magnetic field, was introduced 12 years ago and has now found application in many fields and proved to be a most useful instrument. Meanwhile the need has also been felt for such a simple vacuum gauge for measuring still lower gas pressures. This need has been met by changing the construction of the present gauge. In its new design this instrument is also quite useful as leak detector.

A description of the Philips vacuum gauge was given in this journal 12 years ago [1]). The tube used in this gauge contains two parallel plates of zirconium as cathode and a ring as anode (*fig. 1a*) and is placed between the poles of a permanent magnet with its field perpendicular to the cathode. Owing to the peculiar course of the electrical and magnetic lines of force inside the tube the electrons travel to and fro an immense number of times before reaching the anode, thus making many more collisions with gas molecules than would be the case if they travelled directly from the cathode to the anode. This large number of collisions can be interpreted as being derived from an apparent elevation of the gas pressure.

Consequently it is possible for a gas discharge to take place between cold electrodes at a pressure about 1000 times lower than is possible in a tube of a similar construction without magnetic field. It has been found experimentally that the magnitude of the discharge current is a good measure for the pressure.

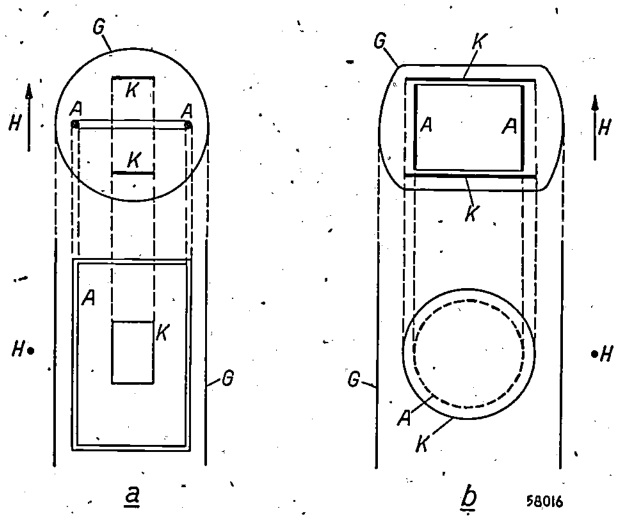


Fig. 1. Diagrammatic representation of the position of the cathode K and the anode A of the Philips vacuum gauge. G represents the glass envelope and H the direction of the magnetic field. *a*) The older design, *b*) the new design.

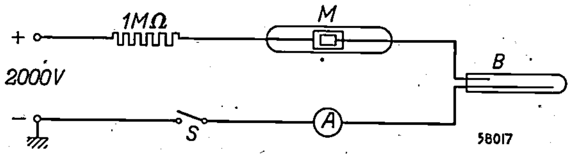


Fig. 2. The circuit diagram of the vacuum gauge M in series with the glow lamp B, acting as current indicator, and a micro-ammeter A. This circuit is applied both in the old design and in the new one.

This vacuum gauge, the circuit diagram of which is given in *fig. 2*, has in recent years been extensively applied in the manufacture of electronic valves and in apparatus in which the vacuum has to be maintained by means of a pump, as is the case with two kinds of apparatus previously described in this journal: the neutron generator [2]) and the electron microscope [3]). It can also be used, as described in literature on the subject, in combination with a "Pirani" manometer [4]) or with a thermocouple manometer [5]), in which case a larger pressure range is covered. Further, this instrument is very suitable for starting a certain signal [6]) or activating a relay [4]) [7]) at the moment that the pressure exceeds a certain limit.

[1]) F. M. Penning, High-vacuum gauges, Philips Techn. Rev. 2, 201-208, 1937.

[2]) F. A. Heyn and A. Bouwers, An apparatus for the transmutation of atomic nuclei, Philips Techn. Rev. 6, 46-53, 1941.

[3]) A. C. Dorsten, W. J. Oosterkamp and J. B. le Poole, An experimental electron microscope for 400 kilovolts, Philips Techn. Rev. 9, 193-201, 1947.

[4]) N. C. Picard, P. C. Smith and S. M. Zollers, A reliable high vacuum gauge and control system, Rev. sci. Instr. 17, 125-129, 1946.

[5]) R. I. Garrod and K. A. Gross, A combined thermocouple and cold-cathode vacuum gauge, J. sci. Instr. 25, 378-383, 1948.

[6]) R. S. Mackey, Non-linear indicator for vacuum gauge, Electronics, Febr. 1948, p. 140.

[7]) H. A. Thomas, T. W. Williams and J. A. Hipple, A mass spectrometer type of leak detector, Rev. sci. Instr. 17, 368-372, 1946; Detecting vacuum leaks electronically, Westinghouse Engineer, July 1946, p. 108.

For such applications the pressure range of 2×10^{-3} to 10^{-5} mm Hg covered by this vacuum gauge is generally sufficient, but in the manufacture of valves which are exceptionally sensitive to gas still lower pressures have to be measured.

The new vacuum gauge

The new vacuum gauge specially designed for lower pressures has a different electrode system (fig. 1*b*). The cathode plates are round discs parallel to the axis of the glass tube, whilst the anode is in the form of a cylindrical jacket perpendicular thereto. Where the cathode plates are mounted the glass tube is pinched. Thanks to this construction the same permanent magnet can be used with the new vacuum gauge as with the old one, whilst the electrical circuit (fig. 2) also remains exactly the same. *Fig. 3* is a photograph of the new gauge. Compared with the old one it has the advantage that the whole of the discharge space is enveloped by the electrodes and the discharge can no longer be affected by the glass wall and the thin layer of metal deposited upon it by sputtering of the cathode.

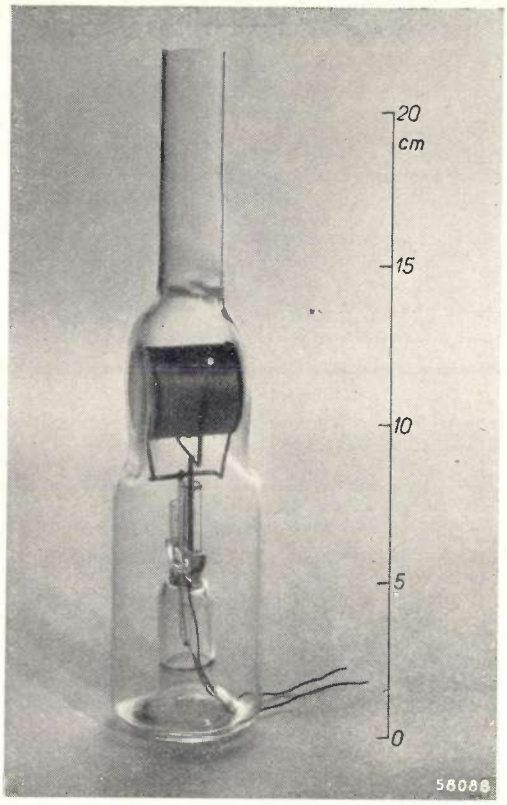


Fig. 3. Photograph of the new vacuum gauge.

In *fig. 4* the current flowing through the vacuum gauge has been plotted as a function of the gas pressure for the voltage (2000 V) and series resistance (1 megohm) usually applying also with the old type. This curve applies for air [8]), its trend depending more or less upon the nature of the gas. A comparison with the analogous curve (dotted line) for the old gauge (fig. 1*a*) shows that the sensitivity of the new vacuum gauge is about ten times greater, due, inter alia, to the larger effective surface area of the cathode. The curve for the new gauge has been drawn as a continuous line down to a pressure of about 10^{-6} mm Hg, but actually the discharge continues to take place at still lower pressures, in which range the sensitivity is represented approximately by the broken line; at such low pressures calibration is very difficult.

In the first description of the Philips vacuum gauge [9]) it was observed that when the pressure is reduced there is sometimes a sudden jump in the intensity of the current. With several specimens of the new design such a jump has been observed round about 10^{-4} mm Hg, thus at the upper limit of the pressure range. Below that pressure there were only small jumps.

Just as with the old construction, for currents between 10 and 1000 mA a small glowlamp (e.g. Philips 4662) can be used as indicator (see fig. 2),

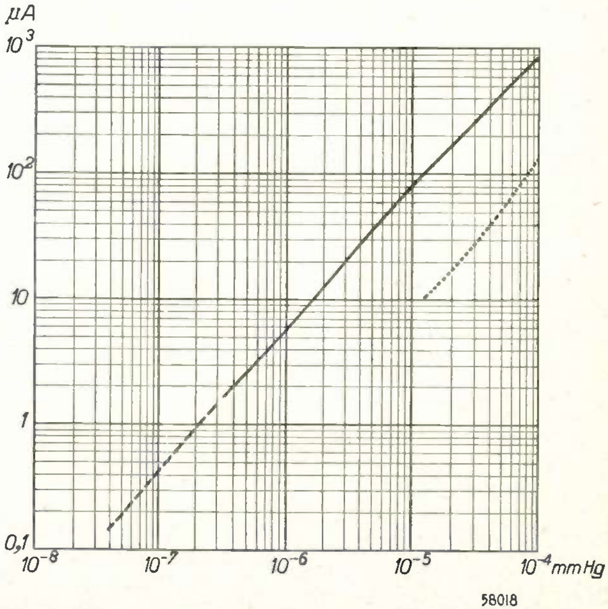


Fig. 4. The continuous line represents the current flowing through the new vacuum gauge plotted as a function of the gas pressure for the usual values of voltage and series resistance (2000 V, 1 megohm). The probable extension of the curve towards lower pressures is represented by the broken line; the dotted line relates to the old vacuum gauge.

[8]) The curve plotted is the average for three different tubes; the greatest deviation was 30 %.

[9]) F. M. Penning, Ein neues Manometer für niedrige Gasdrücke, insbesondere zwischen 10^{-3} und 10^{-5} mm, Physica, The Hague, 4, 71-75, 1937.

but for currents below 10 mA a more sensitive device is needed. A light-spot galvanometer with say 100 scale divisions for 1 μA is highly suitable for this purpose, one scale division corresponding to a difference in pressure of the order of 10^{-8} to 10^{-9} mm Hg. In order to get results at all reproducible in the range below 10^{-5} mm Hg it is of course necessary to degas carefully all glass walls and electrodes, including the vacuum gauge tube itself.

Leak detectors

In any pumping installation the vacuum gauge can also be used for detecting and tracing leaks. To find the origin of a leak in the object to be evacuated (for instancc a transmitting valve or the pumping installation itself) one usually goes over the whole apparatus part for part, replacing the atmospheric air by some other substance. When the leak is reached this substance flows in and the gauge usually registers a change in pressure. As substitute one often uses a liquid, for instance water, and if the apparatus is fitted with a liquid air trap, in which the penetrating water vapour immediately condenses, the pressure in the apparatus will drop to a very low level.

The use of a liquid as substitute for air has its objections. It is not always possible to carry the liquid to the suspected places, and moreover the method becomes very cumbersome when some atmospheric air remains between the liquid and the leak and first has to be pumped away. It is therefore better to use instead of a liquid a gas which can be sprayed upon the suspected places. In recent years this principle has in fact been applied in various ways. For this reason we shall confine our further considerations to gases as substitutes for air. The usual method, whereby a leak is detected through a difference in the pressure registered by the vacuum gauge when the suspected place is sprayed, will be reverted to farther on, after we have first dealt with some other methods.

One can use as indication, for instance, the change in colour of a gas discharge. Hitherto this discharge has usually been brought about by means of a Tesla coil, but this requires a rather high pressure (at least 10^{-3} mm Hg). At lower pressures the change in colour of the discharge can be observed in the Philips vacuum gauge (old design) itself if the tube is provided with a window far enough removed from the discharge to be kept free from any sputtered material (see *fig. 5*). When a leak is sprayed for instance with hydrogen one can clearly see the discharge inside the vacuum gauge change colour.

For smaller leaks other methods have been devised [10]), which were of great importance in nuclear research for testing the extensive vacuum apparatus required for separating U^{235} [11]). For this purpose the apparatus was fitted with a mass spectrograph [7]) sensitive espccially to helium and the vacuum was tested by spraying with this gas. This method has the great advantagc that helium can never occur as an impurity in the apparatus itself, such in contrast, for instance, to hydrogen. In most countries, however, helium is too expensive to be used for this purpose, whilst on the other hand the mass spectrograph is a rather complicated apparatus and the currents have to be amplified before being measured.

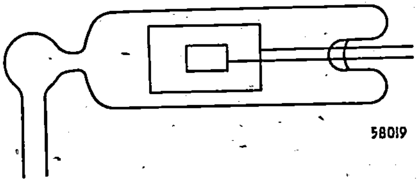


Fig. 5. A vacuum gauge (old design), used as leak detector, provided with a window through which the colour of the discharge can be seen to change when the leak is sprayed with some gas other than air.

Simpler in many respects is a method where advantage is taken of the fact that a heated wall of palladium is only permeable for hydrogen and not for air. With this method hydrogen has therefore to be used as substitute for air. The principle of the method is illustrated in *fig. 6*: the previously evacuated vacuum gauge *M* (e.g. an ionization vacuum gauge) is separated from the apparatus being tested by the palladium tube *P* which can be heated by means of the filament *F*;

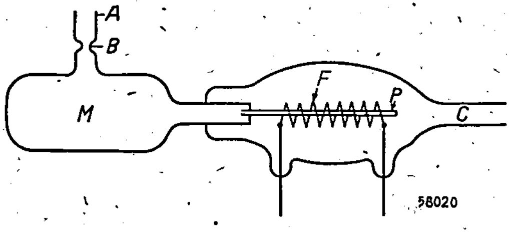


Fig. 6. Principle of a leak detector with palladium wall. *M* is the vacuum gauge which is evacuated via the tube *A*, sealed by fusing at *B* and then connected at *C* to the pump. *P* is the palladium tube which can be heated by means of the filament *F*.

[10]) For a more extensive review see the recent publication of S. Dushman, Scientific foundation of high vacuum technique, John Wiley and Sons, New York 1949.

[11]) R. B. Jacobs and H. F. Zuhr, New developments in vacuum engineering, J. appl. Phys. **18**, 34-48, 1947.

as soon as the jet of hydrogen strikes the leak hydrogen passes through P into M and raises the pressure.

The new vacuum gauge as leak detector

Owing to its high degree of sensitivity the new vacuum gauge described above is highly suitable for detecting small leaks. One can use, for instance, the arrangement shown in fig. 6, where M then represents the new gauge, which is first evacuated via the tube A and then sealed off at B, after which the whole device is connected at C to the pumping installation and the apparatus to be checked.

For tracing a leak the palladium tube is heated with the filament F and current is sent through the vacuum gauge. Since the discharge causes gas to disappear, after some time there will be a low equilibrium pressure, which is not zero because as a rule some hydrogen will still be passing through the heated palladium owing to hydrogen or hydrogen compounds being present in the pumping installation. To localize the leakage the pump is shut off by means of a cock and one goes over the outside of the apparatus with a jet of hydrogen in the manner described above. As soon as the current flowing through the vacuum gauge rises this gives an indication that the leak has been found. If there is any risk of an explosion a mixture of hydrogen with some other gas can be used, but then of course this reduces the sensitivity of the method.

Fig. 7 is an illustration of such a leak detector consisting of a vacuum gauge in the field of a permanent magnet and a bulb containing a palladium tube, the latter in this case being heated by the direct passage of current. The whole device is connected on the right-hand side to the apparatus being tested.

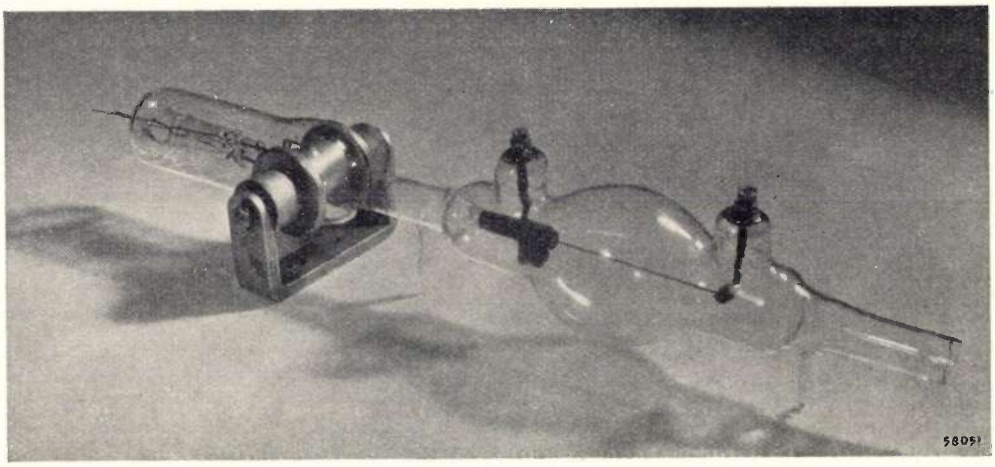


Fig. 7. The new vacuum gauge used as leak detector according to the principle of fig. 6. On the left the new gauge with magnet, on the right the envelope with the palladium tube heated by the direct flow of current via the contacts provided at the top of the envelope.

In *fig. 8* some examples are given of the variations in current as a function of time when a leak in an object (total volume about 2 litres) is exposed alternately to air and to hydrogen. The spraying with hydrogen begins where the arrow points upwards and ceases where the arrow points downwards. It is seen that the pressure continues to rise for some time after the spraying has been stopped. This is due to the fact that the leakage channel is still filled with hydrogen, the pressure ceasing to rise when all the hydrogen in this channel has been replaced by air and diffused through the pumping installation.

For these and subsequent measurements the leak was formed artificially by means of a very narrow glass capillary open to the outside air. Such a capillary is made by first reducing the diameter of the tube in a flame until the air channel inside is just visible, after which the tube is drawn out further and broken at its narrowest part. The leak can be enlarged by breaking off a piece of the capillary, and the tube can be sealed by holding the extremity in the flame for a moment. The passage is previously measured by allowing air to leak in for a considerable length of time. In this way it is possible to make capillaries passing through less than 10^{-7} $cm^3 \times$ mm Hg air per second.

The amount of hydrogen passing through the palladium depends upon the difference in pressure between the two spaces and thus upon the increase of the pressure in the object. For a given leak the sensitivity of this method is therefore roughly inversely proportional to the volume of the object.

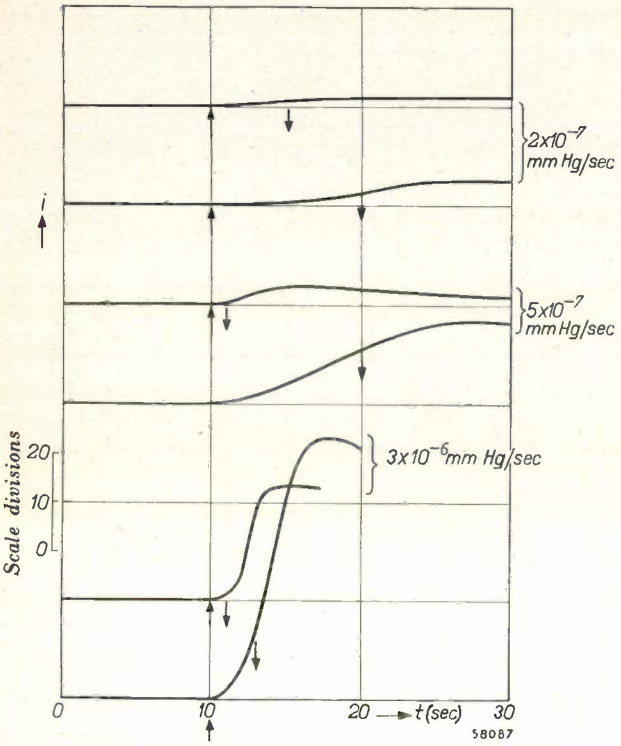


Fig. 8. Some examples of the current variation as function of time when a leak is alternately exposed to air and to hydrogen. Spraying with hydrogen begins at the arrow directed upwards and ceases at the arrow directed downwards. The initial current in air for $0 < t < 10$ sec is indicated at an arbitrary height above the t-axis. An idea of the sensitivity can be formed from the size of the scale divisions of the current meter given in the diagram. To the right of the curves the size of the leak in air is indicated in mm Hg per second.

As already stated, the principle of combining a heated palladium wall with an ionization vacuum gauge with hot cathode has been applied before [12]), the pressure being deduced from the ratio of the ion current to the electron current. The method described here is simpler because there is only one current to be measured and moreover this is greater than the non-amplified ion current in the other method. Another advantage is that when a leak is detected the hydrogen gas that has penetrated into the vacuum gauge can be more easily removed, so that one can pass over more quickly to the next test. In the method with a cold cathode the discharge itself will "burn away" the gas more quickly than when a hot cathode is used.

For the detection of still smaller leaks in valves which can be sealed off from the pumping device another method, without palladium tube, is to be preferred. A vacuum gauge is fused onto the valve to be tested and both of these are then evacuated as well as possible, after which they are sealed off from the pump (see for instance *fig. 9*). The current through the vacuum gauge is then switched on until it reaches a constant value. As already remarked, owing to the discharge a certain amount of gas disappears. In the state of equilibrium the amount burnt away per second by the discharge is obviously equal to the sum of that leaking in and that released from the walls, etc. If the amount of the gas released is negligible some idea of the size of the leak can already be formed from the current flowing through the gauge. On a former occasion [9]) it had been observed that in a well degased tube about 20 i cm^3 × mm Hg air per second is burnt away, i representing the current in amperes. In the case just assumed this is in fact the amount leaking in per second.

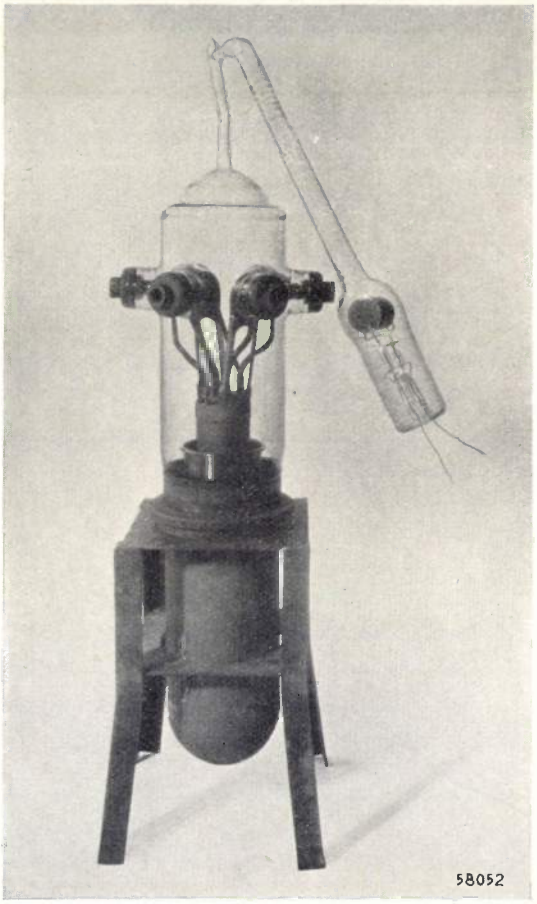


Fig. 9. Photograph of a transmitting valve with vacuum gauge fused on, by which means a leak was found in one of the metal caps for the current lead-ins.

In order to localize the leak one can now spray the valve to be tested with some gas other than air.

[12]) H. Nelson, The hydrogen gauge, an ultrasensitive device for location of air leaks in vacuum-device envelopes, Rev. sci. Instr. **16**, 273-275, 1945.

The current passing through the gauge will generally depend upon the nature of the gas used, because both the rate at which the gas leaks in and that at which the burning away takes place are related thereto, whilst the sensitivity of the instrument also depends upon the kind of gas. Good results have been reached for instance with argon and hydrogen as substitutes. *Fig. 10a* gives some examples of the results obtained with argon. The leak was again in the form of a capillary shaped as described above, and this was directly fused onto the vacuum gauge as indicated in fig. 10*b*. Just as in fig. 8, the arrows indicate the beginning and the end of the spraying of the leak with the gas. When spraying for only one second the current returns to its initial value within a few seconds owing to the argon being burnt away. When spraying for 10 seconds (curves and arrows drawn in dotted lines) the current reaches within a few seconds a new state of equilibrium whereby just as much argon is burnt away as flows in.

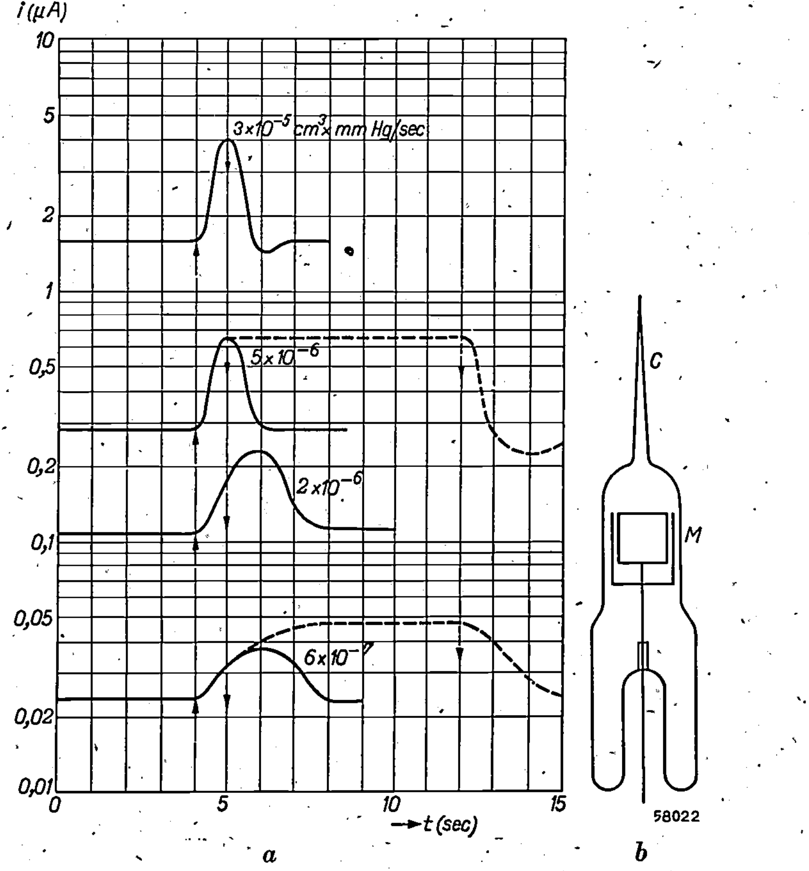


Fig. 10. *a*) Some examples of the detection of leaks in a valve evacuated with the vacuum gauge to which it was fused (the substitute gas was argon). The size of the leak is indicated at the right of the curves in cm³ × mm Hg per second. When no more gas is released from the glass wall and the electrodes then the value at which the current through the gauge reaches an equilibrium, with air around the tube, ($t<4$ sec in the diagram) already gives an idea of the extent of the leakage. *b*) illustrates the manner in which the artificial leak in the form of a capillary *C* was connected to the vacuum gauge.

In this way it is possible to determine extremely small leaks of about 2×10^{-7} cm³ × mm Hg/sec, corresponding to an increase in pressure in the volume of the manometer of 0.0025 mm per month. The sensitivity is several hundred times greater than that given for the mass spectrograph method. An advantage of the latter method, however, is that the inflowing helium can also be measured when there is already a fairly high pressure of other gases in the apparatus. Moreover, to reach the maximum sensitivity with the method described here it is essential that both the object being tested and the manometer together must be well degased, so that the residual pressure is determined only by the leakage.

In most cases the valves to be tested will have a volume of the same order of size or greater than that of the tube of the vacuum gauge, in which case it will take longer for the air to be replaced by the gas injected. *Fig. 11a* gives an example where there is a volume of about 1 litre between the leakage capillary and the vacuum gauge (see also fig. 11*b*). For the same size of leak both the initial value before spraying and the ultimate value after spraying will be the same as in the case of fig. 10. In both cases when the state of equilibrium is reached the amount of gas burnt away by the discharge equals the amount flowing in through the leak. When, however, a volume of 1 litre is placed in between the vacuum gauge and the capillary, the transition from the initial value to the ultimate value takes place much more slowly. This may be seen when comparing fig. 11*a* with fig. 10*a*. In fig. 11*a* it has moreover been indicated by a dotted line what increase in current might be expected if the volume of 1 litre were omitted. With larger volumes it thus takes much longer to detect a leak, but apart from this the sensitivity is just as great as in the case of smaller volumes.

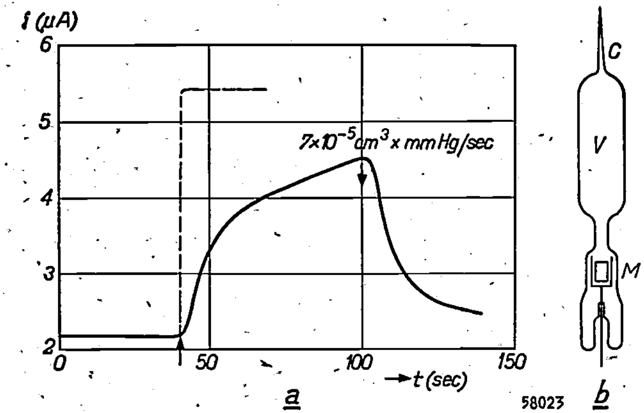


Fig. 11. *a*) An experiment similar to that in fig. 10 but with a volume V of 1 litre between the vacuum gauge and the capillary. The broken line shows the trend to be expected for this leak if the capillary were attached in the manner indicated in fig. 10*b*. *b*) The manner in which the artificial leak (capillary C) was affixed to the vacuum gauge.

Summary. The Philips vacuum gauge previously described in this journal is suitable for gas pressures of 2×10^{-3} to 10^{-5} mm Hg. In order to produce a vacuum gauge based upon the same principle but useful for lower pressures, the electrode system has been replaced by another consisting of two parallel round metal discs as cathode and an anode in the form of a cylindrical jacket. To keep the distance between the pole shoes of the permanent magnet enveloping the glass tube as short as possible this tube has been pinched at that point. These improvements have given the new instrument a sensitivity about ten times as great: it covers a range from 10^{-4} to 10^{-6} mm Hg and can probably be used for pressures even below 10^{-7} mm Hg. Vacuum gauges are also used as leak detectors and some methods followed for this purpose are briefly discussed. It is indicated how the new design of vacuum gauges can best be employed for such an application.

HEATING THE FILAMENTS OF VALVES IN A CASCADE GENERATOR BY MEANS OF HIGH-FREQUENCY CURRENT

by Tj. DOUMA and H. P. J. BREKOO.

621.314.671:621.3.032.44:621.3.023

In a cascade generator the filaments of the valves are at a very high potential with respect to earth. The difficult problem of how to heat these filaments has in principle been ingeniously solved by employing high-frequency current. In the practical application of this method, however, difficulties were encountered which involved the use of highly complicated circuits. Thus in a cascade generator with 12 valves no fewer than 15 tuned circuits were needed. Partly due to the use of "Ferroxcube" for the filament-current transformers it has now been found possible to simplify the circuits appreciably.

A very high D.C. voltage is needed both for generating X-rays and in nuclear physics. For the supplying of such a voltage in the order of 10^5 to 10^6 V for instance, the cascade generator is highly suitable. The circuit of such a generator has been described several times in this journal [1], so that it will suffice here to give a brief explanation in the caption of *fig. 1* representing the basic circuit.

In the designing of a cascade generator one of the problems is how to apply the power required for heating the cathodes of the valves. The most obvious method is to supply each valve from its own transformer connected to the mains. In the case of installations for very high tensions however great difficulties are encountered in insulating the two windings of the transformers from each other particularly in those supplying the uppermost valves (uppermost both in the drawing and in actual execution), which are at the highest potential with respect to earth.

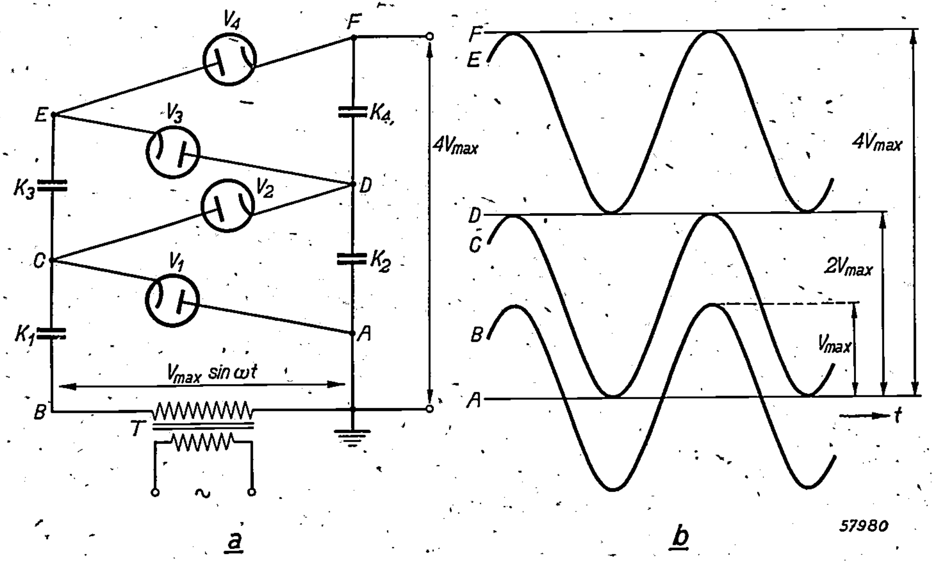


Fig. 1. *a*) Circuit diagram of a cascade generator with four valves ($V_1 \dots V_4$) and four capacitors ($K_1 \dots K_4$). T = transformer; peak value of secondary voltage = $V_{\max}$. The potential at the points $B \dots F$ with respect to the earthed point A has been plotted in (*b*) as a function of the time *t*, with the corresponding lettering. This diagram applies for no load. It is readily seen that the ratio of the highest D.C. voltage obtained (between F and A) to $V_{\max}$ equals the number of valves (in this case four).

Right from the beginning other means have therefore been sought. At first each filament was fed from its own insulated accumulator, later on from a separate dynamo, the dynamos being driven by a single motor by means of long insulating shafts. These solutions, however, were far from ideal.

[1] A modern high-voltage equipment, Philips Techn. Rev. **1**, 6-10, 1936; A. Kunkte, A generator for very high direct current voltage, Philips Techn. Rev. **2**, 161-164, 1937. Sometimes the cascade generator renders good service also for lower voltages, see for instance G. J. Siezen and F. Kerkhof, Projection-television receiver, III. The 25 kV anode voltage supply unit, Philips Techn. Rev. **10**, 125-134, 1948 (No. 5), in which article a further analysis of the working is to be found for a particular case.

Feeding with high-frequency current

It was a big step forward in the right direction when Kuntke evolved the idea of feeding the filaments with a high-frequency current passing through the filaments connected in series via the capacitors of the cascade system itself (see the first of the articles quoted in footnote [1]). The principle of this method is represented in *fig. 2.*

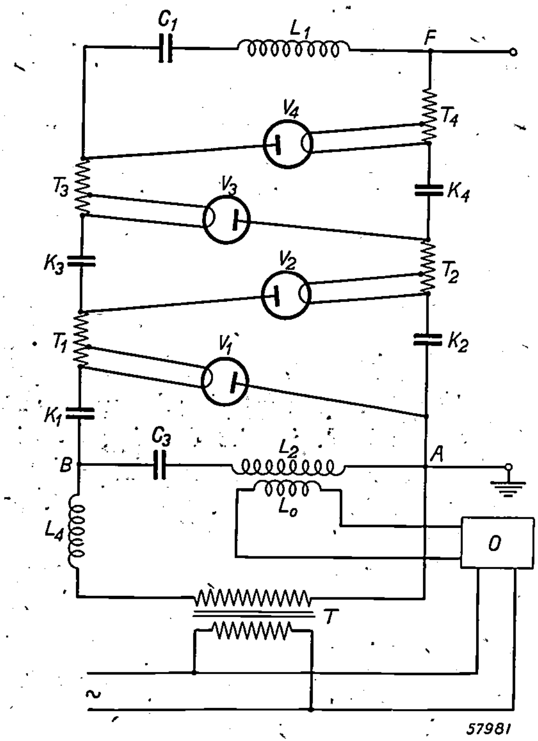


Fig. 2. The filaments of the valves are fed via autotransformers $T_1 \ldots T_4$ with a high-frequency current generated by the oscillator O and circulating through the circuit L_2-C_3-K_1-T_1-K_3-T_3-C_1-L_1-T_4-K_4-T_2-K_2-L_2. The capacitors $K_1 \ldots K_4$ have a negligible impedance to high frequencies, whilst C_1-L_1 by-passes the uppermost valve, and the choke L_4 blocks the path through the transformer T. The capacitor C_3 prevents L_2 and L_4 from forming a short-circuit across T.

This figure likewise shows that filament-current transformers are still employed in this method, since the filament-current itself (3.5 A for a frequently used type of valve) is too high to be circulated without considerable loss. These transformers, however, have the primary winding connected to the secondary, so that here the difficulty of insulation does not arise. At first transformers were used with a ratio of 5:1, so that the primary current amounted to 0.7 A. This current was generated by a valve oscillator working at a frequency between 500 and 750 kc/s.

Attention is drawn to the capacitor C_3 (fig. 2), which prevents the coil L_2 of the h.f. transformer from forming a short circuit for the secondary current (mains frequency) of the main transformer T; further attention is drawn to the choke L_4 preventing the flow of h.f. current through T. There is also a circuit C_1-L_1 tuned approximately to the oscillator frequency and thus providing an h.f. by-passing for the uppermost valve, whilst blocking the direct current.

Objections to the original system

This solution of the problem of heating the filaments, so ingenious in principle, is in practice not so simple as it appears.

The junction points B, C, $D \ldots$ (fig. 1) of a cascade generator are at a high potential with respect to earth. In order to avoid corona discharge it is necessary to enclose each of these points in a rounded metal envelope which is connected to the junction point (*fig. 3*) and which, when viewed from the outside, may nowhere have too small a radius of curvature. In particular the large envelope of the high-voltage equipotential surface — the "hood" — is worthy of note (see *fig. 4*).

Each of these envelopes has a certain capacitance with respect to earth (sometimes almost as much as 100 pF). These capacitances present considerable shunt paths to high frequencies. As a result the filament currents of the valves are not equal and therefore each valve is not fed with the current most favourable for long life.

With a view to equalizing the filament currents as far as possible two steps have been taken to reduce the high-frequency potential of the envelopes, so as to minimize the leakage of current.

The first of these steps consists in tuning each filament circuit to the oscillator frequency with the aid of a series capacitor; this circuit has a quite considerable self-inductance — consisting for the greater part of the stray self-inductance of the filament-current transformer — in which there is a considerable voltage drop. By tuning the circuit this is compensated and as a result the primary potentials, and thus

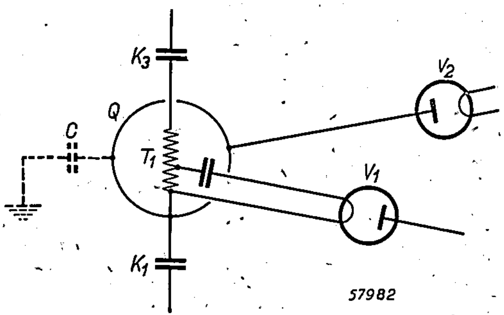


Fig. 3. To avoid corona discharges each of the junction points of the cascade generator which are at a high potential (B and the transformers T_1, $T_2 \ldots$ in fig. 2) is surrounded by a rounded metal envelope Q, as represented here for T_1. These envelopes have a capacitance C to earth which by-passes a portion of the H.F. current.

also the h.f. voltages on the envelopes, are considerably reduced.

The second step consists in the application of a kind of bridge circuit, the principle of which is represented in *fig.* 5. By this means the high-frequency potential of the "hood" with respect to earth can be reduced to zero. It is true that the lower junction point envelopes then have a greater h.f. potential than is the case in the asymmetrical system shown in fig. 2, but this is outweighed by the fact that there is no longer a leakage current from the large hood.

Attention is also drawn to the circuit L_3-C_3 required in this system. The high-frequency current to which it is tuned can easily pass this circuit. The capacitor C_3 serves to block the direct current. C_2 is a variable air capacitor which has to be adjusted when the coupling between L_2 and L_0 is changed.

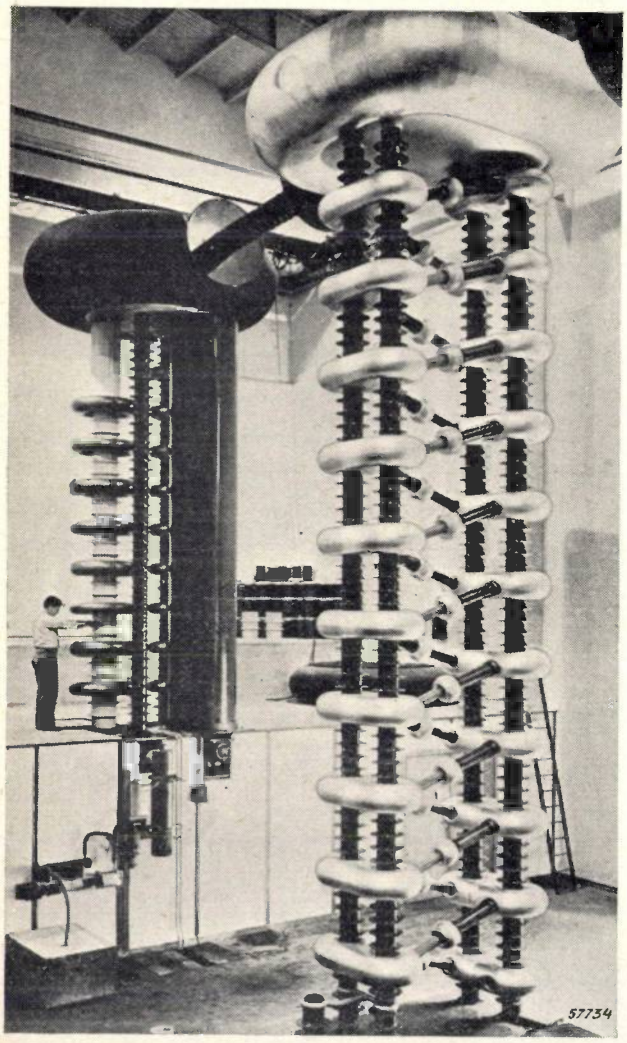


Fig. 4. A 2000 kV cascade generator (on the right) installed in the Cavendish Laboratory at Cambridge (England). It has twenty valves. The geometrical construction corresponds to the form in which the diagram of fig. 2 has been drawn. This installation is connected to a tube (on the left) in which nuclear reactions are brought about. Note the rounded corona shields of the junction points and the large "hood" forming the non-earthed D.C. terminal. The capacitance of these parts causes an unequal distribution of the h.f. filament current of the valves.

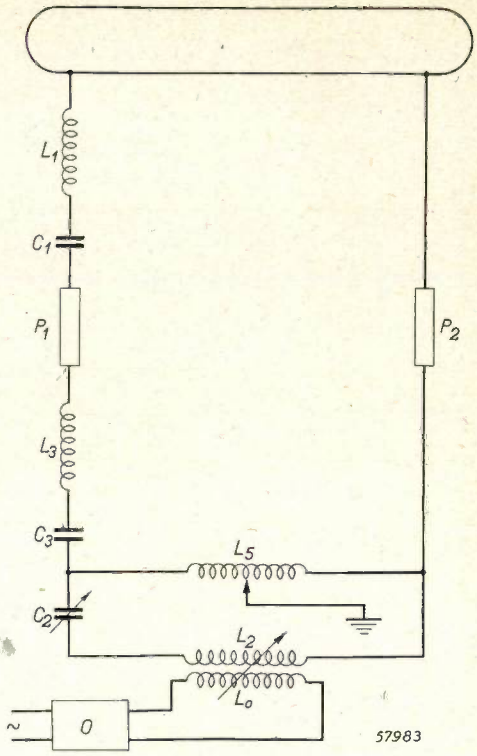


Fig. 5. A circuit allowing the h.f. potential of the "hood" H to be reduced to zero, by means of adjustment of the tapping on the coil L_5. The parts carrying only low-frequency current have been omitted. P_1 and P_2 denote the two columns. The capacitor C_3 is necessary to insulate the bottom of P_1 from earth for direct voltage; the coil L_3 compensates the high-frequency impedance of C_3. The circuit is tuned by means of the variable capacitor C_2. The other letters have the same meaning as in fig. 2.

It will be readily understood that these measures cannot be said to simplify the installation. This is demonstrated most clearly by the diagram in *fig.* 6 representing the system of cascade generators actually constructed. In this system with 12 valves (with which a D.C. voltage of more than 1000 kV can be obtained) there are no fewer than 15 circuits to be tuned: the 12 filament circuits plus the circuits L_1-C_1, L_2-C_2 and L_3-Ç$_3$. Moreover, there is the variable coupling L_0-L_2 and the tapping of L_5. Needless to say, it is no easy matter to adjust all these elements in such a way that the twelve filament currents are kept within certain limits. It is true that, in theory, this adjustment has only to be made once for all, but it does not take much to disturb the balance for proper functioning.

Added to this is the fact that the losses in the two columns (mainly in the high-tension capacitors C_1 and C_3) with a high-frequency current of 0.7 A may be of the same order as the power consumed by the filaments. Thus the efficiency is not favourable.

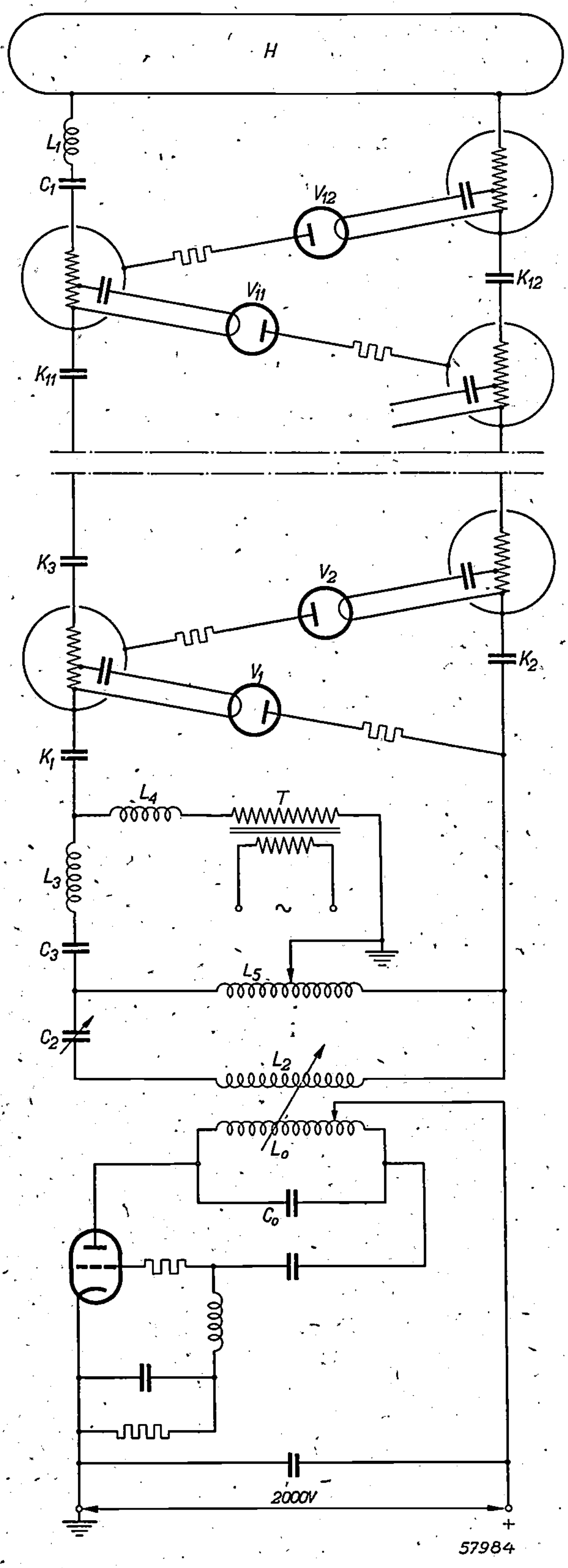


Fig. 6. Superseded circuiting diagram of a cascade generator with h.f. filament heating (12 valves). To compensate inductive voltage drop each filament circuit is tuned with a series capacitor. The circuit is symmetrical (coil L_5, cf. fig. 5). L_0-C_0 = oscillator circuit. The other letters have the same meaning as in fig. 5. Connected in series with each valve is a resistor for limiting current surges which might occur when switching on.

The new system

A new system has been devised which is much simpler, mainly because the tuning of each filament circuit has been dispensed with. Moreover, the losses have been reduced, partly due to the use of a smaller primary current (0.25 A instead of 0.7 A).

These two measures together result in the first place in a higher primary voltage per filament current transformer. This, as we have seen, is undesirable on account of the capacitive leakage currents, which moreover become more important with a primary current of 0.25 A than with one of 0.7 A. This objection has been adequately overcome by adopting two further measures.

In the first place a lower frequency has been chosen, thereby reducing the h.f. voltage drops in the secondary circuit of each filament current transformer. Taking into account the need to be able to introduce the new system into existing installations with as little alteration as possible, we found the most favourable frequency to be approximately 250 kc/s (as mentioned previously, the frequency was 500 to 750 kc/s).

The second measure consists of a drastic reduction of the magnetic leakage of the filament current transformers, which are responsible for a large part of the self-inductance in the secondary circuit. This has been achieved by using a core of "Ferroxcube" [2]) for the transformer instead of one made of iron dust mixed with "Philite" as previously used. This new material has very low magnetic losses (also at high frequencies) and a high permeability (in the order of 1000). The parts from which the core is built up, two E-shaped pieces (see *fig. 7*), can easily be ground so smooth that no appreciable air gap remains in the path of the lines of force. It is due to these two properties of the material that the magnetic leakage of the transformer is so small. This renders the tuning of the filament circuits superfluous.

Fig. 8 shows two complete transformers, one detached and one mounted on the bracket which also carries the valve socket.

The winding ratio is 14:1, corresponding to a primary current of 0.25 A. The loss in the capacitors C_1 and C_3 (fig. 6) is much smaller than the power consumed by the twelve filaments together.

The complete circuit diagram is shown in *fig. 9*. It is also to be seen that the oscillator circuit has been simplified (cf. fig. 6) and that a bridge circuit, making the high-frequency potential of the hood zero, is no longer used. It has been found that,

[2]) J. L. Snoek, Non-metallic magnetic material for high frequencies, Philips Techn. Rev. 8, 353-360, 1946.

thanks to the simplifications introduced, the irregularities in the distribution of current can be made so small, even with an asymmetrical circuit, that a slight correction to each filament current transformer is sufficient to confine the current intensities within the fixed limits. This applies even to the large number of twenty valves as contained in the installation illustrated in fig. 4. If desired, however, the circuit can be made symmetrical (*fig. 10*), in which case, under favourable circumstances, even the individual corrections are unnecessary.

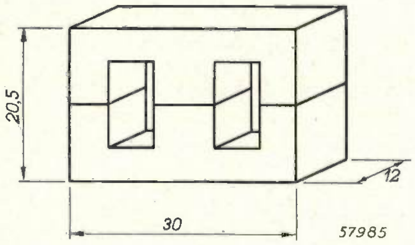


Fig. 7. The new filament-current tranformers have a core consisting of two solid E-shaped pieces of "Ferroxcube", drawn here to actual size: dimensions are in millimeters.

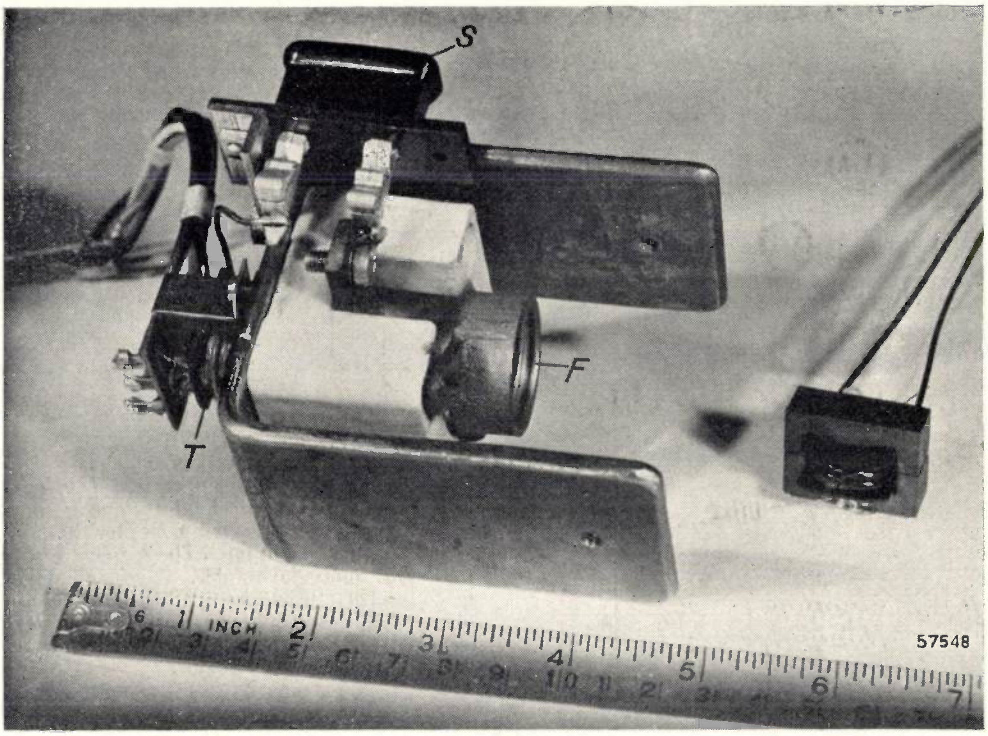


Fig. 8. *Right:* A complete transformer for a secondary power of 2.3 V × 3.5 A = 8 W, frequency 250 kc/s. *Left:* Bracket on which are mounted the filament current transformer *T*, the valve socket *F* and a plug socket with short-circuiting plug *S* which can be replaced by an ammeter for the purpose of checking.

Adjustment

From fig. 9 it is to be seen that the tuning capacitor and the grid resistor of the oscillator can be adjusted. By this means the optimum frequency can be chosen and the output varied by about 10%.

Before the installation is put into use it is necessary to check once for all whether the filament currents are within the fixed tolerances. For this purpose there is a socket at each valve which is normally short-circuited (fig. 8) but to which an ammeter can be connected for testing.

Experience has shown that in the new installation all valves receive approximately the correct filament current if that of the two lowermost valves (to the high-frequency current these are the first and last valves) has been adjusted to the correct value; see *fig. 11*. If necessary any remaining small deviations can be corrected, for instance, by using tappings on the primary winding of the filament current transformers. The mean level of the filament currents — which in fig. 11, curve *c*, is still slightly too high — can be adjusted to the desired value with the aid of the oscillator grid resistor.

Another great practical advantage of this new system is that, thanks to its simplicity, once it has been adjusted it does not easily become detuned.

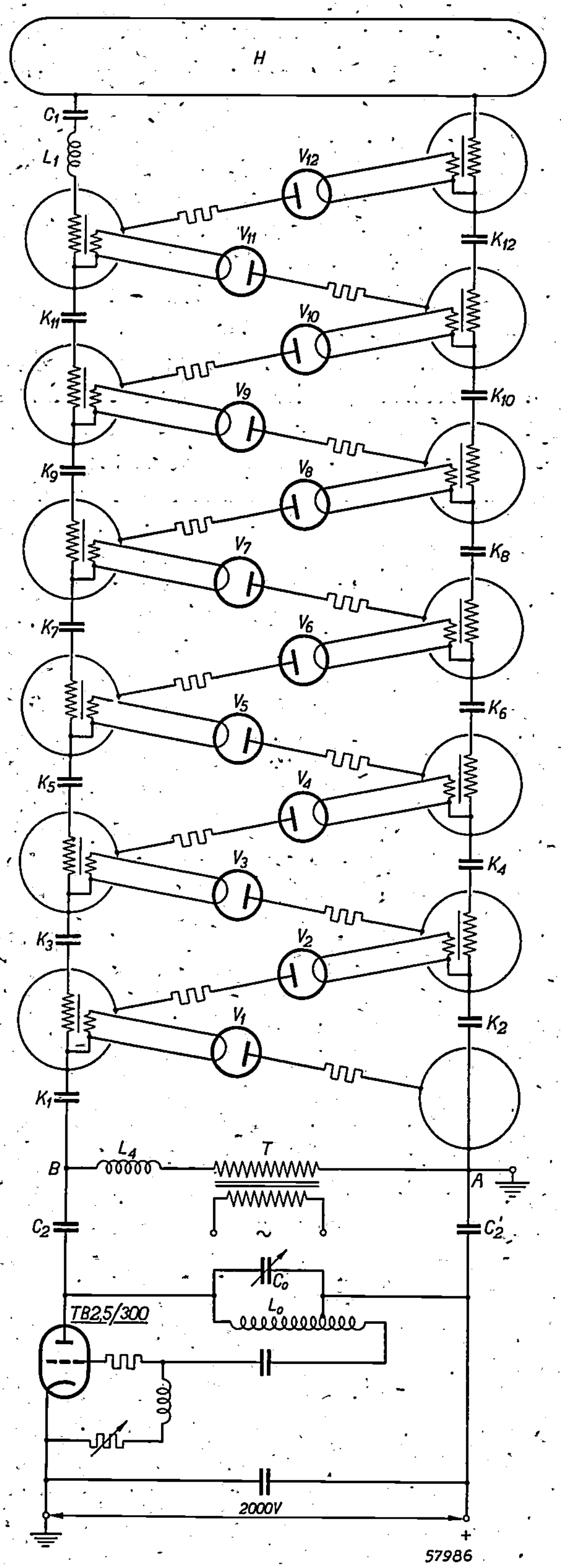


Fig. 9. The new system for feeding the filaments of a cascade generator with h.f. current. The filament circuits are not tuned separately. The oscillator circuit L_0-C_0 is coupled with the cascade generator via two blocking capacitors C_2 and C_2'. The oscillator valve is of the type TB 2.5/300. The output can be varied by about 10% by means of the variable grid resistor.

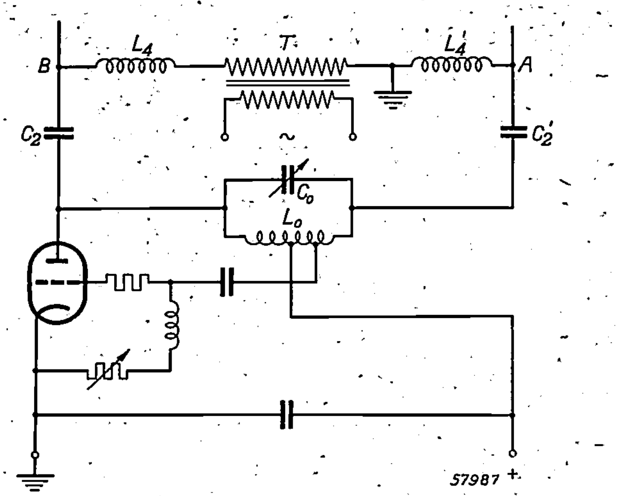


Fig. 10. Oscillator circuit for symmetrical supply. The letters correspond to those in fig. 9. Between A and earth is a second high-frequency choke L_4'.

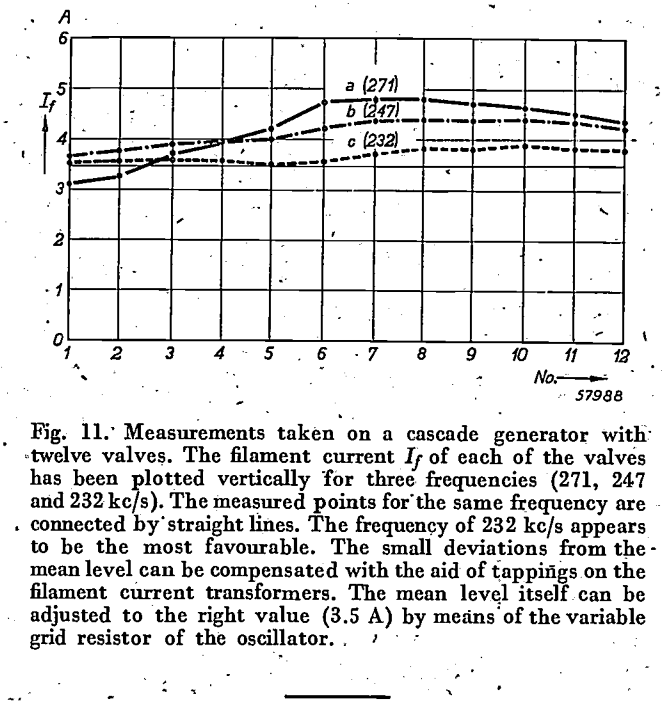
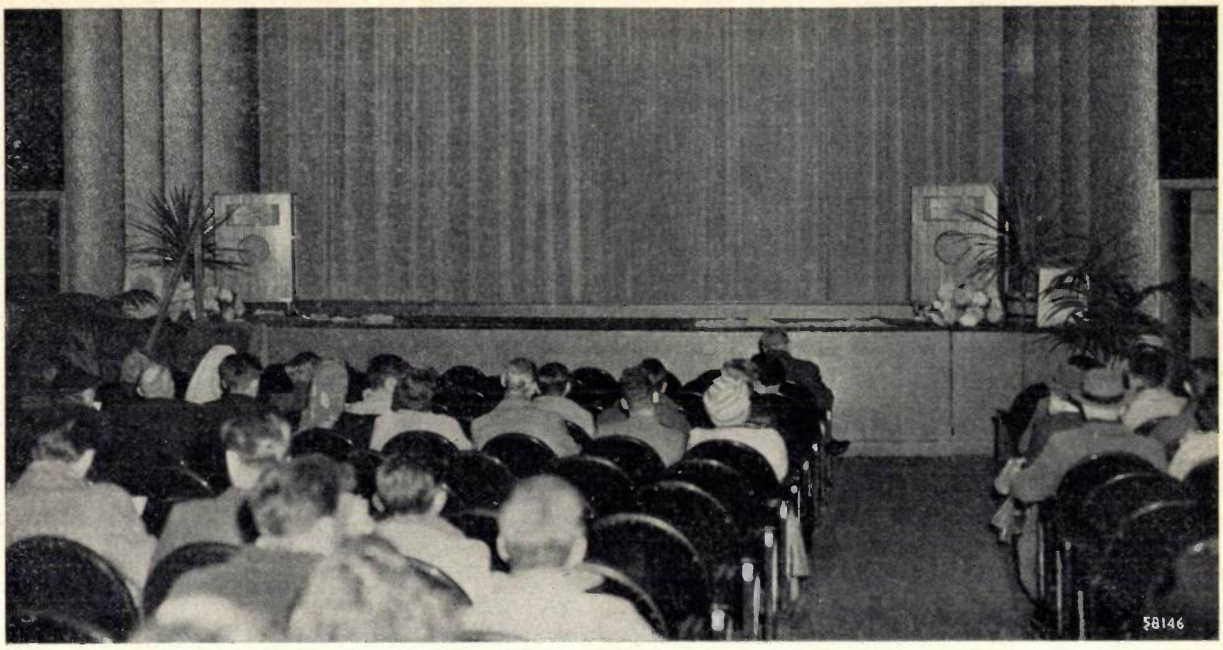


Fig. 11. Measurements taken on a cascade generator with twelve valves. The filament current I_f of each of the valves has been plotted vertically for three frequencies (271, 247 and 232 kc/s). The measured points for the same frequency are connected by straight lines. The frequency of 232 kc/s appears to be the most favourable. The small deviations from the mean level can be compensated with the aid of tappings on the filament current transformers. The mean level itself can be adjusted to the right value (3.5 A) by means of the variable grid resistor of the oscillator.

Summary. When high-frequency current is used for heating the filaments of the valves in a cascade generator one is faced with the difficulty of equalizing the filament currents. The cause of the inequality lies in the capacitance of the metal envelopes placed around the junction points of the cascade generator in order to avoid corona discharges. The lines on which the solution of this problem was at first sought has led to a complicated system in which each filament circuit had to be tuned. In the new system described all these adjustments could be dispensed with, thanks to the low magnetic leakage of the filament current transformers, which are now provided with a core of "Ferroxcube". Thanks to this and other simplifications a system has been devised which, contrary to the old system, is easily adjusted so as to keep all the filament currents within the tolerances allowed and which does not easily become detuned.



STEREOPHONIC MUSIC IN THE CINEMA

534.76:791.45

An experiment was made with stereophonic reproduction of music in a cinema in Eindhoven, the Chicago Theatre, from 29th October to 4th November 1948. In contrast to previous demonstrations given in various places to invited guests or in private circles this experiment was carried out as part of a normal public programme.

In this case stereophonic reproduction was applied for the music played during the interval. It is not likely to be applied for the films themselves for some time because film producers must first decide to have the sound track recorded stereophonically, and although this is already possible technically [1]) it seems that for the present the alterations necessary in the apparatus and the resultant cost will be considered objectionable. Moreover, the improvement to be expected in the case of films will probably be less striking, since any faulty impression of the direction of the sound is more or less predominated and corrected by the strong visual impression of the picture. When music is played in the interval, however, direction is centered entirely upon the sound itself, particularly so when the music is reproduced mechanically, so that the audience is not distracted by the musicians. It was therefore a promising experiment to apply stereophony for improving the reproduction under these circumstances, so as to eliminate as far as possible the mechanical nature of the music and thus afford the audience an opportunity to enjoy the music as much as if they were listening to an actual concert performance [2]).

There is no need to say much here about the technical execution of the experiment. In the photograph one can see how the two loudspeakers were placed either side of the screen. The music given was selected from a series of stereophonic recordings made by the Electro-Acoustic Department of Philips by the magnetic method.

Everyone in the audience was given a questionnaire on which they were invited to answer three questions:

1) Do you think stereophonic music should be regularly given in cinemas? Answer a) yes, b) not interested, c) no.
2) Do you think that the quality of reproduction demonstrated here is better than that which you are used to hearing in cinemas?
   Answer a) yes, a great difference, b) yes, little difference, c) no.

Space was left on the forms for any further remarks.

Of the 7300 odd forms distributed more than 5800 were returned completed and 500 only partly filled in. About 75 % showed their appreciation,

[1]) A possibility has been described by K. de Boer, Stereophonic recording on Philips-Miller film, Philips Techn. Rev. 6, 80-85, 1941; see also R. H. J. Alink, C. J. Dippel and K. J. Keuning, The metal-diazonium system for photographic reproduction, Philips Techn. Rev. 9, 289-300, 1947.

[2]) See the article by R. Vermeulen, Duplication of concerts, Philips Techn. Rev. 10, 169-177, 1948 (No. 6).

whilst less than 3% were definitely averse to the idea. The table below shows how opinions varied (where a question was not answered it was counted as expressing indifference and included under b).

Question	Number of replies		
	a	b	c
1	5764 (79%)	1350	199
2	5548 (76%)	1650	115

In about 800 cases some comments were added, 300 of these expressing enthusiasm, whilst 200 complained that the sound had too sharp a character. In connection with the latter comment it is to be noted that frequencies up to 10,000 c/s were reproduced with less than 2 db attenuation. The public is not accustomed to an unattenuated reproduction of such high notes in mechanical music and there might therefore be a tendency to consider it as being unnatural, unless it were possible to avoid any association with "mechanical music". Other comments made were in respect to the volume of sound, the impression of the stereophonic effect received in unfavourable seats, the selection of the programme, etc.

Although experience teaches that a statistical sounding of the public does not always lead to reliable results, and we shall therefore refrain in this case from giving any comment of our own upon the result of the inquiry, it nevertheless seemed desirable to give publicity to the experiment and its results by means of this communication. An application on a large scale can of course only be expected when a sufficient variety of stereophonic recordings become available for this purpose.

J. P. de VISSER van BLOEMEN

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

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

1845: E. J. W. Verwey, F. de Boer and J. H. van Santen: Cation arrangement in spinels (J. chem. Phys. **16**, 1091-1092, No. 12).

A new calculation of the Madelung constant of the spinel lattice is given, firstly according to Evjen's method and secondly according to that of Ewald. The agreement between the two methods is very satisfactory. The table given is to replace data published previously into which an error of calculation had slipped.

1845a*: G. W. Rathenau: Correspondence on Bowles and Boas's Paper (J. S. Bowles and W. Boas, J. Inst. Metals **74**, 753-755, 1948).

This correspondence contains a few remarks by Dr. G. W. Rathenau made in a discussion of the paper indicated. For full particulars the reader may consult an article by G. W. Rathenau and J. F. H. Custers in Philips Res. Rep. **4**, 241-260, 1949, No. 4.

1846*: P. Cornelius: Het gerationaliseerde Giorgi-stelsel in de electriciteitsleer (T. Ned. Radiog. **14**, 1-19, 1949 No. 1). (The rationalized Giorgi system in electromagnetism; in Dutch.)

Short survey of the rationalized mks system of electrical units and of a number of familiar formulae as expressed in these units (see these abstracts Nos **1803***, **R 103** and **R 111** and Philips Techn. Rev. **10**, 79-86, 1948).

1847: F. A. Kröger and J. E. Hellingman: Chemical proof of the presence of chlorine in blue fluorescent zinc sulphide (J. Electrochem. Soc. **95**, 68-69, 1949, No. 2).

According to the writers blue fluorescent ZnS contains chlorine, which has been added as a flux (e.g. NaCl) or by preparing the phosphor in an atmosphere containing HCl. A chemical determination of the chlorine in ZnS phosphors, giving 2×10^{-5} to 27×10^{-5} gr atoms/mole ZnS, supports this view. This figure is in fair agreement with estimates of the number of centres in ZnS phosphors. The blue luminescence increases with increasing Cl-content.

1848: J. A. Haringx: The cross-spring pivot as a constructional element (Appl. sci. Res. **A 1**, 313-332, 1949. No. 3).

This paper deals with the behaviour of the so-called cross-spring pivot, a flexible element connecting two members of a construction and consisting of two flat springs crossing each other at an angle of, for instance, 90 degrees. To a first approximation the pivot point coincides with the "point of intersection" of the two springs. By an elementary calculation one finds the rigidity of the direction of the force to be transmitted. A second calculation gives us the shift of the pivot point occurring for large angles of deflection, though for a pure bending moment only.

1849: J. D. Fast: Gasvormende stoffen in de bekledingen van lasstaven (Lastechniek **15**, 189-192, 1949, No. 3). (Gas-evolving substances in the coating of welding electrodes; in Dutch).

During arc-welding CO, CO_2, H_2 and H_2O escape from the coating of the welding electrode. The chief task of these gases consists in imparting a great velocity to the drops of molten metal and in causing a good penetration. The gases liberated, however, may also form a real danger, e.g. by causing porosity, fracturing and brittleness. Especially H_2 is discussed in this connection (see these abstracts, No. **1852**).

1850: H. G. Beljers: Measurements on gyromagnetic resonance of a ferrite using cavity resonators (Physica **14**, 629-641, 1949, No. 10).

The phenomenon of gyromagnetic resonance is measured with a resonance cavity, using small spheres and a thin ring made of "Ferroxcube 4", a nickel-zinc ferrite. Samples are placed in a cylindrical cavity at the electric nodal circle of the mode E_{020}. Resonance-frequency shift and quality are measured for various values of the axial magnetic field H.

Some experimental values for the gyromagnetic ratio are deduced. The g-factor was found to be 2.12.

Making some simplifying assumptions, the above-mentioned relation between H and the complex $\Delta\omega$ is calculated and compared with the experimental curve. As could be expected, no quantitative agreement for all values of the magnetic field was found. Some maximum values for the internal damping are deduced from the measurements.

1851*: H. Bremmer: Terrestrial radio waves. Theory of propagation (X + 343 p., 91 fig.; Elsevier Publ. Cy, Inc., New York, Amsterdam, London, Brussels, 1949).

The theoretical problem of the propagation of terrestrial radio waves consists in determining the magnitude of the electromagnetic field, and, more especially, the electric field, if the transmitter is a given source of electromagnetic waves. Whereas this problem is comparatively simple for a transmitter in empty space, it becomes rather complicated when the disturbance caused by a spherical earth is taken into account. The problem is then essentially one of diffraction. In addition the influences of the ionosphere and the refraction in the lower part of the atmosphere have to be discussed.

In chapter II-VI, inclusive, of this book the latter influences are entirely ignored, so that the atmosphere is then imagined to consist of empty space. The results obtained in this way apply in those cases where the field caused by the ionosphere (sky wave) is small as compared with the direct field (ground wave). A solution of Maxwell's equations is obtained, assuming constant ε and conductivity (σ) inside the earth, whereas outside $\varepsilon = 1$ and $\sigma = 0$. The transmitter is imagined as a point source, the field being singular at the transmitter. At the surface of the earth the usual boundary conditions (continuity of tangential components of electrical and magnetic field) apply.

The radio problem is mathematically equivalent to a number of problems in optics, acoustics and elasticity, such as scattering of light (Rayleigh), optical phenomena caused by colloidal particles (Mie, Debye), distribution of light in the rainbow (Airy).

In dealing with the vectorial wave problem the investigation of the six components of the electromagnetic field is replaced by the determination of the Hertzian vector, from which all components may be derived by differentation. In this way the originally vectorial problem is reduced to a scalar problem, in which form the solution applies also to acoustical problems such as: diffraction of sound waves round the human head and the acoustics of a spherical dome-shaped vault, and to the geological problem of the propagation of seismic waves. If the inhomogeneity of the atmosphere is taken into account, as is done in chapters VII-XI, inclusive, the problems also bear a close relation to others in a different field, such as reflection of acoustic waves by the stratosphere and the under-water propagation of acoustic waves. Chapter I (introduction) contains the statement of the problem and its connection with other problems, as outlined above, and further an historical account and a general survey. Part 1, dealing with the theory for a homogeneous atmosphere, consists of the following chapters: II) The series of zonal harmonics for the ground wave; III) Watson's transformations and the residue series; IV) Computations by means of the series of residues; V) Geometric optical approximations; VI) Numerical computations and results. Part 2, dealing with the theory for an inhomogeneous atmosphere, consists of the chapters VII) General considerations; VIII) Geometric optical computations; IX) Computations concerning the field as a sum of "modes"; X) Special applications; XI) Effect of the earth's magnetic field.

In addition figures show the field distribution under various conditions. (See also these abstracts Nos **1790**, **R** 97 and **R 108**.)

1852: J. D. Fast: Waterstof bij het booglassen (Laschtechniek **15**, 220-224, 1949, No. 5). (The role of hydrogen in arc-welding; in Dutch.)

This article deals with the influence of hydrogen on the weld metal. Porosity, fracturing and the occurrence of fish-eyes are extensively discussed (see these abstracts, No. **1849**).

1853: K. F. Niessen: Nodal planes in a perturbed cavity resonator, II (Appl. sci. Res. **B1**, 251-260, 1949, No. 4).

Starting from a field with one nodal plane, parallel to one of the walls of the original resonator, a combination of functions analogous to that described in Part I (see these abstracts No. **1815**) is obtained, fulfilling the conditions in the perturbed resonator. It appears that two of these functions must be multiplied by coefficients of the order δ. For the ratio of the coefficients of order 1 two values are found and consequently two solutions for the perturbed field with two different frequencies. Neither of these solutions returns to the original unperturbed function for $\delta \rightarrow 0$. There follows a brief discussion of the behaviour of the nodal plane.

R 102: D. S. Saxon and R. A. Hutner: Some electronic properties of a one-dimensional crystal model (Philips Res. Rep. **4**, 81-122, 1949, No. 2).

The Kronig-Penney model of a one-dimensional crystal is studied further with the aid of the formalism of the Dirac δ-function. Both a Green's-function method and a scattering-matrix method are used to derive the energy levels and wave functions for the monatomic and the diatomic lattices. The scattering-matrix method is used to study the problems of a single impurity, both substitutionally and interstitially located, and of the coupling between impurities. In the last section the general problem of a solid solution is discussed.

R 103: P. Cornelius and H. C. Hamaker: The rationalized Giorgi system and its consequences (Philips Res. Rep. **4**, 123-142, 1949, No. 2).

It is shown in this paper that, if we interpret electromagnetic phenomena by means of field concepts, if we base our arguments on the simplest type of field, namely the homogeneous field, and if we make use of the pronounced analogy between the current field in a conductor, the electrostatic field in a capacitor, and the magnetic field in a coil, we are then led by a straightforward line of argument to the rationalized Giorgi system, which can thus be explained and introduced in a convenient way. By slightly extending these considerations, material properties, such as polarization, susceptibility, can be defined in a consistent and simple manner. It is pointed out that discussions regarding dimensions, "Grössengleichungen", and fundamental and derived units are generally inconclusive and of no interest in practice (see these abstracts No. **1803**).

R 104: K. F. Niessen: Anomalous skin impedance as a function of the field strength (Philips Res. Rep. **4**, 143-153, 1949, No. 2).

The skin effect in a metal of high conductivity is studied in the case of a strong field of high frequency, using Pippard's original assumption of electrons going to and from the surface only in the normal direction. The curvature of the average electron path by the internal field is taken into account in a very simplistic manner, from which a dependence of the anomalous skin impedance on the field strength may be expected, viz. a decrease of the resistivity and an increase of the reactance for an increasing amplitude of the electric field on the surface.

VOL. 11 No. 5, pp. 133-164 NOVEMBER 1949

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

AN INSTALLATION FOR MULTIPLEX-PULSE MODULATION

by C. J. H. A. STAAL.

621.396.41: 621.396.619.16: 621.396.619.232

For radio-telephone communications very short waves are desired, because for one thing they can easily be beamed; this promotes secrecy and the transmitter can work on a lower power. With these very short waves the ordinary methods of modulation — amplitude and frequency modulation — cannot very well be used. What is required is pulse modulation, whereby pulses are time-modulated or width-modulated. Between these pulses others can be introduced the modulation of which can be employed for transmitting other conversations. This is termed multiplex-pulse modulation. An eight-channel installation working on this principle is described.

Introduction

In certain cases it is found very difficult and consequently highly expensive to lay a cable for a telephone or telegraph link between two places; we have in mind here, for instance, a link with an island or between two places in mountainous country. It may then be much more economical to have recourse to a radio connection which in its properties somewhat resembles a cable link and for which the French have a most typical name, "câble hertzien", (radio waves are sometimes also called Hertzian waves). For the properties of a radio link to resemble those of a cable link the whole of the electromagnetic energy of the transmitter has to be emitted as far as possible in the direction of the receiver, so that as little energy as possible is lost. Furthermore, just as in the case of a cable link, the neccssary secrecy has to be ensured. The transmitters serving for such links are called "link transmitters" [1]). Now the beaming of electromagnetic waves is all the better according as the dimensions of the beaming device employed (e.g. parabolic mirror, horn, Yagi antenna) are greater with respect to the wavelength on which the link transmitter is working. Since the beaming device should preferably be kept as small as possible, for a good directional effect it is best to work on the shortest possible wavelength, viz. centimetric waves. But the use of very short waves has two consequences.

In the first place the application of a Hertzian cable is then limited to cases where the distance between the two places to be linked is not greater than the visual distance. Under certain conditions this may make it necessary to set up the aerials at a very high elevation.

In the second place, with centimetric waves the problem of modulation requires to be solved in a manner different from what is customary with longer waves, where amplitude or frequency modulation of a continuous, sinusoidal carrier is applied. As is known, centimetric waves are generated with the aid of magnetrons or velocity-modulation valves. It appears that these valves cannot very well be used for amplitude or frequency modulation, since it is very difficult to modulate in amplitude without frequency modulation arising at the same time, and vice versa. Moreover, the modulation characteristics are far from linear. When several conversations are being transmitted simultaneously, as is mostly required, a slight deviation from linearity is sufficient to cause troublesome cross-talk between the channels.

One must therefore employ other methods of modulation. With pulse modulation it is possible to overcome these difficulties, while still attaining a signal-to-noise ratio (determining the

[1]) See for instance Philips Techn. Rev. 2, 171-176 and 229-306, 1937; 3, 59-64, 1938; 6, 120-127, 1941.

quality of the link) of the same order as reached with normal amplitude or frequency modulation. What is to be understood by pulse modulation will be made clear below.

We have just mentioned the requirement that it must be possible to transmit several conversations simultaneously. Therefore, just as a single cable becomes, as it were, multicored thanks to the methods of carrier telephony, so it must also be possible for a "Hertzian cable" to be multi-cored.

We shall first begin by dealing briefly with the principles of pulse-modulation methods. One of these methods, so-called pulse time modulation, forms the basis for the system of multiplex pulse modulation which will be described here. We shall confine ourselves to a description of the modulator and of the demodulator. The high-frequency part of the transmitter and receiver will be left out of consideration here.

Principle of pulse modulation

When pulse modulation is applied the transmitter does not produce a continuous carrier (*fig. 1a*) but a sequence of high-frequency wave trains following each other at a certain interval of time (fig. 1*b*). The high-frequency voltage is then modulated by pulses. Typical of this modulation is the fact that both the amplitude and the frequency of the high-frequency alternating voltage are kept constant during the pulses.

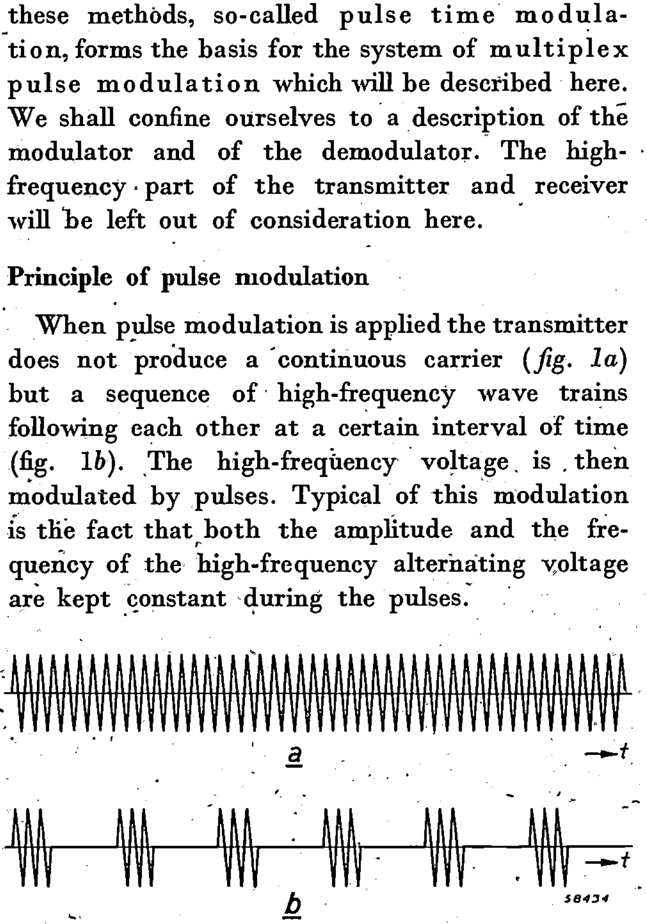


Fig. 1. *a*) A continuous carrier wave as function of the time *t*. *b*) A "carrier wave" consisting of wave trains in the form of rectangular pulses.

So long as the amplitude, width and interval of the pulses remain the same, the wave trains transmit only the repetition frequency of the pulses. To transmit a low-frequency signal one of the said factors is varied in the rhythm of this signal, or in other words the pulses are modulated with the low-frequency signal, which pulses in turn modulate the high-frequency carrier.

With this double modulation the series of pulses acts as a sort of auxiliary carrier. Since the high-frequency part (including the pulse modulation of the H.F. carrier) is beyond the scope of this discussion, the term pulse modulation will refer, throughout the remainder of the article, to the pulse modulation of the transmitted low-frequency signal. The form of the pulses need not in principle be rectangular, as represented in fig. 1, but in this article the pulses are understood to be rectangular.

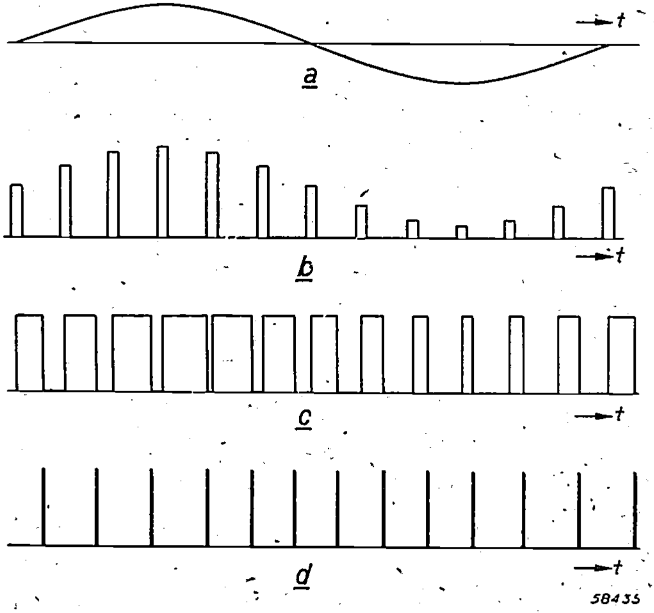


Fig. 2. Different methods of pulse modulation. In (*b*), (*c*) and (*d*) the amplitude, width or time, respectively, of the pulses is modulated in the rhythm of the low-frequency signal drawn in (*a*) as a sinusoidal curve. In (*d*) the pulses have been drawn very narrow for the sake of simplicity. (The time has been plotted on the horizontal axis.)

According as the amplitude of the pulses, their width or their position is varied in the rhythm of the low-frequency signal one speaks of pulse-amplitude, pulse-width or pulse time modulation (see *fig. 2*). For the reasons already indicated above, for very short waves pulse-amplitude modulation cannot be considered, because then also the amplitude of the high-frequency carrier is varied, and this is just what has to be avoided.

It is clear that in this way the low-frequency signal is not transmitted in its entirety but only in "pieces", though in a sufficiently large number to be able to build up again the whole signal at the reception end. Thus the continuous curve representing the amplitude of the signal as a function of time is transmitted in the form of a succession of dashes (see *fig. 3*), each dash corresponding to a certain pulse and its ordinate indicating the time (or width or amplitude) of the pulse.

To characterize a continuous curve unambiguously by dashes, these must lie sufficiently close together, and the less simple the curve the closer the dashes have to be. To put it more precisely, this means that the repetition frequency of the pulses must be great with respect to the highest signal frequency to be transmitted. It appears that good results can already be obtained when the said frequency ratio is 2.5 to 3. In a conversation the highest frequency to be transmitted is roughly equal to 3400 c/s. The repetition frequency must therefore be at least about 10,000 c/s.

Principle of multiplex-pulse modulation

If several conversations, say n in number, are to be transmitted simultaneously with pulse modulation then this is done, in principle, in the following way. A series of equal and equi-distant pulses are produced. Let us suppose that these pulses are numbered in their natural sequence. For the first conversation one modulates (in amplitude, width or time) only those pulses numbered $p + (n + 1), p + 2\ (n + 1), p + 3\ (n + 1)$, etc. ($p$ being a whole number), for the second conversation only the pulses $p + 1, p + 1 + (n + 1), p + 1 + 2\ (n + 1), p + 1 + 3\ (n + 1)$, etc, thus in general for the q-conversation ($1 \leqq q \leqq n$) only the pulses $p + q - 1, p + q - 1 + (n + 1), p + q - 1 + 2\ (n + 1), p + q - 1 + 3\ (n + 1)$. etc.

The pulses numbered $p + n, p + n + (n + 1), p + n + 2\ (n + 1), p + n + 3\ (n + 1)$ etc. are left unmodulated; they serve to indicate on the time axis the place where a new series of n impulses begins (belonging to the n different conversations) and are called synchronizing pulses.

Thus there corresponds to each conversation a specific series of equi-distant pulses with a repetition frequency ($n + 1$) times as small as that of the original series of pulses.

If we draw a comparison with the known method of transmitting several conversations simultaneously as applied in carrier-telephony, it is seen that whereas in carrier-telephony the various "channels" (conversations) are displaced in frequency and laid side by side, with multiplex-pulse modulation they are displaced in time and laid side by side.

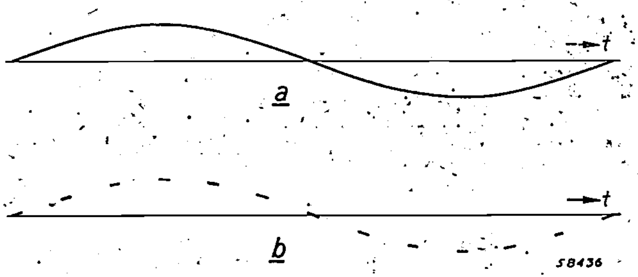


Fig. 3. With every method of pulse modulation a curve (a) representing a signal as a function of time is transmitted in the form of a sequence of dashes (b).

When in carrier-telephony more channels have to be transmitted then the total frequency band is made proportionally wider, whilst with multiplex-pulse modulation the total number of pulses per second, thus the repetition frequency, becomes proportionally greater. The fact is that on the one hand the minimum repetition frequency of pulses belonging to one conversation is fixed — as already stated, it is about 10,000 c/s — whilst on the other hand in the approximately 100 microseconds between two successive pulses belonging to one conversation a pulse of each of the other conversations has to be taken up.

In the case where, as with the system to be described below, one is working for instance with a total of nine channels, namely eight speech channels and a synchronizing channel ($n = 8$), the interval between two successive pulses is 11.1 μsec. and the repetition frequency of the total number of pulses 90 kc/s.

The technical realization of the principle of multiplex-pulse modulation outlined above depends, of course, upon the method of pulse modulation to be applied, but even with one particular method the execution may still vary considerably. In this article we shall direct our attention exclusively to the apparatus for multiplex-pulse modulation which has been developed by Philips and is illustrated in *fig. 4*.

The modulator

The method of pulse-modulation applied

The installation to be described here is based upon pulse-width modulation converted into pulse-time modulation, so that the actual transmission is in time-modulated pulses.

The unmodulated pulses are generated by causing a saw-tooth voltage (*fig. 5a*) to act upon the control grid of a valve. The valve will pass current through so long as the saw-tooth voltage is positive, that is to say in the first half of each saw-tooth cycle (starting to count the cycles from a vertical part of the saw-tooth line). Owing to the flow of grid current, a large portion of the positive part of the saw-tooth voltage appears across a series resistor, so that during this interval the grid voltage itself is limited. Consequently the anode current consists of a series of current pulses (fig. 5b) with the duration of each pulse amounting to half the saw-tooth cycle. When a low-frequency signal is applied to the same control grid (fig. 5c), then the point of intersection of the downward part of the saw-tooth on the horizontal time axis will be displaced to the left

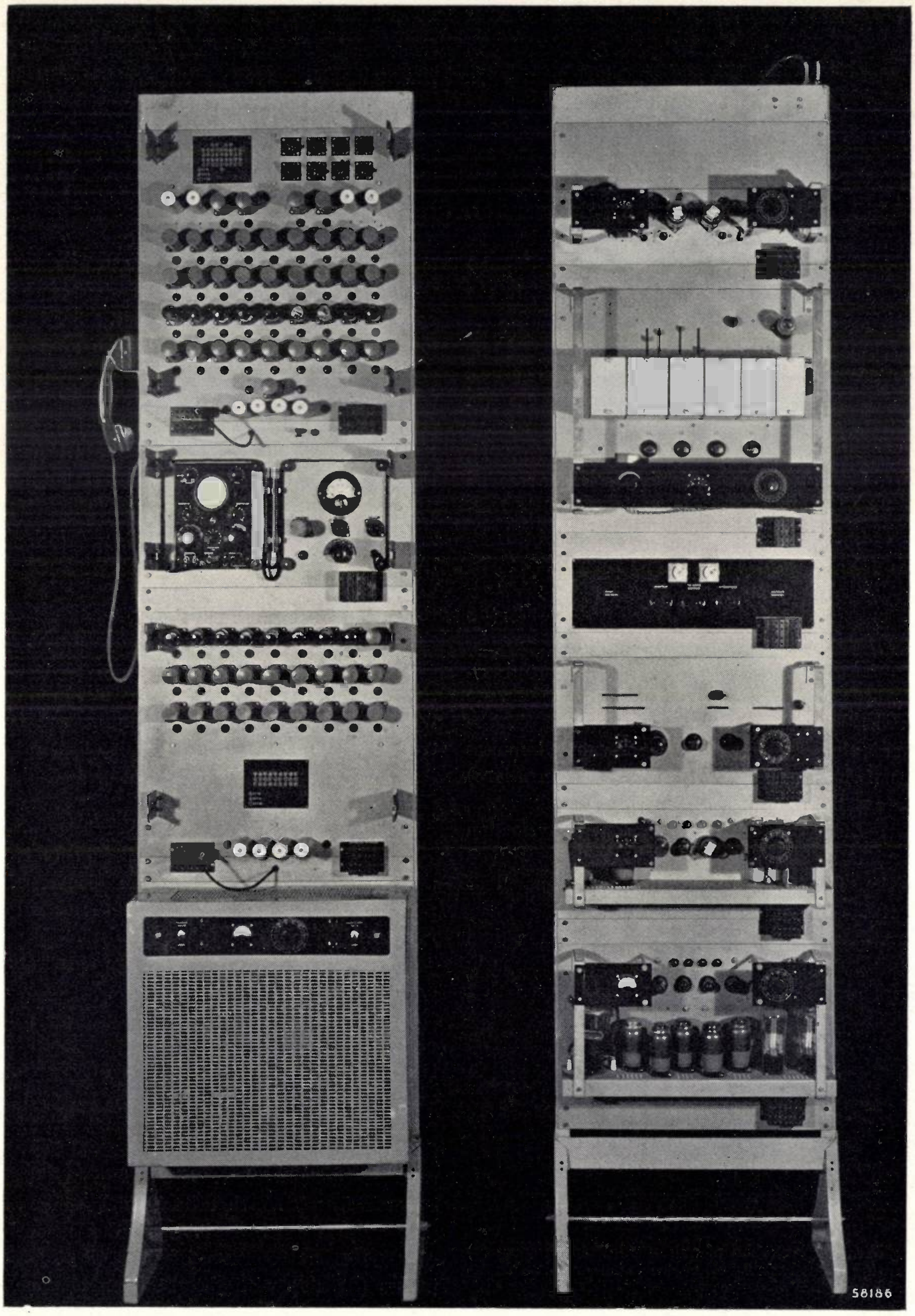


Fig. 4. Transmitting and receiving installation for eight channels with multiplex-pulse modulation. The left-hand bay comprises (from top to bottom) the modulator, a testing panel with oscillograph, the demodulator and the supply unit. The right-hand bay contains the high-frequency part of the apparatus.

or right over a distance proportional to the amplitude of the low-frequency signal; the point of intersection of the vertical part of the saw-tooth on the time axis is not displaced (fig. 5*d*). The valve will therefore interrupt the current somewhat earlier or somewhat later then when there is no low-frequency signal, so that the current pulses will be somewhat shorter or longer (fig. 5*e*) and one thus gets a width-modulation of the pulses.

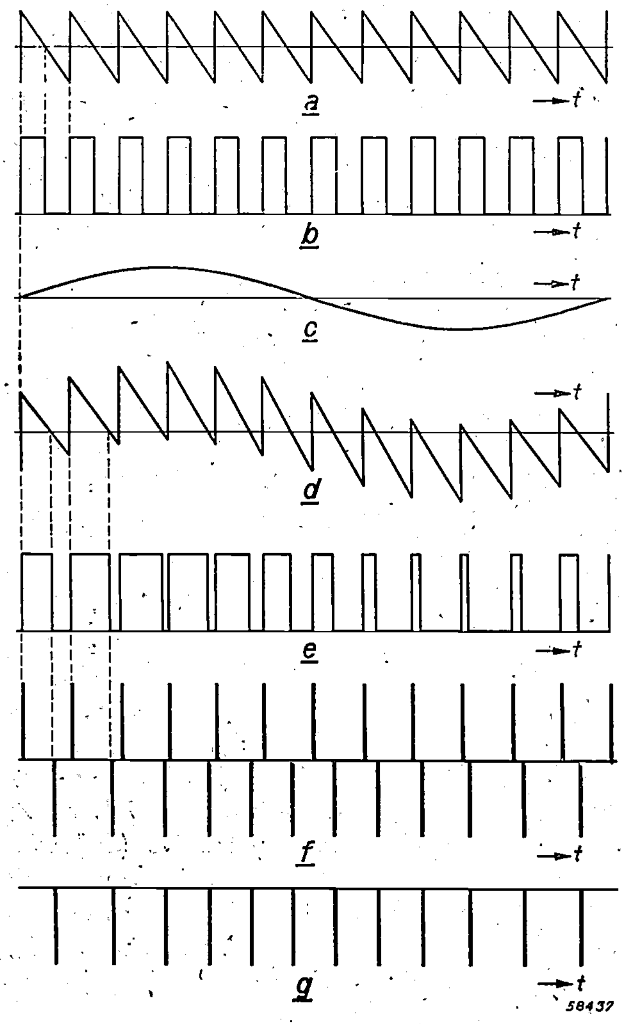


Fig. 5. Conversion of broad width-modulated pulses into narrow time-modulated pulses.
A series of identical equi-distant current pulses (*b*) is obtained by applying a saw-tooth voltage (*a*) to the control grid of a valve. If at the same time a sinusoidal voltage (*c*), i.e. a "signal", is likewise applied to the control grid then the resulting voltage assumes the shape (*d*). The pulses are then modulated in width (*e*). From this series of width-modulated current pulses one can obtain by differentiation a series of narrow positive and negative voltage pulses (*f*). When the positive pulses are suppressed, a series of time-modulated negative pulses remains (*g*).

With the aid of a differentiating network these width-modulated pulses are changed into a series of (much narrower) positive and negative pulses corresponding respectively to the beginning and the end of each of the width-modulated pulses (fig. 5*f*). Now the starting points of these latter pulses are fixed (at fixed intervals of time of 11.1 μsec), and the positive narrow pulses can therefore be cut off without any objection. What then remains are negative narrow pulses varying in time with the low-frequency modulation (fig. 5*g*).

The advantage of transmitting exclusively the narrow pulses lies mainly in the great saving of power that can thereby be obtained, since all the energy otherwise required for transmitting the broad pulses can now be used during the very short duration of the narrow pulses. With a certain average power of the transmitter the peak power in the narrow pulses can thus be made much greater.

Set-up of the modulator for eight channels

The manner in which the modulator works will now be described with reference to the block diagram given in *fig. 6*. Each rectangle in this diagram represents a group of circuit elements. Identical groups of elements are denoted by the same Roman figure (*I . . . IV*). Each of the vertical columns numbered *1* to *8* relates to one of the eight speech channels. The column marked *S* corresponds to the synchronizing channel.

Let us now consider one of the eight columns taken at random. In *I* the modulation signal (conversation) to be transmitted is amplified. In *II* the same takes place with the saw-tooth voltage, which has a frequency of 90 kc/s and is derived from the sinusoidal voltage supplied by the oscillator *A* and delivered to *B*. The superposed modulation and saw-tooth voltages are applied to the control grid of a valve represented in the diagram by a rectangle *IV*. Disregarding for a moment the rectangles *III*, in this way there would be produced in the valve *IV* width-modulated pulses having a frequency of 90 kc/s; the valve *IV* would then be alternately open and closed, on an average (in the absence of modulation even exactly) every 5.55 μsec.

However, a pulsating voltage generated by the circuit *III* (to be described farther on) is applied to a second grid (the screen grid) of the valves *IV*, which are pentodes. The result is that only one of the nine pulses is actually produced each time. These pulses then have a repetition frequency of 10 kc/s, which, as we have seen, is high enough to permit reconstruction of the modulation signal at the receiving end. As will presently be explained, the voltages supplied by the circuits *III* are such that the series of pulses derived from the columns *1 . . . 8* (fig. 6) are mutually displaced in time.

In accordance with the principle of multiplex-pulse modulation described they can therefore be placed side by side in time.

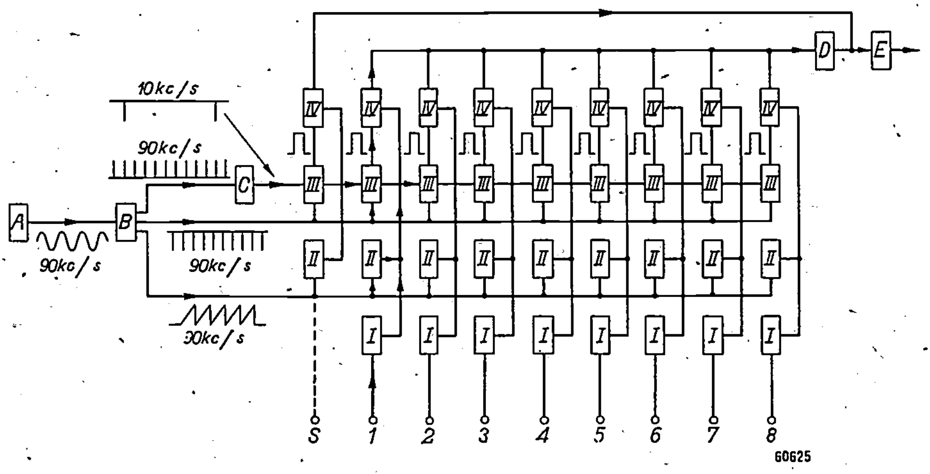


Fig. 6. Block diagram of modulator. A = an oscillator supplying a sinusoidal voltage with a frequency of 90 kc/s. In B this voltage is converted ino a saw-tooth voltage and into a series of pulses with the same frequency. In C the series of pulses with a frequency of 90 kc/s are converted into a series of pulses having a frequency of 10 kc/s. The eight columns correspond to the eight channels. Each column comprises a signal amplifier (*I*), a saw-tooth amplifier (*II*), a channel-gate oscillator (*III*) and a circuit (*IV*) in which the pulses are produced and modulated. The column S corresponds to the synchronizing signal. In D the width-modulated signal pulses are converted into time-modulated pulses. Between D and E the synchronizing pulses are added. In E the whole is amplified and limited.

The result is that the situation at the input of the differentiating circuit D corresponds to what is diagrammatically represented in *fig.* 7*a* (for the absence of modulation): there is a series of broad (average 5.55 μsec) pulses to be modulated in width, occurring in groups of eight (8 channels); between each two successive groups a place (11.1 μsec) is left open for the synchronizing pulse, which is not differentiated, so as to distinguish it from the others. After differentiation in D these broad and width-modulated pulses yield a series of narrow and time-modulated negative pulses and a series of narrow *un*modulated positive pulses (fig. 7*b*). To these there is added the broad synchronizing pulse obtained from column S (fig. 6), so that at the input of the circuit E (fig. 6) we get a series of pulses diagrammatically represented in fig. 7*c*. In E the whole of the input is amplified and limited (the positive pulses being cut off at the same time), after which it is fed to the transmitter.

The channel-gate circuit

We shall now describe a so-called channel-gate circuit (represented in fig. 6 by the rectangles *III*). The object of these channel-gate circuits is to ensure (1) that of the pulses which each of the valves *IV* would otherwise produce only one out of the nine actually arises, or, to express it less correctly, is passed through (hence the name channel-gate); (2) that the pulses derived from the eight channels come to lie side by side displaced in time, as represented in fig. 7*a*, and that the synchronizing pulses subsequently to be added fall in the places left open (fig. 7*c*).

As already stated, the first object is attained by making use of the screen grid of the valves *IV*; the screen-grid voltage is only positive (and constant) during one of the nine cycles of the saw-tooth voltage (90 kc/s), which is applied to the control grid together with the modulation voltage (*fig.* 8*a-c*); during the other eight saw-tooth cycles the screen-grid voltage is zero (fig. 8*d*). The repetition frequency of the width-modulated anode-current pulses is thus 10 kc/s (fig. 8*e*).

The second object — to ensure that the pulses of the eight channels come to lie side by side in time — is achieved by seeing to it that the positive and constant voltage (channel-gate voltage) is applied to the screen grid of a valve *IV* exactly at the right instant and during the right time, viz. 11.1 μsec, the duration of a saw-tooth cycle. In other words, each channel gate opens at the same moment as the preceding one closes, each gate remaining open 11.1 μsec.

This is done by employing a "chain" circuit, whereby the closing of a channel gate causes the

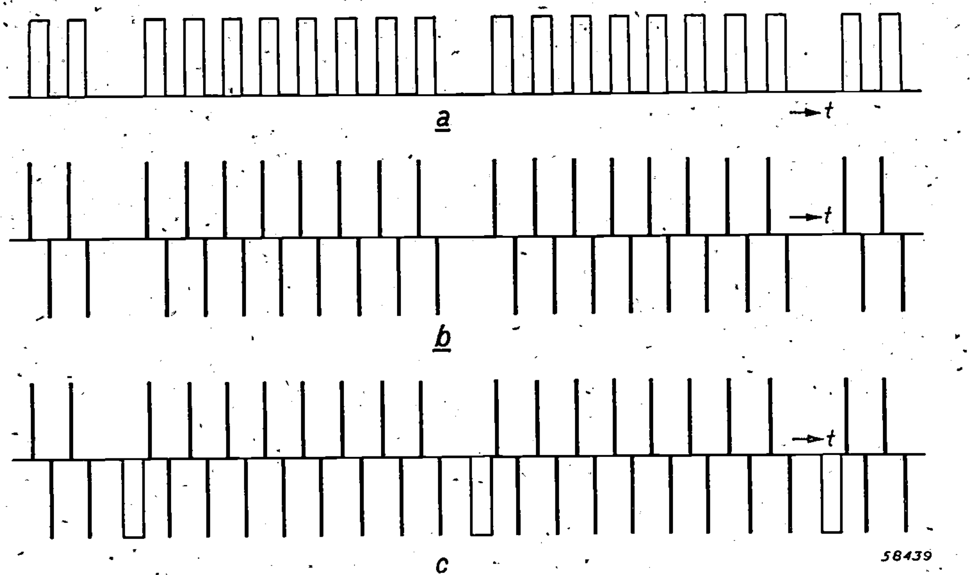


Fig. 7. *a*) The pulses corresponding to all eight channels, as they enter the differentiating circuit *D* (see fig. 6). For the sake of simplicity the pulses are all drawn with the same width. *b*) The same series of pulses after differentiation in *D*. *c*) The same series of pulses to which the series of synchronizing pulses have been added. (This corresponds to the position at the input of the circuit *E* in fig. 6.)

next channel gate to open. This circuit consists of nine (one for each channel and one for the synchronizing signal) identical "channel-gate oscillators" (the nine rectangles denoted by the figure *III* in fig. 6).

The diagram of such a channel-gate oscillator (*fig. 9*) resembles that of the known multi-vibrator of Abraham and Bloch. By slightly modifying the circuit, however, a certain asymmetry has been obtained so that the circuit does not oscillate continuously but has a certain position of rest to which it automatically reverts after a disturbance of the equilibrium.

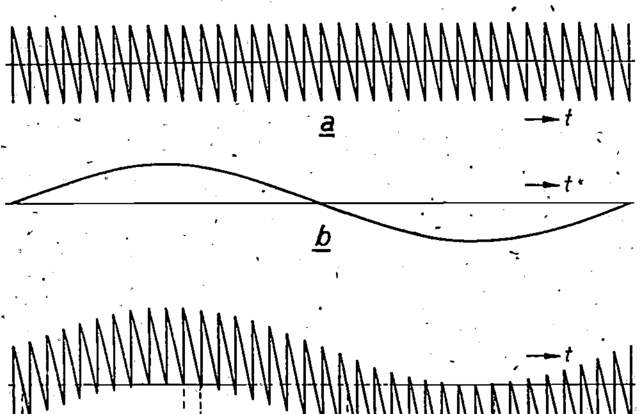


Fig. 8. Illustrating how pulses belonging to one channel are produced. The saw-tooth voltage (*a*) and the signal voltage (*b*) are applied to the control grid of a pentode (denoted by *IV* in fig. 6); superposition of these voltages produces (*c*). The channel-gate voltage (*d*) is applied to the screen grid. The resulting anode-current pulses modulated in width are represented in (*e*).

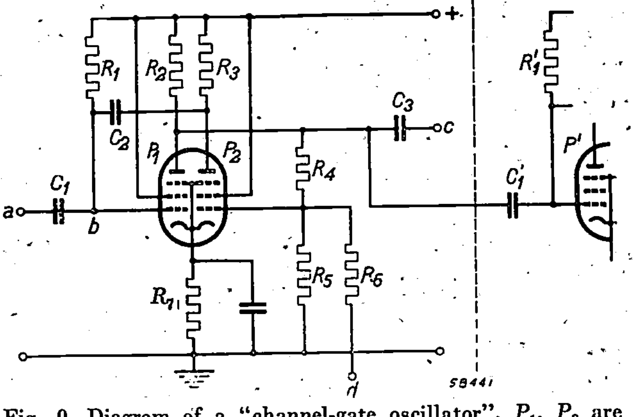


Fig. 9. Diagram of a "channel-gate oscillator". P_1, P_2 are two pentode systems in one envelope (EFF 51). The voltage at the various points (*a*, *b*, *c*, *d*) of the circuit is represented in fig. 10 as a function of time.

We have carried out this circuit with an EFF 51 valve, which contains two pentode systems in one envelope. The control grid of one pentode, P_1, is connected via a high resistance R_1 to the positive pole of the anode voltage source: thus P_1 passes current. As a result the voltage drop across the anode resistor R_2 is great and the anode of P_1 has a low voltage with respect to the cathode. The control grid of the other pentode, P_2, is connected to a tapping on the voltage divider R_4-R_5, which is connected between the anode of P_1 and earth. By a suitable choice of R_4 and R_5 and of the cathode resistor R_7, the control grid of P_2 goes so strongly negative that the pentode P_2 does not pass any current.

When at the instant t_1 the current through P_1 is interrupted by means of a negative pulse applied to the control grid via the capacitor C_1, the voltage at the anode of P_1 rises and thus also the voltage at the control grid of P_2, whereupon P_2 starts to conduct. This causes the voltage at the anode of P_2 to decrease. The voltage across the capacitor C_2 then adapts itself to the changed conditions by means of a charging current through R_1 and R_3, so that the voltage at the control grid of P_1 remains negative even in the absence of the pulse cutting off P_1. This new situation, where P_1 is not conducting and P_2 is conducting, will last until at t_2 the charging current of C_2 has become so small that the voltage across R_1 is not sufficient to keep P_1 cut off. Then P_1 will again begin to pass current and ultimately the initial situation is restored.

Fig. *10a-c* represents the voltage at various points of the circuit as a function of time. From fig. 10*c* it can be seen that the voltage at the anode of P_1 (fig. 9), which voltage is taken off via the capacitor C_3, has the shape required for a "channel gate" for the screen-grid voltage of a pentode *IV*, provided the time t_2 to t_1 is equal to 11.1 μsec and the negative pulse is repeated with a frequency of 10,000 c/s.

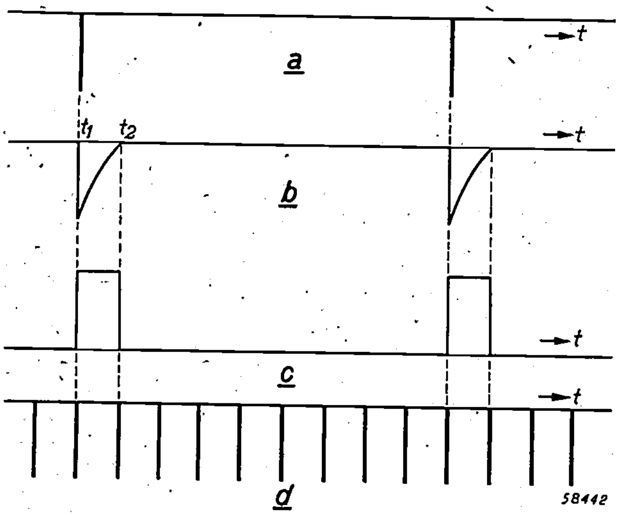


Fig. 10. The voltage as a function of time at the points *a*, *b*, *c*, *d* of the channel-gate circuit (fig. 9). (*a*) represents the pulse activating the circuit, (*b*) the voltage at the control grid of P_1, (*c*) the voltage at the anode of P_1 (the channel-gate voltage), (*d*) the auxiliary pulses fixing the duration of the channel-gate voltages at 11.1 μ sec.

The manner in which nine such circuits are interlinked and bring about the opening and closing of the nine channel gates is as follows. The first circuit — rectangle *III* in column *S* of the block diagram given in fig. 6 — is activated with negative pulses produced in *C* with the repetition frequency of 10,000 c/s. For the activation of the second circuit, 11.1 μsec later, the voltage at the anode of P_1 is "differentiated", so that at the moment that this voltage returns to its original value a negative pulse occurs; this negative pulse is applied to the control grid of P_1' of the second circuit (via the capacitor C_1' corresponding to C_1 of the first circuit). In the same way the second circuit supplies a negative pulse for the third one, the third for the fourth, and so on. This "chain" continues to operate until the last (ninth) circuit has returned to rest and thus the whole channel-gate circuit is at rest. At that moment, 100 μsec after the activation of the first circuit, the latter again receives a negative pulse from *C*.

The "differentiation" of the anode voltage of P_1 is brought about by the elements C_1', R_1', P_1' of the following channel-gate oscillator. Owing to the control-grid current passing through R_1' the control-grid-cathode part of P_1' has a low internal resistance forming together with the capacitor C_1' the differentiating circuit.

It is highly important that all the channel gates remain open exactly the same length of time, namely during exactly one cycle of the saw-tooth voltage, i.e. 11.1 μsec. Now the time during which a channel gate is opened is determined by the *RC*-time of the circuit formed by C_2 and $R_1 + R_3$ (fig. 9). In this circuit there lies a source of errors. In order to keep the channel gate open for exactly the same length of time as the duration of the saw-tooth cycle, a negative pulse with a frequency of 90,000 c/s (fig. 10*d*) is applied to the control grid of the pentode P_2 of each double valve (via the resistor R_6); these pulses coincide with the beginning of the saw-tooth cycles and are produced in the same circuit *B* as produces the saw-tooth voltage (see block diagram, fig. 6). For the greater part of the time P_2 is cut off and these negative auxiliary pulses will have no effect. When, however, current begins to flow through P_2 and if the channel gate should tend to remain open longer than the prescribed 11.1 μsec, then the auxiliary pulse ensures that the current passing through P_2 is interrupted after 11.1 μsec and the circuit again comes to rest.

The demodulator

The anode follower

Let us now see what happens at the receiving end. The high-frequency wave trains are picked up by an aerial in the normal way and led to the high-frequency part of a receiver. After detection this part produces pulses of the same form as those supplied to the transmitter. Now these pulses first have to be separated into series corresponding to the individual low-frequency modulation signals (conversations). Then each series has to be demodulated, that is to say the low-frequency modulation signal has to be reconstructed from each series. We shall now explain how these two stages work in our apparatus, though not in the same order in which they have just been mentioned.

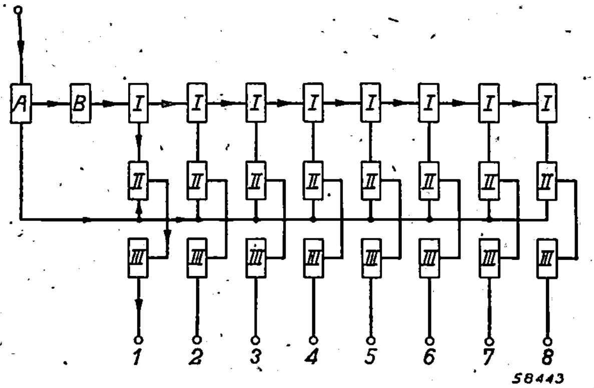


Fig. 11. Block diagram of the demodulator. A = amplifier and limiter; B = circuit in which the synchronizing pulses are separated from the other pulses; I = channel-gate oscillators; II = "anode followers"; III = amplifiers and low-pass filters. The eight columns correspond to the eight channels.

Fig. 11 is a block diagram of the demodulator. Each of the eight columns numbered *1* to *8* corresponds to one channel (conversation). The rectangle *I*, the channel-gate circuit, will be dealt with later. In *II* the channels are separated with the aid of the channel-gate voltages supplied by *I*, whilst at the same time the low-frequency speech frequencies are reconstructed from the separated pulses belonging to one channel. As regards the circuit *III* there is nothing much to be said. It contains a low-frequency amplifier and a low-pass filter, the latter "smoothing out" the signals having the speech frequencies. The output is matched to the telephone line connected to it.

The circuit *II*, however, calls for a more detailed explanation. Once the pulses belonging to one channel have been separated from the others (the manner in which this takes place in *II* will be shown later on), then in principle we could obtain therefrom the low-frequency speech component by filtering out all other frequencies, the repetition frequency and its harmonics, with the aid of a low-pass filter. Since, however, the desired component is only weakly represented in the frequency spectrum of the pulses, such a method requires a great deal of amplification and very selective filters. The same applies for other methods whereby time-modulated pulses are reconstructed into width-modulated pulses or converted into amplitude-modulated pulses. In our installation, however, an entirely different method of demodulation is followed, namely with the aid of a circuit whereby use is made of a secondary-emission valve and which belongs to the group of "anode-follower" circuits (so-called because, with a suitable choice of components, the potential of the auxiliary cathode follows the anode voltage variations).

To explain the action of an anode follower we must consider the characteristics of a secondary emission valve which represent the relation between the current I_{k2} and the voltage V_{k2} of the auxiliary cathode k_2, with the anode voltage V_a as parameter and with a constant voltage on the control and screen grids (*fig. 12*).

When the value of V_{k2} is very small only a few electrons emerge through the openings in the screen grid and reach k_2; in other words, there is a weak current I_{k2}. The direction of this current (electrons striking k_2) will be called positive.

As V_{k2} increases (but for the time being still remaining smaller than V_a) the electrons impinging upon k_2 set up secondary emission, the number of electrons leaving k_2 per second exceeding the number reaching it, so that I_{k2} is reversed. The secondary electrons move towards the anode, which has a higher potential than k_2.

When V_{k2} is allowed to exceed V_a the secondary electrons are unable to reach the anode and return to k_2, so that I_{k2} again becomes positive; actually k_2 then plays the part of an anode. When $V_{k2} = V_a$, $I_{k2} \approx 0$.

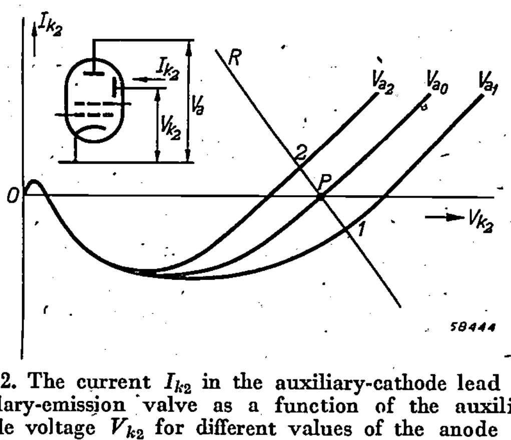


Fig. 12. The current I_{k2} in the auxiliary-cathode lead of a secondary-emission valve as a function of the auxiliary-cathode voltage V_{k2} for different values of the anode voltage V_a. The oblique straight line is the "resistance line".

If a resistor R is connected in series with k_2, and using V_b to denote the supply voltage of the auxiliary cathode circuit, then upon the value of V_a being varied I_{k2} becomes equal to zero when $V_a \approx V_b$. For any other value of V_a (say V_{a1} or V_{a2}) I_{k2} adjusts itself to the point of intersection (*1* or *2* respectively) of the respective characteristic on the "load line", i.e. the straight line drawn with a slope = $1/R$ through the point P on the abscissa with $V_{k2} = V_b$. If R is very large then this line is almost horizontal and the points *1*, *2* fall practically on the abscissa, that is to say V_{k2} is then always approximately equal to V_a. Hence the name "anode follower".

Demodulation of a channel with the aid of the anode follower

The manner in which the anode follower is used for demodulating the pulses is as follows. The circuit [2]) is represented in *fig. 13.* The control grid, the screen grid and the anode of a secondary-emission valve EFP 60 serve as input electrodes, and the auxiliary cathode as output electrode. The whole group of pulses of all the channels is applied to the control grid, a channel-gate voltage (supplied by the circuit *I*, fig. 11) is applied to the screen grid, whilst there is fed to the anode — in addition to a direct voltage — a saw-tooth voltage derived via an integrating network from the channel-gate voltage. The auxiliary cathode is fed via a resistor *R* (shunted by a capacitor *C*) from the same voltage source as the anode and is connected to the output terminal via a blocking capacitor. The channel-gate voltage ensures that the valve is open only for the respective channel and cut off for all other channels. (Where in the further description of this apparatus we speak of pulses only the pulses of one channel are meant.)

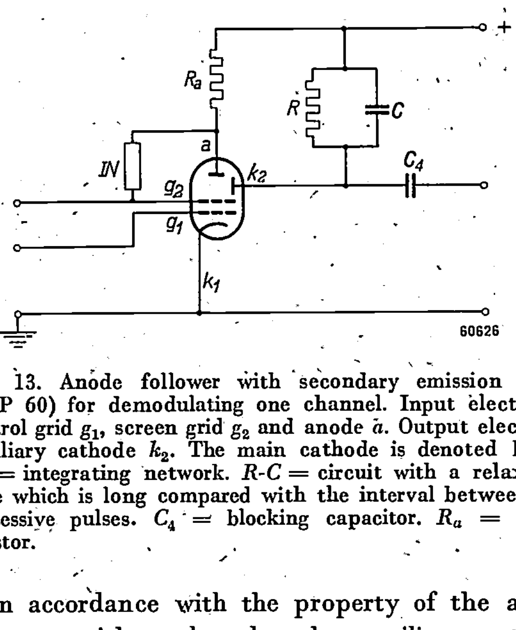


Fig. 13. Anode follower with secondary emission valve (EFP 60) for demodulating one channel. Input electrodes: control grid g_1, screen grid g_2 and anode *a*. Output electrode: auxiliary cathode k_2. The main cathode is denoted by k_1. *IN* = integrating network. *R-C* = circuit with a relaxation time which is long compared with the interval between two successive pulses. C_4 = blocking capacitor. R_a = anode resistor.

In accordance with the property of the anode follower, with each pulse the auxiliary cathode adjusts itself to the instantaneous value of the anode voltage. If there is no modulation then the pulses are in the intermediate position, whereby the anode voltage has just reached half way up the steeply ascending side of the saw-tooth. When modulation is present the pulses move to the left or to the right, with corresponding values of anode voltage lower or higher, as the case may be, in proportion to the displacement of the pulses. The auxiliary cathode follows these fluctuations in the anode voltage, so that the valve acts, as it were, as a switch momentarily connecting the auxiliary cathode to the anode at every pulse. When the pulse has passed, the auxiliary cathode voltage is maintained because the relaxation time *RC* is long compared with the interval of time between the pulses. This is made clear in *fig. 14a-d,* where the bottom curve represents the step-like variation of the output voltage. This voltage is further amplified in *III* (fig. 11).

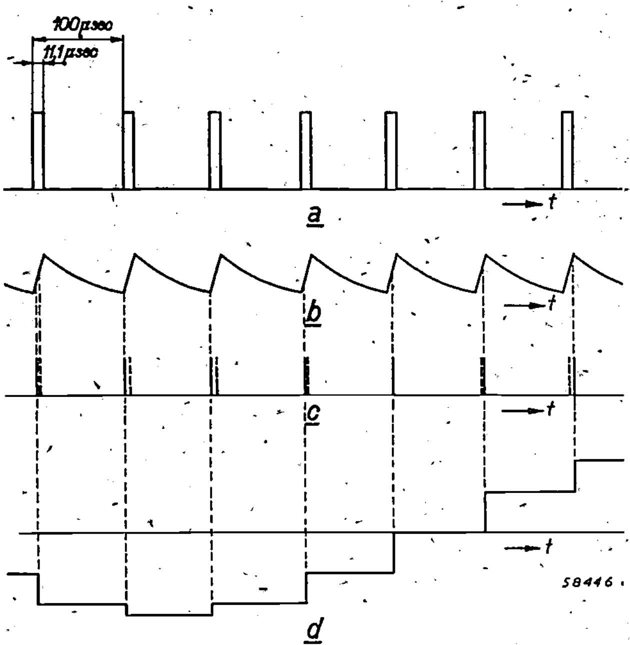


Fig. 14. The voltage as a function of time at various points of the circuit indicated in fig. 13: (*a*) channel-gate voltage at the screen grid, (*b*) saw-tooth voltage at the anode obtained from (*a*) by integration, (*c*) time-modulated signal pulses at the control grid (state of rest indicated by dotted lines), (*d*) output voltage (auxiliary-cathode voltage with the D.C. component removed).

As already observed, in the case of pulse modulation only a few instantaneous values of the low-frequency voltage are transmitted (*fig. 15a*).

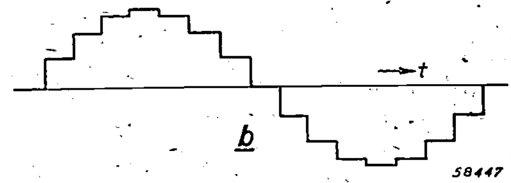


Fig. 15. Of a signal having the shape of a sinusoidal curve only a sinusoidal succession of dashes (see fig. 3) is transmitted such as represented in (*a*). By demodulation a step-like curve (*b*) is obtained therefrom.

[2]) For this circuit we are indebted to Tj. Douma (Philips Research Laboratory).

Since it is not known whether the pulse next in succession corresponds to a larger or a smaller amplitude, the step-like curve illustrated in fig. 15*b* is the best approximation of the signal that can be expected from the curve of successive dashes (fig. 15*a*). The step-like curve, contrary to the pulses, contains the desired low-frequency signal to a high degree. A simple low-pass filter suffices to filter out the undesired higher harmonics, so that there is no need for very selective filters or for any great amplification.

Separation of the eight channels

To complete the description of the functioning of the demodulator we must say something about the manner in which the channel-gate voltages applied to the screen grid of the secondary-emission valves are derived. These voltages are represented in fig. 14*a* as a function of time. Since the installation has been designed for eight channels, there are eight such voltages required, so displaced in time that at any moment only one of them is positive. The same problem has already been encountered in the description of the modulator, and the solution provided there in the channel-gate circuit has likewise been applied for the demodulator. The eight rectangles *I* in the block diagram of fig. 11 represent in fact the same channel-gate circuit as the rectangles *III* in the block diagram of fig. 6 (in the demodulator there is no rectangle corresponding to the synchronizing signal as given for the modulator).

From the description of the channel-gate circuit we have seen that this circuit has to be repeatedly activated with a negative pulse, having a repetition frequency of 10 kc/s. For the demodulator the series of negative pulses required for this purpose is derived from the series of broad synchronization pulses, likewise having the repetition frequency of 10 kc/s [3]). This is done with the aid of a circuit denoted in the block diagram of fig. 11 by the letter *B*. All pulses, including the synchronizing pulses, are fed to this circuit and "differentiated". This differentiation is brought about with the aid of an *RC*-circuit so dimensioned that the *RC*-time is long compared with the duration of the signal pulses but short with respect to the duration of the synchronizing pulses (which are of course much wider). The result is that after the differentiation there is a positive voltage peak corresponding to the end of each synchronizing pulse which is much higher than the positive peaks corresponding to the end of each signal pulse (*fig. 16*) [4]). This makes it possible to separate the voltage peaks belonging to the synchronizing pulses from the other peaks with the aid of a valve, which then supplies the requisite negative pulses. This valve also forms part of the circuit *B*.

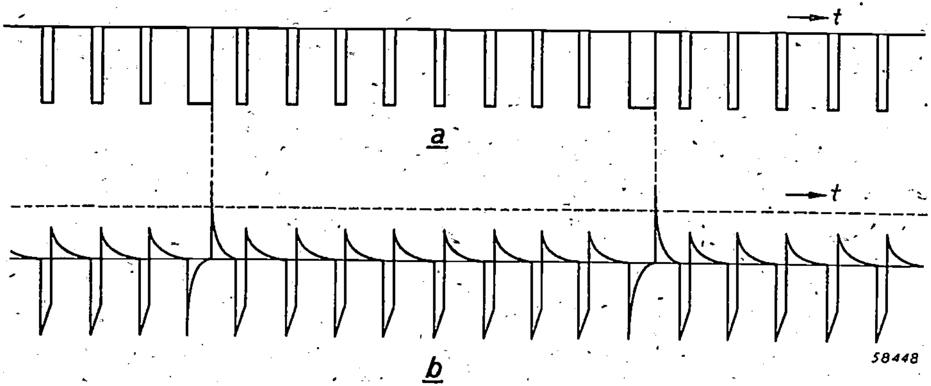


Fig. 16. *a*) The pulses corresponding to all eight channels and the synchronizing pulses, in the form in which they reach the input of the circuit *B* (fig. 11).
b) The same pulses after differentiation with a differentiating network forming part of the circuit *B*.

Signalling

As is known, signalling — the transmitting of the dialling note, the dialling pulses, the "engaged" signal, etc, — is an important aspect of telephone traffic. The technical problem of signalling, which

[3]) It is to be noted that in the demodulator there are no auxiliary pulses with the repetition frequency of 90 kc/s which in the modulator close the channel gates at the right moments. Here the time during which the channel gates are open thus depends entirely upon the *RC*-time of the circuit formed by C_2 and $R_1 + R_3$ (fig. 9). The right adjustment of this time has to be obtained by adjusting, for instance, the capacitor C_2.

[4]) See: Projection-television receiver, V. Synchronization, by J. Haantjes and F. Kerhhof, Philips Techn. Rev. 10, 364-370, 1949 (No. 12).

in carrier telephony is rather difficult [5]), is very easily solved with the form of multiplex-pulse modulation described here. With this method signalling is done by the transmission of direct-current pulses, just as is the case with local telephone traffic. As seen from the description of the demodulator, a fixed displacement of a pulse results in a certain change in the direct voltage at the auxiliary cathode of the anode follower. Consequently the D.C. dialling pulses, displacing a certain signal pulse from the state of rest, will cause analogous direct-voltage pulses at the auxiliary cathode, which can be amplified with a D.C. amplifier and then fed to the respective relays.

Final remarks

It remains to say something about the field of application of the apparatus described here for multiplex-pulse modulation.

Obviously this method of modulation is not likely to be applied in the case of beam transmission that has to form a link in a carrier-telephony system, for this would involve a double apparatus both at the transmitting and at the receiving end, namely at the transmitting end an apparatus for carrier-wave demodulation and one for pulse modulation, and at the receiving end converse apparatus. In such a case therefore a method of transmission will be sought whereby the carrier is transmitted in its entirety [6]). (In general one must then be satisfied with longer and thus less easily beamed waves.) If, on the other hand, there are not too many "detached" conversations to be put through (reaching the transmitter in a form other than that of a carrier-wave band) then the apparatus described here for multiplex-pulse modulation has great advantages, in that it is comparatively simple and inexpensive.

[5]) See for instance: F. A. de Groot, Signalling in carrier telephony, Phiilps Techn. Rev. 8, 168-176. 1946.

[6]) A. van Weel, An experimental transmitter for ultra-short-wave radio-telephony with frequency modulation, Philips Techn. Rev. 8, 121-128 and 193-198, 1946.

Summary. In beam-transmission links difficulties arise which can be met by the application of pulse modulation. In this article a brief explanation is first given of the principles of pulse modulation and of the so-called multiplex-pulse modulation, i.e. the transmission of several telephone channels with the aid of pulse modulation. Next a description is given of a new apparatus for multiplex-pulse modulation with eight channels. In the modulator the succession of pulses belonging to the various channels are obtained by means of a circuit consisting of a chain of channel-gate oscillators. In the demodulator the various channels are separated by a similar chain circuit. Demodulation itself is done with the aid of a circuit employing a special property of the secondary-emission valve (anode follower).

Hennie Wolt.

PIEZO-ELECTRIC MATERIALS

by J. C. B. MISSEL.

537.228.1

The intention is to publish in this periodical some articles giving an impression of the great advance made in the manufacture of quartz-oscillator plates in the U.S.A. during the last war and of the part played therein by the North American Philips Company. As an introduction to these articles, and to others which may be published dealing with more recent developments in this field, it is deemed advisable first to review some of the most important facts and conceptions of piezo-electricity, with particular attention to new piezo-electric materials.

The literature about the theory and practice of the piezo-electric effect has already reached a respectable size. This phenomenon and its application have numerous aspects deserving of consideration, and for the initiated the few pages that will be devoted to the subject here cannot but give the impression of being far from complete. For this the author is to be excused on the ground that this introduction is only intended to make a number of articles to be published on the subject in this periodical understandable without the reader having to consult other literature. For those who wish to go into the subject more deeply some extensive publications are quoted at the end of this article.

Piezo-electricity and its application

The phenomenon of piezo-electricity was discovered by Pierre and Jacques Curie in 1880. They found that under a mechanical load a quartz crystal exhibits electrical charges of opposed polarity on two opposite faces (that is to say it becomes an electrical dipole). Crystals possessing this property always show also the converse effect: when placed in an electric field they exhibit not only electrical polarization but also mechanical deformation. In both cases the effect is linear, that is to say the charges and the deformations are proportional to the causative forces and electrical field strength, and reverse in polarity together with them. Thus the inverse piezo-electric effect can be immediately distinguished from the phenomenon of electrostriction, i.e. the deformation that any substance, crystalline or amorphous, undergoes in an electric field and which varies with the square of the field strength.

The strength of the piezo-electric effect depends mainly upon the material and the direction in which the mechanical load (e.g. a pressure) is applied. The effect is very strong in the case of Seignette salt (Rochelle salt, potassium sodium tartrate): in a certain direction of the crystal a charge can be obtained amounting to about 8×10^{-9} coulomb per newton of the force applied (2.3×10^{-4} e.s.u./dyne) [1]. With quartz one gets in the most favourable direction about 2.10^{-12} C/N.

The practical applications that have been found for the piezo-electric effect are so commonly known that they need only be referred to briefly here. Its most obvious use is for indicating mechanical vibrations by electrical means and, conversely, for generating mechanical vibrations by means of alternating voltages. The first possibility has been realized in the crystal pick-up and the crystal microphone, and the second possibility in supersonic radiators. The powerful supersonic waves emitted by these radiators find practical application for signalling under water, for preparing very finely distributed emulsions and suspensions, and also for tracing defects in large, massive pieces of material. In an entirely different way piezo-electric crystals are used as elements for stabilizing the frequency of radio transmitters, for which purpose during the last war quartz oscillator plates were turned out in tens of millions in the U.S.A. and elsewhere. Finally, there is to be mentioned the use of piezo-crystals as filter elements (transducers) for carrier-telephony and other branches of communication engineering.

The reasons why piezo-crystals are suitable for these purposes will be mentioned below.

Piezo-electric species of crystals

With one single exception piezo-electric effect may be expected in all species of crystals not

[1] Or, in the case of the inverse effect, a deformation of 8×10^{-9} metre per volt applied. From the theory it follows that the proportionality coefficients for the direct and the converse effect are identical; it is easily verified that, as regards the dimensions, there is nothing contradictory in this (C/N is the same dimension as m/V). Thus the substance having the greatest direct effect has also the greatest converse effect.

[2] See, e.g., Philips Techn. Rev. 5, 145, 1940, and 5, 9, 1940 (laryngophone).

exhibiting a centre of symmetry in their crystal lattice. The fact that this effect cannot occur in crystals having a centre of symmetry is easiest understood when examining an elementary cell in the lattice of such a crystal (see *fig. 1*). What the piezo-electric effect amounts to is the accompanying of a deformation of a crystal (thus also of the elementary cell) by a mutual displacement of the centres of gravity of the negative and positive charges carried by the ions in the cell. Let us draw from the centre of the elementary cell a radius vector to one of the ions. If the crystal lattice has a centre of symmetry the opposite and equally long radius vector will end in an ion of the same kind; the centres of gravity referred to therefore both lie in the centre of the cell. For the very reason of the symmetry existing also in the mechanical strength of the crystal in various directions, in the case of a deformation due to external forces the property of the centre of symmetry is maintained; thus the centres of gravity of the positive and the negative charges continue to coincide and there is no piezo-electric effect.

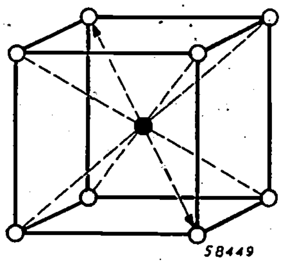


Fig. 1. Simple example of a crystal species with centre of symmetry; elementary cell of the crystal lattice of caesium chloride, CsCl.

The complete lattice is formed by displacing the cell in the direction of the ribs of the cube one rib-length at a time. The white ion may be caesium and the black one chlorine, or vice versa. The lattice is of course electrically neutral, although the elementary cell apparently contains eight white ions and only one black one: each of the white ions belongs simultaneously to eight elementary cells bordering on each other in one corner and may, therefore, be counted as belonging for only one-eight part to the cell drawn. Similar considerations apply for the types of crystals represented in the other illustrations to follow.

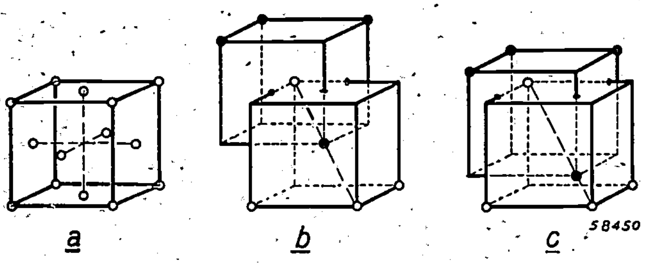


Fig. 2. *a*) Face-centered cubic lattice of one kind of ion.

b) The crystal lattice of the rock-salt type can be constructed by interlacing two lattices according to (*a*), one of "white" ions (e.g. sodium), the other of "black" ions (e.g. chlorine), in such a way that one is displaced, with respect to the other, over half the length of the space diagonal of the elementary cell.

c) The crystal lattice of the sphalerite type is formed similarly to (*b*), but with the "white" lattice (e.g. copper) and the "black" one (e.g. chlorine) displaced with respect to each other over only a quarter of the length of the said space diagonal.

In (*b*) and (*c*) only a few ions of each lattice are shown, for the sake of clarity.

After this somewhat abstract reasoning the reader may feel the need of an example. For this we shall choose alkali-halogenides, which do not exhibit any piezo-electric effect, and cupro-halogenides, which do show that effect. The crystallization of these substances is of the rock-salt type and the sphalerite type respectively. Both crystals contain only two kinds of ions and can be built up from two face-centered cubic lattices of the two kinds of ions interlaced in a certain way. An attempt has been made to illustrate this in *fig. 2*. From the elementary cells drawn in *figs 3a* and *4a* it can be verified that the rock-salt type possesses a centre of symmetry whereas the sphalerite type does not.

Can it be understood how the piezo-electric effect is brought about in the sphalerite type? In the normal state (fig. 4*a*) also with this type the centres of gravity of the "white" and the "black" ions both lie in the centre of the elementary cell. Upon the elementary cell being deformed by a pressure acting in the direction indicated (*P*) all the ions move towards each other in the direction of the arrow and away from each other in the directions perpendicular thereto (fig. 4*b*). Taking into account the fact that the binding forces of all neighbouring ions keep each ion in its position more or less resiliently, it will be clear that as a result of the black ions *A* and *B* approaching each other and the black ions *C* and *D* moving away from each other the white ion *E* will fall deeper between the ions *C* and *D*. This applies similarly for the other three white ions *F*, *G* and *H* in the elementary cell. In the deformed elementary cell (fig. 4*b*) the centre of gravity of the white ions will therefore have dropped with regard to that of the black ions.

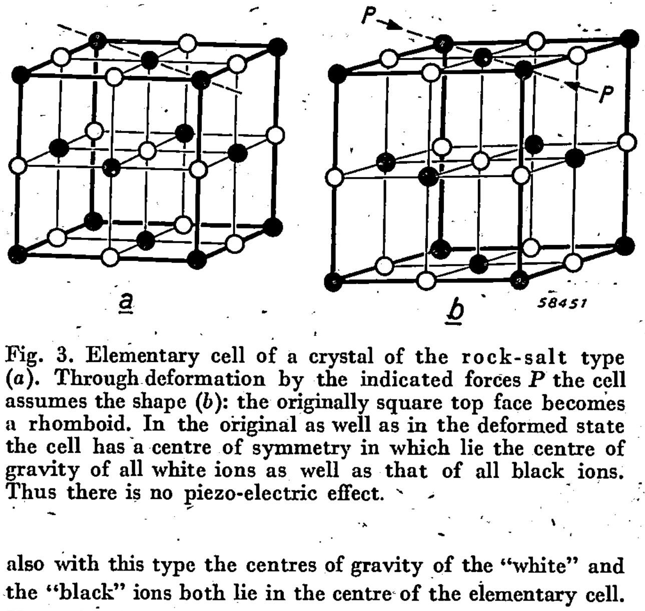


Fig. 3. Elementary cell of a crystal of the rock-salt type (*a*). Through deformation by the indicated forces *P* the cell assumes the shape (*b*): the originally square top face becomes a rhomboid. In the original as well as in the deformed state the cell has a centre of symmetry in which lie the centre of gravity of all white ions as well as that of all black ions. Thus there is no piezo-electric effect.

Here one can see at once how the piezo-electric effect arises by virtue of the absence of a centre of symmetry: if there were also white ions at the points *I*, *K*, *L*, *M* then the same deformation would cause these to be forced upward and the centre of gravity of all white ions would not be displaced.

Such a configuration does actually occur: it is the so-called fluorite type in which, for instance, calcium fluoride crystallizes; see *fig. 5*. As is to be expected, this kind of crystal, which again has a centre of symmetry, is not piezo-electric.

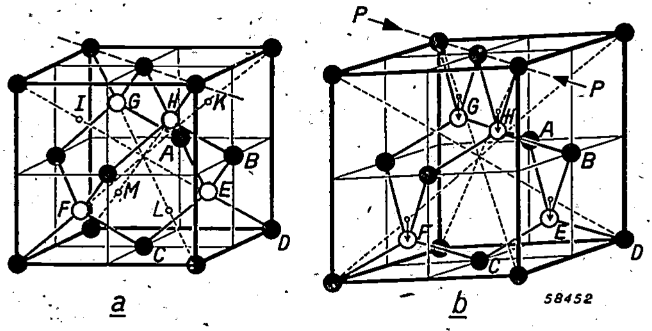


Fig. 4. Elementary cell of a crystal of the sphalerite type (*a*). There is no centre of symmetry. In the case of a deformation similar to that in fig. 3 the cell assumes the shape of (*b*). The white ion *E* has then dropped with respect to the four surrounding black ions *A*, *B*, *C*, *D*. The same applies for the other three white ions *F*, *G*, *H*, so that the centre of gravity of the white ions in the cell now lies lower than that of the black ions (piezo-electric effect).

About 1000 of the 10,000 species of crystals so far identified crystallographically belong to the classes without centre of symmetry and thus exhibit a piezo-electric effect. Unfortunately, there are only a few of these that can be considered for practical use, since in the first place one must have rather large crystals. Of the natural species of crystals only quartz, tourmaline and zinc blende can be considered. Tourmaline crystals of a sufficiently homogeneous structure very seldom occur. Zinc blende is difficult to work with because it readily splits. Thus we are left with quartz as the only useful —,but very useful — natural piezo-crystal. Of the artificial crystals having piezo effect the Seignette salt already mentioned has been known the longest. Some other synthetic piezo-crystals will be mentioned below.

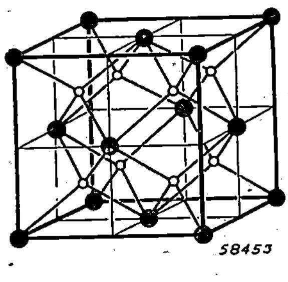


Fig. 5. When the places *I*, *K*, *L*, *M* in fig. 4*a* also become occupied by "white" ions a crystal lattice with a centre of symmetry is again formed, the fluorite type (example: black ions calcium, white ions fluorine; CaF_2). When deformation takes place four white ions drop and four white ions rise, the centre of gravity remaining where it was — hence no piezo-electric effect.

Properties of Seignette salt and quartz as piezo-electric materials

It has been stated above as a practical requirement that one must have piezo-crystals of a sufficiently large size. Of course there are also other criteria of the usefulness of a piezo-crystal. The foremost of these are: the strength of the piezo effect; the mechanical strength, determining what deformations can be allowed; the susceptibility of these properties for such external conditions as temperature, humidity, etc.

The piezo-electric effect of Seignette salt is exceptionally great, but the crystals of this salt can only be used between —30 and 50 °C, they are soluble in water and they have little mechanical strength. Owing to this last-mentioned characteristic the use of Seignette salt is confined mainly to microphones and pick-ups; for such apparatus a forced vibration with greatly divergent frequencies is essential. Consequently the crystal cannot be allowed to act in the vicinity of one of its mechanical resonant frequencies and one has always to do with relatively small deformations which the Seignette salt crystal can still well withstand. It is for the very reason of the smallness of the deformations that the extremely great piezo effect of Seignette salt is essential. As regards its susceptibility to moisture and temperature, in the applications mentioned the conditions can be sufficiently stabilized, so that there are no objections on this account to this use of Seignette salt.

Seignette salt is used also as a generator for mechanical, supersonic vibrations, though as a rule quartz is more suitable for this, especially where it is a matter of large powers. The piezo-electric effect of quartz is, it is true, by no means so strong as that of Seignette salt, but with quartz the permissible deformations are much greater and in this case one can easily get large deformations since in principle only one frequency is to be generated at a time and it can be arranged for this to be a natural frequency of the crystal plate. Waves can thus be produced having a power of about 10 W per cm^2 of the radiating surface, which is much more than can be reached with electrodynamic loudspeaker systems or by means of magneto-striction, which is also sometimes applied.

Quartz likewise possesses excellent properties for other applications of the piezo-electric effect, in particular for the above-mentioned frequency-stabilization of radio transmitters. This we shall explain with reference to the electrical equivalent circuit of a vibrating piezo-electric crystal given in *fig. 6*. The self-inductance L represents the vibrating mass, the capacitance C_1 the mechanical rigidity and C_2 the static capacitance determining the voltages arising in the crystal under the varying charge due to the periodical deformation. The resistance r represents the various causes of damping: internal friction of the crystal, damping due to the supports and the surrounding electrodes, and damping due to possible radiation of acoustic waves. The quantities depend upon the species of crystal, the shape and crystallographic orientation of the crystal plate, and the conditions under which it is used. For quartz one can reckon in practice on values of say $L = 10$ henries, $C_1 = 0.1$ μμF, $C_2 = 20$ μμF, $r = 200$ ohms.

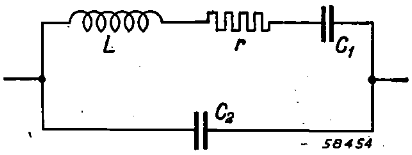


Fig. 6. Electrical equivalent circuit of a piezo crystal vibrating under the influence of an alternating voltage applied to it.

When the crystal is used in the oscillator circuit of a transmitter this circuit can as a rule only oscillate in the frequency range in which the reactance of the crystal is positive. We shall not here write down the formulae for the impedance $R + jX$ of the equivalent circuit (fig. 6) but merely give a diagram of the reactance X as a function of the frequency; see *fig. 7*. This diagram shows a series resonance ($X = 0$) at a frequency f_s and a parallel resonance ($X = \infty$ or at least very large) at a frequency f_p. Between these frequencies $X > 0$ and the circuit is able to oscillate. If C_2 is much greater than C_1, as is the case with quartz, then f_p and f_s lie very close together (in the case of quartz the difference between f_p and f_s is only a few tenths of a percent of f_s) and the frequency at which the circuit oscillates depends almost entirely on the values of L and C_1.

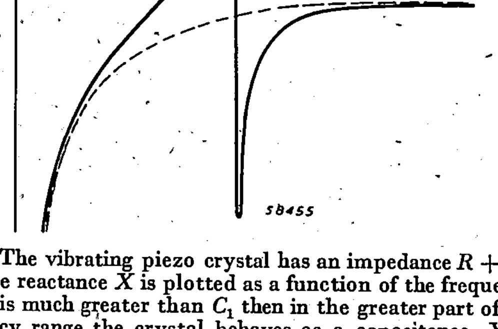


Fig. 7. The vibrating piezo crystal has an impedance $R + jX$. Here the reactance X is plotted as a function of the frequency f. If C_2 is much greater than C_1 then in the greater part of the frequency range the crystal behaves as a capacitance, $X = -1/\omega C_2$ (see the dotted hyperbola). At the frequency f_s a series resonance arises, the impedance becomes real and has the value $R \approx r$ (see fig. 6). At f_p there is a parallel resonance, X becoming very large (in the case $r = 0$, infinite).

When the above-mentioned values of L and C_1 are compared with those of a circuit made from coils and capacitors it is noticed that the desired tuning with a quartz crystal is reached at a high value of L and a small value of C_1. Since r is also fairly small, this means that the quality factor $Q = (1/r)\ \sqrt{L/C_1}$ of the crystal is much higher than that of an LC-circuit. At frequencies of 10 Mc/s normal LC-circuits, which are still sometimes used, have at their best a quality factor $Q \approx 300$, whereas with quartz crystals one can reach a value of about 30,000, and even 200,000 in vacuo (there then being no damping from acoustic radiation). Thanks to this high quality factor the oscillation frequency is very well stabilized.

It is further of great practical importance that the values of L and C, or rather the resonant frequencies, of quartz are highly stable. These can in particular be made highly independent of temperature by cutting the oscillator plate out of a crystal in certain crystallographic directions. In the orientations most used, the so-called AT and BT cuts, the relative change in frequency of the oscillator plate over a range of 100 degrees Centigrade is no more than some hundredths percent, and with the so-called GT cut even no more than 0.0001%. By way of illustration it may be mentioned that with such a crystal placed in a thermostat and forming part of an oscillator circuit having little effect upon the crystal frequency, a clock can be built which measures the time with an error of no more than 1 in 10^9. This is less than the error arising in astronomical time recording due to the irregularities in the rotation of the earth.

This great degree of constancy is also due in part to the fact that quartz is very hard and highly insusceptible to chemical action. In fact, a tolerance of 10^{-5} in the natural frequency of an oscillator plate would already be exceeded if a minute layer some millionths of a millimetre thick were to be chipped off the surface of the plate. Thanks again to its great hardness a quartz plate can be processed with extreme accuracy, so that not only can the highly constant frequency be obtained but it can also be predetermined with a tolerance of say 10^{-5}.

After having summed up all the good properties of quartz, we must point out the restrictions to its use. We have already mentioned the case of microphones and other apparatus not operating with resonant vibrations, for which the piezo effect of quartz is not great enough. To this we may add the application of the piezo effect for transducers in carrier-telephony. Such band-pass filters must have a transmitting range of the width of a speech channel (bandwith about 3400 c/s). This width can be reached with a smaller carrier frequency and/or smaller number of filter elements (crystals) according as the percentage difference between f_p and f_s (see above) of the crystal species is greater. As we have seen, in the case of quartz this difference is small, so that, particularly with low carrier frequencies of, say, 20 kc/s, quartz cannot be used directly.

New piezo-electric materials

The cases mentioned in which quartz is not suitable have led to the development of new piezo-electric materials, a few of which will be referred to below. But also for those cases where quartz would be excellently suitable one cannot, unfortunately, say that the problem of the piezo-electric material has been solved; nature's store of quartz crystals on this earth is not inexhaustive. It is already to be foreseen that the production from the known sources of quartz (mainly Brazil and less so Russia, Japan, Madagascar, and Spitzbergen) will in course of time prove to be too small to meet the demand.

To defer the moment when such an impasse will be reached, means are being sought for making at least the most economical use of the limited stock available, for in this direction there is still much to be desired. At the moment the position is such that the oscillator plates produced represent only 10% by weight of the rough crystals from which they are made, the other 90% being lost mainly through two causes: 1) about 40% is pulverized in the sawing of the crystal into plates, which loss cannot very well be avoided; 2) about 50% of the material does not reach even the sawing stage, because more or less large pieces of practically every natural crystal show all kinds of faults, such as impurities, air inclusions, and particularly the phenomenon of twinning. Where twinning occurs the crystallographic axes in different parts of the crystal are differently orientated, and in an oscillator plate this would result in different parts of the plate reacting to a voltage with different deformations, so that nothing would come of the desired oscillation. Means have now been found [3]), and already applied with some success, to "detwin" a natural crystal, so that a much larger portion of the rough material can be utilized.

Of more importance for the future are the attempts that have been made to produce quartz crystals artificially. It has been found possible to produce mono-crystals of quartz with linear dimensions of 2 or 3 cm in the space of a few weeks' time in autoclaves under a pressure of, for instance, 1000 atmospheres. Similar results have been attained also with lower pressures [4]). It is, however, questionable whether such methods can be developed for large-scale manufacture of a product at a sufficiently low price.

Another course of development is the attempt to find new piezo-electric materials not occurring in nature and which can be synthetized in the form of sufficiently large crystals to be used as substitute for quartz. This brings us to the already-mentioned development of synthetic crystals which are not intended to replace quartz but to supplement it in certain fields of application.

In the search for these new piezo-crystals the rule about the centre of symmetry dealt with in the foregoing pages has been found to be a good guide. Thanks to this rule attention has been particularly centered upon compounds having molecules lacking certain properties of symmetry, since it is to be expected that these compounds will form crystals having no centre of symmetry and thus yield a piezo-electric material. Synthetic piezo-crystals which have been developed and examined by Swiss and American investigators [5]) among others are, for instance, ammonium dihydrogen phosphate (designated by the Bell Telephone Laboratories as ADP), dipotassium tartrate (DKT) and ethylene-diamine tartrate (EDT.) These crystals can be given the desired shapes fairly easily, they have a small internal damping and are not too readily dissociated (they have little or no crystal water). DKT and EDT possess piezo-electric constants 6 to 8 times as great as those of quartz. The ratio of the capacitances C_2/C_1 (see fig. 6) is 4 to 10 times smaller than that of quartz, so that they have wider ranges of resonance. Thanks to

[3]) W. A. Wooster and N. Wooster, Nature 157, 405, 1946.

[4]) I. Franke and M. Huot de Longchamp, C. R. Acad. Sci. Paris **228**, 1136, 1949, No. 13.

[5]) See, i.a., W. Lüdy, Helv. Phys. Acta **12**, 279, 1939; W. P. Mason, New low-coefficient synthetic piezoelectric crystals for use in filters and oscillators, Proc. I.R.E. **35**, 1005-1012, 1947.

these properties the materials mentioned, in particular EDT, have already been applied on a rather large-scale for transducers in carrier-telephony systems. Investigations are being continued, however, because the synthetic crystals so far produced cannot be regarded as equivalent substitutes for quartz. Even when various "temperature-independent" cuts are used their temperature coefficients are still 20 to 30 times, in the most favourable case still 5 times, as high as those of quartz; and for what may be the most important application of the piezo effect, viz. the stabilization of oscillators, it is just a high C_2/C_1 ratio that has to be aimed at, such as obtained with quartz.

Finally, mention should be made of the development of a piezo-electric material of an entirely different type, namely barium titanate. The peculiarity about this material is the fact that it is used in the polycrystalline form. With a polycrystalline material nothing is noticed of the normal piezo-electric effect because the differently orientated crystals neutralize each other's action. There is, therefore, in this case no question of a normal but rather of a so-called induced piezo-electricity. This effect is to be derived from the electro-striction mentioned in the beginning of this article and, consequently, is in principle present in every substance. The phenomenon of electro-striction implies that a deformation x takes place which is proportional to the square of an applied electric field:

$$x = a \cdot E^2$$

(E = electric field strength, a = material constant). When a field E_0 is applied a small change ΔE causes an extra deformation Δx proportional to ΔE, since

$$x + \Delta x = a(E_0 + \Delta E)^2 \approx aE_0^2 + 2aE_0 \cdot \Delta E,$$
$$\Delta x = 2aE_0 \cdot \Delta E.$$

Thus this is equivalent to a normal piezo-electric effect (actually the converse effect, but the direct effect then also occurs). With most substances this effect is extremely small. In the case of barium titanate, however, the constant a is enormously large, millions of times greater than that of other substances, whilst moreover this material is an "electret", that is to say, after removal of the external electric field the material retains a polarisation which can take the place of the external field E_0 [6]). Thanks to these properties practical use can be made of the phenomenon, and in fact crystal pick-ups made with barium titanate are already being manufactured on a rather large scale [7]). The advantage of this material is that it can be processed by ceramic means and thus can be given all sorts of shapes, which is not possible with mono-crystals. Nothing can be predicted at the moment as to how far the possibilities of application of this and similar materials will be extended in the course of time.

[6]) These properties may be compared to those of ferro-magnetic substances. They are also related to another peculiar property of barium titanate, viz. the exceptionally high value of its dielectric constant in a certain temperature range ($\varepsilon \geqq 10{,}000$).

[7]) L. Grant Hector and H. W. Karen, Electronics 21, December 1948, pp. 94-96.

Bibliography:

W. G. Cady, Piezo-electricity. An introduction to the theory and applications of electromechanical phenomena in crystals, Mc Graw Hill, New York 1946 (with extensive bibliography).

R. A. Heising, Quartz crystals for electrical circuits. Their design and manufacture, Van Nostrand, New York 1946.

Summary. This article gives a brief review of the most important facts in the domain of piezo-electricity. In particular the relation is explained between the occurrence of this phenomenon and the symmetry of the crystal. The favourable properties of quartz for various applications of the piezo-electric effect are put forward. Finally mention is made of the search for new piezo-electric materials intended to be used partly as substitutes for quartz and partly for applications for which quartz is unsuitable. Polycrystalline barium titanate is also briefly referred to in this connection.

AN ARRANGEMENT FOR INDICATING PIEZO-ELECTRICITY OF CRYSTALS

by W. G. PERDOK*) and H. van SUCHTELEN.

621.317.36:537.228.1

Bonne wit

The strong piezo-electric effect of a quartz crystal was discovered by the Curie *brothers with comparatively simple means. To indicate the very weak effect of some species of crystals, of which only minute fragments may be available, more sensitive methods are needed. Applied electronics provides such methods and at the same time makes it easy to carry out extensive and systematic investigations.*

There are two sides to the investigation into the piezo-electric behaviour of a substance. In the first place it is needed in the search for new, natural or synthetic materials for the practical application of the piezo-electric effect, the importance of which has been pointed out in the preceding article in this number [1]). In the second place information as to whether a substance is piezo-electric or not provides a guide when investigating the structure of a crystal, since a crystal can only be piezo-electric if it has no centre of symmetry, as likewise explained in the preceding article. Of the 32 classes to which a crystal can belong, the eleven classes with centre of symmetry (and also one of the 21 other classes) are at once eliminated when the crystal is found to be piezo-electric. Moreover, from details of the piezo-electric behaviour it is possible to draw farther-reaching conclusions regarding the crystal structure.

For both these investigations it is useful to have an instrument with which many different substances can be quickly tested for the presence of piezo-electric effect as a first orientation.

It was for this purpose that Giebe and Scheibe designed an apparatus in 1925, which has served as a basis for various modern instruments [2]). The device forming the subject of this article is also based upon the fundamental idea of the Giebe and Scheibe apparatus. The improvement compared with older designs lies in greater sensitivity, clearer indication and easier working [3]).

Principle of the method

Suppose a thin plate sawn from the rough material to be placed between two parallel flat electrodes and electrically connected parallel to a tuning circuit of an oscillator; see fig. 1. If the crystallographic axes in the plate are suitably orientated with respect to the plane of the electrodes then, in the case of a piezo-electric material, a coupling arises between the mechanical vibrations and the electrical oscillations, which can be described with the aid of the equivalent circuit for the crystal (fig. 2). In the larger part of the frequency range the crystal behaves as a capacitor with the capacitance C_2. In a certain, very narrow frequency range, however, a resonance phenomenon occurs, viz. a series resonance, whereby the impedances of L and C_1 neutralize each other, and (at a slightly higher frequency) also a parallel resonance where the impedance of L is equal and opposed to that of

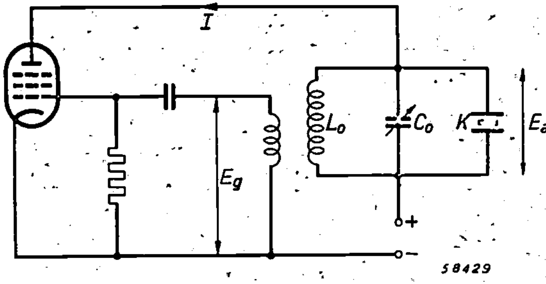


Fig. 1. Oscillator circuit with tuning circuit L_0-C_0 shunted by a crystal between two flat electrodes (K).

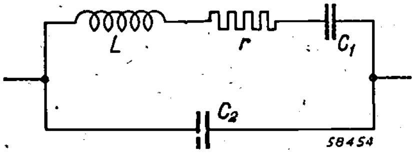


Fig. 2. Equivalent circuit of a vibrating piezo-electric crystal between two electrodes. The static capacitance C_2 is generally much greater than the capacitance C_1 representing the mechanical rigidity.

*) Of the Crystallographic Institute of the Groningen University (Netherlands).

[1]) J. C. B. Missel, Piezo-electric materials, Philips Techn. Rev. 11, 14 -150, 1949 (No. 5).

[2]) E. Giebe and A. Scheibe, Z. Phys. 33, 760, 1925. Further, i.a. S. B. Elings and P. Terpstra, Z. Krist. 67, 279, 1928, E. Burnstein, Rev. sci. Instr. 18, 317, 1947, R. F. Stokes, Amer. Mineralogist 32, 670, 1947.

[3]) The arrangement has been previously described in: W. G. Perdok and H. van Suchtelen, Appl. sci. Res. B1 195-204, 1948, (No. 3).

C_1 and C_2 (see the reactance diagram in fig. 7 of the preceding article). In the case of the series resonance it is as if the admittance of the small resistor r suddenly came to lie in its entirety parallel to the oscillator circuit. Thus this circuit becomes all at once strongly damped.

This behaviour can be taken as an indication that a crystal is piezo-electric. To determine whether this is so, the oscillator circuit is tuned by means of a variable capacitor to various frequencies in succession. When the frequency of the series resonance is reached the amplitude of oscillation suddenly drops, returning to its initial value immediately when the frequency is further changed. This drop in the amplitude is detected and made audible in a loudspeaker.

The resonance phenomenon occurs at different frequencies according to the dimensions of the plate and the orientation of the crystallographic axes with respect to the electrodes. Only with some rather sharply-defined orientations will the effect be quite clear. Now it is not practicable to have to saw every time from the crystal a plate with a different orientation and then scan the frequency range for each plate. Moreover, in many cases there would not even be sufficient material to do this, only some splinters or grains of the crystal being available. This, however, is no objection in principle. It is the essence of the Giebe and Scheibe method that a quantity of splinters of the sample to be tested can be placed between the parallel electrodes and that among these fragments of crystal there will always be some having a suitable orientation to exhibit the effect (hence the effect will generally be noticed at more than one frequency while passing through the range).

The space between the electrodes is, it is true, filled only for a small part by the piezo-electric crystal. The impedance of the crystal is measured, as it were, with the parallel connection and series connection of two constant capacitances (the non-reacting crystals and the air between the crystal and the electrodes). The dips in the impedance measured are therefore only very small, so that in order to indicate the sometimes very weak piezo-electric effects of all sorts of crystals the device has to answer very high demands as regards sensitivity.

Means of increasing the sensitivity

Small dips in the impedance between the electrodes will have a greater effect upon the oscillation according as the impedance of the tuning circuit itself is greater. This impedance is determined for a large part by the internal resistance of the oscillator valve connected to it in parallel. It is therefore advantageous to choose a valve having a high internal resistance, such as a pentode.

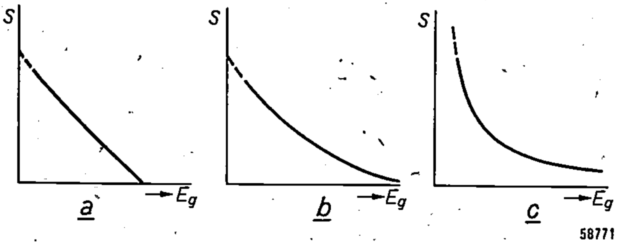


Fig. 3. The relation between the effective mutual conductance S and the alternating grid voltage E_g for different types of valves: *a*) linear, *b*) exponential, *c*) hyperbolic.

Another important feature of the valve to be used is the relation between the effective mutual conductance and the alternating grid voltage; see *fig. 3*. This can be explained as follows:

The effective mutual conductance S is defined by

$$I = S \cdot E_g \, , \quad \ldots \ldots \ldots \quad (1)$$

where E_g is the amplitude of the (sinusoidal) grid voltage and I the amplitude of the first harmonic of the anode current. Since the tuning circuit in the anode circuit always oscillates at its natural frequency, the anode impedance can simply be regarded as a resistance R. The anode current I induces at this resistor an alternating voltage

$$E_a = I \cdot R = S \cdot E_g \cdot R \, ,$$

a fraction of which, $t \cdot E_a$, is fed back to the grid. The circuit will reach a stabilized oscillation when $t \cdot E_a = E_g$, thus when

$$tSR = 1. \quad \ldots \ldots \ldots \quad (2)$$

At given values of t and R the amplitude of the oscillation, or, to express it otherwise, the "working point" on the "S-E_g characteristic" (fig. 3), adjusts itself so that the mutual conductance answers the equation (2) [4]). A variation $\mathrm{d}R/R$ in the anode impedance, brought about when a crystal resonance is passed in the changing of the circuit frequency, will result in an equal change $\mathrm{d}S/S$ in the mutual conductance and thus a displacement of the working point. Now we use as indication for the state of

[4]) In the usual oscillator circuit according to fig. 1, when there is an alternating grid voltage of the amplitude E_g then, owing to grid rectification, a grid bias is automatically obtained which has practically the value E_g. It is due to this that the characteristics in fig. 3 always show a drop in S with increasing E_g. Only with such a variation is the condition according to equation (2) a stable state of equilibrium, the system always tending to revert to this state in the event of any deviations.

oscillation of the circuit the grid voltage E_g (or the proportional alternating anode voltage $E_a = E_g/t$). It will then depend upon the shape of the characteristic in fig. 3 at what working point the greatest variation dE_g takes place for a given relative variation dS/S, or in other words at what strength of oscillation the most sensitive indication of the piezo-electric effect is obtained.

As oscillator valve we have chosen a variable-mu pentode (EF9), the relation S-E_g of which is approximately exponential. A simple calculation shows that with an exponential relation the variation dE_g referred to is independent of the working point. This is very useful in practice because the amplitude of oscillation will naturally vary gradually when a large frequency range is scanned; it would be an awkward complication if, in order to maintain the same sensitivity, the amplitude of the oscillation should have to be repeatedly readjusted, for instance by changing the feedback factor t.

The dips in the alternating anode voltage E_a taking place when passing the resonance frequencies of the crystals are made audible in a loudspeaker by means of a detector circuit, connected to the anode, and an A.F. amplifier.

Under the conditions described here it is advantageous to apply anode detection by means of an amplifying valve instead of the usual simple rectification by a diode. The grid of the amplifying valve is so heavily biased that anode current flows only during a part of each cycle of the alternating grid voltage (*fig. 4*). The amplitude of the anode-current pulses and thus also the mean anode current then depends upon the amplitude of the grid voltage. Thus a dip in the oscillator amplitude causes a corresponding dip in the mean voltage at a coupling resistor in the anode circuit of the detector valve. Thanks to the amplification by the detector valve, this voltage pulse is sufficient to be further amplified in a normal A.F. amplifier, for instance the A.F. part of a radio receiver.

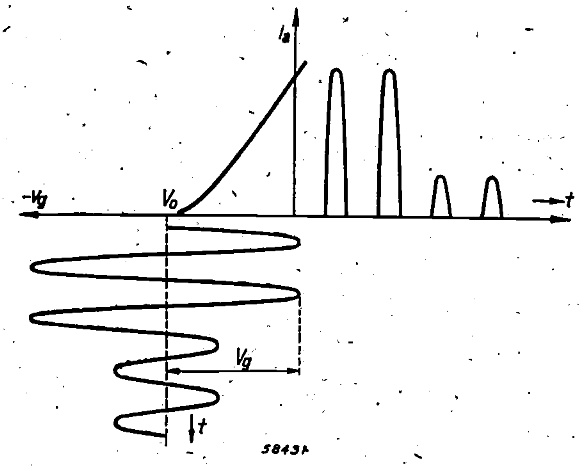


Fig. 4. Anode detection. A dip in the amplitude V_g of the alternating grid voltage is accompanied by a corresponding dip in the amplitude of the anode-current pulses, provided the grid bias V_0 remains constant.

The negative bias for this anode detector, which has to be matched to the oscillator amplitude, is most easily derived from that amplitude itself via a grid capacitor and leak. But of course the bias must not change during the dip in the oscillator amplitude that is to be indicated, for then the effect of the anode detection would for the greater part be lost. Therefore a grid capacitor and leak of a sufficiently high value to ensure a large time constant are used, so that only gradual changes of the oscillator amplitude will affect the grid bias.

In other developments of the Giebe and Scheibe principle there is no separate detector and anode detection takes place in the oscillator itself. In fact, the form of the grid voltage and the anode current of the oscillator valve are in principle identical to those of our detector valve as shown in fig. 4. When passing a crystal frequency one can therefore get directly a D.C. voltage surge at a series resistor in the anode circuit of the oscillator valve. But then we get the effect mentioned above where the automatic adjustment of the grid bias keeps the anode current impulses practically constant; the drop in amplitude of the grid voltage is then mainly manifested in a slight widening of the anode current impulses. The greater sensitivity obtained with a separate detector is thus due in part to the very fact that the grid capacitor and leak can be given a high value, which is not possible with the oscillator valve without adversely affecting the functioning of the oscillator (cf. footnote [4])).

Wobbling tuning

When investigating different kinds of crystals it cannot of course be predicted in what frequency ranges resonances will be found. The oscillator is therefore fitted with exchangeable or variable coils each covering a certain frequency range to be scanned by means of a variable capacitor. As the capacitor is slowly turned a click will be heard in the loudspeaker as soon as a resonance frequency is reached. With a view to making such a point of resonance more noticeable, in earlier designs of apparatus it was recommended to wobble the knob of the capacitor to and fro while gradually turning it farther. Instead of one single click a number of clicks are then heard, but one has to make sure that these sounds are not caused by a "creaking" of the capacitor shaft.

This process has now been very much simplified by an automatic and continuous wobbling of the tuning brought about with the aid of a small vibrating capacitor. This capacitor, consisting of a fixed plate and a movable one, mounted on a leaf spring kept in vibration at a frequency of 100 c/s electromagnetically, is shunted across the tuning capacitor and causes the total tuning capacitance to vary periodically by a few pF. One can now steadily turn the tuning capacitor in one direction, and when a resonance is reached a rattling noise with the fundamental frequency of 100 c/s is heard in the loudspeaker.

With the gradual change of the tuning capacitance also the impedance of the circuit gradually changes, but this is not noticed in the loudspeaker. However, the rapid periodical variation caused by the vibrating capacitor, although still bearing a much more continuous character than the dip when passing a resonance, can indeed be heard as a hum in the loudspeaker, particularly when the tuning capacitor is turned almost to its minimum. The effect sought might be more or less masked by this hum. To avoid this, the signal obtained after detection is passed through a relatively narrow band-pass filter tuned to the (somewhat arbitrarily chosen) frequency of 600 c/s, which strongly attenuates both lower and higher frequencies. Behind the filter the variation of the oscillation amplitude with 100 c/s, which variation is roughly sinusoidal, is practically unnoticeable, whilst also noise and other interferences are effectively suppressed, whereas the essential part of the rattling signal (caused by a number of higher harmonics of 100 c/s, e.g. 500, 600, 700 and 800 c/s) is passed through.

Practical construction

The whole of the circuit, based on the principles described above, is represented in *fig. 5*. Here the crystal holder and the wobbling capacitor with its drive are seen connected parallel to the tuning circuit. The oscillator valve is inductively fed back to the tuning coil. Actually there are five interchangeable tuning coils with their feed-back coils, covering a wavelength range from about 10 m to 1000 m.

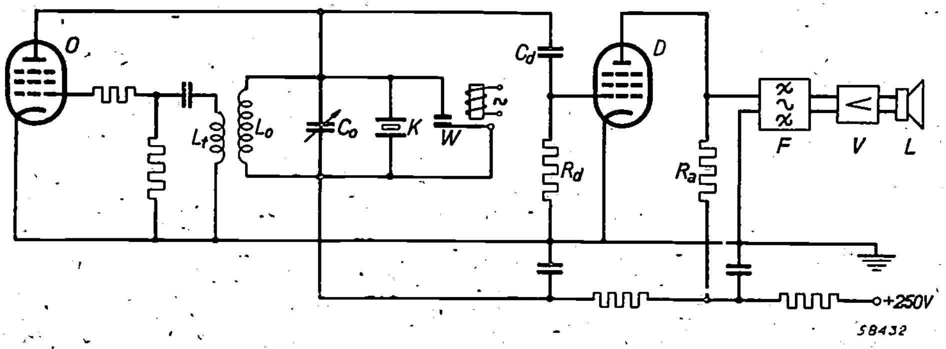


Fig. 5. Circuit diagram of the whole apparatus (somewhat simplified). O oscillator valve, L_0-C_0 tuning circuit, K crystal holder, W wobbling capacitor, D detector valve for anode detection, with coupling capacitor, C_d and grid-leak R_d. The signal to be made audible is taken from R_a. F band-pass filter for 600 c/s, V audio-frequency amplifier, L loudspeaker.

The coupling capacitor C_d between the oscillator valve and the detector valve has the fairly high value of 50,000 pF. Thus with a grid-leak resistor of 2 MΩ the RC-time of the coupling is 0.1 sec., which is long enough to avoid any noticeable change in the grid bias of the detector when a resonance is reached.

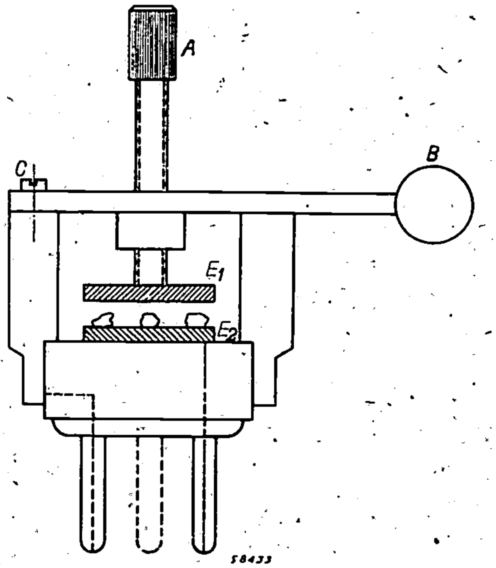


Fig. 6. Crystal holder with electrodes E_1 and E_2 mounted on a normal valve base (in this case with only two connecting pins in use). By turning the screw A the upper electrode E_1 is lowered until it is immediately above the crystals (it must not touch them). In this way the sensitivity is increased to the utmost. The holder is opened by turning aside the cover together with the upper electrode, the cover being rotatable about the vertical axis C by means of the knob B.

Formcr designs of apparatus working on the Giebe and Scheibe principle have all been built for battery supply, at least as regards the oscillator feed. This was necessary because otherwise the very weak clicking in the loudspeaker would have been drowned in the hum, in itself also weak, caused by the A.C. mains frequency of 50 c/s. With our set-up, owing to the measures taken for increasing the sensitivity, the indication in the loudspeaker is so clear that the hum can be ignored. The apparatus is therefore fed entirely from the mains, which of course makes it very much easier to use.

Finally, *fig. 6* gives an idea of the crystal holder, which is mounted on a valve base. It is very convenient to be able to remove this holder for cleaning and for refilling with samples.

This construction has proved to be sufficiently sensitive to be able to detect quite clearly the weakest piezo-electric effects, such as those of sodium santoninate and melinophane, which hitherto could only be detected with difficulty.

Summary. According to the Giebe and Scheibe method the piezo-electric effect of a crystal is indicated by placing a number of fragments of crystal between two flat electrodes and connecting the latter parallel to the tuning circuit of an oscillator. The dips in the oscillation amplitudc when the natural frequencies of the crystals are reached are made audible in a loudspeaker.

By employing a variable-mu pentode as oscillator valve it has been possible to ensure reasonably large dips in the amplitude, whilst the sensitivity of this indication is also practically independent of the oscillation amplitude. Sensitivity is further increased by detecting the dips in the anode voltage of the oscillator with a separate amplifying valve. The tuning frequency of the oscillator is made to wobble over a small frequency range by means of a small vibrating capacitor. When a crystal frequency is reached, instead of a single click a rattling tone is heard in the loudspeaker, which is much less liable to escape notice.

A MODEL FOR STUDYING ELECTROMAGNETIC WAVES IN RECTANGULAR WAVE GUIDES

by K. S. KNOL and G. DIEMER.

538.566.5:621.392.26:621.317.35

Wave guides are metal tubes used for transmitting electromagnetic energy of very short wavelengths over short distances, for instance from one part of a transmitter to another. In the case of guides having a rectangular cross section, the wave phenomena occurring can be studied with the aid of a mechanical model in the form of a vibrating membrane. Such a model is not only of value as an ingenious means of demonstrating all sorts of properties of wave guides but it is also of immediate practical importance for the designing of wave guides, since it enables one to determine easily the effect of varying certain parameters.

The membrane as a means of studying electrostatic fields

A well-known means of studying two-dimensional electrostatic fields consists of a membrane in the form of a sheet of rubber stretched in a frame with a constant tension on which the boundary conditions of the electrostatic field being studied are imitated [1]). These conditions are imitated by imparting to the sheet of rubber, with the aid of rigid rods, constant deflections proportional to the potential of the electrodes of the electrostatic problem. Provided the deflections are small, the shape then assumed by the sheet (a minimum surface) corresponds to the potential variation sought, as shown in the article quoted in footnote [1]).

In that article it was found that the sum of the forces acting upon a surface element $dxdy$ of the sheet equals

$$s\left(\frac{\partial^2 h}{\partial x^2}+\frac{\partial^2 h}{\partial y^2}\right)\mathrm{d}x\,\mathrm{d}y,$$

in which s is the tension in the sheet, h the deflection from the original (flat) state of equilibrium, and x and y are Cartesian co-ordinates in the plane of that state of equilibrium. In the new state of equilibrium this sum of forces is everywhere zero, so that one arrives at the differential equation of Laplace

$$\frac{\partial^2 h}{\partial x^2}+\frac{\partial^2 h}{\partial y^2}=0. \quad \ldots\ldots \quad (1)$$

It is known that for the electrostatic potential V in free space the same equation holds, in the general case with three co-ordinates:

$$\frac{\partial^2 V}{\partial x^2}+\frac{\partial^2 V}{\partial y^2}+\frac{\partial^2 V}{\partial z^2}=0.$$

When, however, one has to deal with a case where the potential is not dependent upon one of the co-ordinates, say upon z, so that the problem can be formulated with only two Cartesian co-ordinates, this equation becomes

$$\frac{\partial^2 V}{\partial x^2}+\frac{\partial^2 V}{\partial y^2}=0,$$

from which it follows that the potential V is proportional to the deflection of the sheet h which satisfies equation (1).

In the article referred to, the use of this model for studying electron paths in radio valves is explained. It will now be shown that a somewhat modified form of this model lends itself to the study of electromagnetic waves; in this case the membrane is vibrated instead of being brought to a state of equilibrium.

The vibrating membrane as a means of studying electromagnetic waves

Vibrations of the membrane are governed by the same differential equation as electromagnetic waves in a space with conductive walls, again provided these waves can be formulated with only two Cartesian co-ordinates. This will be made clear from what follows:

When the membrane is in vibration, then at any arbitrary moment the sum of the forces acting upon a surface element $dxdy$ is not zero but equal to the product of mass and acceleration of the element. If ϱ is the mass of the sheet per unit area and t the time, then one finds from the equality just mentioned, after dividing by $dxdy$,

$$s\left(\frac{\partial^2 h}{\partial x^2}+\frac{\partial^2 h}{\partial y^2}\right)=\varrho\,\frac{\partial^2 h}{\partial t^2}.$$

Putting $s/\varrho = v^2$ and replacing y by z to agree

[1]) P. H. J. A. Kleynen, The motion of an electron in two-dimensional electrostatic fields, Philips Techn. Rev. 2, 338-345, 1937.

with the article in this journal [2]) on electromagnetic waves, we arrive at

$$\frac{\partial^2 h}{\partial x^2}+\frac{\partial^2 h}{\partial z^2}-\frac{1}{v^2}\frac{\partial^2 h}{\partial t^2}=0. \quad . \quad . \quad (2)$$

This is the well-known wave equation. Solutions of this equation are progressive waves, for example of the form:

$$h = H \sin \omega \left(t \pm \frac{z}{v}\right)$$

($H =$ amplitude, $\omega =$ angular frequency), as can be seen by substitution in equation (2). The term v is the velocity at which the wave is propagated across the membrane.

In the case of a plane electromagnetic wave in an unbounded homogeneous and isotropic medium (with magnetic permeability μ, dielectric constant ε and electric conductivity σ) for the components $H_x, H_y, H_z, E_x, E_y, E_z$ of the magnetic and electric field strength there are six simultaneous differential equations [3]). In two of these there occurs, for instance, the component E_y, viz:

$$\frac{\partial E_y}{\partial z} = \mu \frac{\partial H_x}{\partial t}$$

and

$$\frac{\partial H_x}{\partial z} = \sigma E_y + \varepsilon \frac{\partial E_y}{\partial t}.$$

Eliminating H_x and putting $\sigma = 0$ (non-conducting medium) and $\varepsilon\mu = 1/v^2$, then

$$\frac{\partial^2 E_y}{\partial z^2}-\frac{1}{v^2}\frac{\partial^2 E_y}{\partial t^2}=0. \quad . \quad . \quad (3)$$

This is the equation for the component E_y of a plane wave travelling in the z-direction with a velocity of propagation v, which in vacuum is equal to $c \approx 3{\cdot}10^8$ m/sec. Analogous equations hold for the other components of the electric field strength E and likewise for the three components of the magnetic field strength H.

By a linear superposition of waves travelling in the x- and z-directions, for a two-dimensional case where the quantities do not depend upon y, we arrive at the equation

$$\frac{\partial^2 F}{\partial x^2}+\frac{\partial^2 F}{\partial z^2}-\frac{1}{v^2}\frac{\partial^2 F}{\partial t^2}=0\,, \quad . \quad . \quad (4)$$

in which F may represent any component of E or H. This wave equation is quite analogous to the equation (2) found above for the vibrating membrane. Thus the term F corresponds to the deflection h of the membrane. This means that the vibrating membrane can serve as a model for studying electromagnetic waves in which the field quantities can be expressed in terms of only two Cartesian co-ordinates, since only two such co-ordinates are available on the membrane. This condition is satisfied in the case of transverse electric (TE) waves [4]), a wave form with which one is almost exclusively concerned in rectangular wave guides.

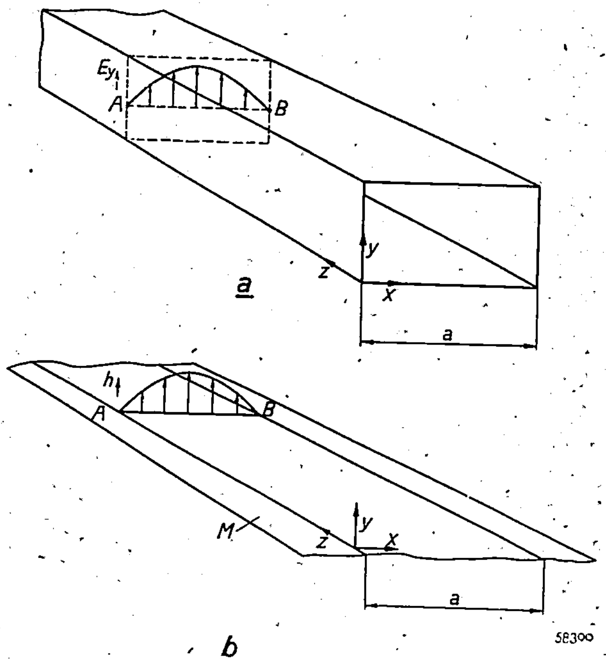


Fig. 1. *a*) Rectangular wave guide in which a transverse electric wave is propagated. The only component E_y of the electric field strength along a line AB (parallel to the x axis) at a given moment, in the case of the simplest form, follows a sine AB (the straight line AB runs parallel to the x-axis), independently of the distance y between AB and the x-z plane. *b*) Corresponding to the field strength E_y in the wave guide is the deflection h of a transverse vibrating membrane M clamped along two parallel straight lines.

Fig. 1a is a diagrammatic representation of a rectangular wave guide, with the co-ordinate axes indicated. For the components of the electric field strength in a TE-wave, according to the equations (29) and (30) of the article quoted in footnote [2]), we have

$$\left.\begin{aligned} &E_x = E_z = 0\,, \;.\;.\;.\;.\;.\;.\;.\;.\;.\;.\;.\;. \\ &E_y = E_{\max}\sin\frac{m\pi x}{a}\cdot\exp j\omega\left\{t-\frac{z}{c}\sqrt{1-\left(\frac{m\lambda}{2a}\right)^2}\right\}, \end{aligned}\right\} \quad (5)$$

[2]) W. Opechowski, Electromagnetic waves in wave guides, Philips Techn. Rev. **10**, 13-25, 1948 (No. 1).

[3]) See e.g. (5a)... (6c) in the article (page 16) quoted in footnote [2]).

[4]) See, e.g., the article (page 22) quoted in footnote [2]).

in which $E_{\max}$ is the maximum amplitude of E_y, a the width of the wave guide, ω the angular frequency, $\lambda = 2\pi c/\omega$ the free-space wavelength and $m = 1, 2, 3,\ldots$ In a given cross section (z = constant) and at a given time t it thus follows that the only component, E_y, depends upon x according to $\sin (m\pi x/a)$. Since E_y is independent of y this is true, with the same proportionality factor, for any height y above the plane x-z. In fig. 1a the simplest wave form is represented, namely that for $m = 1$, whereby the width a of the wave guide is covered by one half-cycle of the sine representing E_y as a function of x. Both with this mode and with that where $m = 2, 3, \ldots$ one finds $E_y = 0$ for $x = 0$ and for $x = a$. This is in agreement with the boundary condition which requires that the component of the electric field in a plane of a vertical wall of the wave guide is everywhere zero.

Corresponding boundary conditions have to be satisfied with the vibrating membrane. That is to say, the amplitude of the membrane must be kept zero along two parallel lines (fig. 1b). This requirement is easily met by clamping the membrane along such lines, as we shall see when we come to the description of the apparatus.

Two cases should be mentioned in which the waves cannot be described with two Cartesian space coordinates and consequently cannot be studied with the aid of this model.

In the first place there are the transverse magnetic (TM) waves in a guide of rectangular cross section. As explained in the article quoted in footnote [2]) (p. 23), in a rectangular wave guide only certain TM-waves can be propagated, which are of a much more complex structure than the TE-waves according to eq. (5). This implies, inter alia, that in the case of such TM-waves the field components depend not only upon x and z but also upon y.

Neither is the model suitable for studying circular wave guides. In a circular guide, where cylindrical co-ordinates (r, φ, z) will obviously be used, wave forms are possible in which the field components are independent of φ. This, it is true, leaves only two co-ordinates (r, z), but these do not correspond to the Cartesian co-ordinates on the membrane.

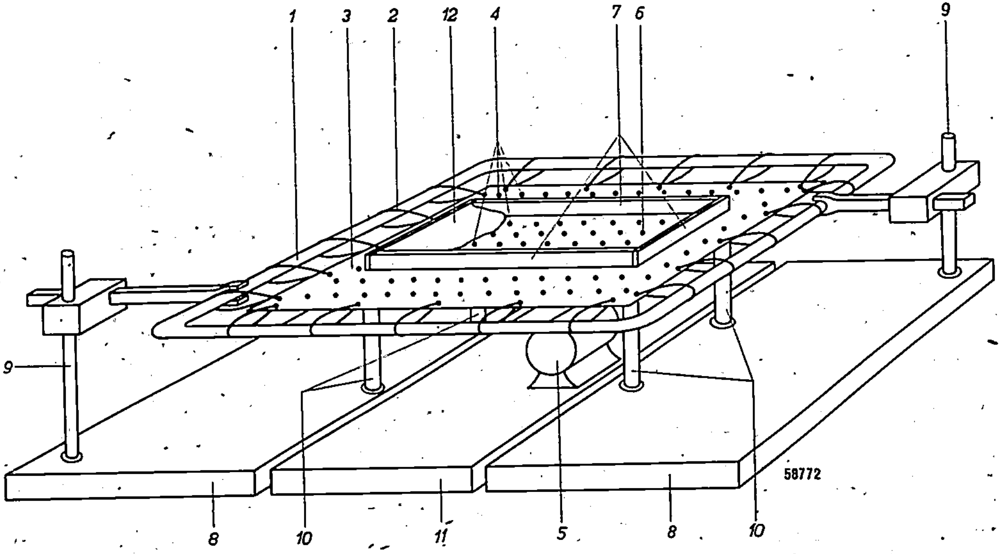


Fig. 2. Sketch of the mechanical model for studying two-dimensional waves. *1* = frame of steel tubes on which a sheet of rubber *3* is stretched by cords *2*, the rubber thus serving as a membrane. *4* = white dots making the movement of the membrane easily visible. *5* = motor driving the membrane at the point *6*. *7* = plane-ground bars (a second set is underneath the membrane) between which the membrane is clamped according to the boundaries of the wave guide being imitated. *8* = tables on which the frame *1* (via the rods *9*) and the bars *7* (via the stands *10*) rest. *11* = heavy iron table with vibration-free mounting, carrying the motor *5*. *12* = wedge-shaped pad of cotton-wool closing the wave guide and making it free of reflections.

Construction of the model with vibrating membrane

Fig. 2 is a sketch of the set-up arranged by us. A sheet of rubber is stretched in a strong frame with the aid of cords. The strain in all directions is about 10 %. Plane-ground bars representing the sides of the wave guide hold the rubber tight top and bottom. The "drive" is underneath the rubber at the place denoted by the figure *6*: by means of the motor *5* and an eccentric, a vertical rod is moved up and down; between the sheet and the rod is a piece of sponge-rubber. (Owing to the eccentric the movement of the rod is not absolutely sinusoidal, but the harmonics are not troublesome.) The motor stands on a separate, heavy, iron table (11) resting upon rubber blocks, thus avoiding parasitic couplings between the motor and the membrane. The motor is a direct-current motor, with its speed and thus the frequency of the membrane controlled by means of the armature voltage.

At the two short sides of the rectangle formed by the bars *7* reflections can take place just as would be the case in a wave guide closed by a conducting plate. Reflections at the side nearest to the driving point *6* are not troublesome if that point is suitably chosen; the result at some distance to the left of *6* (in fig. 2) is then always a wave coming from the right. If a wave guide is to be imitated in which, as is normally the case, the energy is consumed at the other end, then one must avoid reflection at the side of the membrane removed from the driving point. A suitable means of absorption has been found to consist in a pad of cotton-wool laid upon the membrane (*12* in fig. 2) with the thin end directed towards the driving point.

We now come to the practical question as to how the waves on the membrane can best be observed and their amplitude measured.

The oldest method is that of Makinson[5], whereby fine sand is scattered over the horizontal membrane. As in the well-known Chladni sound figures, the grains of sand roll away from the antinodes and ultimately collect near the nodal lines. This method, however, gives only a qualitative picture and is not suitable for investigating progressive waves, while furthermore there is the practical objection that in course of time all the sand moves towards the walls (the only places continuing to be absolutely at rest). For these reasons, therefore, we tried to find better solutions and actually found them in the two following methods:

1) A rectangular co-ordinate system of white dots was painted on the sheet of rubber (see fig. 2). As the sheet vibrates, those dots which are not on nodal lines look like small rods whose length is proportional to the deflection. This gives a much more graphical picture than figures in sand. Moreover, the vibrating membrane can be photographed and by using a sufficiently long exposure and measuring the small white lines produced by the vibrating dots one can arrive at more or less quantitative results.

A valuable aid in observing the waves on the membrane is stroboscopic exposure[6] at a frequency differing slightly from the frequency at which the membrane vibrates.

2) The amplitude at any point of the sheet of rubber can be more accurately determined by placing over that point a minute metal leaf spring mounted via an insulating cylinder on a micrometer screw at a short distance from the end of the screw (*fig. 3*). The screw is so adjusted that the vibrating sheet of rubber just presses the spring up against the end of the screw, thereby making and breaking an electrical contact, which can be made audible with the aid of an amplifier and a loudspeaker. The position of the micrometer screw in which contact is just made is a measure of the amplitude at that particular spot.

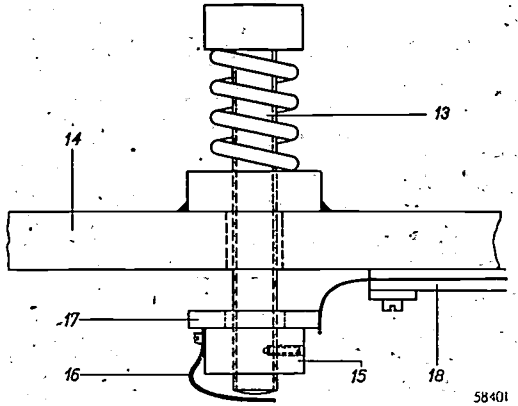


Fig. 3. Device for measuring the amplitude of the membrane at any point. *13* = micrometer screw. *13* is screwed into a nut soldered onto the plane-ground bar *14*, which slides along the bars 7 (see fig. 2). *15* = insulating cylinder, *16* = leaf spring, *17* = contact ring, *18* = wire for current feeding. The screw is placed over the point where the amplitude is to be measured and so adjusted that the vibrating membrane periodically touches the spring and just causes it to make contact with the screw, this being made audible by means of a signalling arrangement.

This latter method is particularly useful for determining the standing-wave ratio, i.e. the ratio of the maximum and minimum amplitudes along the axis of the wave guide or, respectively, of the membrane. This ratio is a measure of the reflection taking place in the wave guide[7].

[5] R. E. B. Makinson, A mechanical analogy for transverse electric waves in a guide of rectangular section, J. sci. Instr. **24**, 189-190, 1947.

[6] The stroboscope used (GM 5500) is described in Philips Techn. Rev. 8, 25-32, 1946.

[7] For Lecher systems (transmission lines) the standing-wave ratio — though this term was not introduced there — was mentioned implicitly in the article by J. M. van Hofweegen: Impedance measurements with a non-tuned Lecher system, Philips Techn. Rev. 8, 278-286, 1946, in particular in fig. 2 and eq. (20).

Some properties of rectangular wave guides that can be demonstrated with the membrane.

Critical wavelength

From the equation (5) given above for the component E_y of a TE-wave an important property of rectangular wave guides can be deduced. For wavelengths which are less than a critical wavelength $\lambda_{cr} = 2a/m$ these guides behave quite differently from the way they do for wavelengths greater than λ_{cr}. Only for $\lambda \leqq \lambda_{cr}$ is $\sqrt{1-(m\lambda/2a)^2}$ real and thus the exponent purely imaginary, which corresponds to a wave travelling in the z-direction [8]). If on the other hand $\lambda > \lambda_{cr}$ then the exponent is complex, i.e. the amplitude of E_y diminishes exponentially in the z-direction. In practice use is made of both these possibilities. Where a wave guide has to transmit energy it should be so dimensioned that $\lambda < \lambda_{cr}$. The case where $\lambda > \lambda_{cr}$ is found in wave guides serving as attenuators, in which the change in field strength between two points along the axis of the guide is accurately known as a function of the distance (z) between them.

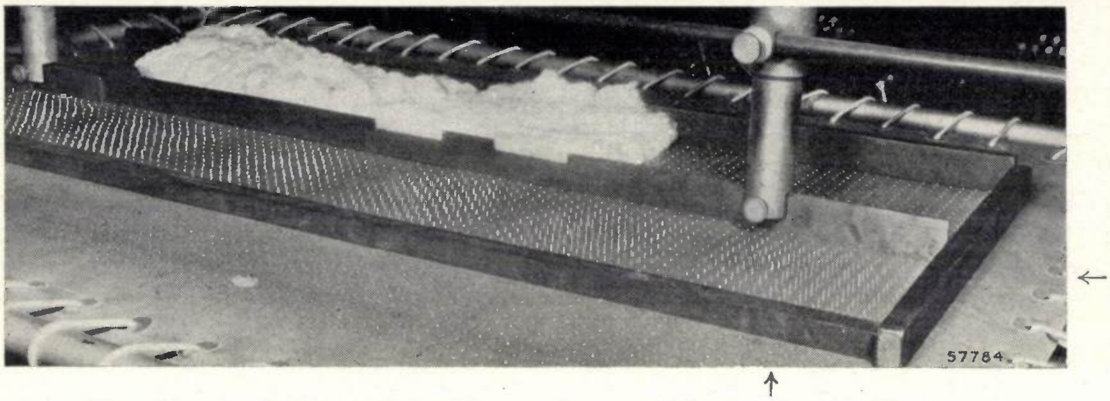


Fig. 4. Photograph of the membrane driven at a frequency higher than the critical frequency ($\lambda < \lambda_{cr}$). The arrows indicate the point at which the membrane is driven. Owing to reflection at the left-hand end (not visible in the picture) standing waves arise, the nodes and antinodes of which are clearly seen. In the background is a pad of cotton-wool which may be used for damping.

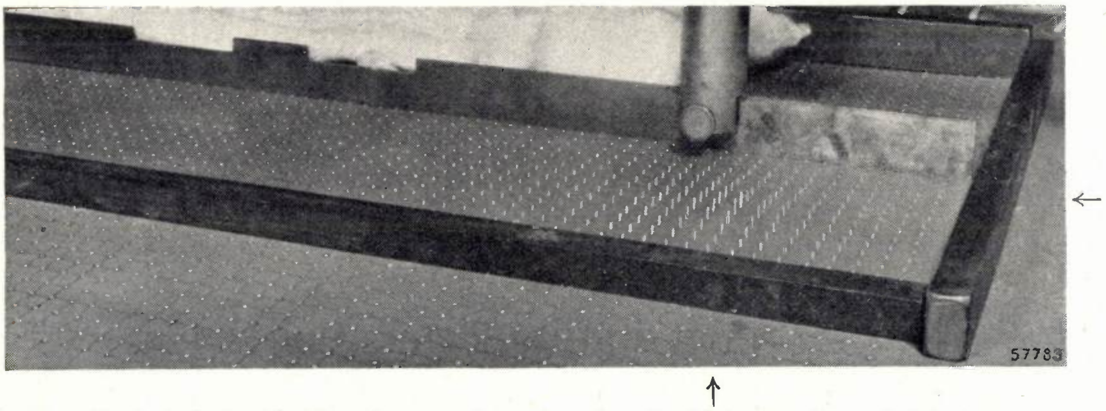


Fig. 5. As in fig. 4 but for a frequency lower than the critical frequency ($\lambda > \lambda_{cr}$). A wave cannot penetrate into the "wave guide" in this case; to the left of the driving point the amplitude rapidly decreases.

Figs 4 and 5 give two photographs of the membrane demonstrating this property. In fig. 4 the frequency has been so chosen that $\lambda < \lambda_{cr}$. It will be seen that a wave phenomenon occurs over the entire length of the "wave guide". The membrane was blocked at the end with a rigid rod (not visible in the photograph) and therefore corresponded to a wave guide closed by a conducting plate. The reflection gave rise to standing waves and the nodes and antinodes are clearly seen in fig. 4. It will also be seen that there is only one half wave across the width (due to the manner of drive), so that $m = 1$, just as in fig. 1.

[8]) In the article quoted in footnote [2]), page 24, left-hand column, third line from formula (31), $\lambda > \lambda_{cr}$ should be replaced by $\lambda < \lambda_{cr}$.

Fig. 5 relates to a lower frequency where $\lambda > \lambda_{cr}$. It is seen that the amplitude is practically zero even at a short distance to the left of the driving point.

From $\lambda_{cr} = 2a/m = v/f_{cr}$, where f_{cr} is the critical frequency, it follows that $f_{cr} = mv/2a$. The velocity of propagation v in the sheet of rubber is about 10 m/sec, so that with a width $a = 0.25$ m and the simplest mode ($m = 1$) for the membrane $f_{cr} \approx 20$ per/sec, $\lambda_{cr} = 0.50$ m.

Directional coupler

The second test to be described is on a directional coupler. This is illustrated in *fig. 6*, where A represents a wave guide (or in this case a membrane) driven at one end, at C, by $\lambda < \lambda_{cr}$. At the other end (W_A), as we shall first assume, the energy of the waves is absorbed, so that progressive waves are maintained. Side by side with this main wave guide is an auxiliary wave guide (B) coupled with A by two or more holes through which energy is transmitted from A to B. B is closed with a non-reflecting termination at the same end as A. In general, therefore, vibrations set up in B are propagated in both directions. If, however, the centre-to-centre distance of the holes is just $\lambda/4$, then in B the only possible direction of propagation is the same as in A (hence the name directional coupler). The reason for this is that two waves are propagated in the opposite direction in B. Since these are derived from two holes $\lambda/4$ apart, there is a phase difference of $2 \times \lambda/4$, i.e. 180°, and the waves cancel. According to the diagram in fig. 6, therefore, in the auxiliary wave guide there are only waves travelling to the left of the holes, and these are absorbed at W_B. To the right of the holes there is no energy at all, so that a detector placed at D does not indicate.

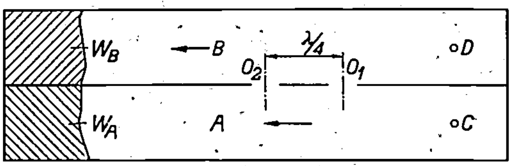


Fig. 6. Main wave guide A and auxiliary wave guide B, coupled via two holes (O_1, O_2) at a distance $\lambda/4$ (directional coupler). C = driving point, D = detector. W_A, W_B = damping material. In B waves are propagated only in the same direction as in A. When the absorption at W_A is absolute then in A waves travel only from C to W_A and in B only from the holes towards W_B; D receives no energy. If the absorption at W_A is not complete then in A waves also travel from W_A to C, setting up waves in B travelling from the holes in the direction of D; in that case the detector indicates.

The situation is different, however, when some reflection takes place in the main wave guide at W_A. In this wave guide and consequently in the auxiliary wave guide A too, there are then also waves travelling from left to right. In that case the detector D does indicate, and the quantity detected is a measure for the reflection at W_A. Thus a directional coupler can be used for investigating whether the energy in a wave guide does actually travel exclusively in the desired direction.

Figs 7 and 8 show how the action of a directional coupler can be demonstrated with the aid of the membrane. For this purpose the membrane is divided by clamping bars into two sections, that in the front, which is driven, corresponding to the main wave guide A (fig. 6) and that at the back corresponding to the auxiliary wave guide B. The coupling between the two sections is formed by two holes in the common intermediate bar between the sections. The frequency is so chosen that the centre-to-centre distance of the holes is $\lambda/4$. Absorption at the end of the second section is obtained with a pad of cotton-wool.

In the case depicted in fig. 7 a pad of cotton-wool was likewise laid on the membrane at the end of the section representing the main wave guide, so that there were only waves travelling from right to left. It is seen that there are also waves in the "auxiliary wave guide", but only to the left of the holes; to the right the membrane remains at rest. When, however, the absorbent material is removed from the front section then, owing to the reflection at the closing bar, waves are also propagated from left to right, and similarly in the second section. Consequently the right-hand part of the second section is set in motion, as can be seen in fig. 8. At the same time nodal lines are observed in the first section, so that standing waves arise as a consequence of reflection. (Between the driving point and openings a travelling wave, energizing the second section, is superposed upon these standing waves.)

Bevelling the bend in a wave guide

Where wave guides are used it is not usually possible to manage with simple straight pipes and one or more bends are unavoidable. This, however, calls for some care so as to avoid as far as possible the reflection of energy in the bend. In the case of a rectangular wave guide the simplest solution, from the constructional point of view, is to make a bend bevelled off on the outside. The question then arises as to what is the most favourable value of the parameter d/d_0 (see *fig. 9*).

A problem such as this can be solved very well with the aid of the membrane, since in the model

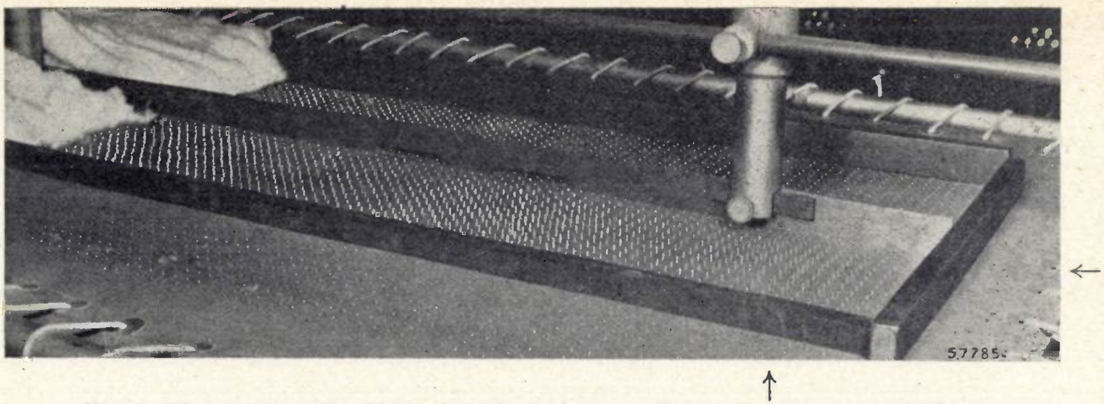


Fig. 7. Directional coupler imitated on the membrane. The front section corresponds to *A* (fig. 6) and the back one to *B*. In the intermediate wall are the two holes providing the coupling. To the left are the damping pads. Owing to the damping at the end of the front section the waves in this section travel from right to left only. Those in the second section travel in the same direction to the left of the holes. To the right (at *D* in fig. 6) the rear section remains at rest.

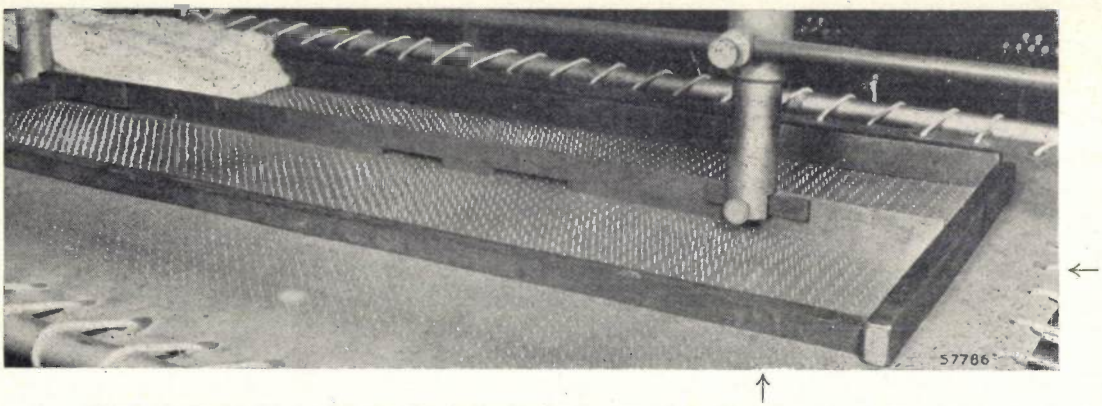


Fig. 8. As in fig. 7 but with no absorption in the front section. Standing waves now occur in the front section and in the rear section to the right of the holes. At the point *D* (fig. 6) a fairly strong vibration is set up; the amplitude is a measure of the reflection at the end of the front section.

the value of the parameter can easily be varied and the result more easily observed than is possible in an actual wave guide.

On the membrane we have imitated the shape of a wave guide bent at an angle of 90 degrees. With the aid of the aforementioned contact spring and micrometer screw the standing-wave ratio *r* has been measured, as a function of $d/d_0 = d/a\sqrt{2}$, of the wave patterns formed between the point of drive and the bend at different frequencies. The free end of the "wave guide" was closed with a non-reflecting termination.

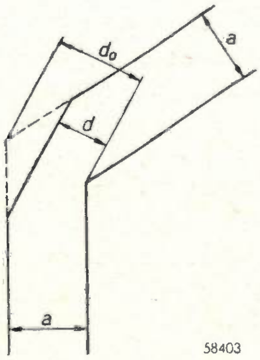


Fig. 9. When a wave guide has to be bent the bend should be bevelled off to prevent reflection, the ratio d/d_0 being given a certain optimum value.

In *fig. 10* $1/r$ is plotted as a function of d/d_0 for three frequencies. The smaller the standing-wave ratio — thus the greater $1/r$ — the less reflection takes place and thus the less disturbing is the bend in the wave guide. As may be seen from fig. 10, the optimum value of d/d_0 lies between 0.6 and 0.7, in agreement with the value found by entirely different means [9]).

The foregoing examples will have illustrated clearly enough that a mechanical model such as described here may be of great value for studying wave phenomena in rectangular wave guides. Naturally this also applies for cavity resonators

[9]) See, e.g., G. L. Ragan, Microwave transmission circuits, Massachusetts Institute of Technology, New York 1948, page 207.

and other configurations, provided the phenomena can be described with two Cartesian co-ordinates.

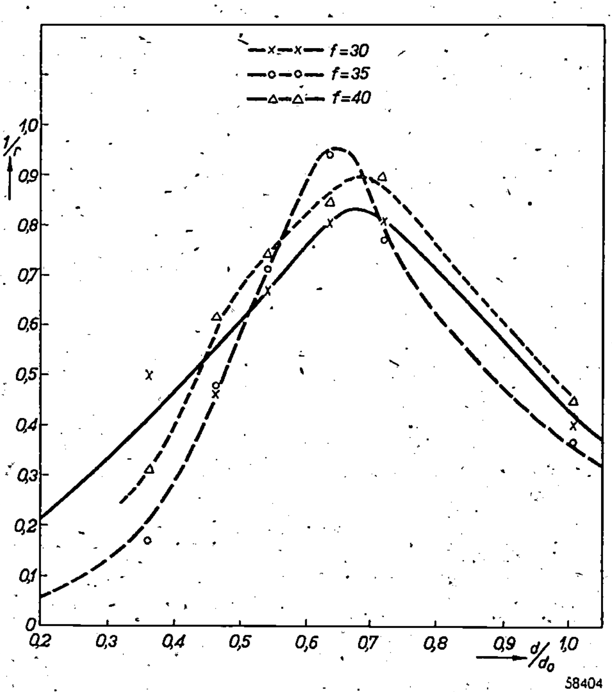


Fig. 10. On the membrane a wave guide bent 90 degrees is imitated and the standing-wave ratio r (in the absence of reflection this equals 1) is measured as a function of d/d_0 (fig. 9). The curves are related to the frequencies 30, 35 and 40 c/s. The optimum value of d/d_0 lies between 0.6 and 0.7.

It seems likely to us that electrical or mechanical vibrations in non-homogeneous media can also be investigated with the aid of such a model. The inhomogeneities of the medium would then have to be imitated in the model by varying the tension in the membrane and/or its density according to a certain function of the co-ordinates x and z. Many cases where calculations are too difficult could then be investigated by this means.

Summary. A mechanical model is described with which the properties of rectangular wave guides can be studied. It consists of a membrane in the form of a sheet of rubber stretched on a frame and caused to vibrate in a suitable manner. By clamping the sheet between flat bars, boundary conditions can be created corresponding to those of the wave guide. Absorption and reflection at the end of the wave guide can be imitated on the membrane by laying cotton wool on it and by clamping it. The membrane is driven by a direct-current motor via an eccentric. The movement of the membrane can easily be followed visually (especially when applying stroboscopic exposure) by the use of a network of white dots painted on the membrane. The amplitude of the vibration at any point can be measured with the aid of a simple instrument. With this model one can demonstrate and investigate the phenomenon of critical wavelength, the properties of a directional coupler and the most favourable bevelling of a bend in a wave guide. The value of the model described lies especially in its pictorial presentation and the ease with which various parameters can be altered. In principle it is also possible to make the model suitable for studying the propagation of waves in non-homogeneous media.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATION OF THE N.V. PHILIPS' GLOEILAMPENFABRIEKEN

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

R 105: G. Diemer and K. S. Knol: Frequency conversion by phase variation (Philips Res. Rep. 4, 161-167, 1949, No. 3).

An improvement of Herold's method of phase reversal resulting in a still higher value of the conversion transconductance is studied theoretically. It is proved that the conversion transconductance can equal the maximum value of the high-frequency transconductance. This improvement by a factor $\pi/2$ in conversion transconductance over Herold's method is achieved by varying in a suitable way the phase of the transconductance with time. This is theoretically the highest value of the conversion transconductance that can be reached in any design. The theory has been corroborated by experiments with a tube for low frequencies.

R 106: A. van der Ziel and K. S. Knol: On the power gain and the bandwidth of feedback amplifier stages (Philips Res. Rep. 4, 168-178, 1949, No. 3).

The influence of feedback upon the power gain of an amplifier stage is discussed and it is shown that when for a given value of the feedback the limit of stability is reached, the power gain remains finite. The formulae are applied to the case of a grounded-grid stage and it is shown that rather small output losses may already give rise to a considerable decrease in gain. The influence of feedback upon the band width of an amplifier stage is also discussed. The results are again applied to a grounded-grid triode stage. The product of the band width B and the power gain g is plotted against g (with the coupling of the output circuit to the load as a parameter) for various values of the feedback capacitance and the output losses. It is shown that an amplifier stage having such an amount of feedback that it can just give oscillations for proper output coupling can only give a very high values of the power gain g for very small band widths. A considerable improvement in bandwidth is then possible by increasing the amount of feedback.

R 107: C. J. Bouwkamp: On the effective length of a linear transmitting antenna (Philips Res. Rep. 4, 179-188, 1949, No. 3).

The concept of effective length, proposed by King for the cylindrical, centre-driven antenna is extended. The new definition may be applied to any current distribution in the antenna. As an illustrative example a top-loaded antenna is discussed in detail.

R 108: H. Bremmer: Some remarks on ionospheric double refraction, II (Philips Res. Rep. 4, 189-205, 1949, No. 3).

Maxwell's equations for plane-wave propagation through a stratified anisotropic medium are reduced to a system of four ordinary first-order equations. These equations are very suitable for the derivation of W.K.B. approximations. The general theory given is worked out in detail for the special case of a stratified isotropic medium. In the last section the computation of field strengths in the case of ionospheric reflections is discussed (see these abstracts, No. **1851**).

R 109: J. A. Haringx: On highly compressible helical springs and rubber rods and their application for vibration-free mountings, III (Philips Res. Rep. 4, 206-220, 1949, No. 3).

This paper (see Nos **R 91** and **R 97**) deals with the elastic stability and the lateral rigidity of solid rubber rods under axial compression. The various calculations are based upon equations derived in the preceding papers for the elastic prismatic rod replacing the helical spring, though here some additional difficulties arise from the effect on the rigidity against shearing, caused by the inevitable vaulting of the initially flat normal cross sections of the rod. This effect, however, is of minor importance for rods of circular or rectangular cross section, so that in these cases a solution of the respective problems can be given. As to the buckling phenomenon a satisfactory agreement is found between the results of the present calculation and those of experiments by Kosten. The minor discrepancies occurring are probably due to the limited validity of the formulae introduced to describe the elastic behaviour of highly compressed rubber material.

If the rubber rods are vulcanized or affixed in some other way to metal pieces, at their ends a thin layer of material is to be regarded as fully incompressible, though it must be borne in mind that this layer retains its elasticity in respect of shearing. The corresponding effect upon the lateral rigidity is taken into account.

R 110: W. Elenbaas: High-pressure rare-gas discharges (Philips Res. Rep. 4, 221-231, 1949, No. 3).

The light output, the total radiation and the electrical gradient have been measured for A, Kr and Xe high-pressure discharges of the wall-stabilized type. The characteristics of these discharges have been calculated assuming the discharges to be in temperature equilibrium and the radiation to be mainly a recombination continuum. The light output, the total radiation and the electrical gradient have been derived assuming a constant energy per unit of frequency (as calculated by Unsöld) in the frequency band from $\nu = 0$ to $\nu = \nu_m = eV_i/h$. In the calculation of the electrical gradient the influence of the positive ions on the mobility of the electrons has to be considered. The influence of the atoms and that of the positive ions have been experimentally separated and the magnitude of both has been found in accordance with the calculations.

R 111: P. Cornelius: Proposals and recommendations concerning the definitions and units of electromagnetic quantities (Philips Res. Rep. 4, 232-237, 1949, No. 3).

To complete the arguments of a previous article (see No. **R 103**) the writer ventures to bring forward a set of proposals by which, in his opinion, many questions still in dispute may be finally settled.

VOL. 11 No. 6, pp. 165-192 DECEMBER 1949

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

HEATING BY HIGH-FREQUENCY FIELDS

I. INDUCTION HEATING

by E. C. WITSENBURG *).

621.364.156:621.785

The heating of a metal by means of an electric current induced inside it is by no means new; attempts to put this method into practical application date as far back as the beginning of this century. Shortly after 1920 induction heating with high-frequency current was introduced in the manufacture of radio valves. It was not until some years later (1926-1929) that further theoretical and practical research led to a better insight being obtained, upon which is based the rapid development of induction heating since 1940. Since the frequencies and the power required for this method of heating are the same as those used in radio transmitters, this development is due to the great progress made in radio engineering during the last 20 years. — Although the theory, which allows of the most favourable frequency being chosen and the apparatus suitably dimensioned, now no longer holds any secrets, the possibilities of application have not yet been fully investigated. The author mentions as an example a number of processes (surface hardening, annealing, soldering, melting) where induction heating has already opened up important perspectives for the metal-working industry.

A further article will deal with capacitive heating, which in a certain sense forms the complement of induction heating: whereas the latter is confined to conducting materials, capacitive heating, which is based upon dielectric losses, is applied to non-conducting or poorly conducting materials.

Heating is a treatment that has to be applied for innumerable processes in industry to give the product the properties required. In by far the majority of cases the heat needed is generated externally, either by burning or by some other chemical process, or by an electric current flowing through a separate resistance wire or forming an arc. The heat is carried to the point to be heated by conduction, convection and/or radiation.

Apart from this indirect heating, however, direct heating is also possible, the heat then being generated inside the material to be heated. If this material is electrically conductive (metal, graphite) the heat can then be generated by sending an electric current through it. Mostly, however, the shape of the object (or of the part of it to be heated) does not lend itself to connection to a source of electric current. Some of the few examples of such are resistance-welding and the heating to which the filaments of incandescent lamps and radio valves are submitted during their processing. In many cases the only practical way to generate the heating current in the material is the method of induction, thus with the aid of an alternating magnetic field. Hence the name given to the process to be dealt with in this article: induction heating.

Heat can be generated not only in conducting materials but also in insulating media, thanks to the phenomenon of dielectric losses. These losses arise when the material is situated in an alternating electric field, thus for instance between the plates of a capacitor connected to an alternating voltage. Hence the name capacitive heating.

Thanks to these two methods of heating many

*) Philips Telecommunication Industry, formerly N.S.F., Hilversum (Netherlands).

cases of thermal treatment in industry have been made possible which by a more conventional method would be less efficient or even impossible. The improvement in quality and increased production reached in many cases with induction or capacitive heating have already led to extensive application of both these methods, whilst it is to be expected that this industrial field of application will be considerably expanded in the years to come. Experience will then show more clearly which thermal treatments lend themselves best to induction or to capacitive heating and which can best be carried out by the means hitherto more commonly applied, since, as with every process, induction and capacitive heating methods naturally have their limitations too.

Apart from the fact that in induction heating as well as in capacitive heating the heat is generated *inside* the material, these two methods have another point in common, in that both with capacitive heating and also in by far the most cases with induction heating satisfactory results are only possible by employing high frequencies, which can only be obtained in a suitable manner by the methods of radio engineering. Otherwise there is such a great difference between induction and capacitive heating — if only by reason of the nature of the object that can be treated (conductors as opposed to non-conductors) — that they are best dealt with in separate articles. Induction heating will be dealt with first.

Heating processes which have in many cascs been greatly improved by the introduction of high-frequency inductive heating are: hardening, annealing, brazing and soldering, and melting. Still further possibilities are being sought in all directions, as evidenced by the extensive literature being published nowadays on these subjects. The great interest being taken in this matter might give the erroneous impression that this is an entirely new technique. However, already shortly after 1920 the Philips factories at Eindhoven, among others, were applying induction heating on an industrial scale for the heating (degassing) of the electrodes of radio valves. Since 1935 the same works have been using an induction furnace [1]) for melting magnet steel and other alloys. In both these cases the high-frequency current is generated by means of a valve oscillator. For certain applications, however, lower frequencies suffice, in which case the power may alternatively be supplied by rotary generators.

Some brief theoretical considerations may serve to explain the characteristic results to be achieved with induction heating. Before proceeding to discuss some of its applications a very simple example will therefore be given to elucidate the most typical phenomena.

General comments on induction heating

Let us consider what happens when a wire is wound round a steel spindle of say 2 cm diameter (*fig. 1*) and the ends of the coil formed by these turns of wire (called the work coil) are connected to a generator supplying an alternating current with a frequency of say 500,000 c/s [2]). The current flowing through the work coil sets up a magnetic field which induces currents in the rod. Due to the I^2R losses these currents supply the heat required. In accordance with the known law of induction the currents flowing in the rod counteract the magnetic field from which they originated. From this it follows, as the first typical property of induction heating, that the heat generated is mainly confined to that part of the workpiece immediately opposite the work coil. Thus, given suitable dimensions and position of the coil, a workpiece can be partially heated, and in many cases this will be found to be one of the main advantages of this method of heating.

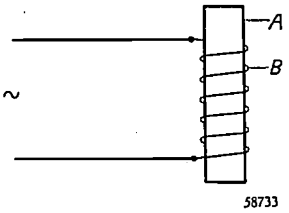


Fig. 1. A rod A is heated by the current induced in it by the alternating current in the work coil B.

There is also another special point about the induced current. The current density is greatest at the surface of the rod, rapidly diminishing inwards approximately according to the formula

$$J_x = J_0 e^{-\frac{x}{\delta}},$$

in which J_x represents the current density at a

[1]) For a description of this installation see Philips Techn. Rev. 1, 53-59, 1936.

[2]) A similar case, but not with the object of heating, was recently dealt with in this periodical: P. Zijlstra, An apparatus for detecting superficial cracks in wires, Philips Techn. Rev. 11, 12-15, 1949 (No. 1).

depth x, J_0 the density at the surface and δ the so-called penetration depth, i.e. the depth at which $J_x = J_0/e$. About 83% of the total heat is generated in the outermost layer of the rod of the thickness δ.

The depth of penetration depends upon the specific resistance ϱ and the relative permeability μ_r of the material of the rod and also upon the frequency f. It can be calculated with the aid of the formula

$$\delta = \frac{1}{2\pi}\sqrt{\frac{\varrho \cdot 10^7}{\mu_r f}} \approx 503\sqrt{\frac{\varrho}{\mu_r f}} \text{ meters}. \quad (1)$$

(ϱ in Ω·m, f in c/s.)

As regards the relative permeability of ferromagnetic materials it is to be noted that on account of the very high field strength usual with inductive heating there is considerable saturation, so that the value of μ_r is always rather small (in the order of 100 or even 10). When the material is heated to above the Curie point (for iron 770 °C) the ferromagnetic properties disappear and the value of $\mu_r = 1$.

In *fig. 2* the penetration depth has been plotted as a function of the frequency for several materials. From this it is to be seen that in the case of the rod in question (iron in the hot state) the heat is generated mainly in a skin 0.85 mm thick (at the chosen frequency of 500,000 c/s.)

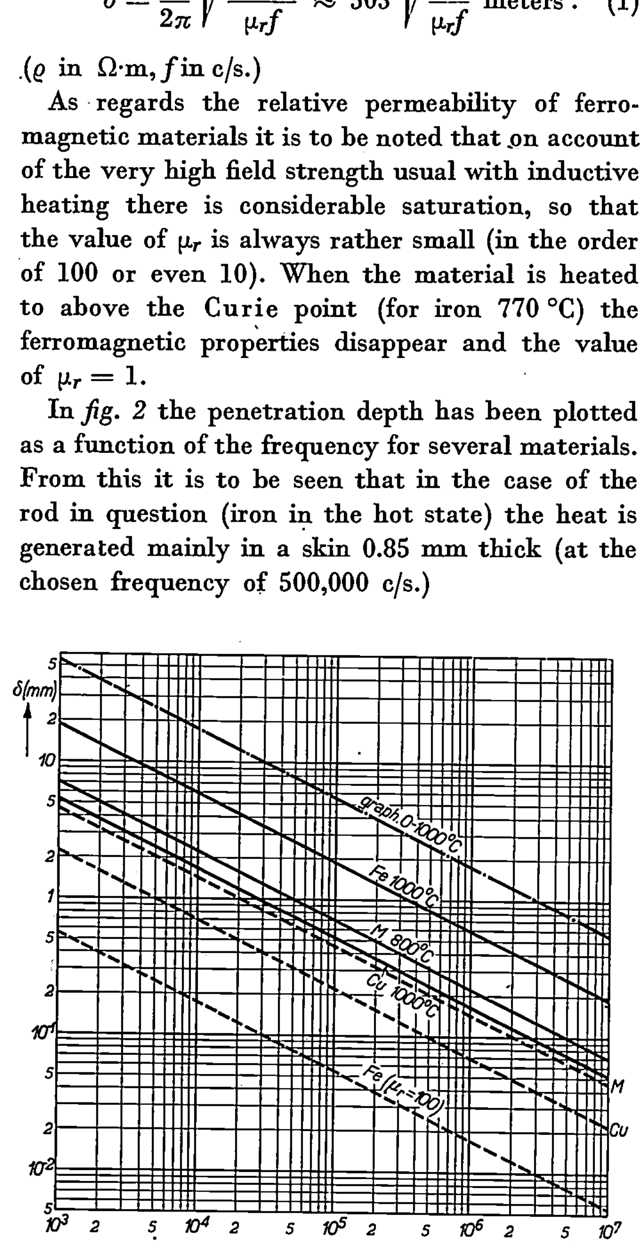


Fig. 2. The penetration depth δ as a function of the frequency f for several materials, according to formula (1). *Graph* = graphite, *Fe* = iron, *Cu* = copper, *M* = brass. The dotted lines apply for room temperature, the others for 800-1000 °C. The line for graphite is practically independent of the temperature. In the case of cold iron it has been taken that $\mu_r = 100$; for iron at 1000 °C $\mu_r = 1$.

This heat is used in the first place for heating the skin, but as soon as the temperature of the skin rises above that of the surroundings heat begins to flow away in two directions: through the rod inwards by conduction, and to a less degree outwards by radiation and convection. Mathematically it is difficult to determine exactly what happens, but from the heat conductivity and the coefficient of radiation of the material (in this case steel) an idea can be formed of magnitude of the two thermal flows. For instance, in iron a temperature difference of 100 °C per mm gives rise to a thermal current of about 7 W/mm². And at a temperature of 1000 °C at the surface the radiation amounts to 0.15 W/mm². We shall revert to the matter of heat dissipation when dealing with surface hardening.

Another point to be noted is that, if the shape of the workpiece makes it desirable, the work coil can also be placed inside the workpiece, an example of which will be given presently.

Efficiency of the work coil

Part of the power supplied to the work coil is lost owing to the resistance of the coil. Thus we can speak of the efficiency of the work coil as representing the ratio η of the power P_2 reaching the charge with respect to the power supplied $P_0 = P_1 + P_2$, with P_1 denoting the power lost in the work coil. Thus

$$\eta = \frac{P_2}{P_1 + P_2}. \quad \dots\dots \quad (2)$$

It is not difficult to understand that, within certain limits, the efficiency will be greater according as the specific resistance (ϱ_1) of the charge is higher with respect to that (ϱ_2) of the material of which the work coil is made, and also according as the charge is more closely enveloped by this coil. In addition to these factors there is another less obvious one bearing upon the efficiency of the work coil, namely the ratio of the diameter of the charge to the penetration depth δ_2 in the rod. A simplified theoretical calculation yields the following formula for the efficiency of the work coil:

$$\eta = \frac{1}{1 + \frac{D^2}{d^2}\left(1 + 6.25\,\frac{{\delta_2}^2}{d^2}\right)\sqrt{\frac{\varrho_1}{\mu_r\,\varrho_2}}}, \quad . \quad (3)$$

in which D = the diameter of the work coil and d = the diameter of the charge.

This formula can be derived in the following way [3].

In the equivalent circuit (*fig. 3*) R_1 and L_1 represent the resistance and the self-inductance of the work coil, and R_2 and L_2 the resistance and inductance of the charge, the latter being coupled to the work coil via the mutual inductance M.

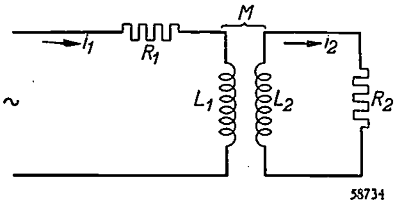


Fig. 3. Equivalent circuit for the situation illustrated in fig. 1. R_1, L_1 = resistance and inductance of the work coil, R_2, L_2 = resistance and inductance of the workpiece, M = mutual inductance.

If I_1 is the R.M.S. value of the primary current and I_2 that of the secondary current then the output is:

$$P_2 = I_2^2 R_2 = n^2 I_1^2 R_2,$$

where $n = I_2/I_1$ represents the current-transformation ratio.

The power P_1 lost in the work coil is

$$P_1 = I_1^2 R_1,$$

so that the efficiency of the work coil is:

$$\eta = \frac{P_2}{P_1 + P_2} = \frac{n^2 R_2}{R_1 + n^2 R_2} = \frac{1}{1 + \frac{R_1}{R_2} \cdot \frac{1}{n^2}} . \quad . \quad . \quad (4)$$

We shall now consider separately the factors R_1/R_2 and $1/n^2$ occurring in the denominator of the last fraction, for a simple case taken very schematically.

Let us assume that the work coil consists of one turn of strip with such a width (H in *fig. 4*) that the field inside this coil can be regarded as being homogeneous. A rod-shaped workpiece (length $h < H$) is placed inside this homogeneous field.

The current in the work coil can be imagined as being concentrated in a thin layer (thickness δ_1) on the inside. The resistance of the work coil is therefore $R_1 = \varrho_1 \, . \, \pi D/H\delta_1$. Similarly, the resistance of the rod is $R_2 = \varrho_2 \, . \, \pi d/h\delta_2$.

For the sake of simplicity it is further assumed that the rod is similar to the work coil, thus that $d/h = D/H$. In that case

$$\frac{R_1}{R_2} = \frac{\varrho_1}{\varrho_2} \cdot \frac{\delta_2}{\delta_1} .$$

Using formula (1) and taking $\mu_r = 1$ for the relative permeability of the work coil, consisting of non-ferromagnetic material, we arrive at

$$\frac{R_1}{R_2} = \sqrt{\frac{\varrho_1}{\varrho_2 \, \mu_r}} . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

There now remains the factor $1/n^2$. For the secondary circuit (fig. 3) we have:

$$i_2 \, (R_2 + \mathrm{j}\omega L_2) + i_1 \cdot \mathrm{j}\omega M = 0,$$

where i_1 and i_2 represent the complex values of the two currents and $\omega = 2\pi f$. Thus

$$\frac{i_1}{i_2} = \frac{R_2 + \mathrm{j}\omega L_2}{-\mathrm{j}\omega M} .$$

$1/n^2$ is the square of the absolute value of this expression:

$$\frac{1}{n^2} = \frac{R_2^2 + \omega^2 L_2^2}{\omega^2 M^2} = \frac{\frac{R_2^2}{\omega^2 L_2^2} + 1}{k^2 \frac{L_1}{L_2}} .$$

in which $k^2 = M^2/L_1 L_2$, for which we can write approximately d^3/D^3.

Substituting:

$$R_2 = \varrho_2 \cdot \pi d/h\delta_2,$$
$$L_2 = \pi\mu F d^2/4h,$$
$$L_1 = \pi\mu F D^2/4H$$

(in which $\mu = \mu_0 \mu_r = 4\pi \cdot 10^{-7} \cdot \mu_r$ and F is a factor depending upon $D/H = d/h$) we find after some conversion:

$$\frac{1}{n^2} = \frac{D^2}{d^2} \left(1 + \frac{\varrho_2^2 \cdot 10^7}{\pi^2 F^2 f^2 \mu_r^2 d^2}\right) = \frac{D^2}{d^2} \left(1 + \frac{4}{F^2} \cdot \frac{\delta_2^2}{d^2}\right) . \quad (6)$$

Substituting (5) and (6) in (4) and with $F = 0.8$ (corresponding to $D/H = 0.6$), we arrive at the equation (3) for the efficiency η.

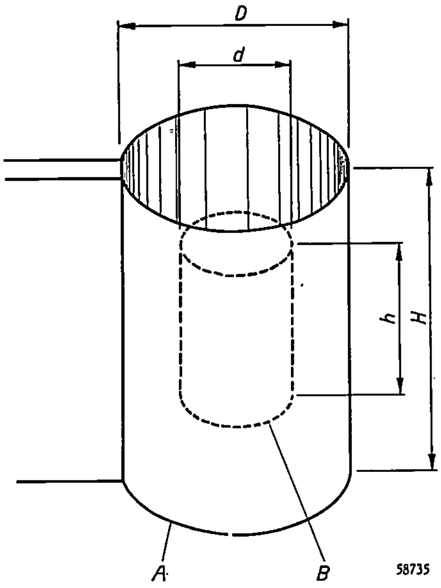


Fig. 4. Sketch of the set-up for which the efficiency of the work coil is calculated. A = work coil of one turn (diameter D, width H), B = rod-shaped workpiece (diameter d, length h). It is assumed that H/D is so great that the field inside the coil may be regarded as being homogeneous, and that $H/D = h/d$.

Although formula (3) applies only for a very simple case, we can nevertheless draw some conclusions from it also for other cases in regard to the steps that have to be taken in order to get a high efficiency, in other words, to get a small value for the denominator of the right-hand member of (3). The second term in this denominator consists of three factors, which will be dealt with separately.

1) D^2/d^2. For high efficiency the difference between D and d has to be chosen as small as possible; hence the work coil has to be made to fit as

[3] See also: E. C. Witsenburg, High-frequency inductive heating, (Dutch) T. Ned. Radiog. **12**, 201-211, 1947.

closely as possible round the workpiece (or, inversely, the workpiece has to fit as closely as possible round the coil).

2) $(1 + 6.25\ \delta_2{}^2/d^2)$. Just as is the case with the first factor, this is also greater than 1. Allowing as the maximum value that can still usually be reached, say, 1.1, then we must have $d/\delta_2 \geqq \sqrt{62.5} \approx 8$. In other, words, the frequency has to be chosen so high that the penetration depth is not greater than about one eighth of the diameter of the workpiece. Substituting for δ_2 the value according to (1), we get the following equation for the minimum frequency f_{min} required to reach a high efficiency:

$$f_{\text{min}} = 503^2 \cdot 62.5 \cdot \frac{\varrho_2}{\mu_r d^2} = 16 \cdot 10^6 \cdot \frac{\varrho_2}{\mu_r d^2}\ \text{c/s} \qquad (7)$$

(ϱ_2 in Ω·m, d in m).

Taking as an example the case of iron heated to 1000 °C, with $\varrho_2 = 1.4 \times 10^{-6}$ Ω·m and $\mu_r = 1$ (1000 °C is above the Curie point), then $f_{\text{min}} \approx 22.5/d^2$. From this we find the following values for various diameters of the rod:

d	f_{min}
500 mm (≈ 20 ")	90 c/s
100 mm (≈ 4 ")	2 250 c/s
10 mm (≈ 0.4")	225 000 c/s
5 mm (≈ 0.2")	900 000 c/s

If one chooses $f > f_{\text{min}}$ the efficiency of the work coil is not appreciably increased, whilst the efficiency of the generator supplying the high-frequency current is reduced. Therefore f_{min} is about the optimum frequency at which one can work.

In *fig. 5* the frequency f_{min} is plotted as a function of d for several materials.

3) $\sqrt{\varrho_1/\varrho_2\mu_r}$. As regards this factor, with a given charge we only know the value of ϱ_1. As was to be expected, in order to reach a high efficiency the work coil must be of high conductivity, thus of copper. For the same reason it has to be kept cold, and that is why it is often made in the form of a copper tube through which cooling water is circulated.

When the workpiece is likewise of copper and is still cold then the factor $\sqrt{\varrho_1/\varrho_2\mu_r} = 1$ and the efficiency remains less than 50%. In most cases, however, $\varrho_2\mu_r$ is much greater than ϱ_1, and higher values of η are therefore reached, even as much as 96%, as shown by *fig. 6*. In fig. 6 η has been plotted as a function of D/d on the assumption that the frequency equals f_{min}. For non-ferromagnetic materials with a positive temperature coefficient of the resistance, such as copper, the efficiency rises with the temperature. In the case of iron, on the other hand, the efficiency drops as soon as the temperature exceeds the Curie point, since μ_r then decreases more quickly than ϱ_2 increases.

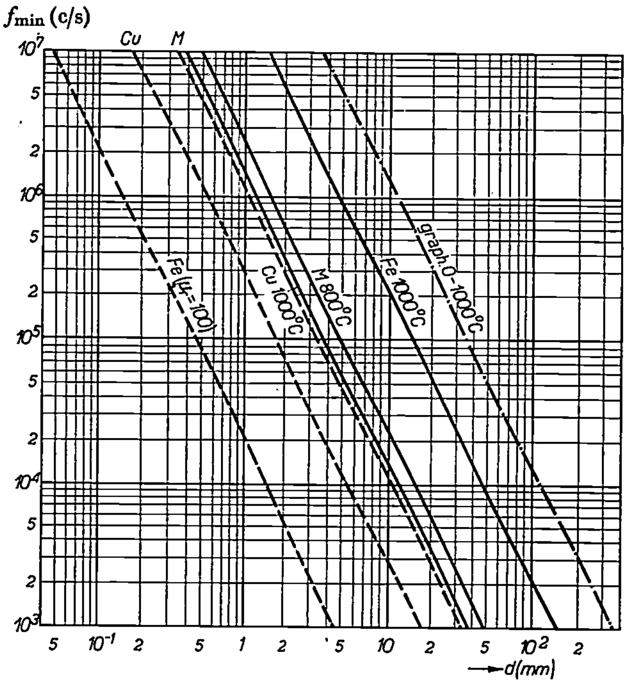


Fig. 5. The lowest frequency f_{min} at which the efficiency is not unnecessarily low, as a function of the diameter d of the workpiece, for graphite (*graph*), iron (*Fe*), copper (*Cu*) and brass (*M*), at room temperature (dotted lines) and at 800-1000 °C (fully drawn lines). For cold iron it has been taken that $\mu_r = 100$.

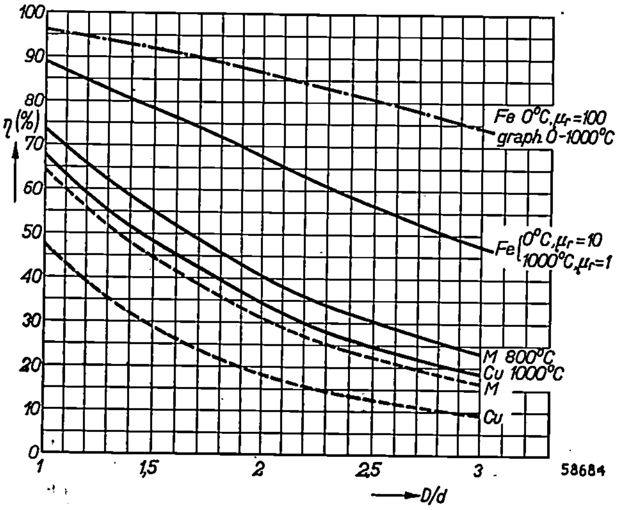


Fig. 6. Efficiency η of the work coil as a function of D/d (D = diameter of the work coil, d = diameter of the workpiece) for the same materials as in figs 2 and 5, at 0 °C (dotted lines) and at 800-1000 °C (fully drawn lines). For cold iron a curve has been plotted for $\mu_r = 100$ and another for $\mu_r = 10$. The frequency is $= f_{\text{min}}$.

Electrodynamic forces acting upon the charge

We shall conclude our general note with a consideration of the electrodynamic forces exercised upon the charge by the magnetic field.

Acting upon each point of the work coil and of the charge is a force which is proportional to the modulus of the vectorial product of the current density and the flux density at that point and which is perpendicular to the plane in which these two vectors lie. In the simple case sketched in fig. 1, given perfect symmetry, the forces acting upon the charge are directed radially inwards and cancel each other, so that the workpiece as a whole is not subject to any force. Where there are deviations from symmetry, however, such as occur, for instance, while the charge is being placed in the work coil, a (repelling) force is indeed exercised upon the workpiece and under certain circumstances this may prove to be troublesome. From what follows, where for the sake of simplicity the charge is assumed to be in a homogeneous field, it will be seen that this force becomes smaller as the frequency chosen is higher.

We have already seen that the depth of penetration of the current in the charge is proportional to $1/\sqrt{f}$ (see (1)), so that the resistance R_2 of the charge is proportional to $\sqrt{f}$. For a given power $I_2{}^2R_2$ in the charge $I_2{}^2$ must therefore be proportional to $1/\sqrt{f}$. The electrodynamic force acting upon the charge is proportional to the product of I_2 and the flux density B, so that if B is proportional to I_2 the force is proportional to $I_2{}^2$, thus proportional to $1/\sqrt{f}$; with a given power the force is inversely proportional to the square root of the frequency. Thus, for instance, with a frequency of 500,000 c/s the force, for a given power, is 100 times smaller than that with a frequency of 50 c/s.

This low force at high frequencies is a great advantage in such processes as soldering, to which we shall revert later. Possibly this may lead one to choose a frequency even higher than what is actually required to reach a high efficiency with a given workpiece.

Generators for high-frequency induction heating

From fig. 5 we have seen that in the case of large workpieces of a material that is not so highly conductive comparatively low frequencies suffice, in the order of 500 to 10,000 c/s. In the case of objects no more than a few millimetres in size, however, frequencies are required of the order of 1,000,000 c/s. Now fig. 2 shows that at the low frequencies just mentioned the penetration depth is comparatively great. Therefore, if for a certain heat treatment, particularly of iron, a small penetration depth is desired (e.g. 1 mm or less), as is often the case, then one should use a frequency in the order of 500,000 c/s.

The low frequencies can be generated by rotary converters of the type formerly used for wireless telegraphy on very long wavelengths. We shall not discuss these machines here, but a few words will be devoted to the valve generators with which, as is known, much higher frequencies can be generated.

The most important elements of such a valve generator are a triode and an oscillatory circuit. The circuit is caused to oscillate by means of feedback. The work coil is taken up in the self-inductance branch of the circuit, either direct or via a transformer. The latter is often necessary to bring about the optimum matching, i.e. to transform the very low resistance of the workpiece to a value at which the transmitting valve is correctly loaded.

At the resonant frequency the resistance r in series with one of the branches of an oscillating circuit (*fig. 7a*) is equivalent to a resistance R_p of the value Q^2r parallel to the circuit (Q = quality factor). If, with a suitable value of Q (about 30), R_p has to be made of magnitude required for the transmitting valve (e.g. 1000 Ω) then r must be of the order of 1 Ω. Now as a rule the resistance R_2 of the workpiece is much less than this. In the simple case considered in the foregoing $R_2 = \varrho_2 \cdot \pi d/h\delta_2$. Where $f = f_{\min}$, $d \approx 8\delta_2$ and thus $R_2 \approx 8\pi\varrho_2/h$. With, for instance, $h = 5$ cm we find for iron at a temperature of 1000 °C: $R_2 \approx 0.7$ mΩ. To transform this value to $r = 1\Omega$ a work coil with $\sqrt{r/R_2} \approx 38$ turns would be required, and usually this is not practicable. The desired transformation is then brought about with the aid of an intermediate transformer (T in fig. 7*b*), which then allows of a work coil being used consisting of only one turn. This results in the heating being highly concentrated and for that reason the transformer T is called a concentrator.

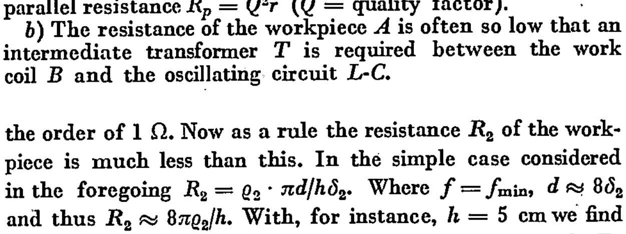


Fig. 7. *a*) At the rcsonant frequency of an oscillatory circuit *L-C* the resistance r in one of the branches is equivalent to a parallel resistance $R_p = Q^2r$ (Q = quality factor).

b) The resistance of the workpiece *A* is often so low that an intermediate transformer *T* is required between the work coil *B* and the oscillating circuit *L-C*.

Since the resistance of most materials changes appreciably with the temperature, in order to maintain the optimum conditions repeated matchings would in fact have to be made in the course of the heating process. For melting processes, lasting say 10 minutes, this can be done by connecting capacitors in parallel to the work coil

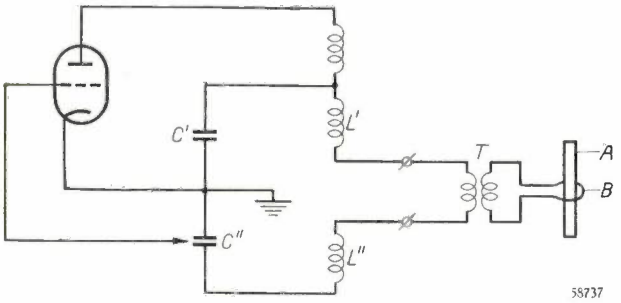


Fig. 8. Basic circuit diagram of a generator for high-frequency induction heating (Colpitts circuit). C'-C''-L'-L'' = symmetrically designed oscillatory circuit, T = intermediate transformer, B = work coil, A = charge.

step by step [4]). In many other cases, however, the heating process takes no more than a few seconds, so that manual control is impracticable. Although automatic control would be possible, a fixed adjustment averaging approximately the optimum adjustments for the cold and the hot state is usually considered sufficient.

In order to avoid the complication of a feedback coil in the generator (for constructional reasons it is often difficult to use one) it is usually preferred to employ the Colpitts circuit (*fig. 8*), in which the alternating grid voltage is derived from the capacitive branch of the oscillatory circuit.

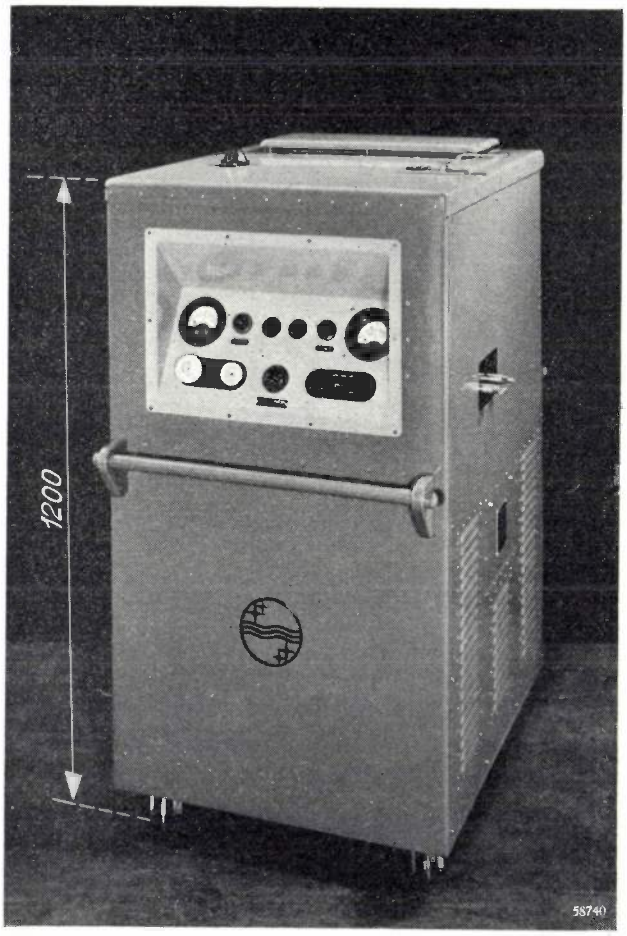


Fig. 9. Generator for high-frequency induction heating, type SFG 136/00. On the right the connections for the work coil. Maximum power in the workpiece 2 kW, frequency 1.2 Mc/s. The apparatus is equipped with one transmitting valve TB 3/2000 and two rectifying valves DCG 5/5000. Height indicated in mm.

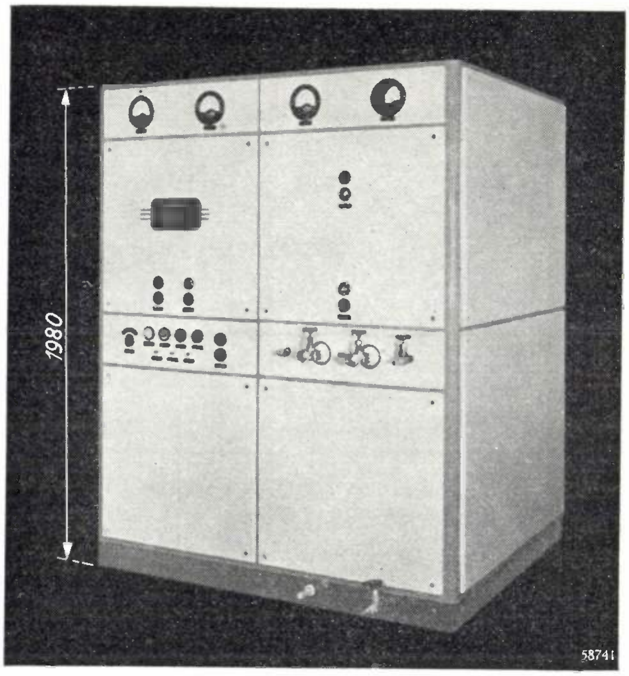


Fig. 10. Generator for high-frequency induction heating, type SFG 134/00. On the right panel the connections for the work coil, with water cooling. Maximum power in the workpiece 20 kW, frequency 450 kc/s. The apparatus is equipped with two valves TA 12/20 connected in parallel and three rectifying valves DCG 12/30. Height in mm.

The output is controlled by means of the D.C. voltage used for feeding the anode circuit of the transmitting valve. This voltage is derived from a rectifier and is variable either in stages (tappings on the secondary of the supply transformer) or — as is mostly required — continuously. For the latter method there are various means available: a variable transformer (with sliding contact) or an induction regulator connected between the mains and the supply transformer, or else a system of control with the aid of rectifying valves provided with a control grid (relay valves or thyratrons) [5]).

Figs. 9 and *10* give photographs of generators for

[4]) See page 55 of the article quoted in footnote [1]).

[5]) See, e.g.: J. G. W. Mulder and H. L. van der Horst, A controllable rectifier unit for 20,000 V and 18 A, Philips Techn. Rev. **1**, 161-165, 1936.

induction heating, for powers of 2 kW and 20 kW. These generators have a fixed frequency of 1.2 and 0.45 Mc/s respectively, which according to fig. 5 are high enough for heating objects of a diameter down to a few millimetres with a reasonable efficiency and a small depth of penetration.

Applications of induction haeting

We shall now describe some applications of induction heating where this method yields exceptionally favourable results.

Surface hardening

Many parts of machinery, such as spindles and gear-wheels, and also cutting tools, for instance screw taps, are subject to wear and at the same time are subjected to great forces. The steel used for making these parts and tools can be hardened to give it better wearing properties, but this also makes it brittle. To withstand any great force, however, the steel must be tough, but then it is also soft. It is not possible to combine hardness with toughness throughout the whole of the material, but in fact this is not essential. It is sufficient if the material is hard on the surface where it is subject to wear, and everywhere else tough. To reach this distribution of properties the case-hardening process is traditionally employed: the workpiece is made of a non-hardening steel and the surface layer is subjected to the action of carbon, thus forming in that layer a steel that can be hardened. For this case-hardening a lengthy process of annealing is essential, and in that process undesired alterations may take place in the structure of the material and there are apt to be deviations from the specified dimensions.

A much better solution of the problem of surface hardening is offered by the process of high-frequency induction heating. With this method heat is generated in a layer on the surface, just where it is needed. It is now only necessary for the outermost layer of the non-hardened, tough carbon steel (hardness about 15 Rockwell) of the workpiece to be heated to the hardening temperature so quickly that the deeper layers, receiving heat only through conduction, are kept at a much lower temperature, after which the outermost layer is rapidly cooled. This layer then has a hardness of about 60 Rockwell. In the process of heating the power supplied has to be raised high enough to exceed amply the dissipation by conduction. This means that the surface density of the power supplied must be of the order of 1 kW/cm^2. Such powers and even higher are actually employed in induction heating (the highest values correspond to the thinnest hardened layers).

No other means are known for reaching such a high concentration of heat, excepting the electric arc as used for welding, but this cannot be sufficiently controlled for the hardening of metals. The transmission of heat by radiation or direct contact with, for instance, an acetylene flame is absolutely inadequate.

The concentration that can be reached in practice with induction heating is limited by the losses in the work coil. Therefore in order to get very high concentrations one must take into account the already discussed factors governing the efficiency

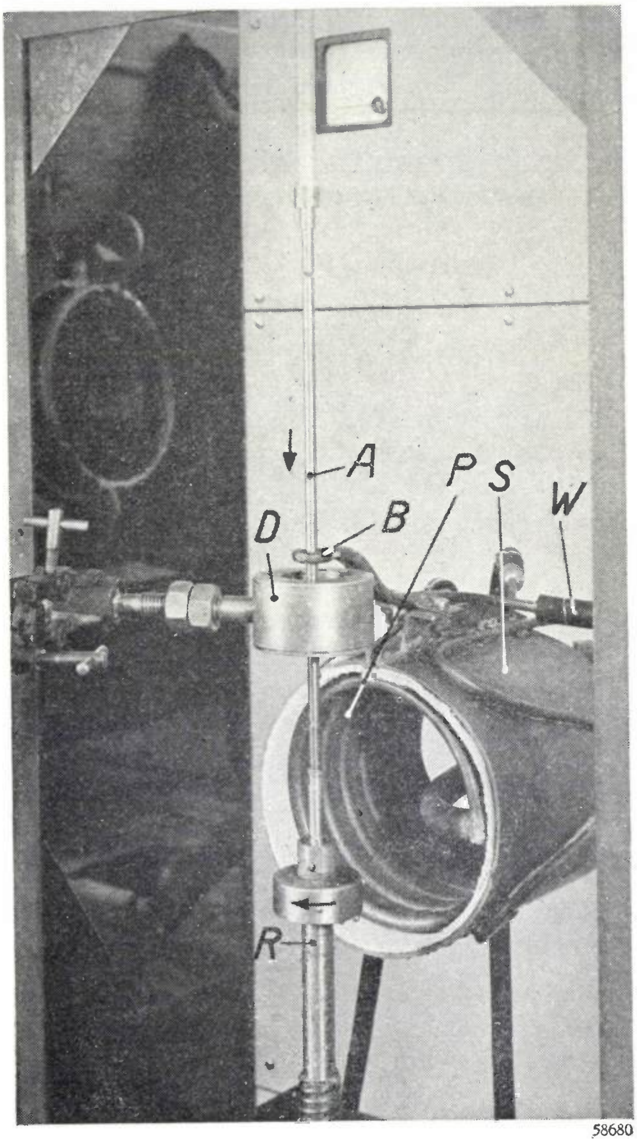


Fig. 11. Hardening of a carbon-steel spindle (A). B = work coil of one turn, P = primary, S = secondary coil of the intermediate transformer (T in fig. 8), D = water spray for quenching the spindle, moving downwards at a rate of 2 m/min. and at the same time rotated by the shaft R. W = cooling-water supply for the work coil.

of the work coil and arrange for this coil to be well cooled with water (so as to carry off the loss heat and thus keep the resistance of the coil low).

The process of surface hardening will now be illustrated with three examples. *Fig. 11* shows a set-up for hardening small spindles (6 mm in diameter) made of carbon steel. These are passed through a work coil at an axial speed of 2 m/min. Immediately underneath this coil is a water-spray for quenching. During this process the spindle is rotated in order to neutralize the effect of small asymmetries.

After the treatment the spindles have an outer layer 0.5 mm thick with a hardness of 64 Rockwell, whilst the hardness of the core remains unchanged. Thus the core retains its toughness, and it is due to this that the spindle shows very little warp (owing to inhomogeneity of the base material): it is at most 0.07 mm over a length of 300 mm.

The set-up of *fig. 12* is used for hardening a ring of carbon steel on the inside. The work coil is lowered into the ring from above. The ring lies on a rotating bush, again to neutralize asymmetry. The treatment is carried out with a 20 kW plant and takes no more than a few seconds. The result is a layer of 0.8 mm with a hardness of 60 Rockwell, whilst at 2 mm depth the hardness is 15 Rockwell.

The experimental set-up of *fig. 13* serves for the hardening of endless band-saws, which are passed under the work coil at a rate of 2 m/min. This

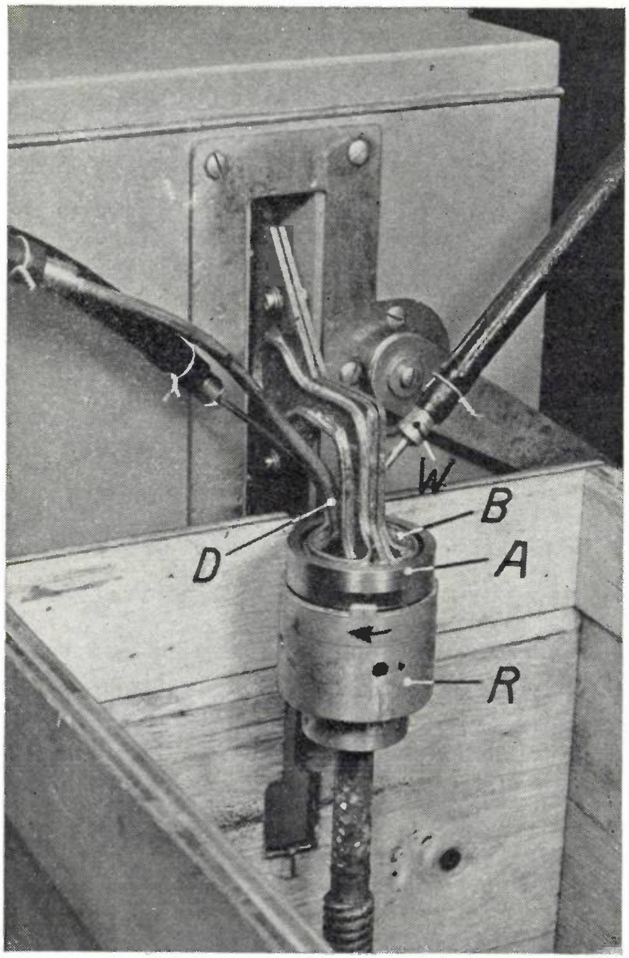


Fig. 12. Hardening the inside of a carbon-steel ring (*A*). The letters *B*, *D*, *R* and *W* have the same meaning as in fig. 11.

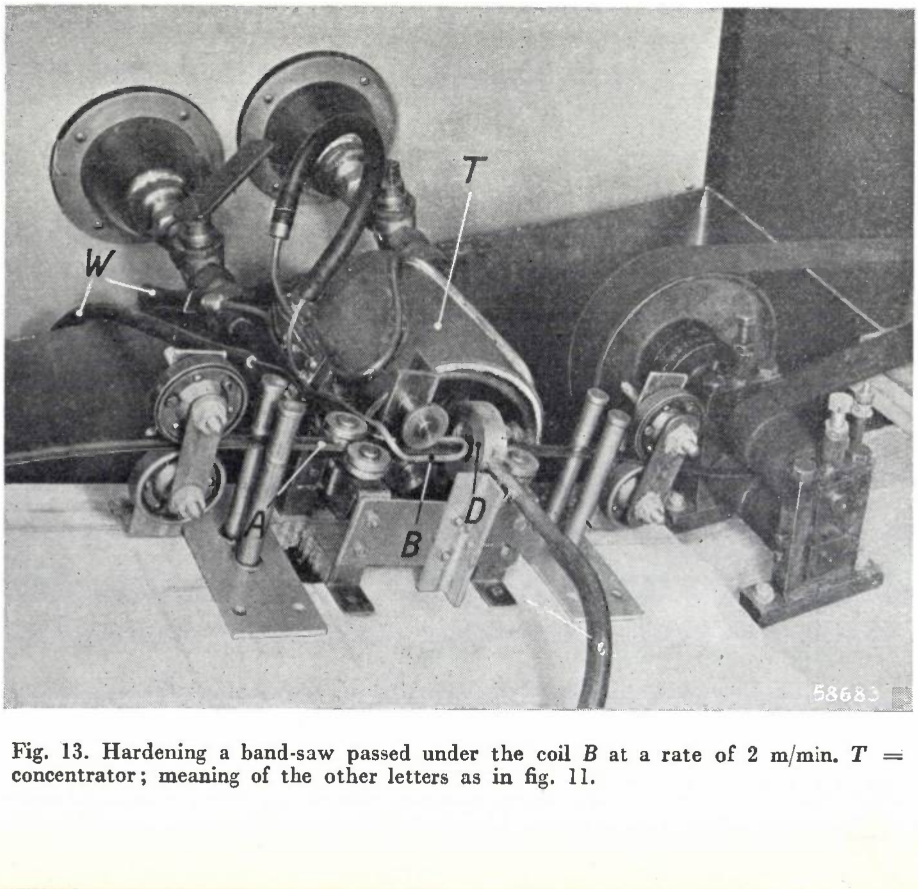


Fig. 13. Hardening a band-saw passed under the coil *B* at a rate of 2 m/min. *T* = concentrator; meaning of the other letters as in fig. 11.

coil is so shaped that both the toothed and the plain sides of the band are heated, to avoid warping.

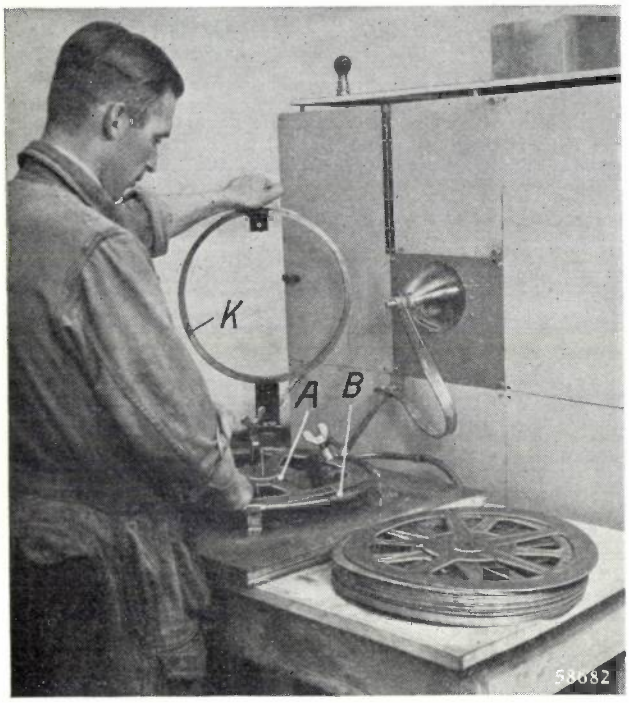


Fig. 14. Removing stresses from the rim of punched discs (flanges of film spools) by annealing. A = disc about to be treated. B = work coil. K = hinged ring for holding the disc.

Along the edge a hardness of 64 Rockwell is reached, at 1.5 mm away from the edge 60 Rockwell, whilst in the middle the hardness of 15 Rockwell remains unchanged.

Annealing

In the case of processes like forcing and deep-drawing, which mostly take place in several steps, often the workpiece becomes too hard and has to be annealed. Particularly when this has to be done locally, the method of induction heating has its advantages. As an example may be mentioned the beading of the rim of reflectors. *Fig. 14* gives another example, where the rim of punched discs (flanges of film spools) is inductively heated for 7 seconds to neutralize the stresses due to the punching.

Brazing and soldering

Induction heating can be carried out so quickly that hardly any oxidation of the workpiece arises. This makes this method highly suitable both for brazing and for soldering, for which purposes it is in fact already being applied on a large scale. Brazed joints requiring several minutes' heating with a flame can now, thanks to the method of induction heating, be made in a few seconds.

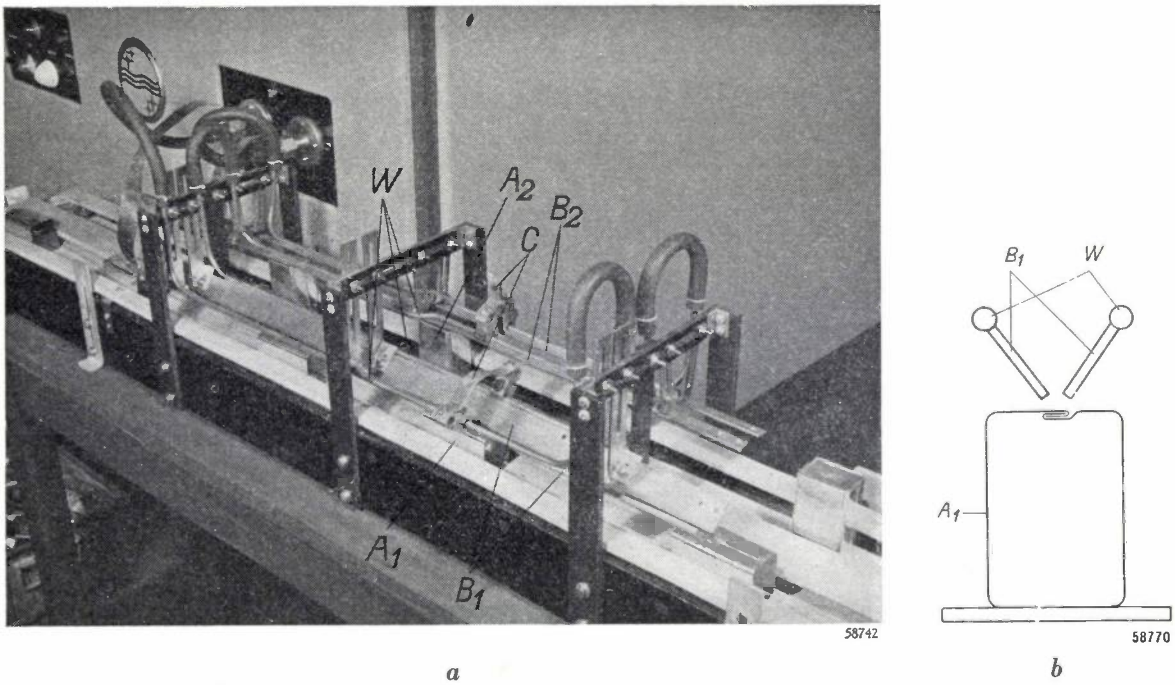


Fig. 15. *a*) Soldering small tins fed underneath the work coil by a moving belt. The tins A_1 are soldered along the longitudinal seam and at A_2 the lid is soldered on. The work coil for A_1 consists of the slanting strips B_1 (see the cross section in sketch *b*), that for A_2 consisting of the horizontal strips B_2. The work coils are cooled with water flowing through a tube (W). The slanting position of B_1 gives a better distribution of current (concentrated at the narrow sides of the strips facing the workpiece). The length of the work coils can be adjusted with movable clamps C.

Another point of importance in a process like brazing or soldering is the already mentioned fact that the forces acting upon the workpiece are small, provided a sufficiently high frequency is used. It is then a very simple matter to place the workpiece inside the coil; often it can be done by hand, or the workpieces can be fed into the heating apparatus by a moving belt without — although they may be light — being displaced by the electrodynamic forces.

Fig. 15 shows a set-up for the soldering of small tins. For particulars see the explanation given in the caption.

Melting

Owing to the short melting times that can be reached with induction heating (e.g. 10 to 15 minutes) this method is excellently suited for the melting of alloys where an exact composition is of importance. With such a short melting time the melt has little opportunity to be affected by the atmosphere and any volatile components have little chance of escaping.

In the article referred to in footnote [1]) a description is given of the plant used at Eindhoven for melting magnet steel and other alloys. The capacity of this plant is 200 kg steel per hour. The transmitting valve (type TA 20/250) has an output of 250 kW and works at a frequency of 5000 to 10,000 c/s. Similar installations have meanwhile been taken into use elsewhere.

These examples should suffice for the present, but, as already stated, the possibilities of induction heating are by no means exhausted and new applications are being sought in all directions. It can therefore be taken as a certainty that more and more use will be made of this new technique in industry.

Summary. A feature which induction heating has in common with capacitive heating is the fact that the heat is generated inside the object, contrary to the case with most of the other methods of heating. A formula is derived, for a simple case, for determining the efficiency of the work coil through which the current passes that induces the heating current in the workpieces. From this formula it appears, inter alia, that in order to work with a good efficiency the frequency has to be chosen sufficiently high to ensure that the penetration depth in the workpiece is not greater than about one eighth of the diameter of the workpiece. This small depth of penetration confines the generation of heat to a thin layer on the surface, a fact which is turned to good account in many applications. It is also shown that with a given power the electrodynamic forces acting upon the workpiece are smaller according as the frequency chosen is higher. After a brief discussion of valve generators for induction heating, some applications are dealt with: surface hardening, annealing, brazing and soldering, and melting.

LOT INSPECTION BY SAMPLING

by H. C. HAMAKER.

620.113.2:658.573

If for some reason the reader should be interested in the number of printer's errors in this journal he would not be likely to go punctiliously through all the pages, but he would take a sample. Sampling procedures are in fact quite common; everyone uses them, often without much thought. Where, however, mass-production processes have to be checked by means of samples, large sums of money may be involved in carrying out the sampling procedure and in utilizing its results. Hence it is of value to obtain an insight into the various factors playing a part in the organization and operation of such lot-by-lot inspection tests.

In the mass-producing factory we are often faced with the problem of testing an inspection lot[1]) of components or of finished articles in order to decide whether they come up to specification and thus may serve their intended purpose. In a great many such cases it is not economically justifiable to inspect every item in the lot, so that one has to resort to the taking of a sample. The question then arises how this sample should be taken, how large its size should be, and what requirements it should satisfy, in order that it can be decided with a reasonable degree of certainy whether the inspection lot as a whole is "good" or "bad". Some fundamental aspects of this problem will be discussed below.

We shall confine our considerations to inspection lots consisting of a number of concrete units, as for example an inspection lot of screws, resistors or incandescent lamps, which are inspected according to "attributes"; that is each item is classified as either "good" or "bad", and it is accordingly "accepted" or "rejected". An item upon inspection found to be defective is called a "reject". The percentage of defective items contained in an inspection lot is called the "percentage defective".

Each inspection lot contains a certain unknown percentage defective, and it is desired to provide from a sample some useful information concerning this percentage. The problem is how this can best be done.

Fundamentals; the operating characteristic

To cast our arguments in a concrete form let us consider a practical example. We shall suppose that we have received from an outside supplier a lot of, say, 100,000 small rivets. A sound riveted joint requires that a rivet fits properly into its hole, and consequently some trouble will be experienced when the rivet is burred. Let us assume that if the lot contains not more than 5% rivets with burs this is not serious, but that a higher percentage causes an undesirable delay in the riveting work. It is therefore specified that to be accepted a lot may not contain more than 5% of burred rivets. Whether this condition is satisfied or not has clearly to be decided on the basis of a sample, because inspecting each rivet separately would require an unwarranted amount of labour.

The simplest procedure is that of "single sampling": we take out of the lot a sample of n_0 rivets which we inspect to see how many of them are burred. The number n_0 is called the "sample size".

If reliable conclusions are to be drawn from the sample concerning the lot as a whole, we have first of all to see to it that the sample is taken with care. It is, for instance, conceivable that the rivets in the lot have come from different machines producing different percentages of burrs. If the lot has been delivered in one sack it may well be that the percentage of burrs near the bottom is greater than in the top layers, so that a sample taken for the sake of convenience from the top alone is not representative of the lot as a whole. It is therefore of essential importance that the sample should be drawn from different levels. The ideal procedure would be to extract one rivet at a time, to shake the sack thoroughly before extracting the next one, and so on until the sample is complete. Mostly, however, this is not practicable and we shall have to approximate our ideal by taking, say, five or ten sub-samples from different spots and combine these to make up our final sample for inspection. In what follows it will always tacitly be assumed that such precautions have been taken.

[1]) The term "inspection lot" or briefly "lot" will be used throughout this article for a collection of items accepted or rejected as a whole on the basis of a sampling plan.

Let us now suppose that in a random sample of 200 rivets 6 have been found to have burrs, that is 3%. In judging the lot on the basis of this datum two lines of argument are open.

Since the sample contains 3% of rejects the lot is obviously not free of defectives. It is also evident that the consignment is not likely to contain more than, say, 20% of defectives, for it would then be highly improbable that our sample contained only 3% of rejects. Thus we are clearly able to form some sort of estimate of the percentage defective in the lot. This percentage will lie somewhere around 3%, but as only a comparatively small sample has been inspected it may not be concluded that this percentage is exactly 3%. Within what limits, then, is it reasonable to suppose the percentage defective to lie and what confidence should we place on such a judgment?

Though mathematical statistics furnish a clear answer to this question, we shall not consider this aspect of our problem in detail, because it does not directly bear on the problems encountered in the factory. There it is desired to lay down clear rules such as the following: take a sample of size n_0 accept the inspection lot when it contains no more than c_0 defectives and reject the lot when this number is exceeded. The figure c_0 is commonly called the "acceptance number". Now for a proper understanding of the meaning and consequences of such a rule it is appropriate to look at the problem from a somewhat different angle than has been done above.

To this end we no longer ask what conclusions can be drawn from the sample regarding the unknown percentage defective in the inspection lot, but instead we enquire into the results that would be reached when an inspection lot with a prescribed percentage defective were repeatedly subjected to a certain "sampling procedure" (for example to a single sampling inspection specified by a sample size n_0 and an acceptance number c_0).

This can easily be settled by experiment. We take, for instance, some thousands of small pieces of cardboard, identical in shape, a small percentage of which has been marked as "defective" and then thoroughly mixed with the rest. From this artificially constructed inspection lot with known percentage defective we then proceed to draw a series of random samples in each of which the number of rejects is determined. The outcome of such an experiment has been recorded in *table I*, where the various columns refer to inspection lots containing different percentages of defectives.

Table I. *a*) The numbers of rejects observed in 10 successive samples taken from the same inspection lot. The figures in the various colums give the results for different percentages defective in this inspection lot. *b*) The result of judging by the sample when 6 is the maximum number of permissible rejects. *c*) The probability of acceptance with the sampling plan considered ($n_0 = 200$, $c_0 = 6$).

Percentage defective	1%	2%	3%	4%	5%	6%
a) Number of rejects in each sample	1 1 1 1 0 3 0 2 1 2	4 9 2 3 5 8 4 6 4 6	6 5 6 7 10 9 7 15 9 4	5 6 13 6 6 8 9 2 9 7	14 14 8 9 10 9 14 11 8 5	8 11 16 6 11 9 14 8 19 14
b) Accepted Rejected	10× 0×	8× 2×	4× 6×	5× 5×	1× 9×	1× 9×
c) Probability of acceptance	0.995	0.889	0.606	0.313	0.130	0.046

Tables I assumes, it is true, that ten samples were drawn from one and the same inspection lot, which is never done in practice. The same results would, however, have been obtained if the ten samples were drawn from ten different inspection lots each of them containing the same percentage defective, and, interpreted in this way, table I corresponds more closely to the conditions prevailing in the factory. In the course of time a large number of inspection lots will be submitted for inspection differing one from the other in the percentage of defectives they contain. Amongst these now and again inspection lots will occur with a percentage defective of 5% and the data recorded in the fifth column of table I may then be considered as the numbers of rejects observed in the successive samples drawn from these specific lots.

If, as we assumed above, inspection lots containing more than 5% of burred rivets are unacceptable, table I shows at once that it would be wrong to prescribe that there should be not more than 5% of rejects in the sample, or in other words that there should be no more than 10 rejects in a sample of 200. From the last column we see that, under such a specification, out of ten lots with 6% defectives four would still have been accepted, the number of rejects being less than 10. Consequently to satisfy our demands the requirements laid down for the sample should be more severe.

Let us therefore see what happens if only 3% of defectives are permitted in the sample, so that among 200 items inspected there should be no more than 6 rejects. The results then obtained have been recorded at the bottom of table I. The ten inspection lots with 1% defectives are all accepted, two of those containing 2% defectives are rejected, and so on; with one exception all the lots with 5 or 6% are rejected, which answers our requirements much better than the previous case.

The vagaries of chance are also manifest: of the inspection lots with 3% defective six are rejected, whereas of those with 4% defective only five are rejected.

The whole procedure is very much like a lottery. For each inspection lot a draw is taken which leads to acceptance if it yields a prize, but leads to rejection if it is a blank. Evidently the chance of drawing a prize diminishes as the percentage defective in the inspection lot increases.

Theoretically these principles are expressed by assigning to each inspection lot a certain "probability of acceptance", P. In accordance with the theory of probability P is so defined that it is a number between 0 and 1. A probability of acceptance of 0.30 signifies that if the corresponding lot were repeatedly subjected to the sampling procedure it would be accepted on the average in 30% of the cases and would be rejected in 70%. Once the sampling plan — i.e. the detailed specification of the sampling inspection procedure — has been fixed the probabilities of acceptance can be computed by application of the theory of probability.

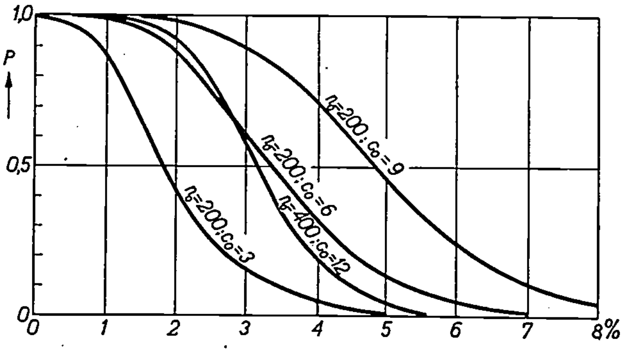


Fig. 1. Operating characteristic of some single sampling plans, i.e. the probability of acceptance plotted as a function of the percentage defective in the inspection lot; n_0 is the sample size, c_0 the acceptance number, or the number of rejects allowed.

For a single sampling plan with a sample size $n_0 = 200$ and an acceptance number $c_0 = 6$ the probabilities of acceptance thus obtained have been entered in the bottom row of table I. By plotting the probability of acceptance as a function of the percentage defective we obtain curves which are called the "operating characteristics"; some examples are represented in *fig. 1*. Each sampling inspection plan possesses its operating characteristics which are invariably curves of the same shape, though they may differ in location and slope.

If, keeping the sample size constant, we increase the acceptance number the operating characteristic shifts towards higher percentages, as would indeed be expected. For a sample size $n_0 = 200$ and an acceptance number $c_0 = 6$, 3% of rejects are permitted in the sample and the operating characteristic is seen to be steepest in the neighbourhood of this value. If we increase the size of the sample, while leaving the percentage of permissible rejects unaltered, the operating characteristic remains at about the same place but assumes a steeper slope, as is evidenced by the curve for $n_0 = 400$, $c_0 = 12$. In the extreme case, if the entire lot were subjected to inspection, we know the percentage defective with precision and we can decide with certainty whether it is higher or lower than the maximum value permitted. The operating characteristic then consists of a horizontal part at $P = 1$ for percentages defective. less than the permitted maximum, and a horizontal line at $P = 0$, for higher percentages, these two horizontal segments being interconnected by a perpendicular located at the permitted maximum.

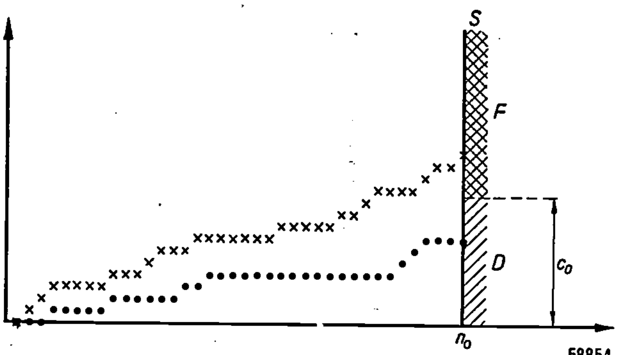
This is of course a purely theoretical case, because inspection of the entire lot would enable us to remove all the rejects it contains, so that we should be left with a lot without defectives. In this connection it should also be mentioned that a similar simplification has been introduced in plotting the curves in fig. 1, it having been assumed that the sample comprises only a small fraction of the inspection lot, so that the number of rejects detected and consequently removed from the sample can be ignored. These, however, are details which do not affect the general trend of our argument.

The purpose of these arguments was merely to show that the operating characteristic usefully portrays the practical performances of a sampling inspection plan. Hence these characteristics will play an important part in our further considerations.

Different sampling plans

So far we have spoken of sampling plans as the concrete specifications for taking a sample and acting upon the result of its inspection; we have only considered single sampling plans characterised by a sample size n_0 and an acceptance number c_0. All simple sampling plans will henceforth collectively be designated as the "single sampling system".

Apart from the single sampling system other systems have been developed and successfully applied, the most important being the "double" and the "sequential" sampling systems. These possess the advantage of requiring on the average fewer observations than the single system, which is, however, partly offset by their being somewhat more complicated. Later on we shall return to the problem of choosing between these different methods, confining ourselves for the moment to explaining the principle on which they are based. This can best be achieved by means of a simple graphic representation.



58854

Fig. 2. The "random walk" diagram. The number of items inspected is plotted along the abscissa, the number of rejects observed among them along the ordinate. For a particular inspection lot the result of inspection is represented by the path indicated by the dots, for another lot by the crosses. The procedure of single sampling is illustrated by the screen S, erected in the point n_0. When the random walk ends in segment D the lot is accepted, when it ends in F the lot is rejected, c_0 being the acceptance number.

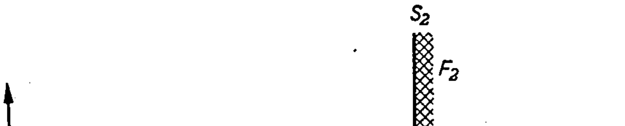
Along the abscissa of a set of rectangular axes we plot the total number of items inspected and along the ordinate the number of rejects found among them, see *fig. 2*. Starting from the origin we go in this diagram one step to the right for every good item inspected, and one step to the right and one step upward for every defective item. While inspection is going on we thus obtain step by step a detailed picture of the results obtained, as exemplified by the dots and crosses in fig. 2; such a path is called a "random walk", the entire graph being denoted as the "random walk diagram".

When several samples taken from the same lot are inspected the corresponding random walks will not, of course, be identical, but on the average they will run in the same direction. On the other hand, the greater the percentage defective of an inspection lot, the greater will be the average number of rejects in a sample drawn from it, and consequently the steeper the average slope of the corresponding random walks; thus the crosses in fig. 2 refer to a lot with a higher percentage defective than the dots.

This random walk diagram lends itself to a simple and easy illustration of the different sampling systems. Single sampling, for instance, is represented by erecting in the point n_0 on the abscissa a perpendicular divided into two segments D and F according to the acceptance number c_0. After inspecting a sample of size n_0 the random walk ends in some point on this perpendicular; if this point lies in D the lot is accepted, if in F it is rejected.

From single sampling we now pass on to "double sampling" by using two screens, S_1 and S_2, as in *fig. 3*, the first with an opening E in it. We now perform a sort of pre-selection. We start by taking a first sample of size n_1, which brings us to some point on S_1; if our random walk ends in D_1 the number of rejects is so low that the lot may at once be accepted, and if we end in F_1 the number of rejects is so high that the lot may be rejected. But if after the first sample we end somewhere in the opening E the case is considered as a doubtful one, so that it is desirable to collect some further data before deciding; we then take a second sample n_2 and base our final decision upon the total number of rejects in both samples together, according to the segments D_2 and F_2 of the second screen S_2.

Compared with single sampling, double sampling offers two advantages. Firstly many of us are inclined to give an inspection lot a second chance when the results of a first sample have not been entirely satisfactory, and this natural tendency is met by the double sampling principle. Secondly, decidedly good or decidedly bad lots are already brought to light by the first sample and this leads to a saving in the average number of observations. Against this we must set the drawback of a more



58855

Fig. 3. The principle of double sampling depicted by two screens S_1 and S_2. S_1, corresponding to a sample size n_1, affords a pre-selection. Only those inspection lots for which the random walk after the first sample ends in the opening E, fixed by two criteria c_1 and c_2, are subjected to a further scrutiny, the result being considered as doubtful. This is done by taking a second sample of size n_2, and judging according to the screen S_2 and the criterion c_3.

complicated specification manifest from the larger number of data required: two sample sizes, n_1 and n_2, and three criteria c_1, c_2 and c_3 (see fig. 3).

Obviously by the same principles we might go farther still and sucessively develop three-fold, four-fold and in general m-fold sampling systems. Ultimately this leads automatically to the sequential sampling system which is represented by two parallel boundary lines dividing the random walk diagram into three regions D, E and F (*fig. 4*).

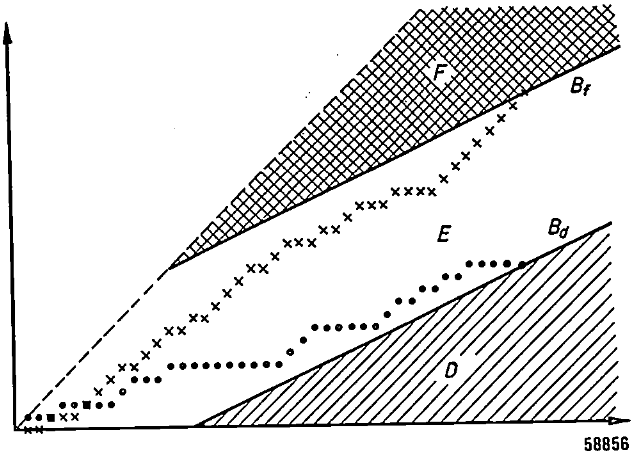


Fig. 4. The method of sequential sampling. When the random walk crosses the line B_d the lot is accepted, but when it crosses B_f the lot is rejected. (Since the number of rejects observed can never be greater than the number of units inspected, the random walk can never cross the line passing through the origin at an angle of 45°. This applies equally to figs 2 and 3.)

The regions D and F can be conceived as being the sum of an infinite number of screens such as S_1 in fig. 3. Inspection is continued as long as the random walks lie completely inside region E, but is stopped as soon as one of the two boundary lines B_d or B_f is crossed. If it is B_d the lot is accepted, if B_f it is rejected. When using the sequential method inspection is stopped as soon as sufficient data have been collected to reach a decision. This is the fundamental view from which sequential sampling has been developed, and from our above arguments it follows logically that by sequential sampling a still greater saving in the average number of observations can be obtained than with double or multiple sampling.

The denomination "sequential" expresses that inspection is carried out sequentially, that is step by step; after each step it is decided whether the lot can be accepted or rejected, or whether we have to go on with inspecting another item. Another characteristic of this method is that the number of items to be examined is undetermined at the outset and depends on the results obtained while inspection is proceeding [2]).

In the foregoing the operating characteristic has been explained with reference to the single sampling system, but these characteristics can just as well be constructed for double and sequential systems, and this always produces curves of the same shape as those of fig. 1. It is possible to devise single, double and sequential sampling plans the characteristics of which are practically coincident. Such systems will on the average yield the same results as regards acceptance or rejection; they may for that reason be said to be "equivalent" in practical performance, but this equivalency does not hold for the average size of the sample. On the contrary our above arguments lead us to expect that a double plan will require a smaller number of observations than its equivalent single sampling plan, this number being still lower in the case of sequential sampling. Thus a systematic comparison of equivalent sampling plans will enable us to decide to what extent the double or sequential system is of advantage; this principle will be worked out in detail in a subsequent article.

One might perhaps be inclined to conclude that sequential sampling, requiring the smallest number of observations, is always to be preferred, but this is by no means the case. Sequential sampling is less simple from an administrative point of view, and when the observations are relatively cheap this is a drawback that may well outweigh the advantage of a smaller number of data required. Moreover, in choosing a sampling plan most suitable for a practical problem many other factors play a part, as will now be discussed.

The practical choice of a sampling plan

In situations where lot-by-lot inspection may be practicable we can usually distinguish a "producer" and "consumer"; the consumer may be another factory, another department in the same works using the products delivered as components of a more complicated assembly, or a shopkeeper retailing to the public.

Both producer and consumer will lay down their requirements: the former demands that not too many "good" lots shall be rejected by the sampling inspection, while the latter demands

[2]) Double sampling was systematically applied for the first time by H.F. Dodge and H. G. Romig, whilst sequential sampling was developed during the last war by A. Wald in the U.S.A. and by G. A. Barnard in England. See H. F. Dodge and H. G. Romig, Single and double sampling inspection tables, Bell system techn. J. 20, 1-61, 1941; A. Wald, Sequential analysis, Wiley and Son, New York 1947; Ann. Math. Stat. 16, 117-186, 1945. G. A. Barnard, J. Roy. Stat. Soc. Suppl. 8, 1-27, 1946.

that not too many "bad" lots shall be accepted. In choosing a sampling plan attempts will be made to meet these somewhat opposing requirements.

Under favourable circumstances it may happen that all the lots produced are acceptable to the consumer, none being "bad", whereas under unfavourable conditions they may all happen to be "bad". In such cases lot-by-lot inspection is of no avail; the lots can either be delivered without any inspection at all, or they must all be examined for the full 100%. Consequently lot-by-lot inspection may be particularly useful where among a large number of good inspection lots we have reason to expect the occasional occurrence of a bad one. We are then in a position, by means of a sample, to discriminate good from bad and to pick out and correct the majority of the bad inspection lots in time.

Looked at in this way the results that can be achieved evidently depend on the frequency with which "bad" inspection lots occur; if no more than 1 in a 100 of the lots is bad the usefulness of lot-by-lot inspection will probably be less than when this ratio is 1 in 10. Another influencing factor is the difference between the "good" and the "bad" inspection lots. If this difference is relatively large it is comparatively easy to distinguish good from bad, but if the difference is only slight a satisfactory discrimination may require a sampling plan with a very steep operating characteristic needing very large samples. These are factors largely beyond our control, subject to changes owing to the continuous introduction of improvements in the production process, and which cannot easily be determined in a concrete way, but which are nevertheless of decisive importance in determining the proper choice of a sampling inspection procedure. As a rule we have to rely upon our "experience".

The cost of inspection is another point to be considered. If we are checking a diameter by means of a gauge or visually inspecting whether a product is burred or not, the cost of an observation is comparatively low and the size of a sample is of small concern. If, on the other hand, inspection is "destructive", as in testing a breakdown voltage or mechanical strength, every unit inspected is a loss and the number of items inspected must of necessity be limited; in such cases the use of double or sequential sampling may be advisable. Yet another situation is encountered in life tests where the use of a small sample is obviously desirable, but where sequential sampling must be ruled out since the inspection of one unit after another would take far too much time.

Finally it must be borne in mind that the object of detecting and correcting "bad" inspection lots is to avoid losses or damage that might result if they were passed on undetected. In this respect, too, practical circumstances may greatly diverge. Resistors or capacitors of inferior electrical quality lead to the production of badly functioning radio sets, and the tracing and correcting of the faulty elements involves considerable expense; only a very small percentage of such defects can be tolerated. But if the same components should lack, for example, one of the two terminal leads, this would be immediately noticed in the assembly shop and little harm is done. Similar considerations apply to products delivered to retailers. If among a consignment of incandescent lamps sold to the public the bulbs or a certain percentage of them are not in proper alignment with the base this will hardly be noticed, but if a high percentage should have a sub-normal life this will sooner or later be noticed and the customers will go elsewhere for their future supplies.

All in all we see that a number of economic factors play their part in the choice of a suitable sampling plan, factors that may differ widely from case to case, and the importance of which cannot be expressed in exact numerical values and which we must consequently evaluate on the basis of our practical experience.

From these considerations some important conclusions may be drawn as to the choice of a sampling plan. The introduction of a lot-by-lot inspection procedure can be regarded as a kind of insurance. The cost of the regularly recurring inspection constitutes the premium paid to cover ourselves against the losses resulting from failure to detect inferior products in time. But, as pointed out, we usually lack the precise data needed for an accurate evaluation of the premium to be paid, that is of the size of the sample. This at once implies that the finding of a suitable sampling plan is largely a question of experimental investigation. At the outset we have to go by the complaints concerning the quality of the goods supplied. Taking into account the consumer's requirements and the quality the producer claims as reasonably producable, we then make a preliminary choice in which we may be guided by our knowledge of the operating characteristics corresponding to various sampling plans. Practice will then have to show whether a wise choice has been made, this being judged by the degree in which complaints diminish or cease altogether. If such is not the case the sampling plan must be modified. The experience gained in the foregoing experimental stage usually indicates the direction of the change required.

From the impossibility of determining the various economic factors with accuracy it also follows that the conditions to be satisfied in practice are not very stringent. If, for instance, a sampling plan possessing one of the operating characteristics of fig. 1 proves to be satisfactory, another plan with a characteristic deviating somewhat to the right or to the left, or being somewhat steeper or flatter, will be equally convenient. Theoretically there would, of course, be certain differences, but these will only become manifest from carefully kept records regarding the frequency and the nature of the complaints that still occur from time to time.

Taking a broad view of our problem we can now distinguish two different problems. First we have to specify the operating characteristic that is desired, then we have to construct a sampling plan satisfying this demand. These two problems cannot be completely separated one from the other, because in selecting a characteristic we have to take into account the number of observations required to obtain the curve selected. Nevertheless it is convenient to proceed in two stages as indicated.

Since no great precision is needed it has also been found sufficient to specify an operating characteristic by two numerical data (parameters), which simplifies our two problems to:

a) choosing the value of these parameters, and
b) finding a sampling plan corresponding to the values chosen.

In following this procedure it will be of great importance to adopt a set of parameters meeting the practical needs of the factory as far as possible. In this respect uniformity has not been reached, different sets of parameters having been introduced in the various sampling inspection tables so far developed. These sets of parameters and methods of constructing sampling plans when their values have been prescribed will form the subject of a subsequent article.

Summary. The quality of an inspection lot of mass-produced articles may often effectively be judged by means of a sample, the simplest method being that of single sampling characterized by a sample size n_0 and an acceptance number c_0 (number of rejects allowed in the sample). By plotting the probability of acceptance as a function of the percentage defective in the inspection lot the "operating characteristic" is obtained which is shown to represent the practical performance of a sampling method in a satisfactory manner. The principles of double and sequential sampling are next discussed, procedures which, though more complicated, possess the advantage of requiring a smaller number of observations than single sampling. It is indicated how the relative merits of these methods can be evaluated. Finally the factors, mainly economic, determining our choice of a sampling plan in particular cases are considered. For example the harm caused by "bad" lots not detected in time plays an important part. The choosing of a sampling plan is divided into two distinct stages; (a) the choice of an operating characteristic, and (b) the construction of a sampling plan possessing this characteristic. The latter problem, which is of a more purely mathematical nature, will form the main subject of a second paper.

THE FERRO-ELECTRICITY OF TITANATES

by G. H. JONKER and J. H. VAN SANTEN.

621.319.412.4

Some chemical compounds, among which Rochelle salt (potassium sodium tartrate) has been known the longest, when used as a dielectric between the plates of a capacitor show a remarkable behaviour similar to that of ferromagnetic materials. It has recently been found that certain titanates, of which barium titanate is an example, show this "ferro-electric" property. The behaviour of these latter substances can be explained to a certain extent in theory. Owing to the high values of their permittivities these titanates are employed in the construction of small capacitors of relatively large capacitance.

Introduction

Ferromagnetic materials such as iron, iron alloys and nickel, etc. are known to behave in the following way. At temperatures below a certain critical temperature T_c (the Curie temperature) these substances show a spontaneous magnetization. As the temperature rises this magnetization is reduced, first slowly and then gradually quicker until it becomes zero when $T = T_c$.

A given piece of ferromagnetic material usually contains a large number of small domains (Weiss domains) each with its own spontaneous magnetization. Since this magnetization may be directed in all sorts of ways, in the absence of an external magnetic field the total magnetization is as a rule zero or at least much smaller than what corresponds to the spontaneous magnetization of a single domain.

When an external magnetic field H is applied a reorientation of the spontaneous magnetization takes place in the various domains, as a consequence of which, even in a relatively weak field, the material as a whole suddenly shows a rather strong magnetization and thus assumes a considerable magnetic moment J per unit of volume. The relation between this magnetic moment J and the external field H is generally non-linear and not well defined. Starting from an originally non-magnetic sample of ferromagnetic material, as H increases from zero to high values there is usually at first a sharp increase and later on a less sharp increase of J until finally J reaches its ultimate saturation value J_s. As the value of H is reduced to zero again, J is also reduced, but the values of J with decreasing H are generally higher than those found with increasing value of H, so that when H has dropped to zero J shows a certain residual value (remanence) J_r (*fig. 1a*).

If H is made to oscillate periodically between a high positive and a high negative value then the point representing the state of the material in an H-J diagram describes a so-called hysteresis loop (fig. 1*b*) [1]). On traversing the hysteresis loop once a conversion of energy into heat takes place in the material equal to $\oint H\mathrm{d}J$ per unit of volume (hysteresis losses).

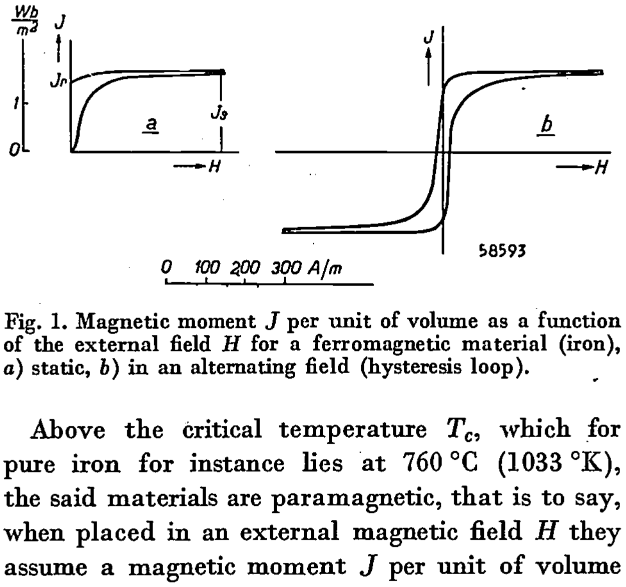


Fig. 1. Magnetic moment J per unit of volume as a function of the external field H for a ferromagnetic material (iron), *a*) static, *b*) in an alternating field (hysteresis loop).

Above the critical temperature T_c, which for pure iron for instance lies at 760 °C (1033 °K), the said materials are paramagnetic, that is to say, when placed in an external magnetic field H they assume a magnetic moment J per unit of volume proportional to H:

$$J = \varkappa H \,, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

in which [2])

$$\varkappa = (\mu_r - 1)\,\mu_0 \,. \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

The factor μ_r depends upon the temperature T

[1]) See, for instance, J. J. Went, Philips Techn. Rev. **10**, 246-254, 1948, No. 8.

[2]) For the sake of simplicity the material is assumed to be in the shape of a ring wound with wire (toroid). In such a form no complications arise owing to "demagnetizing forces". μ_r is the relative permeability with respect to vacuum; $\mu_0 = 4\pi/10^7$ henry/m. In this article the formulae are written according to the rationalized Giorgi system of electrical units with absolute volt and ampere. For further particulars about this system see P. Cornelius, Philips Techn. Rev. **10**, 79-86, 1948, No. 3 and Philips Res. Rep. **4**, 232-237, 1949, No. 3.

according to the equation:

$$\mu_r - 1 = \frac{A}{T - T_c}, \; (T > T_c) \; . \; . \; . \; . \quad (3)$$

in which A represents a constant. It can be seen that μ_r increases sharply as T approaches the temperature T_c, so that T_c may indeed be regarded as a critical temperature.

Now a certain analogy exists between a magnetic material placed in an external magnetic field and a dielectric material placed in an external electric field.

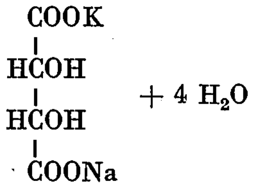
In a flat capacitor, where the space between the plates is filled with a dielectric, there is usually a proportionality between the electric moment P per unit of volume and the field strength E, viz.

$$P = \chi E \, , \; . \; . \; . \; . \; . \; . \; . \; . \quad (4)$$

in which [3])

$$\chi = (\varepsilon_r - 1)\, \varepsilon_0 \, . \; . \; . \; . \; . \; . \; . \; . \quad (5)$$

Similar to the magnetic case, a medium which behaves according to (4), where χ denotes a constant independent of the field strength, could be said to be "para-electric". As a rule χ will be a function of the temperature. For most materials χ changes but little with T, but some are known where χ depends very strongly upon the temperature and shows at a certain temperature ϑ a high peak value. Some of these substances show at temperatures lower than ϑ a behaviour similar to that of ferromagnetic materials, in that P and E bear no longer a linear relation to each other, whilst at the same time hysteresis occurs. Such materials are said to have an "electric Curie point" and are generally called ferro-electric. The material first (in 1918) found to possess this property is potassium sodium tartrate or Rochelle salt:



This substance has a Curie point at $T = 300\,°\mathrm{K}$ (27 °C), below which hysteresis occurs. Other materials showing a similar abnormal behaviour are KH_2PO_4 and allied compounds such as KH_2AsO_4, RbH_2PO_4, etc. In the case of KH_2PO_4 the Curie point lies at 124 °K. The hysteresis (P-E) diagrams for Rochelle salt and for KH_2PO_4 are given in *figs 2* and *3* for a number of different temperatures.

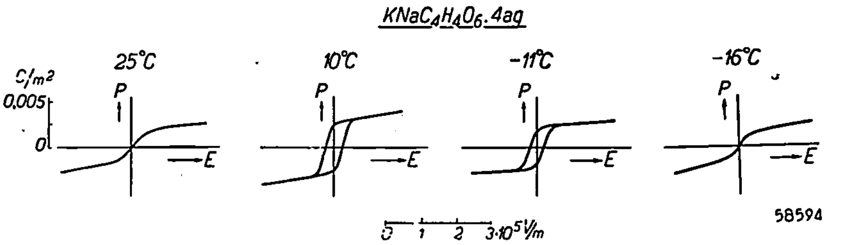


Fig. 2. The relation between P and E for Rochelle salt at different temperatures (taken from J. Hablützel, Helv. Phys. Acta 12, 489, 1939).

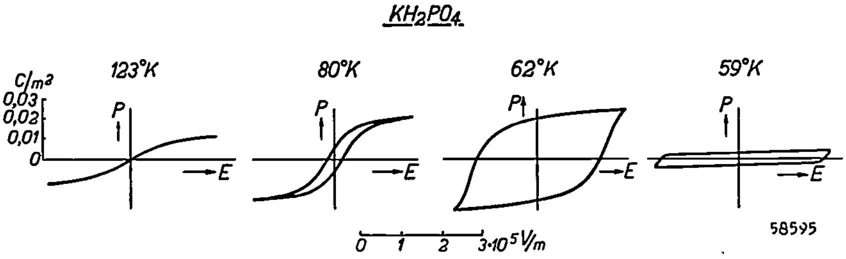


Fig. 3. The relation between P and E for KH_2PO_4 at different temperatures (taken from G. Busch and E. Ganz, Helv. Phys. Acta 15, 501, 1942).

[3]) ε_r is the so-called relative permittivity of the medium. For a vacuum the value of this constant is 1 and for other media mostly >1. The value of ε_0 is $10^7/4\pi c^2 = 8.85 \cdot 10^{-12}$ farad/m. The relation between P, E and the charge per unit of surface (electrical flux density) D is given by $D = \varepsilon_0 E + P$.

Fig. 2 shows that in the case of Rochelle salt at lower temperatures the non-linear relation between P and E and the hysteresis loop begin to disappear. It is said, therefore, that this material has not only the common or "upper" Curie point (27 °C) but also a "lower" Curie point (—19.5 °C). Something similar occurs with KH_2PO_4, see fig. 3, but here the case is somewhat different. At 58 °K the non-linear behaviour disappears but in this case the straight line indicating the relation between P and E has to be regarded as a hysteresis loop flattened in the vertical sense. We shall not, however, enter into these details, but will confine ourselves exclusively to the common or upper Curie point.

The causes of ferromagnetism and of ferro-electricity

The theory of magnetism is based on the conception that the electrons, owing to their describing orbits in the atoms and also rotating around their axis ("spinning"), are elementary magnetic dipoles. In the case of paramagnetism these elementary magnets (for the sake of simplicity denoted as spin vectors) are orientated by an external field, whilst the thermal movement tends to disturb this order. The spontaneous magnetization in the case of ferromagnetism is brought about by a gain in energy obtained when all the spin vectors are directed parallel to each other. This is not due to the magnetic interaction between the elementary magnets (it is too small for that) but arises from an additional gain in energy which can only be accounted for by quantum mechanics. According to this theory, which we cannot go into here, it can be understood more or less why it is that this additional interaction is sufficient to lead to ferromagnetism in the case of some materials and not so with others.

The behaviour of Rochelle salt must likewise be ascribed to dipoles present in the material, though in this case these are of an electrical nature. These electrical dipoles arise not from the properties of the electrons but from the arrangement of the ions in the crystal. The crystal lattice of Rochelle salt contains H^+-ions placed asymmetrically between negative oxygen ions. This results in the presence of electric dipoles. Now these dipoles have the tendency to orientate each other and in this case the energy of the dipole forces is indeed sufficient to account for this orientating effect. Thus it is that the spontaneous polarization in this case is of a nature different from the magnetization in the magnetic case.

Barium titanate and allied compounds

We shall now discuss a group of compounds whose behaviour corresponds in many respects to that of Rochelle salt but is easier to discuss because these compounds have a very simple (cubic) crystal structure. In the years 1940-1945 it was found [4]) that compounds of TiO_2 with oxides of bivalent metals, for instance $BaTiO_3$ and $PbTiO_3$ and mixed crystals of these, show as a function of the temperature a high peak in the value of ε_r, whilst other similar compounds such as $SrTiO_3$ and $CaTiO_3$, as well as TiO_2, show a sharp increase of ε_r with decreasing temperature as far as the absolute zero point [5]).

The titanates referred to above are prepared by ceramic methods, mixtures of TiO_2 and, for instance, $BaCO_3$ being ground together, moulded to the desired shape and fired at 1300-1400 °C. X-ray crystallographic examination shows that a complete reaction takes place. The preparation of $PbTiO_3$ is somewhat more difficult because in the process of sintering some of the PbO evaporates. It is due to the great stability of these ceramic mateiials that use can be made of their very high permittivities for practical purposes, whereas this is not easily possible with Rochelle salt and allied materials. Something more will be said about this at the end of this paper.

The permittivity of titanates as a function of temperature

The behaviour of the permittivity of $BaTiO_3$ as a function of T is rather complex. As is to be seen from *fig. 4*, ε_r for $BaTiO_3$ has a very high maximum ($\varepsilon_r \approx 10{,}000$) at $T = 396$ °K (123 °C) and two smaller maxima in the neighbourhood of

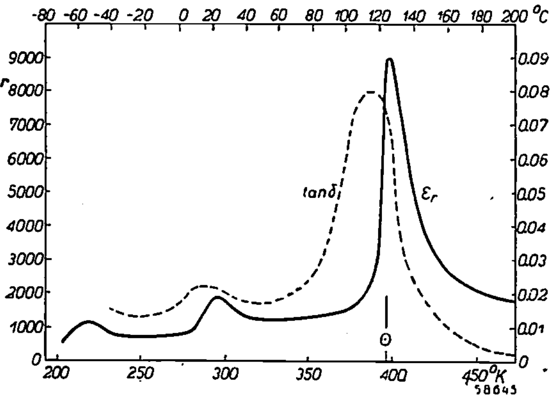


Fig. 4. The permittivity ε_r and the power factor tan δ of $BaTiO_3$ as functions of the temperature, measured at a low field strength (1000 V/m).

[4]) B. Wul, Dielectric constant of some titanates, Nature **156**, 480, 1945; Willis Jackson and W. Reddish, High permittivity crystalline aggregates, Nature **156**, 717, 1945; E. Wainer, High titania dielectrics, Trans. Electrochem. Soc. **89**, 331-356, 1946; A. von Hippel. R. G. Breckenridge, F. G. Chesley and L. Tisza, High dielectric constant ceramics, Indust. Eng. Chem. **38**, 1097-1109, 1946.

[5]) E. J. W. Verwey and R. D. Bügel, Philips Techn. Rev. **10**, 231-238, 1948, No. 8.

$T = 10$ °C and -70 °C. The main maximum is greatly affected, as far as its height and width are concerned, by the degree of sintering and the purity of the material. The temperature ϑ of the main maximum forms the boundary between two temperature regions in which the electrical and other properties are very different.

In the range $T > \vartheta$ the temperature coefficient of the permittivity is negative. In this range ε_r can be represented as a function of T by the simple formula

$$\varepsilon_r = \frac{A'}{T - \vartheta}, \quad \dots \dots \dots \quad (6)$$

in which A' is a constant. This formula (6) is entirely analogous to formula (3) which applies for the magnetic case. (The fact that in formula (6) we have ε_r instead of $\varepsilon_r - 1$ is of no importance because we are here considering dielectrics of which $\varepsilon_r \gg 1$.)

The other titanates and TiO_2 show the same behaviour in a certain range of temperature; the value of the constant A' is always about 1.0×10^5, but the value of ϑ differs from case to case, being 396 °K for $BaTiO_3$, 800 °K for $PbTiO_3$ and about 0 °K for $SrTiO_3$. A similar formula applies for $CaTiO_3$ and TiO_2 but with a negative value of ϑ. Since T is always positive, throughout the whole temperature range ($T >$ 0 °K) $SrTiO_3$, $CaTiO_3$ and TiO_2 therefore behave in the same way as $BaTiO_3$ above 396 °K.

All the above titanates have the same crystal structure. They can be used for producing mixed crystals in all proportions by sintering suitable mixtures together. An examination for instance of mixed crystals of $BaTiO_3$ and $SrTiO_3$ shows that a practically linear relation exists between the mixing proportions and the peak temperature ϑ (*fig. 5*).

Thus by starting with mixtures of $BaTiO_3$ and $SrTiO_3$ the peak temperature can be made to lie anywhere between 396 ° K and 0 °K, so that it is possible, for instance, to study and apply the properties in the neighbourhood of $T = \vartheta$ at room temperature.

Formula (6) holds equally well for the mixed crystals. The constant A' has practically the same value also in this case.

Obviously, near the maximum of ε_r ($T = \vartheta$) formula (6) no longer holds because it would lead to the value $\varepsilon_r = \infty$. But within the range from a few degrees above $T = \vartheta$ up to the highest temperatures to be considered formula (6) does indeed hold good. Provided that T is chosen sufficiently high above ϑ then in the range of $T > \vartheta$ the dielectric losses are fairly small. The temperature coefficient of the dielectric constant is more strongly negative as T approaches ϑ, and there the losses are greatly increased.

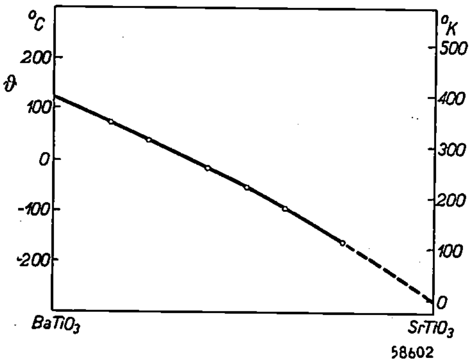


Fig. 5. Peak temperature ϑ as function of the proportions of $BaTiO_3$ and $SrTiO_3$ in mixed crystals.

Electrical behaviour for $T < \vartheta$.

For the titanates having a positive value of ϑ and thus a permittivity showing a high maximum (see fig. 4) the dielectric properties in the range $T < \vartheta$ are entirely different from those in the range $T > \vartheta$. The temperature coefficient in the range $T < \vartheta$ may be positive or negative, and, taken absolutely, generally smaller than in the range $T > \vartheta$. In the range $T < \vartheta$ the losses are considerable. This is connected with the fact that there appears to be a non-linear relation between P and E and hysteresis is found to occur. Because of these facts, in this temperature range these substances strongly resemble ferromagnetic materials and Rochelle salt.

When the material is used as a dielectric between the plates of a capacitor under the conditions mentioned above the charge Q of the capacitor is no longer proportional to the potential difference V between the electrodes. Instead of a simple

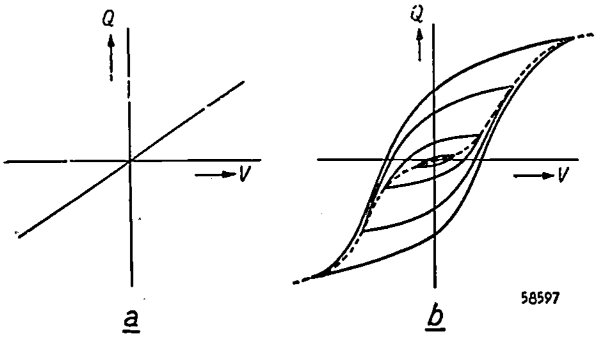


Fig. 6. The charge Q of a capacitor with $BaTiO_3$ as dielectric, as a function of the potential difference V, *a*) in the range $T > \vartheta$, *b*) in the range $T < \vartheta$. The dotted line represents (schematically) the initial curve.

Q-V figure [6]) as represented in *fig. 6a* we then get a hysteresis loop (fig. 6*b*).

It is not easy to determine the hysteresis figure statically because the charges arising, in so far as they are present at the crystal surface, mostly leak away quickly. Measurements are therefore usually taken with the aid of alternating voltages of different amplitudes. The initial curve can be determined by assuming that the ends of all hysteresis loops lie on this curve (see the dotted line in fig. 6*b*). It is most easily determined by the method to be mentioned farther on, where use is made of a cathode-ray oscillograph.

The initial curve of a titanate is drawn separately in *fig. 7*. The initial slope is of the same order as the slope found with the highest attainable voltages. The charge and thus the polarization of the dielectric can be regarded as consisting of a part linearly dependent upon the field strength and another part which becomes saturated at a certain field strength.

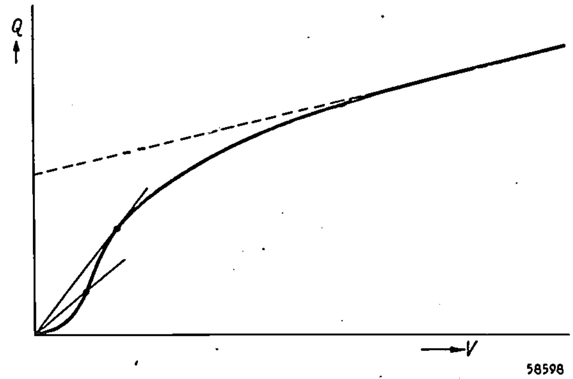


Fig. 7. The initial curve of a titanate (schematically).

As a result of the contribution yielded by the part that becomes saturated, the effective permittivity, determined by the ratio of the amplitudes of the charge and voltage, first rises sharply with increasing amplitude of the alternating voltage and then gradually drops until the ultimate value is of the same order as the initial value (*fig. 8*).

In contrast to the effective permittivity the reversible permittivity (slope of a small hysteresis loop starting from an arbitrary working point) is relatively independent upon the circumstances. It is approximately equal to the initial value of the permittivity (slope of a small hysteresis loop described around the origin, $Q = 0$, $V = 0$).

As is already to be deduced from fig. 6*b*, in the range $T < \vartheta$ the ε_r-T curve of fig. 4 is dependent upon the alternating voltage used. This curve represents the initial value and applies for a small field strength (< 1000 V/m). When measurements are carried out at higher voltages we get the effective permittivity and thus the part of the curve in the range $T < \vartheta$ changes considerably (*fig. 9*), whilst also the losses are greatly increased.

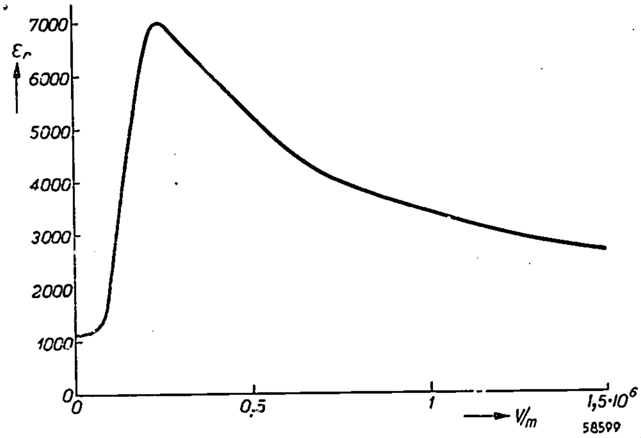


Fig. 8. The effective permittivity of a titanate mixed crystal ((87.5 % Ba, 12.5 % Sr) TiO_3, $T = 20$ °C) as a function of the amplitude of the alternating voltage.

From this it follows that at lower temperatures a higher field strength is needed to reach the maximum ε_r. This is to be seen also from *fig. 10*, where the initial curve has been drawn for two temperatures.

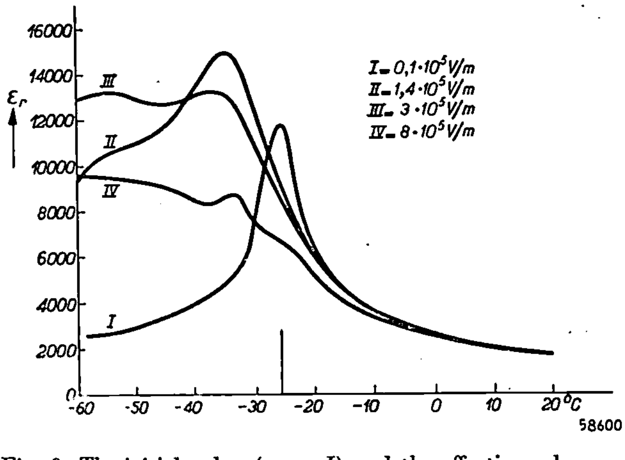


Fig. 9. The initial value (curve I) and the effective value ε_{eff} of the permittivity for different amplitudes of the alternating voltage as functions of T (for the mixed crystal (55 % Ba, 45 % Sr) TiO_3).

Method of investigating the dielectric behaviour

This investigation is made largely with the aid of a cathode-ray oscillograph, by means of which pictures can be obtained of the hysteresis loop. As regards the circuit reference is made to *fig. 11* and the explanation given in the caption.

[6]) The potential difference V is proportional to E and the charge Q is proportional to D, thus to ε_r. P is proportional to $\varepsilon_r - 1$ (see footnote [3])) but in our case invariably $\varepsilon_r \gg 1$. Therefore the Q-V diagram is practically equal to the P-E diagram.

The capacitor C_1 (fig. 11) containing the material to be investigated is placed in a bath of paraffin that can be heated to different temperatures. In this way it can be clearly seen how the hysteresis loop changes into a linear characteristic at $T = \vartheta$. Photographic recordings of this are given in *fig. 12*. In *fig. 13*, in addition to a voltage of 50 c/s also a small voltage with a frequency which is a multiple of 50 c/s (in the case in question 1750 c/s) is applied to the plates of the capacitor, making the reversible permittivity visible. At the point of intersection of the axes there is a loop with a very small amplitude, the slope of which corresponds to the initial value of the permittivity.

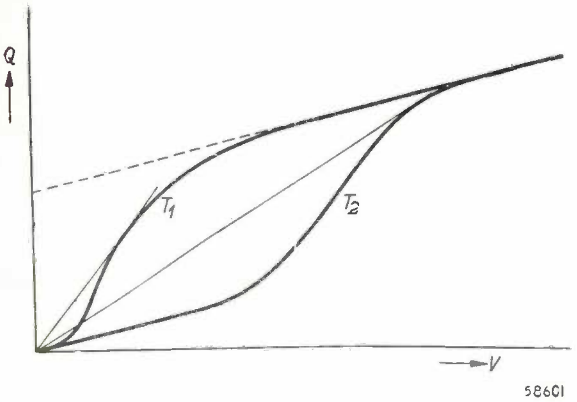


Fig. 10. The initial curve of a titanate (schematically) at two temperatures ($T_1 > T_2$).

Fig. 13 clearly shows that there is little difference between the slope of the initial loop, that of the reversible loops and the final slope of the large hysteresis loop.

Explanation of the exceptional behaviour of titanates

What are the causes of the remarkable behaviour of the dielectric constant of titanates? Let us first consider the region $T > \vartheta$. According to Clausius-Mosotti and Lorentz:

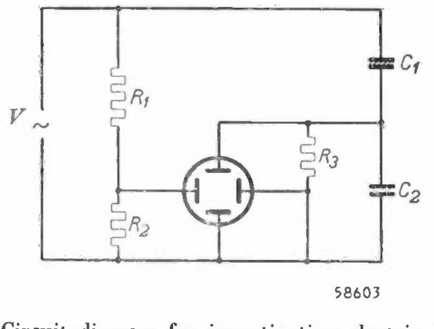


Fig. 11. Circuit diagram for investigating electric hysteresis (C. B. Sawyer and C. H. Tower, Phys. Rev. **35**, 269, 1939). The horizontal deflection voltage is directly proportional to the total voltage V and thus practically also proportional to the voltage at the capacitor C_1 with the substance under examination, since $C_2 \gg C_1$ (C_2 is an auxiliary capacitor with normal dielectric). The vertical deflection voltage V_2 is proportional to the charge of C_2 and thus also proportional to the charge of C_1, since $Q_2 = Q_1$.

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{1}{3\varepsilon_0} N\alpha \quad . \; . \; . \; . \; . \; . \; . \quad (7)$$

where N represents the number of titanate "molecules" per unit of volume and α the sum of the

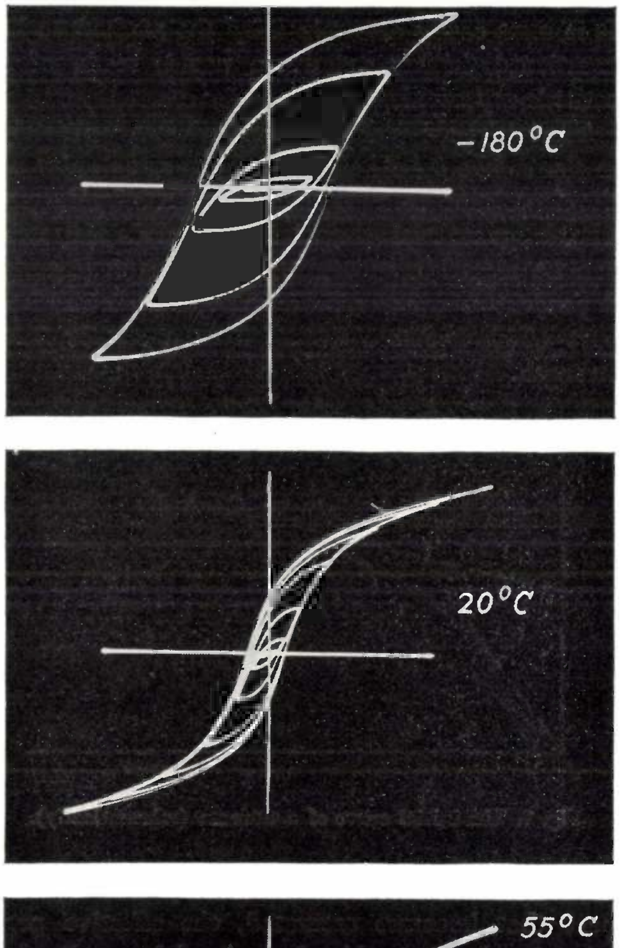


Fig. 12. Photographs of the Q-V diagram of a mixed crystal (79 % Ba, 21 % Sr) TiO_3 at different temperatures and with different amplitudes of voltage, recorded with the cathode-ray oscillograph. At −180 °C the hysteresis loop is very broad, at 20 °C narrower, whilst at the transitional point $\vartheta \approx$ 55 °C the two branches practically coincide, a straight characteristic being obtained at a somewhat higher temperature.

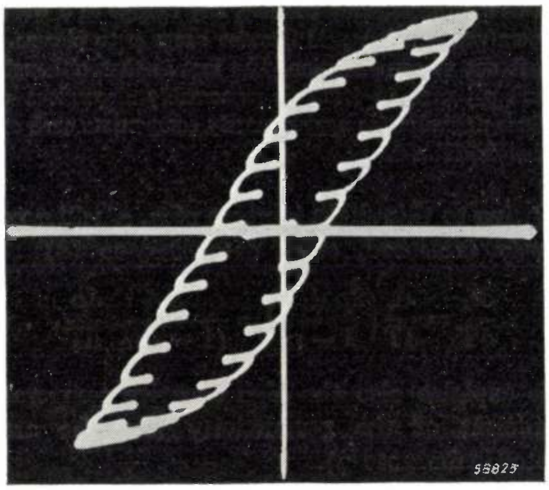


Fig. 13. Photograph of a hysteresis loop of $BaTiO_3$ with parasitic loops the slope of which denotes the reversible permittivity; furthermore, at the point of intersection of the axes a small loop is to be seen showing the initial value of the permittivity.

polarizabilities of all the ions in the molecule [7]). Substituting p for the term $(1/3\ \varepsilon_0)Na$ we get

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = p\,, \quad \text{or} \quad \varepsilon_r = \frac{2p+1}{1-p}. \quad . \quad . \quad . \quad (8)$$

For most solids the values of p are between 0.3 and 0.9. If $p = 1$ the ε_r would be infinite. Now the abnormal behaviour of titanates is to be accounted for by the fact that for these materials the value of p happens to differ but little from unity.

In the article referred to in footnote [5]) it has already been pointed out that the quantity p can be considered as being built up from various contributions according to the formulae

$$a = a_e + a_i + \dots, \quad . \quad . \quad . \quad . \quad . \quad (9a)$$

$$p = p_e + p_i + \dots, \quad . \quad . \quad . \quad . \quad . \quad (9b)$$

in which $p_e = ({}^1\!/_3\ \varepsilon_0)Na_e$ is the contribution of the electrons and $p_i = ({}^1\!/_3\ \varepsilon_0)Na_i$ that of the ions.

In addition to the contribution of a_e towards a, resulting from the deformation of the atomic electron clouds under the influence of an electric field (electronic polarization), and the contribution a_i which follows from the relative displacement of the ions in a crystal under the influence of an external field, account has also to be taken of the possibility of a contribution towards the polarization due to the presence of fixed dipoles in the crystal which are orientated under the influence of the field. As will be explained farther on, in the range $T < \vartheta$ it is to be assumed that dipoles are present in $BaTiO_3$. To account for the behaviour of titanates in the range $T > \vartheta$,however, there is *no* need to assume the presence of permanent dipoles, contrary to the case with Rochelle salt, where dipoles play a part both in the range of $T < \vartheta$ and in that of $T > \vartheta$.

Another important point is the following: in addition to the contributions towards a derived from the electron and from the ions, in the case of titanates, even disregarding the presence of permanent dipoles, there is an additional contribution due to the fact that although the structure of the lattice is cubic there are still some ions in the lattice (O^{2-} ions) which are not cubic-symmetrically surrounded by other ions. It can be proved [8]) that this results in an increase of a and thus leads to an additional increase of the value of p, so that one may write $a = a_e + a_i + a_{as}$, in which a_{as} represents the correction term for the asymmetrical environment, and thus also $p = p_e + p_i + p_{as}$.

The contribution p_e can be calculated separately, since according to Lorentz

$$p_e = \frac{n_0^{\,2} - 1}{n_0^{\,2} + 2}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

where n_0 represents the optical index of refraction extrapolated to infinitely low frequencies. For $BaTiO_3$, for instance, this contribution towards p has the value 0.63, so that a contribution of $p_i + p_{as}$ of 0.37 is required to give $p = 1$.

The high index of refraction is related to the high density (large number of easily polarizable O^{2-} ions in a small volume) and to the fact that in the case of $BaTiO_3$ and allied compounds the crystal absorption (responsible for refraction and dispersion) takes place at fairly low frequencies.

In order to understand the high value of p_i, the contribution due to the displacement of ions under the influence of an external electric field, it is necessary to investigate the crystal structure of the titanates more deeply.

Crystal structure of titanates

The compounds $CaTiO_3$, $SrTiO_3$, $BaTiO_3$, $PbTiO_3$ show the perovskite structure, so-called after the mineral perovskite $CaTiO_3$ (see *fig. 14*). The cubic

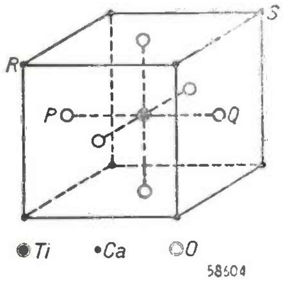


Fig. 14. Unit cell of perovskite ($CaTiO_3$).

[7]) The formula written in this way applies when Giorgi units are employed (a... (A·sec·m²/V)). In cgs units we get on the right-hand side $(4\pi/3)\ Na$ (a...(cm³)).

[8]) G. H. Jonker and J. H. van Santen, Properties of barium titanate in connection with its crystal structure, Science **109**, 632-635, 1949. See also J. H. van Santen and W. Opechowski, Physica **14**, 545-552, 1948, No. 10.

unit cell has Ca-ions at the corners, O-ions at the face-centre and a Ti-ion at the cube centre.

Goldschmidt [9]) has already made an exhaustive study of this structure and found that compounds with the formula ABO_3 have to satisfy a certain condition to be able to crystallize in the perovskite lattice. This condition is that the ions must have suitable dimensions, which in geometrical terms means that the length of the unit cell calculated from the radii of the ions along the line *PQ* must equal the length calculated from the diagonal *RS*; in other words

$$r_A + r_O = (r_B + r_O)\sqrt{2}\,,$$

in which O represents the oxygen ion, B the titanium ion and A the alkaline-earth ion.

Goldschmidt describes this condition with a parameter

$$t = \frac{r_A + r_O}{(r_B + r_O)\sqrt{2}}\,,$$

which for the perovskite structure has to be approximately equal to 1. This is not a stringent requirement; among the known perovskites t varies between 0.80 and 1.14. If, however, t is less than 0.80 or greater than 1.14 the compounds ABO_3 crystallize according to a different type of crystal.

If t answers the requirement $0.8 < t < 1.14$ the structure may still differ from the purely cubic type of fig. 14 owing to a small deformation of the lattice (tetragonal and monoclinic perovskite). Nevertheless it appears that at least in the tetragonal case at sufficiently high temperatures the structure becomes purely cubic. This is for instance the case with $BaTiO_3$ for $T > 396$ °K. Since $t > 1$, which means that $r_{Ba} + r_O > (r_{Ti} + r_O)\sqrt{2}$, the Ti-ion has at its disposal a large space in the crystal. It is therefore easily displaced and is capable of yielding a larger contribution towards p_i than is the case with other titanates. There are a number of mixed crystals of $BaTiO_3$ with other perovskites which have the same property.

The point of transition

Let us now consider what takes place at the point of transition, taking again $BaTiO_3$ at a temperature above 396 °K. Let, for instance, p equal 0.99 ($\varepsilon_r \approx 3000$, $T = 420$ °K). When the temperature decreases the density of the crystal and thus N will increase, and with that also p.

Although theoretically possible, any change in a with the temperature, which would likewise cause p to change, can be left out of consideration. The fact is that although a_e, a_i and a_{as} each in itself is presumably dependent upon T, these dependencies approximately compensate each other, so that $da/dT \approx 0$.

Owing to this slight increase of p, ε_r will now be greatly increased, since according to (8)

$$\frac{d\varepsilon_r}{dT} = \frac{d}{dT}\left(\frac{2p+1}{1-p}\right) = \frac{3}{(1-p)^2}\cdot\frac{dp}{dT}. \quad (11)$$

This explains also the high negative temperature coefficient of the permittivity (see the article quoted in footnote [5])).

With increasing ε_r a smaller external field is required to bring about a certain polarization of the crystal. Where $p = 1$ even an infinitesimally small external field is sufficient to cause a finite polarization. This becomes manifest in the fact that the state of the crystal is then unstable; in other words, when $p = 1$ the cubic arrangement from which we started is no longer the most stable state. Spontaneous polarization takes place, accompanied by the formation of dipoles (with simultaneous displacement of all Ti-ions) and a small deformation of the originally cubic unit cell.

Thus the temperature at which $p = 1$ represents a crystallographic point of transition. Below this temperature $BaTiO_3$ does indeed still retain its perovskite character but the unit cell is then no longer purely cubic, being slightly distorted tetragonally and assuming the shape of a four-sided prism. This is also apparent from *fig. 15*, where the axial ratio of the unit cell is represented as a function of T.

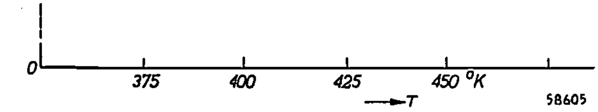


Fig. 15. Variation of the axial ratio of $BaTiO_3$ (according to measurements taken by M. G. Harwood, P. Popper and D. F. Rushman, Nature **160**, 58, 1947).

It appears that in this transition an originally cubic monocrystal mostly changes into an agglomerate of tetragonal domains. Each of these domains is spontaneously polarized.

[9]) See for instance, H. D. Megaw, Proc. Phys. Soc. **58**, 133-152, 1946.

The cause of hysteresis

In a crystal in which the spontaneously polarized domains described above occur there will be no outward evidence of an electric field, because according to the laws of probability the direction of the polarization will be distributed over the various possibilities. Upon an electric field being applied not only will the normal mechanism of polarization come into play but there will also be an expansion of some domains at the cost of others, thus giving rise to a volume polarization. This polarization continues to exist when the external field is removed, so that the phenomenon of remanence arises and in the case of alternating fields a hysteresis loop is described.

A crystal treated in this way may, therefore, show a macroscopic moment; in other words it may be an "electret". Actually anything of this nature will not be noticeable for long because charges will very soon flow along the surface and compensate the field of the internal charges. Inside the crystal, however, the state of polarization may indeed continue to exist. One of its results is the occurrence of a so-called induced piezoelcctric effect [10]).

Since the area $\oint E\mathrm{d}P$ of the hysteresis loop is proportional to the work performed by the electric forces there will obviously be great losses in the temperature range in which hysteresis occurs, particularly if the material is exposed to a strong alternating field.

Application in capacitors

In the application of $BaTiO_3$ and mixtures thereof with other titanates in capacitors, with a view to utilizing the high value of ε_r, the great dependence of the dielectric constant upon the temperature is found to be a drawback. In pure $BaTiO_3$ ε_r is only fairly constant between 30 °C and 80 °C, having a value of about 1200. What is desired, however, is to utilize the very much higher values reached round about the transition temperature; as stated above; by using mixed crystals of titanates this transition point can be made to lie at practically any temperature desired.

The fact still remains however that the very high values of ε_r desired occur only in a small temperature range and that within that range ε_r is greatly influenced by the temperature.

This can be remedied in the first place by admixing foreign materials to the titanate mixtures. The width of the peak, which in pure titanates is in the order of 20-30 °C, can then be raised to 40-50 °C, though at a slight cost of the height of the maximum.

The range within which ε_r has a high value can be further extended by combining two suitable kinds of mixed crystals with different proportions of $SrTiO_3$ and $BaTiO_3$, each placed as a dielectric in a separate capacitor. The most economical way is to connect these capacitors in parallel. It depends entirely upon the requirements to be made of such a composite capacitor as to how far the peaks of the two mixed crystals are to be separated (*fig. 16*).

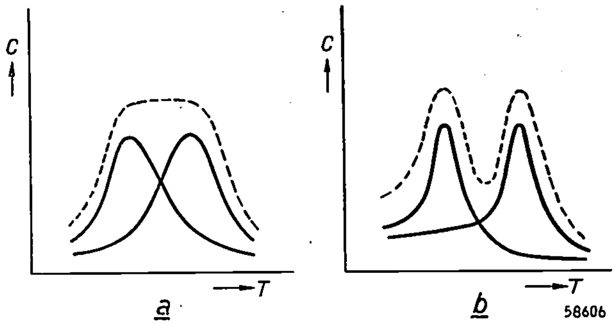


Fig. 16. Schematic representation of the result of capacitors with different peak temperatures being connected in parallel. *a*) Fairly wide peaks close together, *b*) narrower peaks farther apart.

In practice one may choose the tubular type commonly used for ceramic capacitors, where the tube is made partly of one mixture and partly of the other (see *fig. 17*). Of course one cannot then avoid having to use titanates in a temperature range in which the hysteresis phenomena and high

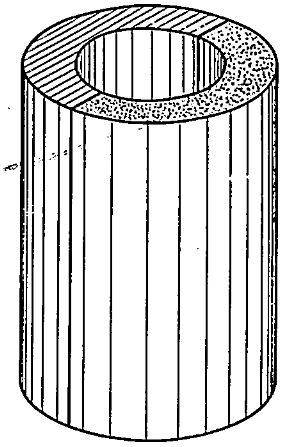


Fig. 17. Filling of a tubular capacitor in which two mixtures of titanates are placed parallel to each other.

[10]) See for instance J. C. B. Missel, Philips Techn. Rev. **11**, 145-150, 1949, No. 5.

losses occur. Consequently the possibilities of application of these capacitors are somewhat limited and they are not suitable for use in tuned circuits, though they can be used, for instance, as coupling capacitors.

Summary: The most important properties of ferromagnetic materials are briefly discussed in connection with the fact that in an electric field the behaviour of certain chemical compounds such as $KNaC_4O_6$ $4H_2O$. (Rochelle salt) and KH_2PO_4 very closely resembles the phenomenon of ferromagnetism. Particularly of importance is the high value of ε_r, similar to that of μ_r in the case of ferromagnetic materials. Further, in analogy with ferromagnetic materials, below a certain critical temperature ε_r becomes dependent upon the field strength and hysteresis occurs. Such is found to be the behaviour of barium titanate and allied compounds. In the latter case, owing to the comparatively simple crystal structure, these phenomena are easily explained theoretically. The theoretical explanation of the behaviour of titanates is given in broad lines, whilst at the same time it is shown how the very high value of the permittivity found with these materials can be utilized for the construction of capacitors required to possess a fairly constant and large capacitance within a certain temperature range, whilst being of the smallest possible dimensions.

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

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

1854: K. F. Niessen: Nodal planes in a perturbed cavity resonator, III (Appl. sci. Res. **B 1**, 284-298, 1949, No. 4).

In choosing in the unperturbed resonator a vibration having two nodal planes (one perpendicular and one parallel to the movable wall) the mathematics required is of quite another nature than in Part II. Again a combination of functions is chosen, the coefficients of which have to be determined. Now it appears that all these coefficients, except that belonging to the original function, must be of the order of δ, just as was found in Part I. There is only one possible solution and its nodal figure is permanent, being no function of time. Instead of the two intersecting nodal planes $y = a/2$ and $z = a/2$, existing for $\delta = 0$, we obtain for $\delta \geqslant 0$ two sheets, the intersection of which with a plane perpendicular to the x-axis resembles an orthogonal hyperbola and the difference between the new and the original nodal system is relatively large in the neighbourhood of the central axis of the resonator. The use of this effect for the construction of special resonators containing a "side room" and a bar is shown. (See also abstracts Nos **1815** and **1853**.)

1855: Th. J. Weijers: Toelaatbare onderlinge storing van twee in frequentie gemoduleerde omroepzenders (T. Ned. Radiog. **14**, 61-72, 1949, No. 3). (Permissible mutual interference of two F.M. broadcast transmitters; in Dutch.)

In order to obtain data about the selectivity of an f.m. broadcast receiver, the measurements have to be made according to the two-signal method. The results of these measurements, taken on three different receivers, are given and discussed. An interesting result is that, for an f.m. broadcast service in a given area, the required frequency band is reduced considerably if, by repeating the same frequency band at a sufficient distance, the central frequencies of the second set of transmitters are shifted over 50 kc/s or more.

1856: A. Bril: On the saturation of fluorescence with cathode-ray excitation (Physica 's Grav. **15**, 361-379, 1949, No. 3/4).

Measurements of the light output as a function of the beam current under stationary cathode-ray excitation are performed for zinc silicate activated with various concentrations of Mn or Ti and for zinc sulphide activated with ZnCl and with AgCl. The intensity of fluorescence saturates at high current densities. Under continuous electron bombardment the saturation depends on the activator content, on the presence of quenchers and on the decay time of the fluorescence. The concentration of centers in ZnS-ZnCl is estimated. In the case of a spot of high current density scanning a definite area, the saturation does not depend on the decay time of fluorescence if the time in which the electron bcam passes a certain point of this area is small with respect to the decay time.

VOL. 11 No. 7, pp. 193-220 JANUARY 1950

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

AN X-RAY APPARATUS FOR THERAPEUTIC TREATMENTS

by B. COMBEE and J. FRANSEN.

621.386.1:615.849

X-ray treatment is still for the present one of the most important means of combating malignant growths. For such treatment there is a tendency to use also extremely hard rays, obtained with tensions of millions of volts, in order to get a more effective distribution of the dosage in the body. For institutions with ample funds at their disposal new and larger installations are being developed for this purpose. In the field of the more conventional apparatus, working with voltages up to some hundreds of kilovolts, development is, however, also proceeding apace: by the application of modern methods of construction efforts are being made to produce this apparatus in a form enabling the medical practitioner to apply X-ray treatment more quickly, more efficiently and with accurate doses. An example of such a modern design is the Philips "Compactix" therapeutic apparatus.

Introduction

The requirement usually made of X-ray apparatus for medical treatment is that there must be a high degree of freedom of movement of the beam of rays, so that they can be directed upon the patient from all sides. In the conventional arrangement the X-ray tube was therefore mounted on a stand which permitted various degrees of freedom of movement, and the high tension required was supplied to the tube from a fixed high-tension generator via flexible cables. During the last decade other therapeutic equipment has been developed, based on a different principle of construction, the high-tension generator and the X-ray tube being built together to form one unit. This unit, when mounted on a suitable stand, is likewise freely movable, whilst the constructionally difficult cable connection has been eliminated. In this article a description will be given of the "Compactix" therapeutic apparatus which has been built by Philips on this principle and which is intended for a tube voltage up to 200 kV_{max} with an average tube current up to 10 mA. *Fig. 1* gives a general view of the installation.

For extremely high tensions, say one million volts, the building together of the tube and the generator is the only practicable solution if one is to follow the now generally accepted principle of complete protection of the user against the high tension without sacrificing the freedom of movement of the beam. This principle of the building together of tube and generator was in fact first applied for very small X-ray units with relatively low tension and low power, constructed for certain diagnostic purposes [1]). With such a small apparatus the high-tension cables would have formed a far too large and heavy part of the whole. For installations with tensions up to 200 or 300 kV, which are generally used for the treatment of deep-seated tumours, however, both methods of construction, with and without cables, have their merits.

For the construction of the "Compactix" apparatus new means were sought for providing for insulation against high tension, cooling, and the protection of the user against scattered X-rays. The satisfactory solution found for these problems made it possible to build a tube and generator unit comparatively small in size and light in weight, so that an easily adjustable apparatus could be obtained which enables the medical

[1]) Such an apparatus was demonstrated already in 1933 by A. Bouwers. See A. Bouwers, Modern X-ray development, Brit. J. Radiology 7, 21, 1934; further also Philips Techn. Rev. 6, 225, 1941 and 10, 221, 1949.

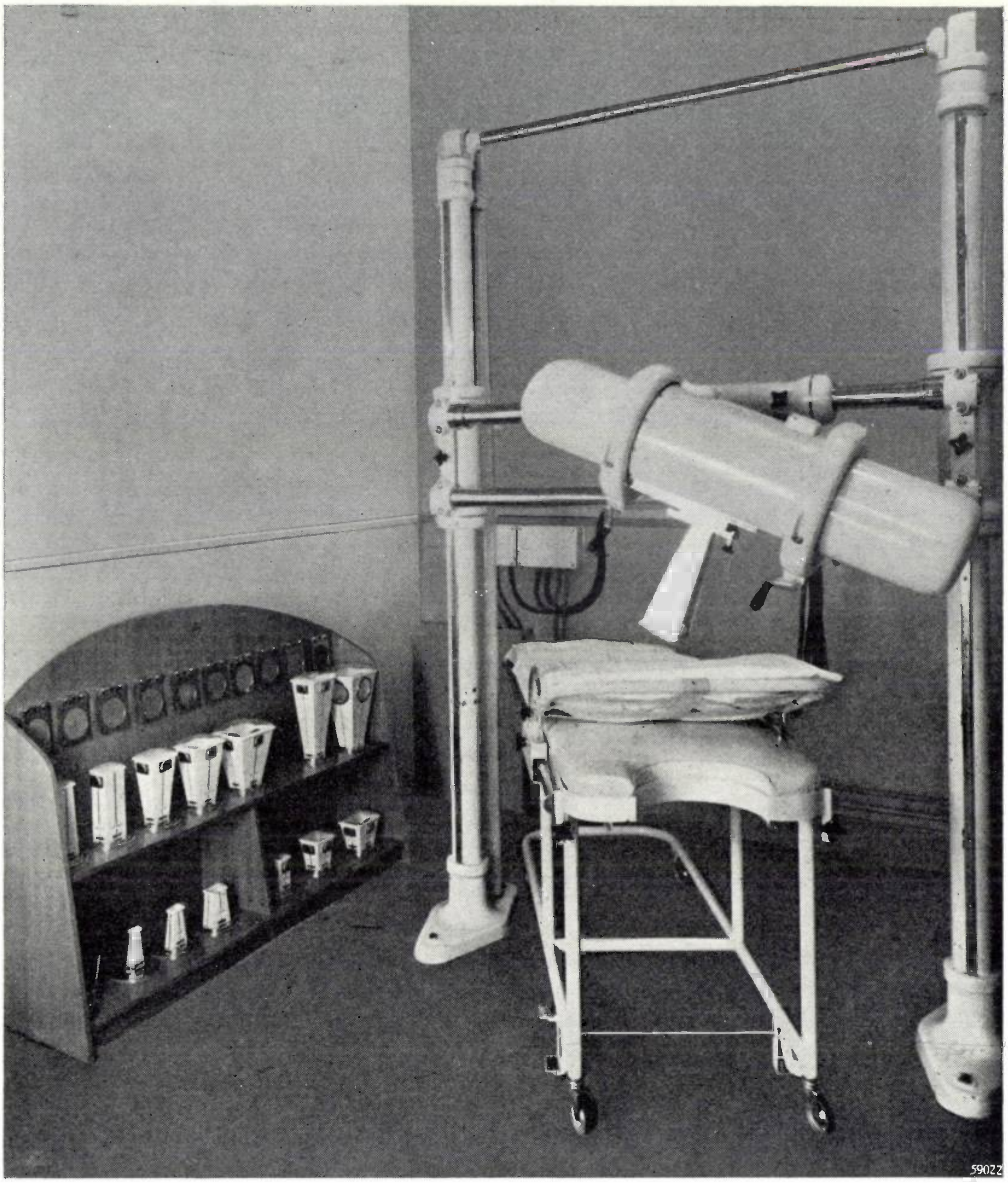


Fig. 1. The "Compactix" apparatus for X-ray therapy with stand and movable couch. In the cabinet on the left are the applicators and filters belonging to this apparatus. The apparatus is operated from a control desk in an adjoining room, where the operator can observe the patient through a lead-glass window.

practitioner to apply any treatment under the most favourable conditions.

In addition to the constructional problems mentioned, which will be dealt with when describing the X-ray tube and generator, some attention will also be devoted in the following pages to various new devices employed in the circuit of the apparatus, and the operating desk. The article will be concluded with some remarks about the accessories to be used, such as filters and applicators (cones), and about the distribution of dosage that can be obtained.

The X-ray tube

In *fig. 2* a photograph of the X-ray tube is reproduced; *fig. 3* is a somewhat simplified cross-sectional drawing in which also the most important dimensions are given. In the drawing one can see the tungsten target *P* which is mounted in the slanting face of the copper anode block and upon which the electrons emitted by the filament *F* are focused. It is in the comparatively sharp slope of the anode (anode angle $\alpha = 32°$, as compared with 15°-20° in most diagnostic tubes) that the specific character of therapeutic tube is imme-

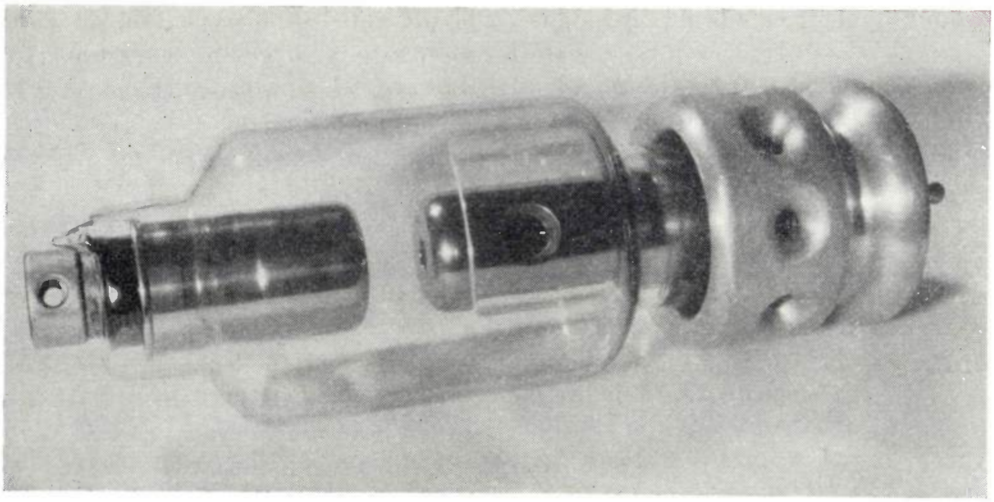


Fig. 2. X-ray tube of the "Compactix" apparatus. It is of hard glass and immersed in oil. The radiator on the right-hand end transmits the heat from the anode to the oil. The holes in the radiator allow the oil to flow through freely, so that there are no dead spots around the point where the anode is fused into the tube.

diately manifest: from the X-rays emitted from P in all directions it is thereby possible to make efficient use of a cone (R) with a large vertex angle, so that large fields can be irradiated at a relatively short distance.

The tube is immersed in oil in an earthed metal shield. It is due to the oil insulation that a tube of such a short length (short distance between the two poles) suffices for the high tension of 200 kV_{max}. The oil serves at the same time for cooling [2]). A radiator with a large surface area is mounted on the end of the anode protruding through the bulb. Around this radiator at a distance of 4.5 cm, inside the shield, are cooling-water spirals which are earthed and can therefore be connected to the water mains. Thus the heat developed on the focus and conducted through the anode body of the radiator is transmitted to the surrounding oil and then carried off by convection in the freely circulating oil to the water-cooled pipes. With this simple method of cooling a power of 1.4 kW can be applied continuously without any difficulty.

If, instead of letting the oil circulate freely, forced cooling is applied, by injecting the oil into the hollow anode, greater powers can be dissipated, but then there is the disadvantage of having to use an oil pump, with all its attendant noise, and in our case it was desired to avoid this.

[2]) This method of construction, applied to diagnostic tubes and also to a therapeutic tube resembling in broad lines that described here, has already been described in detail in this journal: J. H. van der Tuuk, Hard-glass X-ray tubes in oil. Philips Techn. Rev. 6, 309-375, 1941.

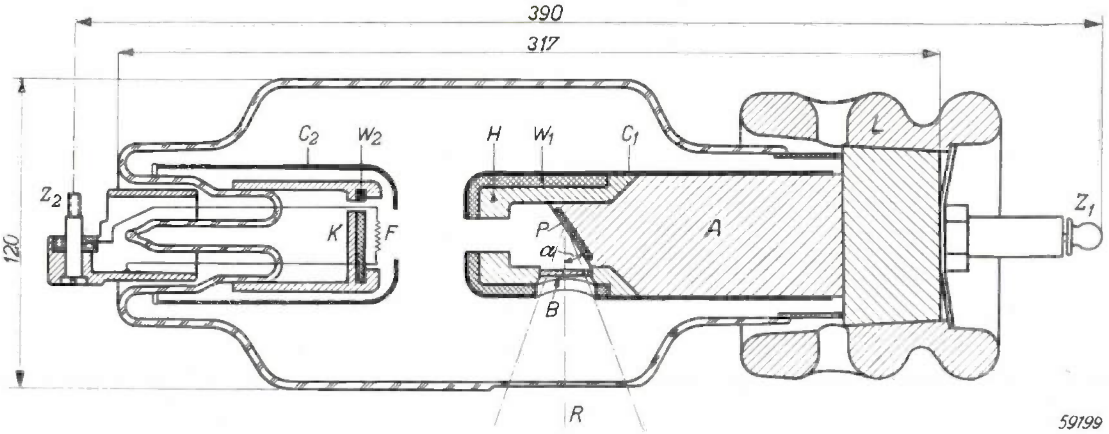


Fig. 3. Cross section of the X-ray tube (simplified). K cathode with filament F, A anode with tungsten target P, H anode hood for intercepting secondary electrons, B beryllium window allowing the X-ray beam R to pass through, W_1 tungsten hood and W_2 tungsten disc for absorption of the undesired X-rays, C_1 and C_2 polished chrome-iron caps, L radiator, Z_1 and Z_2 poles for connection to the high tension up to 200 kV_{max}.

The factor limiting the power in our case is the temperature of the oil at the surface of the radiator: if it should become too high then the oil would lose some of its insulating properties. The temperature of the anode at the point of fusion, which in other constructions sometimes sets the limit to the power, in our case remains well below the maximum permissible level of about 200 °C. Elsewhere, namely in the focus, the anode is of course hotter, notwithstanding the good heat conduction of the copper. The specific focus load, however, is comparatively small (30 W/mm^2), so that the focus temperature also remains well below the permissible value (e.g. 500 °C). Therefore the problem of the primary heat transfer of the focus, which is of great importance in heavily loaded diagnostic tubes [3]), does not arise here. It is true that the specific focus load has been made greater for this tube than is usually the case with therapeutic tubes: the focus is comparatively small (5 mm × 10 mm, projected size 5 mm × 5 mm) so as to make the apparatus also suitable for the examination of materials, where the sharpness of the shadow pictures produced depends upon the size of focus.

For therapeutic purposes the size of the focus is not of primary importance, but here, too, limited dimensions of the focus are favourable because the volume irradiated can be more sharply defined (less penumbra round the edge of the diaphragm).

Mounted on the anode is an "anode hood" taking up the secondary electrons emitted by the focus on the anode. This protects the glass wall of the tube against an electron bombardment, so that no dangerous wall charges can arise, which may lead to disruption. The hood is made of copper and is in good thermal contact with the anode, so that the energy from the secondary electrons is quickly transmitted to the anode block. Over this copper hood a second hood made of tungsten is fitted, the purpose of which is to prevent X-ray radiation in undesired directions. Over this tungsten hood and the whole of the anode block is a polished chrome-iron cap to prevent "cold emission" of electrons from the outer wall of the anode during the half cycle that the latter is at a negative potential (the tube is fed with an alternating voltage). Furthermore, any impurities that may be released from the anode parts during the pumping process or while the apparatus is working, and which should be prevented from settling on the wall of the tube, are precipitated on the inner wall of this chrome-iron cap, which remains comparatively cold.

Something more has to be said about the absorption of undesired X-rays. This is a point that has always received a great deal of attention in the development of Philips X-ray apparatus. We would recall, for instance, the development of the "Metalix" tubes in 1923 [4]), when for the first time a construction was materialized which gave full protection of the user against X-rays emitted outside the effective cone (some years later the system was extended to give full protection also against contact with live parts). This idea was kept in the foreground also in the designing of the "Compactix" apparatus. In order to attenuate the 200 kV irradiation so that in the room where the operating staff is working the tolerance dose is not exceeded (in various countries a limit of about 0.05 to 0.1 roentgen per day is laid down), a screen of lead 4 to 5 mm thick is required round the source of the rays. Owing to its great weight such a thick layer of lead placed on the outside of the shield of the apparatus would hamper the easy handling aimed at. By applying an absorbing layer inside the tube itself at a short distance round the focus its weight has been reduced to a few hundred grams.

Of course lead cannot be used inside the tube, because at the temperature at which the tube is degassed it would melt. Absorption of the rays is therefore mainly brought about with tungsten, the atomic number of which (74) is only slightly lower than that of lead (82), so that the necessary layer of tungsten has about the same weight as an equivalent layer of lead. The copper of the anode hood and the chrome-iron of the outer cap do not provide much absorption as they are only equivalent to about 0.4 mm lead. The total lead equivalent of the three layers amounts to 4 mm lead. One cannot entirely do without an absorbent layer in the shield of the apparatus on account of the secondary X-rays formed in the oil and on the inner wall of the shield, which also have to be rendered harmless; a layer of lead 1 mm thick will provide sufficient protection against secondary radiation. The total absorption of undesired radiation then corresponds to 5 mm of lead.

In the anode hood is an opening to let in the primary electrons and another opening to let the effective X-ray cone pass out. The first opening

[3]) See for instance J. H. van der Tuuk, Philips Techn. Rev. 3, 292, 1938 and 8, 33, 1946.

[4]) See for instance A. Brouwers, A new X-ray tube, Physica, 4, 137, 1924.

allows not only a small part of the secondary electrons to escape but also undesired X-rays. The latter are absorbed by a tungsten disc mounted in the cathode for that purpose. The secondary electrons tending to escape through the second opening are stopped by a small disc of beryllium placed in this opening. Thanks to the very low atomic number (4) of this metal the useful X-rays are only very slightly weakened. In order that these rays should be weakened as little as possible on their further outward passage, the glass wall of the tube has been ground where the beam passes through, as may be clearly seen in figs 2 and 3. The surrounding layer of oil is also greatly reduced in the beam by a cap of "Philite" placed in the shield. Thus the inherent filtration of the apparatus (see the last chapter) is reduced to an equivalent of about 2 mm aluminium at 80 kV_{max}.

The high-tension generator

The desire to keep the high-tension generator as small and as light as possible led to the choice of an alternating current for supplying the X-ray tube. In the main the generator therefore consists only of a high-tension transformer and a filament transformer. The input voltage may be any mains voltage between 200 and 250 V, with a frequency of 40, 50 or 60 c/s.

In order to minimize as far as possible the difficulties of high-tension insulation, the X-ray tube is fed with a voltage which is symmetrical with respect to earth. For this purpose two separate high-tension transformers are employed, each of which, working in antiphase, supplies half the high tension (100 kV_{max}) to one pole of the tube. Owing to this arrangement it has been possible to divide the weight of the generator, one of the two transformers being placed at each end of the cylindrical tank in which the apparatus is mounted; see *fig. 4*. Thus the focus of the X-ray tube placed in the middle of the cylinder lies approximately in the centre of gravity of the whole apparatus, this greatly facilitating adjustments (see below).

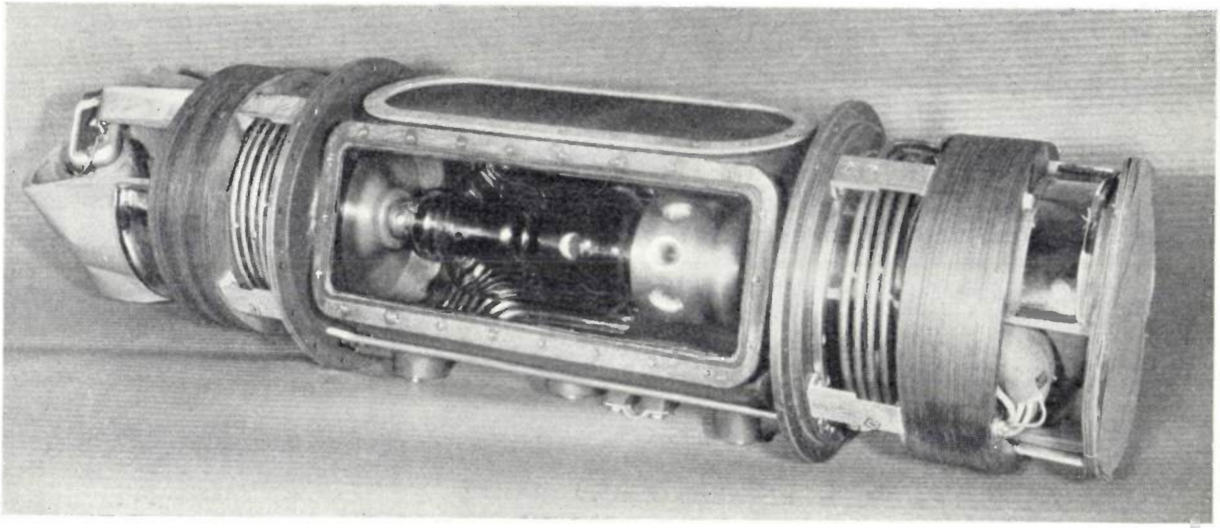


Fig. 4. The X-ray tube, the high-tension transformers (on the right and on the left) and the filament transformer (on the extreme left) are built together into one unit which is mounted in a cylindrical tank. The tube compartment has a detachable lid. Cooling water spirals are contained in all three compartments.

The X-ray tube itself functions as rectifier, passing current only during the half cycle when the high tension at the anode is positive and the filament negative. Owing to the voltage drop caused by the current in the transformer windings, etc., the tube voltage in this half cycle may be up to 10% lower than that in the other half cycle. In order that the insulation requirements should not be unnecessarily severe on account of the non-conducting half cycle, so called inverse-voltage suppression is applied: a combination of a selenium rectifier and a resistor is connected in series with each of the two primary windings (see *fig. 5*). During the working half cycle the primary current flows through the selenium rectifiers with practically no resistance, whereas during the other half cycle it has to pass through the resistors, so that the voltage drop in the resistors reduces the high tension generated.

These resistors perform at the same time an important function in the event of momentary disturbances in the X-ray tube. The stray self-inductance of the high-tension transformers together with the earth capacitances, always present in the secondary windings, may form oscillatory circuits tending to produce dangerous overvoltages in the event of a surge in the tube. These overvoltages are strongly suppressed by the damping action of the resistors in the primary circuits.

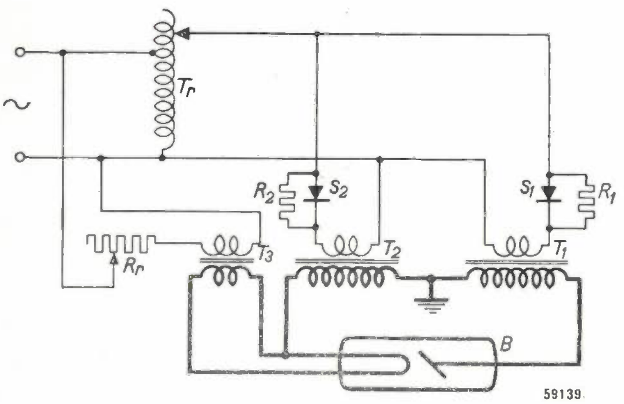


Fig. 5. Circuit of the high-tension generator. In the primary circuits of the two high-tension transformers T_1 and T_2 "inverse voltage suppression" has been applied (selenium rectifiers and resistors S_1, R_1 and S_2, R_2). *B* X-ray tube, T_3 filament transformer, T_r variable ratio transformer and R_r variable series resistor for adjusting respectively the tube voltage and tube current (to be discussed later).

The compact construction of the transformers has been made possible by using as insulating material paper impregnated with oil, a combination which has proved to be reliable and capable of withstanding heavy loads in the manufacture of capacitors and cables. One of the two high-tension transformers is illustrated in *fig. 6*. The circular shape of the yoke has been chosen so that the transformer can be placed in a cylindrical tank. The primary winding is made in two parts placed round the core on either side of the secondary winding; thus optimum use is made of the space available and at the same time a favourable distribution is obtained of the electric field between the windings of the secondary coil and the surroundings. The paper insulation is so shaped that the electric lines of force run as far as possible perpendicular to the surface of the paper. At the end of the secondary coil, to which the pole of the X-ray tube is connected, a metal screen is incorporated in the paper insulation (so-called potential screen) which helps to keep the maximum field strength between this high-tension pole and the earthed yoke small and at the same time reduces the capacitance of the outermost windings of the secondary coil with respect to earth. This latter point is of importance with a view to distributing overvoltages arising from possible tube disturbances evenly over the secondary winding.

From fig. 4 and the simplified cross-sectional drawing in *fig. 7* it may be seen how the transformers and the X-ray tube are located in the tank. The compartment for the tube formed by a central casting of light metal is entirely separated from the two adjacent compartments containing the transformers. As a result of this separation of the compartments the oil for the transformers is not called upon to help in carrying off the heat from the X-ray tube and thus is kept cooler, whilst when the X-ray tube reaches the end of its useful life it can be replaced without the transformer oil being exposed to the air.

In this case the problem of the thermal expansion of the oil has been solved in the following way. The oil in the tube compartment is able to expand on account of an oil-proof rubber diaphragm having been placed in one of the walls (to be seen at the top of fig. 4). The two transformer compartments are interconnected by a conduit passing along the outside of the tube compartment and thus need only one common expansion system. This has been provided by metal expansion bellows fitted at the anode end of the anode transformer compartment. The corresponding space in the compartment at the other end offers room for the

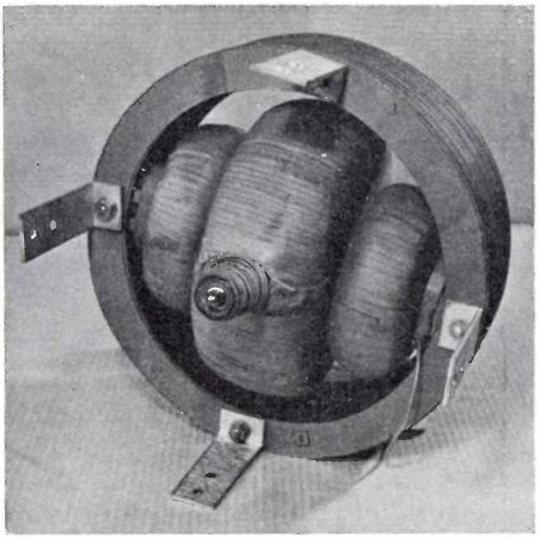


Fig. 6. One of the two high-tension transformers, with circular yoke, oil-impregnated paper insulation, and a primary winding made in two parts and placed either side of the secondary winding. The high-tension pole of the secondary winding, which is connected to one of the poles of the X-ray tube, protrudes a little.

filament transformer, so that with the compact filling of the cylinder with the various parts the apparatus remains symmetrical.

Since a connection has to be made to the water mains for the cooling of the X-ray tube, it involves no extra complication to put some cooling spirals in the transformer compartments, this being advantageous for keeping the transformers as small as possible. The total losses in these transformers amount to about 200 W.

through an angle of 240°. In both these movements the whole apparatus, and also the focus of the X-ray tube, revolves about the centre of gravity.

The bridge part of the stand runs on ball bearings and is counterbalanced with weights moving up and down inside the vertical columns on which it runs, and can be moved in vertical direction. Further, the axis of the suspension arm can be moved in and out and the arm can be traversed horizontally along the tubes of the bridge.

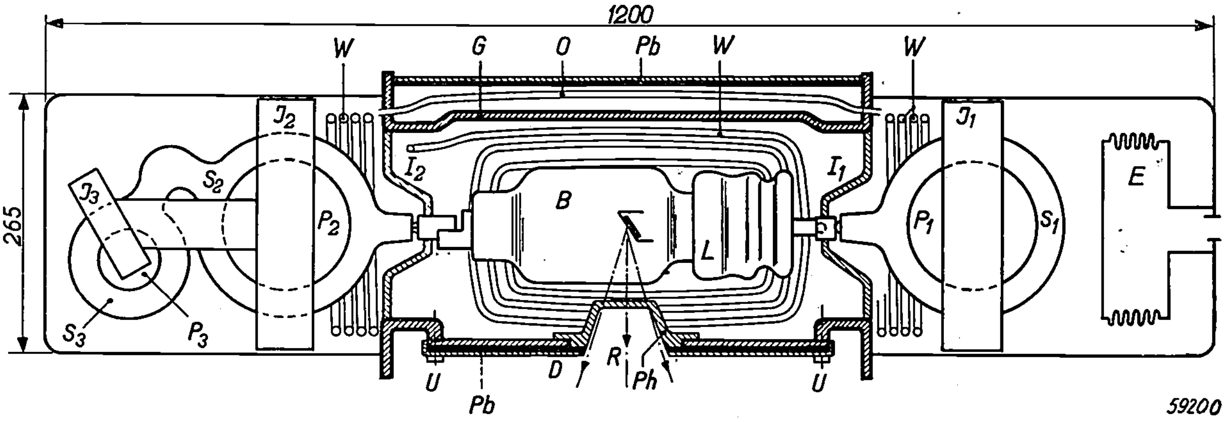


Fig. 7. Cross section of the tank of the "Compactix" apparatus (simplified). *G* central casting, *B* X-ray tube, I_1-I_2 "Philite" cups with lead-ins for the high tension. J_1, P_1, S_1 yoke, primary and secondary windings of the anode transformer, J_2, P_2, S_2 ditto of the cathode transformer, J_3, P_3, S_3 ditto of the filament transformer, *L* radiator, *W* cooling-water spirals in the three compartments, *O* connecting pipe between the two transformer compartments, *E* expansion bellows for the oil, *D* detachable lid fixed with screws *U*, *Ph* "Philite" cap where the X-ray cone *R* emerges, *Pb* lead jacket 1 mm thick in the shield.

Suspension of the apparatus

Fig. 1 shows the cylindrical tank suspended from a stand. The tank weighs in all 125 kg. (275 lbs). It is carried on an arm consisting of two rings joined together by two steel tubes. The arm is mounted in a bearing in the bridge of the stand in such a way that it can be turned around the horizontal axis. The movement around this transverse axis, which enables the tank to be set more or less upright, is brought about by means of a handle in one of the rings (the left-hand one in fig. 1). The cylinder can be turned 90° in the direction in which the cathode of the X-ray tube comes uppermost, and 45° in the opposite direction; it has been necessary to limit the latter movement in order to maintain adequate cooling of the anode by convection in the oil (the tube cannot be allowed to stand upright with the anode radiator uppermost). By means of a second handle (the right-hand one in fig. 1) the tank can be turned in the two rings about its own longitudinal axis

Fig. 8 shows how the tank can be moved, with the five degrees of freedom of movement mentioned. It can be seen that the practitioner is offered every desired possibility of adjustment. It is quite easy, for instance, to direct the rays in an upward direction (so-called under-couch treatment).

Stabilization, adjustment and measurement of voltage and current

The first essential for the success of X-ray treatment is exact dosage. To this end the practitioner must be able to rely upon the measuring instruments recording exactly the tube current and tube voltage, which determine the intensity and hardness of the X-rays generated, whilst it is also necessary that the current and voltage are not subject to fluctuations while the treatment is being applied.

Let us first consider the measurement. Connected to the earthed side of the secondary winding of each high-tension transformer is a moving-coil instrument which measures directly the rectified current passing through the X-ray tube; see *fig. 9*. The object of using two meters for measuring this current is to avoid wrong dosage being applied in the event of one of the meters becoming defective, a safety measure that is prescribed in several countries.

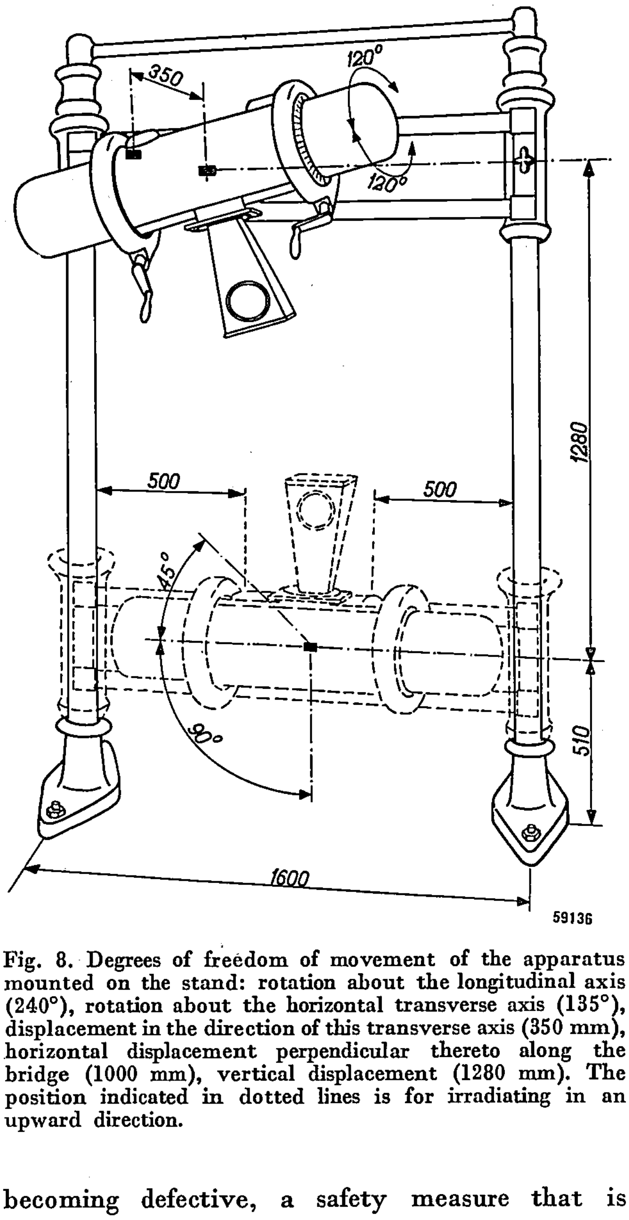


Fig. 8. Degrees of freedom of movement of the apparatus mounted on the stand: rotation about the longitudinal axis (240°), rotation about the horizontal transverse axis (135°), displacement in the direction of this transverse axis (350 mm), horizontal displacement perpendicular thereto along the bridge (1000 mm), vertical displacement (1280 mm). The position indicated in dotted lines is for irradiating in an upward direction.

Direct measurement of the voltage across the X-ray tube is not possible in the high-tension part, so that only the primary voltage of the high-tension transformers can be measured; multiplication by the transformation ratio gives the secondary voltage that would arise under no load. This no-load voltage, however, is the sum of the actual tube voltage and the voltage drop in the transformers on load, the drop being proportional to the tube current. A direct indication of the tube voltage can be obtained, independent of variations due to the voltage drop, with the aid of a measuring circuit, as shown in fig. 9. Across the voltmeter kV is the primary voltage, reduced by a correction voltage produced by the small transformer T_4 connected to the resistor R_4. This correction voltage is proportional to the primary current from the high-tension transformers flowing through T_4, and is thus practically proportional to the tube current. Consequently, by properly adjusting R_4, the voltmeter can be calibrated direct in kilovolts applied to the tube.

The adjustment of the tube current and tube voltage to the values desired for the radiation is usually done by means of a variable resistor in series with the primary winding of the filament transformer and by means of a variable ratio transformer feeding the primary winding of the high-tension transformer (cf. fig. 5). In principle the same method has been followed here, but with the addition of means for stabilizing the adjusted values of current and voltage within narrow limits with respect to mains voltage fluctuations and other possible disturbances. One has to reckon with mains fluctuations up to 10%. Without stabilization the tube voltage varies in proportion to the mains voltage, and the tube current in proportion to a higher power of the mains voltage, owing to the strong temperature-dependence of the emission of the filament. Furthermore, the emission may also vary to a certain extent in consequence of the filament absorbing impurities that may be released from the anode. This goes to show how important it is to have a properly stabilized current and voltage, the more so when it is borne in mind that the intensity of the radiation, which is in fact the point at issue, is proportional to the tube current and roughly proportional to the square of the tube voltage.

As regards the tube current, the stabilizing system employed corresponds in principle to one that has already been previously described [5]). The "resistance" of a pentode (ratio of anode voltage to anode current) is taken up in the primary circuit of the filament transformer via an isolating transformer; see fig. 9. This resistance is dependent upon the control-grid voltage. The voltage drop produced across the resistor R_5 by the current passing through the X-ray tube, reduced by a highly constant reference voltage E_0, is applied to the control grid. Thus any fluctuation in the tube current reacts via the pentode upon the filament voltage of the X-ray tube, and in such a direction and to such an extent that the mean

[5]) Philips Techn. Rev. 8, 8, 1946.

tube current is kept constant within about 1%. (Actually the voltage across the resistor R_5 consists of pulses, due to the X-ray tube being fed with an alternating current. Before being applied to the control grid of the pentode, however, this pulsatory voltage is smoothed by an R-C circuit, not shown in the illustration. Thus the stabilization acts only upon the mean value of the tube current.)

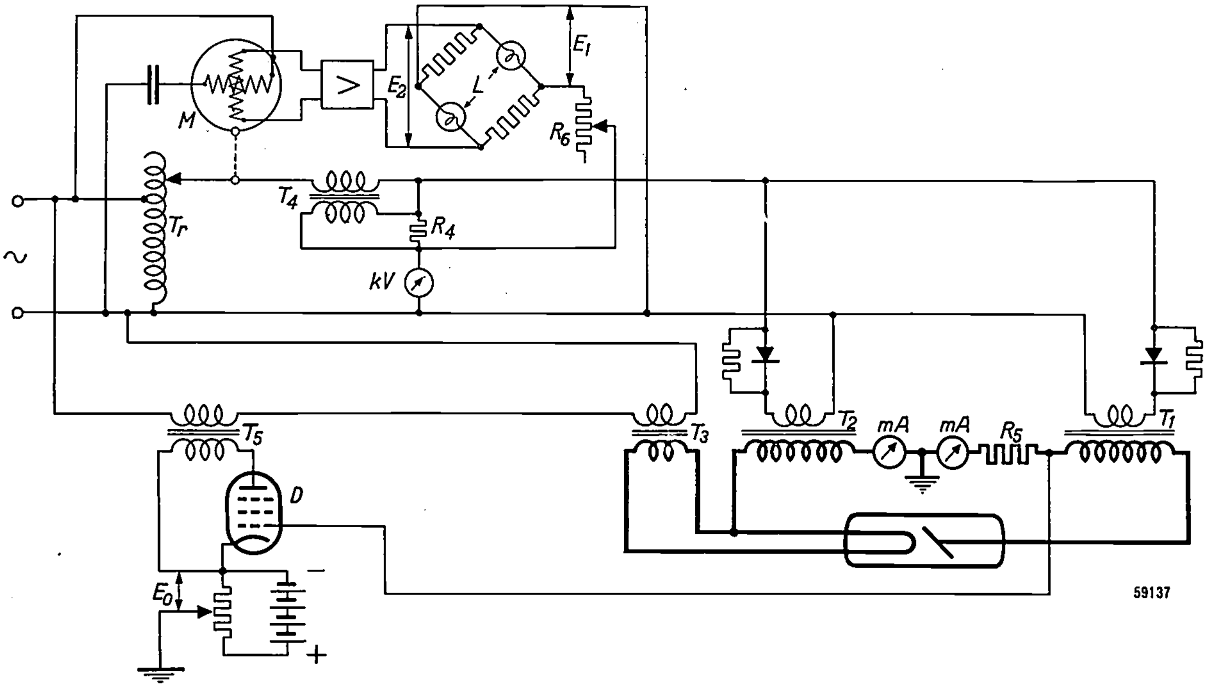


Fig. 9. Circuit diagram for measuring, controlling and stabilizing the tube current and tube voltage. The tube current is measured with the two moving-coil meters "mA", the tube voltage with the voltmeter "kV" calibrated in kilovolts.

For stabilization of the tube current: D pentode, the "resistance" of which is connected in series with the filament transformer T_3 via the isolating transformer T_5; R_5 resistor across which the tube current sets up a voltage which is applied to D as control-grid voltage; E_0 adjustable reference voltage, for the sake of simplicity indicated as being derived from a battery (actually it is supplied by a small rectifier with stabilized output connected to the mains).

For stabilization of the tube voltage: L two incandescent lamps which together with two normal resistors form a bridge, E_1 input voltage on this bridge proportional to the tube voltage, E_2 output alternating voltage the phase of which is reversed when E_1 passes through the rated value, M asynchronous motor driving the brush gear of the variable-ratio transformer T_r for the primary voltage of the high-tension transformers T_1 and T_2.

The reference voltage E_0 having been made adjustable, it is possible to determine at what voltage across R_5 equilibrium is reached, and thus by means of the same mechanism the desired control of the tube current is obtained, this being adjusted as desired between 2 and 10 mA.

For stabilizing the high tension somewhat stronger measures than a single electronic valve are needed. For this purpose a small asynchronous motor is employed for driving the brush gear of the above-mentioned variable ratio transformer for feeding the high-tension transformers. A servo-mechanism causes the motor to rotate in one direction or the other according to whether the tube voltage is higher or lower than a predetermined value. How this regulating circuit works can be explained with reference to fig. 9. Two normal resistors and two small incandescent lamps form the four branches of a bridge circuit. The lamps function as resistors having a value which varies steeply with the applied alternating voltage E_1 (owing to the fairly large fluctuation of the filament temperature). As a result the bridge is only balanced when E_1 has a certain value, that is to say, at a certain value of the tube voltage; the output voltage E_2 is then zero. When the value of E_1 varies, an alternating voltage $E_2 \neq 0$ is obtained which, when E_1 is too high, is in antiphase with the. voltage obtained at too low a value of E_1. The voltage E_2 is amplified and applied to one of the two field windings of the afore-mentioned asynchronous motor, whilst the other field winding is connected to the mains (via a capacitor for the required phase shift). Since the direction of rotation of the motor is reversed when the phase of E_2 changes, by means of this mechanism any deviation of E_1 from a predetermined value can be made to bring about a counteracting change in the variable ratio transformer. In practice it has been found possible to stabilize the high tension in this way to within a variation of 0.9%, which is sufficient for the object in view.

As the diagram shows, the regulating voltage E_1 is obtained in the same way as the voltage for the voltmeter calibrated in kV. Thus also E_1 is corrected for the voltage drop in the high-tension transformers and is indeed made proportional to the tube voltage. The tube voltage (in the conducting half cycle) is therefore kept constant, even when the adjustment of the tube current is changed. With the aid of the resistor R_6 the tube voltage at which the regulating circuit comes to rest can be adjusted as desired between 80 and 200 kV$_{\text{max}}$.

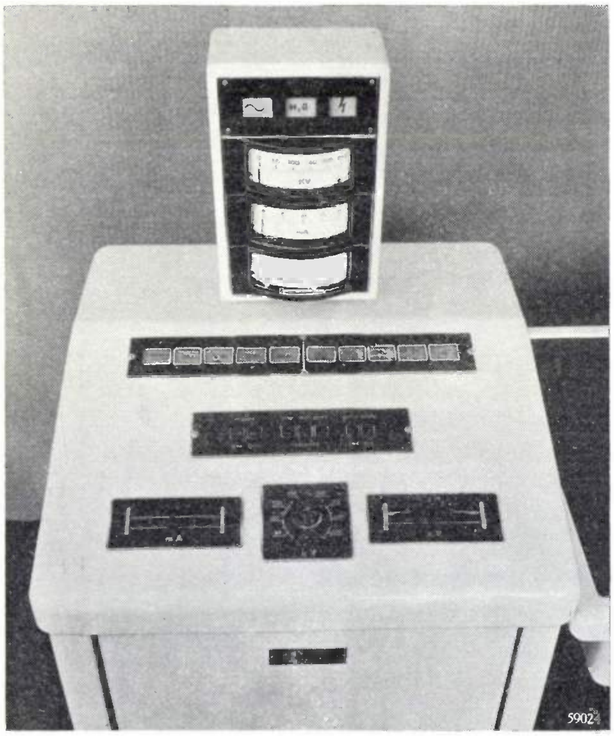


Fig. 10. Top view of the control desk. Bottom row of controls: Control disc for the tube current (on the left), control knob for the tube voltage (in the middle), and on the right a control disc which is coupled direct to the variable ratio transformer for the tube voltage but which normally does not need operation by hand, Next row: On the left keys for switching the mains voltage on and off, and on the right keys for switching the high tension on and off, as also for selecting three standard ray qualities. Upper row: 10 pilot lamps indicating which filter is in use. In the "meter tower": kV-meter, two mA-meters and three signals with green, white and red lights to indicate respectively whether the mains voltage, the water cooling and the high tension are switched on.

The circuit described in this chapter, together with some other accessories, is contained in a control desk, a close-up view of which is given in *fig. 10.* The functions of the various knobs, meters, etc. seen on the top panel are explained in the legend. There is one particular point to be discussed which is connected with the stabilization of the tube voltage. In the middle of the panel on top of the desk is a row of keys, of which the first two from the left serve for switching the mains voltage on and off, whilst the last two on the right serve for switching the high tension on and off, which can then be freely chosen. The three keys in the middle of the row, however, serve for selecting any one of three standard X-ray qualities; in each case the tube-voltage regulation is adjusted to a certain value fixed before delivery of the apparatus, and at the same time a locking device is brought into action which only allows the high tension to be applied to the X-ray tube when a certain filter of the series of ten supplied with the apparatus (see the last chapter) is interposed. For example, these three keys may correspond to the following three voltages and added filters:
80 kV$_{\text{max}}$ and 0.5 mm Al (superficial therapy)
140 kV$_{\text{max}}$ and 3.0 mm Al (intermediate therapy)
200 kV$_{\text{max}}$ and 0.5 mm Cu + 1.0 mm Al (deep therapy).

This automatic selection of the treatment technique has only been made possible by the automatic stabilization of the tube voltage described here. It may be very convenient for the operator, since he already has to determine so many variables for each treatment: tube voltage, tube current, irradiation time, filter, focus-skin distance, size of the field to be irradiated, and the direction of the beam with respect to the object.

Applicators and filters

The focus-skin distance and size of field chosen can easily be obtained with the aid of one of the series of 15 applicators (cones), supplied with the installation (see fig. 1), two of which are shown in *fig. 11.* Such an applicator [6]) contains, in the circular flange with which it is attached to the apparatus, a lead diaphragm which confines the useful beam to the dimensions required for the desired field. The applicators have three different lengths corresponding to the most commonly used

[6]) The applicators have been made according to the directions of Dr. G. J. van der Plaats, Maastricht.

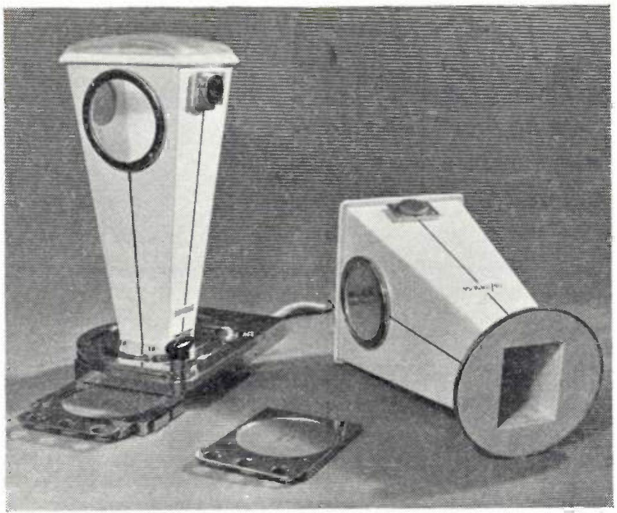


Fig. 11. Two applicators, both for a focus-skin distance of 50 cm; the one on the right, turned over to show the lead diaphragm in the flange, gives a field of 15 cm × 20 cm; the upright applicator on the left, attached to the filter-holder, which is normally screwed onto the tank in front of the window, gives a field of 10 cm × 15 cm. In the filter-holder is a filter pushed half-way in, whilst another filter lies beside it.

focus-skin distances of 30, 40 and 50 cm. They can be placed direct against the patient's sk.n with the free end, which is closed with a convex plastic cap; if necessary compression of the patient is used, the right focus-skin distance being retained. The vertex angle of the sides of each applicator is so chosen that every point of the field to be irradiated (and only of that field) faces the whole of the focus. Points of the penumbra of the diaphragm still face (through the wall of the applicators) part of the focus and would thus also be irradiated, though with reduced intensity. To avoid this the sides of the applicators are lined with lead. The largest have loteral windows, shut off with lead glass, and openings for inserting an ionization chamber for measuring the dosage.

Mention has already been made of the use of certain filters in the beam. As may be well known, the object of these filters is to increase the penetration of the rays applied, by suppressing the soft components in the X-ray spectrum; see *fig. 12*. Of course the radiation cannot be made any harder than that which corresponds to the short-wave limit of the spectrum, this being determined by the tube voltage. Moreover, with a given voltage one cannot choose too heavy a filter, because the filter also weakens the hard components

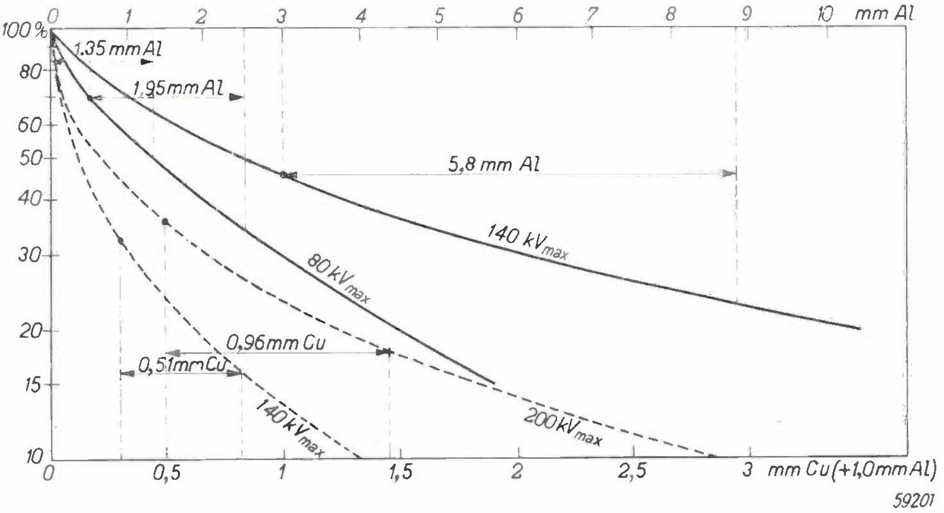


Fig. 12. When the rays from the "Compactix" apparatus, produced with tube voltages of 80, 140 and 200 kV$_{max}$, are passed through a layer of aluminium or copper their intensity is reduced according to the curves given (fully-drawn lines Al, dotted lines Cu + 1.0 mm Al). It is seen that by passing through the layer the rays are not only weakened but also become harder: the greater the thickness of the layer (filter) the smaller is the slope of the curve. The degree of penetration is denoted by the "half-value layer", i.e. the layer of Al or Cu which reduces the intensity of the already filtered rays to one half; up to 140 kV aluminium filters are used and the half-value layer is given in millimetres Al, whilst above 140 kV copper filters are used and the half-value layer is expressed in millimetres Cu. The non-filtered radiation from the "Compactix" apparatus produced with 80 kV$_{max}$ has a "half-value layer" H = 1.35 mm Al (for different kinds of therapeutic apparatus this value varies according to the form of voltage applied and the inherent filtration). With an extra filter of 0.5 mm Al the value of H rises to 1.95 mm Al. With 140 kV$_{max}$ and a 3.0 mm Al filter H = 5.8 mm Al, with 410 kV$_{max}$ and a 0.3 mm Cu + 1.0 Al filter H = 0.51 mm Cu. Finally, with 200 kV$_{max}$ and a 0.5 mm Cu + 1.0 mm Al filter H = 0.96 mm Cu.

and the intensity must be kept sufficient. For each tube voltage, therefore, only a few of the series of normal filter values can be considered. For use with the "Compactix" apparatus a series of ten different filters are supplied with which all the usual combinations of voltages (up to 200 kV) and filters can be obtained; see the table given below.

Table I. Filters for the "Compactix" apparatus. A filter has to be inserted for any treatment, even when for a particular treatment the "inherent filtration" of the apparatus would be considered sufficient; in the latter case filter number 1 has to be used, thus avoiding the possibility of the user forgetting to select a filter and thus by accident working without an extra filter. Filter number 10 is used for cutting off the radiation when the operating staff have to be near the apparatus while it is working. Filter No. 9 is the well-known Thoraeus filter, which has a smaller total absorption than an equivalent copper filter, as regards penetration of the rays.

No.	Composition and thickness
1	0.0 mm —
2	0.5 mm Al
3	1.0 mm Al
4	2.0 mm Al
5	3.0 mm Al
6	0.3 mm Cu + 1.0 mm Al
7	0.5 mm Cu + 1.0 mm Al
8	1.0 mm Cu + 1.0 mm Al
9	0.4 mm Sn + 0.25 mm Cu + 1.0 Al
10	5.0 mm Pb

Each filter is in a metal frame which slides into the slot of the filter-holder (see fig. 11). This frame, which is screwed onto the tank window, contains a click-knob mechanism which holds the filter frame in place when it is pushed in (against a pressure spring) or releases it when the knob is pressed. Further the filter-holder contains a set of contacts which are closed by the filter frames in different combinations and which perform two functions: firstly, the interlocking of the high tension, which cannot be switched on if one should have forgotten to insert a filter; secondly, the signalling to the control desk, each filter causing one of a row of ten pilot lamps to light up so that the operator can see which filter has been inserted.

We shall briefly explain the influence of the penetration of the rays (ray quality). When treating affections of the skin one has to save the underlying tissues and thus use soft rays which are almost entirely absorbed in the outermost layers of the body. For this superficial therapy one therefore chooses the lowest voltage, 80 kV$_{max}$, without filter, or, to be more exact, with filter No. 1, which does not contain any absorbent material, so that only the inherent filtration of the apparatus is operative. On the other hand, when treating deep-seated tumours hard rays are desired in order to get a reasonably large "depth dose", that is to say a high ratio of the intensity at the given depth underneath the skin to the intensity on the skin itself (in normal cases, owing to the absorption in the body and the fact that the intensity decreases according to the square of the distance as the latter becomes shorter, this ratio of intensity is less than 1). For such a case of deep therapy, therefore, one will choose a high tube voltage and a heavy filter, higher and heavier according to the depth at which the object lies below the skin. The obvious question is what improvement is reached in the depth quotient with increasing penetration of the rays. An answer to this question is given in *fig. 13*, where the depth dose at three different depths is plotted as a function of the penetration for a certain focus-skin distance and a certain size of field. As measure for the hardness use has been made of the "half-value layer" defined in the legend of fig. 12. We see that up to about 0.8 mm Cu the depth dose rises sharply. With higher half-value layers the further gain in the depth dose is relatively little and of scarcely any consequence compared with the much greater gain that can be obtained in many cases by compression of the tissues, thus artificially

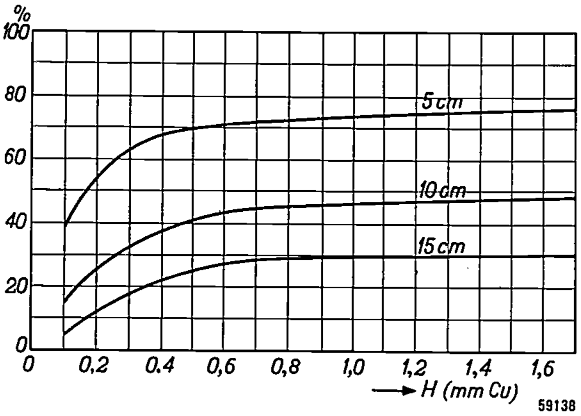


Fig. 13. The depth dose at three different depths underneath the skin as a function of the half-value layer H of the radiation. The focus-skin distance, which affects the depth dose according to the square of the decrease in intensity, is taken here to be 50 cm. The size of the field likewise affects the depth dose, since the secondary radiation at depths comparable with the field dimensions tends to equalize the intensity; these curves were plotted for a field of 400 cm². (Derived from L. Greber and K. Nitzge, Tabellen zur Dosierung der Röntgenstrahlen, Urban & Schwarzenberg, Berlin and Vienna, 1930.)

reducing the depth of the tumour. The maximum half-value layer that can be reached in practice with the "Compactix" is 0.96 mm Cu, obtained with 200 kV_{max} and a filter of 0.5 mm Cu + 1.0 mm Al, whereby the dosage ratc with 10 mA and at a distance of 50 cm amounts to 20 roentgen per minute.

Only with tube voltages of an entirely different order, for instance of some millions of volts, essentially greater depth doses, even considerably greater than 1 can be obtained. The secondary X-rays and electrons generated in the human body, and which contribute towards the total dosage, become more and more directed forward according as the tube voltage is increased. Thus in the dosage curve, plotted as a function of the depth below the skin, a maximum is found at a certain depth. Such a maximum is already to be found at voltages of 200 kV but it is still little pronounced and lies at no more than about 1 cm below the surface of the skin; with very high voltages the maximum is fairly high and may come to lie at depths of 10 cm or more.

Summary. A description is given of the "Compactix" apparatus for X-ray therapy, which works with alternating voltages up to 200 kV_{max} and a tube current up to 10 mA. The high-tension generator is mounted together with the X-ray tube in a cylindrical tank 1.20 m long and weighing 125 kg (275 lbs). The tank is mounted on a stand which permits five degrees of freedom of movement and is easy to adjust for any type of treatment. Some of the constructional features are: separation of the high-tension transformer into two uuits for 100 kV placed at either end of the X-ray tube, so that the focus of the tube lies in the centre of gravity of thc tank; there are three separate compartments in the tank, for the tube and the two transformers; the tube is insulated and cooled with freely circulating oil; the transformers are insulated with paper impregnated in oil; the oil in all three compartments is cooled by means of earthed cooling-water spirals which are directly connected to the water mains, thus dispensing with oil pumps or suchlike; the operator is fully protected against X-rays outside the effective applicator by a tungsten hood on the anode and a layer of lead 1 mm thick in the shield, with a total lead equivalent of 5 mm lead; high-tension transformer supply with "inverse voltage suppression"; tube-voltage measuring with automatic correction for the effect of the tube current; stabilization of tube current and tube voltage within about 1%, with special regulating circuits directly controlled by the current or the voltage; on the control desk, in addition to the normal current and voltage regulators, there are three keys for selecting three standard ray qualities, and signalling of the filter in use. Supplied with the apparatus is a series of 15 applicators, making it easy to limit the focus-skin distance and the size of field to be irradiated, and also a series of 10 filters. The maximum half-value layer obtained with a tube voltage of 200 kV_{max} and the filter of 0.5 mm copper plus 1.0 mm aluminium amounts to 0.96 mm Cu. With a tube current of 10 mA the dosage intensity at 50 cm distance is then 20 roentgen per minute.

Erratum

In the article by H. Rinia, D. Kleis and M. van Tol, Experimental transmitting and receiving equipment for high-speed facsimile transmission, I. General, published in Philips Techn. Review **10**, 189-195, 1949 (No. 7) the term photostat paper was erroneously mentioned on page 193; this should read dye-line paper.

A MILLIVOLTMETER FOR THE FREQUENCY RANGE FROM 1000 TO 30 × 10⁶ c/s

by H. J. LINDENHOVIUS, G. ARBELET and J. C. van der BREGGEN.

621.317.725: 621.3.027.213.7:
621.3.027.4: 621.3.029.45: 621.3.029.5

For taking measurements of electrical apparatus in which voltages of widely divergent amplitude and frequency occur, such as in radio transmitters and receivers, carrier-telephony installations, etc., there is need for a voltmeter of a high but reducible sensitivity and having a frequency range extending from audio-frequencies to radio-frequencies of some tens of Mc/s. The measuring instrument described here, which in essence consists of a variable attenuator, an amplifier, a rectifier and a moving-coil meter, has been designed to provide for this need.

For measuring alternating voltages according to a system which has been known for a long time the voltage is rectified (with the aid of a diode or a crystal detector) and this rectified voltage is measured with a moving-coil instrument. One of the advantages of such a system is that within very wide limits the reading can be made independent of the frequency; steps can be taken, for instance, to cover a frequency range from 20 c/s to about 500 × 10^6 c/s. On the other hand, however, meters working according to this system are relatively insensitive: for full-scale deflection a voltage is required of the order of at least 1 V (we shall revert to this point later). It is obvious that in order to be able to measure also smaller voltages an amplifier should be connected in front of the meter. This idea has been put into practice in instruments which give full deflection at only a fraction of 1 mV, but with the instruments so far placed on the market this very high sensitivity has been obtained at the cost of the frequency range, which then covers a bandwidth of only about 10,000 c/s.

Here an instrument will be described (type GM 6006) in which the amplifier is of such a construction as to pass a very wide frequency range (from 1000 to 30×10^6 c/s), whilst the full deflection is obtained with an input voltage of 1 mV. The amplification amounts to about 1500 and in order to keep it independent of the frequency in such a wide frequency range the amplification per stage has to be small, as will be seen later on, so that a rather large number of stages are required. This involves problems such as stabilization of the amplification, reduction of noise and the designing of a simple device with which the instrument can be calibrated at any moment.

Fig. 1 is a photograph of the instrument in question; it is provided with an attenuator — to be described farther on — which makes it possible to measure voltages up to 1 kV.

The amplifier

Upper limit of the frequency range

The amplifier has six stages, each equipped with an EF 42 pentode. Except for some minor points the first five stages are identical. The sixth stage, the load of which differs from that of the preceding ones and which is therefore built differently, will not be considered for the moment.

In *fig. 2a* a diagram is given of one of the first five amplifying stages (omitting the direct voltage sources). In fig. 2*b* we have the same diagram but showing only those elements which essentially determine the amplification at high frequencies, viz: the valve, the anode resistor R, the reactance coil L (the object of this will be shown later), and the capacitances C_1 and C_2 consisting respectively of the output capacitance of the preceding valve and the input capacitance of the following valve, plus some stray capacitance.

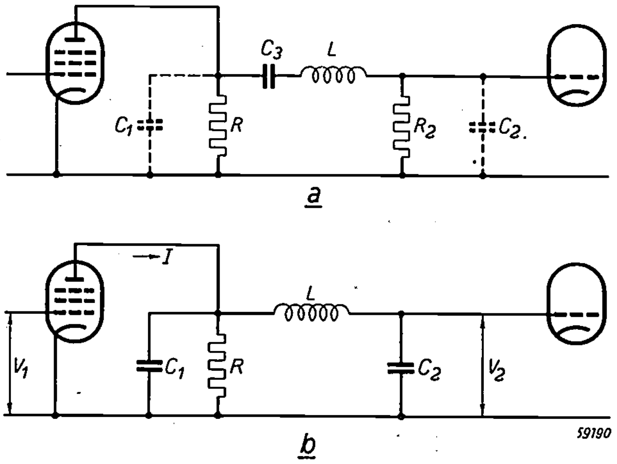


Fig. 2. *a*) One of the first five amplifying stages. R anode resistor. R_2 = grid resistor, C_1 = output capacitance of the preceding valve plus some stray capacitance, C_2 = input capacitance of the following valve plus some stray capacitance, C_3 = coupling capacitor, L = reactance coil for raising the upper frequency limit. *b*) The same but omitting the elements which do not affect the amplification at high frequencies.

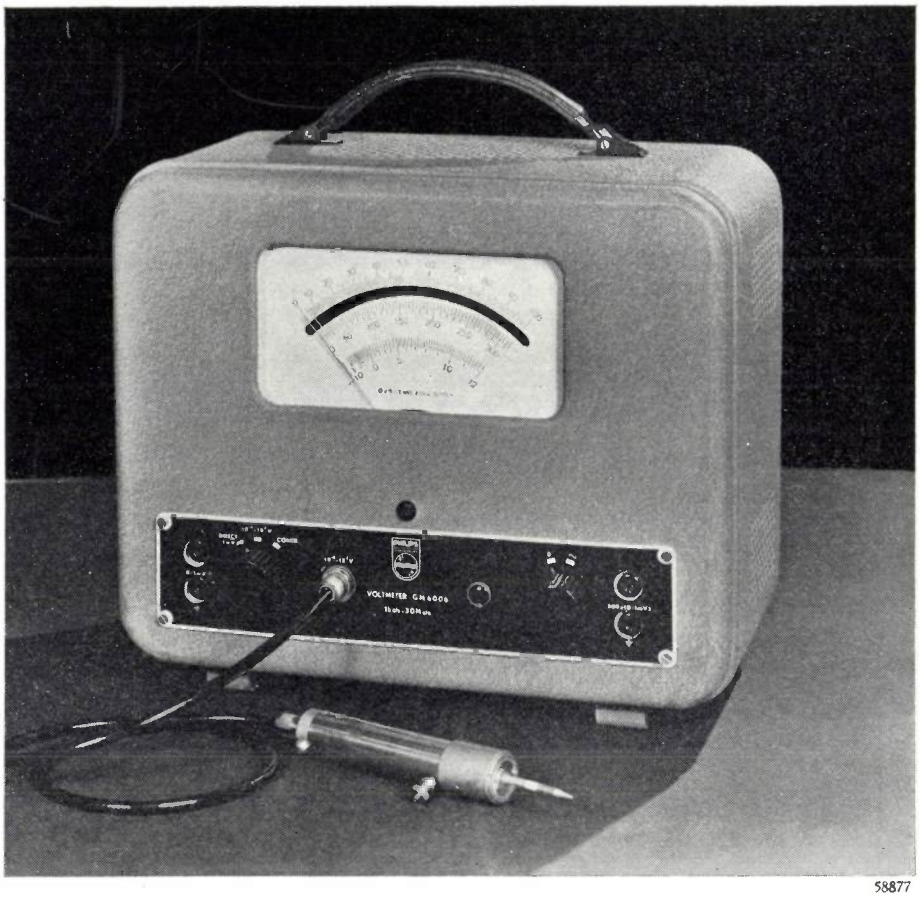


Fig. 1. Millivoltmeter type GM 6006, with the attenuator in the foreground. On the control panel from left to right: terminal sockets for an input voltage of max. 1 mV, change-over switch, plug socket to which the attenuator is connected, pilot lamp, mains switch, and terminal sockets from which a voltage can be taken (max. 0.5 V) that is 500 times as great as the input voltage. The change-over switch has three positions: (1) "Direct" (attenuator out of action, for voltages up to 1 mV), (2) "10^{-3}-10^3 V" (with attenuator) and (3) "Contr." (calibrating position).

The top and middle scales differ mutually by a factor $\sqrt{10}$; it depends upon the position of the attenuator which of these scales has to be used and by how many factors of 10 the reading has to be multiplied to give the voltage to be measured in millivolts or volts. The bottom scale is calibrated in decibels; 0 db lies at 0.775 V (corresponding to 1 mW in a resistance of 600 ohms).

For the case where $L = 0$ the relation between the alternating voltage V_2 (with angular frequency ω) at the grid of the right-hand valve and the anode alternating current I of the left-hand valve (V_2 and I in absolute values) is

$$V_2 = \frac{R}{\sqrt{1 + (\omega C_p R)^2}} \cdot I,$$

where $C_p = C_1 + C_2$.

Taking V_1 as representing the alternating voltage at the grid of the left-hand valve and S the mutual conductance of that valve, then $I = SV_1$, so that when we put $\omega C_p R = x$ the amplification V_2/V_1 of this stage is:

$$\frac{V_2}{V_1} = \frac{SR}{\sqrt{1 + x^2}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

In *fig. 3* the dotted curve represents $V_2/V_1SR = 1/\sqrt{1 + x^2}$ as a function of x, the latter being a measure of the frequency. It shows the well-known phenomenon of the amplification diminishing as the frequency is raised, due to the capacitance parallel to R.

Allowing for a moment a drop of say 5% per stage (which for six stages means a drop of 27% in the total amplification), then the dotted curve in fig. 3 shows that this limit is reached at $x = \omega C_p \mathrm{R} = 0.33$. To get a high upper frequency limit C_p and R have to be kept small. Now even with a proper choice and judicious mounting of the components C_p cannot be made smaller than about 20 pF, whilst the value of R is governed by the nominal amplification per stage ($= SR$) and the mutual conductance. The latter, in the case of the "Rimlock" pentode EF 42 employed, amounts to a maximum of 9.5 mA/V, but with a view to having some reserve and the possibility of adjustment these valves are adjusted for $S =$ 8.0-8.5 mA/V. Thus for an amplification of say 3 per stage — a lower factor is not likely to be desired — an anode resistance of 370 ohms is required. For the upper limit $f_{\max}$ of the frequency range we then find:

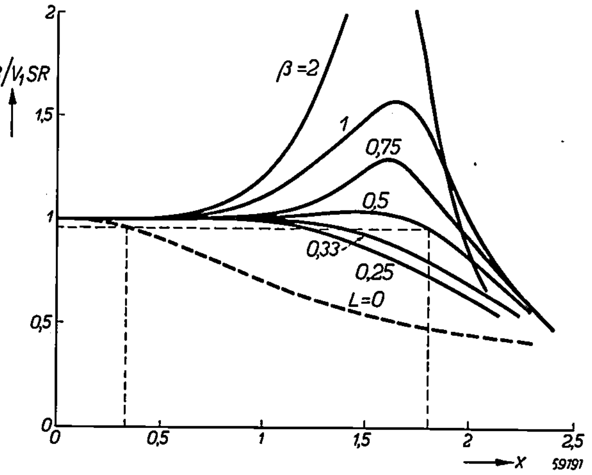


Fig. 3. V_2/V_1SR as a function of $x = \omega C_pR$. The dotted curve applies for $L = 0$ (formula (1), the full lines applying for L according to formula (3), with $\beta = C_1/C_2$ as parameter (formula (4)). On a certain scale these curves represent the amplification per stage (V_2/V_1) as a function of the frequency. The best characteristic is obtained with $\beta \approx 0.5$.

$$f_{\max} = 0.33/(2\pi \times 20 \times 10^{-12} \times 370) \text{ c/s} \approx 7 \text{ Mc/s}.$$

A much higher frequency limit can be reached by adding a reactance coil with a suitably chosen inductance L (fig. 2). A simple calculation gives for the system shown in fig. 2:

$$\frac{V_2}{V_1} = \frac{SR}{\sqrt{(1-\omega^2LC_2)^2 + (1-\omega^2LC_s)^2 \cdot \omega^2 {C_p}^2R^2}}, \quad (2)$$

where $C_s = C_1C_2/(C_1 + C_2) = C_1C_2/C_p$. Underneath the root sign we have, in addition to the constant 1, a term with ω^2, one with ω^4 and another with ω^6. If ω is made to rise from a low value then it is in general mainly the term with ω^2 which causes the greatest change in the denominator of (2). This term is:

$$({C_p}^2R^2 - 2LC_2)\omega^2 .$$

It can be made to disappear by ensuring that ${C_p}^2R^2 = 2LC_2$, thus by choosing

$$L = \frac{{C_p}^2R^2}{2C_2} = \frac{1+\beta}{2} C_p R^2 , \quad . \quad . \quad . \quad (3)$$

where $\beta = C_1/C_2$. With this value for L eq. (2) becomes

$$\frac{V_2}{V_1} = \frac{SR}{\sqrt{1 + (\frac{1}{4} - \frac{\beta}{1+\beta})x^4 + \frac{1}{4}(\frac{\beta}{1+\beta})^2x^6}}, \quad . \quad (4)$$

in which again $x = \omega C_pR$.

The full lines in fig. 3 represent V_2/V_1SR as a function of x for several values of the parameter β. It is seen that the amplification is now constant within 5% up to a value of x amounting to about 1.8 when $\beta \approx 0.5$. In other words, by a suitable choice of L and of C_1/C_2 one can reach, with the same nominal amplification per stage (3) and the same valves, a frequency limit more than five times as high as without the reactance coil L, namely about 38 Mc/s.

It is therefore a question of getting the right ratio of the capacitances C_1 and C_2. As already stated, these consist mainly of the output capacitance of the preceding valve (4.5 pF) and the input capacitance of the following valve (9.5 pF) respectively. To this must be added the capacitance of the valve-holder and of some components, so that we find that $C_1 = 7$ pF and $C_2 = 10$ pF. Thus the ratio $\beta = 7/15 = 0.47$ happens to be very favourable.

If, having regard to the tolerance, we do not go to the extreme value of x (≈ 1.8) but to 1.6, and putting the upper frequency limit at 30 Mc/s, then with $C_p = 7 + 15 = 22$ pF we get for R a valve of 400 ohms. For the corresponding value of L equation (3) gives $L = 2.5$ μH. The amplification per stage amounts to $SR = 3.25$.

We now come to the last stage, to which is connected — instead of the input of an EF 42 valve as in the preceding stages — the rectifying circuit, which has a much lower capacitance, viz. $C_2 = 3.5$ pF instead of 15 pF. If this stage were arranged in the same way as the preceding ones then we should have $\beta = 2$ (C_1 is again 7 pF) and, as can be seen from fig. 3, this would cause a high peak in the frequency characteristic. C_2 could be raised to 15 pF by shunting a small capacitor across it, thus making conditions equal to those in the preceding stages. However, there is a better method that can be followed, by means of which a greater amplification can be obtained. For this purpose the cricuit of the sixth stage is slightly altered: the coil L is placed in front of the resistor R, now denoted by R_1 (*fig. 4*, compare fig. 2). As can be readily checked, the formulae (2), (3) and (4) also hold for this case if C_1 and C_2 are interchanged (that is to say in (2) C_2 has to be replaced by C_1; C_p and C_s remain unaltered, and β is to be taken as C_2/C_1). We can now again use the curves of fig. 3, particularly that having the most favourable shape ($\beta \approx 0.5$).

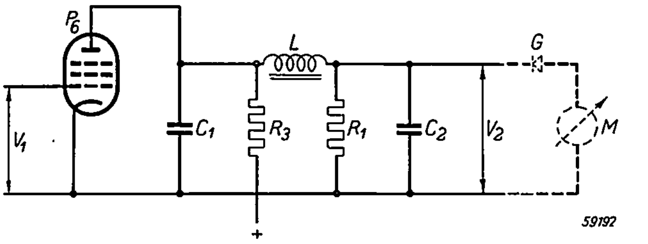


Fig. 4. To get a favourable value of β (fig. 3) in the network following the last amplifying valve (P_6) the reactance coil L precedes the resistor R, now termed R_1 (cf. fig. 2). R_3 = the resistor via which the anode is fed. The dotted lines indicate the rectifying unit (detector G, moving-coil meter M).

Taking again 1.6 as the highest value of x, then with 30 Mc/s as the frequency limit and $C_p = 7 + 3.5 = 10.5$ pF we find $R_1 = 800$ ohms. The amplification of the last stage, in which the EF 42 valve is also adjusted to a mutual conductance of about 8 mA/V, then amounts to $SR_1 \approx 6.5$.

In two respects the measures taken to keep the amplification constant up to the highest possible frequencies need supplementing.

1) The value of the capacitances C_1 and C_2, the ratio of which greatly affects the form of the frequency characteristics, cannot be governed very accurately, so that as a rule one stage will be working with a somewhat too high value of β and the other with a somewhat too low value. This could be remedied by adding trimmers. In practice it has been found that one. trimmer suffices (across C_2 in the last stage). As an inevitable result C_p becomes somewhat greater than the above-mentioned value of 10.5 pF, and for that reason a smaller value has been chosen for R_1, viz. 510 ohms. Thus the amplification of the sixth stage becomes $SR_1 = 4.2$; the total amplification is therefore $3.25^5 \times 4.2 = 1500$.

It has also been found desirable to make the self-inductance L of the last stage adjustable. This coil has therefore been provided with a sliding core of "Ferroxcube", a material of high permeability and low losses [1]).

2) It has already been remarked in passing that a drop of 5% in the amplification per stage results in a drop of more than 25% in the total amplification (six stages). We shall now remove any impression that it may have been left at that. To avoid this drop, which would become particularly noticeable in the frequency range between 20 and 30 Mc/s, a coil with variable self-inductance L_1 has been incorporated in one of the anode circuits in series with the resistor R. The amplification of this stage is then approximately given by formula (4) if R is replaced in the numerator by $\sqrt{R^2 + \omega^2 L_1^2}$. The value of L_1 can be so chosen that as the frequency rises the numerator increases at a higher rate than the denominator, so that a rising frequency characteristic is obtained. Within a limited frequency range the decline in the amplification of the other stages can thus be compensated.

This measure has been taken in only one stage for the following reasons. If a self-inductance L_1 were to be placed in series with R in each of the six stages (it would then have to compensate a drop of only about 5% in the amplification) it would have to be so impractically small that it would hardly be possible to make these coils with the required accuracy.

This coil L_1 likewise has a sliding core of "Ferroxcube", with which the self-inductance can be adjusted to the optimum value.

The result of all this is to be seen from *fig. 5*, curve *I*, representing the deflection of the instrument as a function of the frequency with a constant input voltage. The small fluctuations in the characteristic between 10 and 30 Mc/s are due to small differences in the value of β in the various stages.

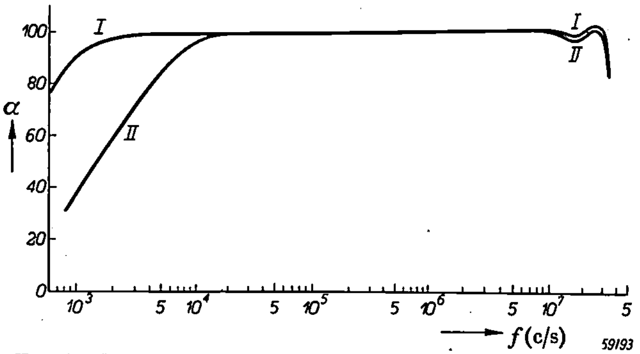


Fig. 5. The deflection α of the moving-coil meter as a function of the frequency f of the constant, sinusoidal, input voltage, I without attenuator, II with attenuator (in all positions).

The fact that at low frequencies the characteristic does not follow a horizontal line like that of fig. 3 is explained in the following paragraph.

Lower limit of the frequency range

The circuit elements essentially determining the shape of the frequency characteristic at low frequencies are — as will be seen — the coupling capacitors C_3 and C_4 and the resistors R_2 and R_3 (see *fig. 6*, obtained from fig. 2*a* by omitting the elements which are of no importance for the low frequencies, such as the capacitances C_1 and C_2 and the self-inductance L).

[1]) J. L. Snoek, Non-metallic magnetic material for high frequencies, Philips Techn. Review. 8, 353-360, 1946.

C_3 and R_2 — a combination which occurs at the control grid of each of the six valves — form a voltage divider which gives for the ratio (in absolute value) of the voltages V_2 and V_1 (fig. 6):

$$\frac{V_2}{V_1}=\frac{RR_2}{R+R_2}\cdot S\cdot\left[\frac{1}{\sqrt{1+\left\{\frac{1}{\omega C_3(R+R_2)}\right\}^2}}\right]. \quad (5)$$

For frequencies at which $\omega C_3\ (R+R_2) \gg 1$, the term between the square brackets is practically 1, but for lower frequencies it is less than 1 and is frequency-dependent.

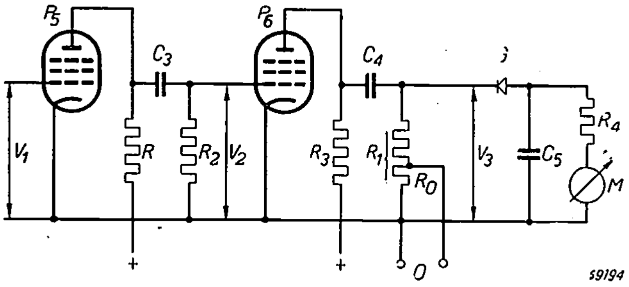


Fig. 6. The last and last but one amplifying stages and the rectifying circuit (the direct-current sources and the elements which are only of importance at high frequencies have been omitted). P_5 = fifth, P_6 = sixth amplifying valve, C_3 and C_4 = coupling capacitors, C_5 = smoothing capacitor, G = crystal detector, R and R_1 = anode resistors, R_2 = grid leak, R_3 = anode supply resistor of P_6, R_4 = series resistor of the moving-coil meter M, R_0 = the part of R_1 to which the terminals O are connected. The relaxation times C_3R_2 and C_4R_3 determine the shape of the frequency characteristic at low frequencies.

The lower frequency limit, at which V_2/V_1 has a still admissible value less than 1, can in principle be fixed as low as desired by giving C_3 and $R+R_2$ sufficently large values. Neither of these, however, can be raised indefinitely. C_3 is limited because a larger capacitance requires a capacitor of larger dimensions having a greater capacitance with respect to the surroundings; in that case an excessive lowering of the lower frequency limit would be accompanied by a lowering of the upper frequency limit. And as regards $R+R_2$ the value of this is limited by the maximum permissible value of the grid resistance R_2 (1 megohm in the case of the valve EF 42); the resistance R, which, as we have seen, amounts to no more than 400 ohms, is unimportant in this respect.

Similar considerations hold for the last stage. Again denoting the input voltage of the last valve by V_2 and the output voltage (across R_1, fig. 6) by V_3, then, in analogy with formula (5), we have:

$$\frac{V_3}{V_2}=\frac{R_1R_3}{R_1+R_3}\cdot S\cdot\frac{1}{\sqrt{1+\left\{\frac{1}{\omega C_4(R_1+R_3)}\right\}^2}},$$

where R_1, R_3 and C_4 take the place of R, R_2 and C_3. R_1 has the afore-mentioned value of about 500 ohms and is thus of little consequence in comparison with the resistor R_3 via which the anode of P_6 is fed, the value of which may be for instance 10,000 ohms but not much more in view of the D.C. voltage drop occurring in this resistor.

In the main, therefore, the deviation from the nominal amplification at low frequencies is determined by the relaxation times C_3R_2 and C_4R_3. With

C_3 in the order of 1000 pF,
C_4 in the order of 0.1 μF,
R_2 in the order of 1 MΩ and
R_3 in the order of 10 000 Ω

the apparatus has the characteristic represented by curve I in fig. 5, which at 1000 c/s deviates from the nominal level by only a few percent.

Stabilizing the amplification

The total amplification is proportional to the continued product of the mutual conductance of the six amplifying valves. Without special measures being taken the amplification would therefore depend greatly upon factors affecting the mutual conductance, i.e. the amplitude of the supply voltages and the emissive power of the cathodes.

To get a practically constant amplification we have started from the fact that the mutual conductance of a valve is strongly correlated with the mean value of the cathode current, so that as far as the mutual conductance is concerned it is immaterial, to a first approximation, whether a certain mean cathode current is obtained with a strongly negative control-grid voltage and a high screen-grid voltage or with a slightly negative control-grid voltage and a low screen-grid voltage. It is therefore a matter of keeping the mean cathode current constant. With this in view a strong direct-current feedback has been provided (24-fold per stage) by means of cathode resistors of exceptionally large value (2700 ohms; with the EF 42 resistors of 160 ohms are normally used). Each of the cathode resistors is shunted by a large capacitor, so that for alternating voltages with frequencies higher than 1000 c/s the feedback is inactive and, as a consequence, in the measuring range the alternating-voltage amplication is not thereby reduced.

Since the voltage drop in these cathode resistors is much greater than the bias required in the control-grid circuits, an adequate positive direct voltage is also applied to these circuits so that the right bias is obtained. This positive direct voltage is derived from the supply unit via a potentiometer and kept constant with a stabilizing tube (type 85 A1).

As a result of these measures the cathode current of the EF 42 valves and thus also their mutual conductance undergo but little change as a consequence of fluctuations in the supply voltage or the ageing or replacement of the valves. With a mains voltage fluctuation of 5% the variation in the total amplification is likewise 5% (in the opposite sense). This remaining inconstancy of the amplification is due to the fact that the point from which we started, as already remarked, is only approximately correct: the mutual conductance depends mainly but not exclusively upon the mean cathode current.

The potentiometer just referred to, which is adjustable by means of a small screw sunk into one of the side panels of the cabinet, allows the calibration to be restored, if necessary, by adjusting the bias in all six grid circuits simultaneously when calibrating. In the event, for instance, of one of the valves being replaced by another with a 5% higher mutual conductance, then the corresponding increase of the amplification is compensated by reducing the mutual conductance of the other five valves by 1%; this requires only a slight alteration of the bias.

Noise

In the designing of an amplifier with such a wide frequency band as that in the present case attention has to be given to the noise.

In an amplifier there are two kinds of elements forming a source of noise: resistors and valves. Let us first take the valves.

For the R.M.S. value V of the noise voltage at the input of a valve in a frequency interval Δf we have [2]):

$$V = \sqrt{4\,k\,T\,R_{\mathrm{eq}} \cdot \Delta f},$$

where k is the Boltzmann constant $= 1.38 \times 10^{-23}$ J/°K, T is the temperature in °K and R_{eq} is the equivalent noise resistance of the valve at room temperature. For the EF 42 valve $R_{\mathrm{eq}} = 750$ ohms. With $\Delta f = 30 \times 10^6$ c/s we find $V = 19$ μV. Thus the noise of the valve in the first stage can be reckoned with as a noise voltage of 19 μV at the input of the first stage; likewise the noise of the second stage can be reckoned with as a noise voltage of the same value at the input of this second stage. The latter, in the case of a three-fold amplification per stage, is equivalent to $19/3 = 6.3$ μV at the input of the first stage. Thus the noise of these two valves together corresponds to a noise voltage of $\sqrt{19^2 + 6.3^2} = 20$ μV at the input of the first stage. The noise contribution of the third and later valves is so small that it can well be ignored.

Similarly, in regard to the noise of the resistors in the amplifier, in the first instance we need only reckon with the resistance at the input of the amplifier. This resistance (the grid resistance of the first valve) amounts to 1 megohm, which in itself would yield a considerable noise voltage were it not for the fact that the input capacitance of the amplifier (about 20 pF) is shunted across the resistance. This capacitance together with the resistance forms a voltage divider which particularly attenuates the noise-voltage components with high frequencies. It is due to this that the remaining noise voltage originating in this resistance is small compared with the voltage of 20 μV for which the valves are responsible. The noise voltage of 20 μV causes a deflection of the meter which is only 2% of the full deflection (1 mV); this is so little that it can easily be compensated by sending through the moving-coil meter a weak counteracting direct current, which is drawn from the supply unit.

Use of the amplifier for other purposes

Since the amplifier with a characteristic like that of fig. 5 (*I*) can render good service for all sorts of purposes, the amplifier of the meter GM 6006 is provided with output terminals (*O*, fig. 6) from which a part of the output voltage can be taken. With an input voltage of 1 mV the voltage at *O* is 0.5 V. The resistance R_0 between the terminals *O* has a low value (180 ohms), so that fairly low impedances can be connected at *O* without causing any appreciable drop in the voltage. The meter continues to indicate the value of the input voltage.

The rectifying circuit

Choice of the detector

As already stated, the output voltage from the amplifier is rectified and the resultant direct voltage is fed to a moving-coil meter (fig. 6). A smoothing capacitor (C_5) is used, so that it is actually the peak value of the alternating voltage that is measured. The scale, however, is calibrated for the root-mean-square value of a sinusoidal voltage. Also when the voltage applied is not sinusoidal the peak value of this voltage is $\sqrt{2}$ times that of the meter reading.

[2]) See for instance M. Ziegler, Noise in amplifiers contributed by the valves, Philips Techn. Rev. 2, 329-333, 1937.

For the detection either a diode or a crystal detector (e.g. with germanium crystal) can be used. The latter has been chosen for the following reason.

The mere fact of a diode occurring in a circuit immediately causes some current to flow through that circuit even if no voltage source is connected to it. The value of this zero current is greatly dependent upon the temperature of the cathode (and thus upon the mains voltage.) It is therefore not easy to compensate the zero current without repeated readjustment. Only when there is a fairly large voltage in the diode circuit (at least 3 to 4 V for full deflection) is little trouble experienced from the zero current. With a crystal detector, on the other hand, there is no zero current. Nevertheless also in this case it is advisable to work with not too low a voltage, with a view to the reproducibility of the characteristic at small voltages. For full deflection 1.5 V is sufficient. With the same sensitivity the amplification in this case can thus be two to three times as small as in the case of a diode.

Overloading

With an input voltage of 1 mV at the amplifier a current of 100 μA flows through the moving-coil meter. When the input voltage is increased a state of overloading of the last valve is very soon reached, where the output voltage rises less than proportionately with the input voltage and finally approaches a constant level. The rectifier current thereby increases up to about 400 μA, a value which both the meter and the crystal can well withstand (in contrast to a thermocouple, for instance, which, although it is otherwise a useful instrument for converting the output alternating current of an amplifier into a direct current, has the great disadvantage of not being able to withstand overloading). There is, therefore, no risk of the instrument being damaged through overloading, unless it were to be connected direct to a voltage of several hundreds of volts without an attenuator, when the input resistor might become too hot or the input grid capacitor break down; this does not occur, however, until the overload exceeds a hundred thousand times the nominal rating!

Internal calibration

With the left-hand switch (fig. 1) in the position "Contr." an alternating voltage of 1 mV is put across the input of the amplifier, so that there should then be a full deflection of the meter. Any deviations can be corrected by means of the method already described for adjusting the bias of the amplifying valves.

The calibrating voltage of 1 mV is obtained by causing the amplifier to oscillate, this being done by connecting to the output a tuned circuit and feeding back to the input of the amplifier (*fig.* 7) an exactly fixed fraction of the voltage of that circuit in the right phase. When the amplifier is switched on there arises in the tuned circuit an increasing alternating voltage which is checked by a diode in series with a direct voltage source (threshold voltage) shunted across that circuit. So long as the peak of the alternating voltage is smaller than the threshold voltage the parallel branch has no effect, but as soon as the peak exceeds the threshold voltage the tuned circuit is heavily damped. The result is that the voltage amplitude of the circuit is limited to (practically) the threshold voltage. The latter (about 7 V) is obtained by voltage-division from the constant voltage of the stabilizing tube already mentioned (85 A1).

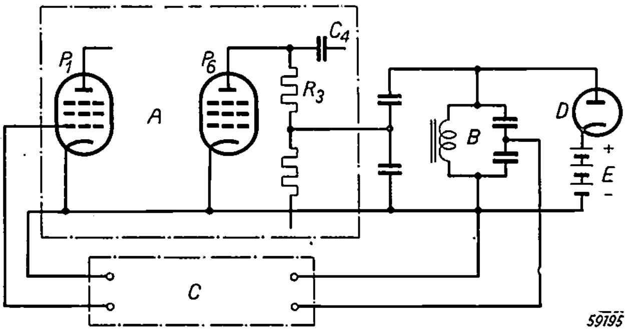


Fig. 7. Calibrating circuit. Part of the voltage across the tuned circuit *B* which is connected to the output of the amplifier *A* is fed back via a network *C* to the input of the amplifier *A*. This feedback makes the circuit oscillate. Through the action of the diode *D* the amplitude of the voltage across *B* is limited to the constant value *E*, which is so chosen that the input voltage of *A* is exactly 1 mV. R_3 and C_4 are as in fig. 6.

The frequency of the oscillation is about 5000 c/s, thus being in a range in which the amplification does not depend upon frequency (see fig. 5).

The tuned circuit has a high quality factor (impedance 3 megohms); the coil has a core of "Ferroxcube".

The attenuator

For measuring voltages higher than 1 mV use is made of an attenuator preceding the amplifier. The impedance of this attenuator together with the input impedance of the amplifier forms a voltage divider. For frequencies above 25,000 c/s the latter impedance is almost purely capacitive. By using a capacitor for the attenuator one therefore gets an attenuation which in this range is practically independent of the frequency. Below 25,000 c/s, however, the lower the frequency the more the input impedance of the amplifier assumes the character of a resistance, so that the attenuation then becomes frequency-dependent (see curve *II*, fig. 5).

The attenuator of the voltmeter GM 6006 (*fig. 8*) consists of a capacitor which is variable in twelve steps and with which the following degrees of sensitivity can be obtained: full deflection at 1 mV, 10 mV, 31.6 mV, 0.1 V, 0.316 V, 1 V, 3.16 V, 10 V, 31.6 V, 100 V, 316 V, 1 kV. The size of the steps is therefore a factor $\sqrt{10}$, except for the first step, the size of which is a factor 10. The attenuator is made in the form of a sliding capacitor and is mounted in a "probe" connected to the amplifier by means of a cable.

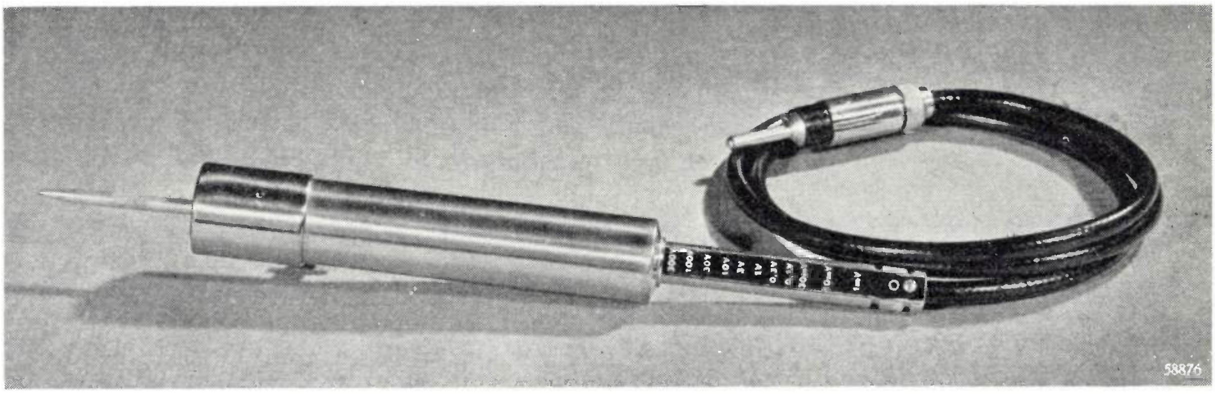


Fig. 8. Capacitive attenuator with cable. The rod slides in and out to vary the sensitivity. The maximum sensitivity corresponds to 1 mV and the minimum to 1 kV for full deflection. Stops hold the rod in each of the 12 positions.

The fixed electrode of the sliding capacitor is connected to an external contact pin and the sliding electrode to the core of the cable. The sliding electrode is fixed in each of the 12 positions by a stopping device. The position is read from a graduated scale following the movement of the sliding electrode (fig. 8; the number 31.6 is rounded off on this scale to 30, 0.316 is rounded off to 0.30, and so on).

The aim has been to construct the capacitor in such a way that the successive positions each differing in sensitivity by a factor $\sqrt{10}$ [3]) are obtained by equal movements of the slide, since then the influence of a given inaccuracy is constant at each stop. This implies that the capacitance between the electrodes must be an exponential function of the distance between those electrodes. This has been reached in the following way.

If two disc-shaped electrodes are placed in an earthed tube (*fig. 9*) and the distance l between the electrodes is at least several times the diameter D of the tube, then as l increases the capacitance between the electrodes diminishes approximately exponentially.

The configuration illustrated in fig. 9 may be regarded as a wave guide. Waves which are longer than a certain critical wavelength cannot be propagated through a wave guide [4]); the amplitude of such waves decreases exponentially with the depth to which the wave has penetrated into the pipe [5]).

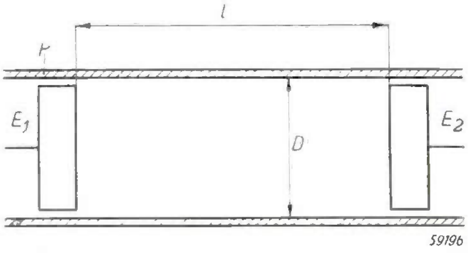


Fig. 9. The capacitance between two electrodes (E_1, E_2) in an earthed tube P is to a good approximation an exponential function of the distance l, provided l is at least several times as large as D.

To change the capacitance by a factor 10 an axial displacement of 0.48 D is required. When, however, the electrodes are brought so close together that $l \approx D$, or $l < D$, the capacitance is also strongly influenced by the shape of the electrodes.

[3]) The position for 1 mV, where the electrodes are short-circuited, is not considered here.

[4]) See for instance W. Opechowski, Electromagnetic waves in wave guides, II, Philips Techn. Rev. **10**. 46-54, 1948 (No. 2), in particular page 52.

[5]) A mechanical model with which this can be demonstrated for rectangular wave guides was recently described by K. S. Knol and G. Diemer in "A model for studying electromagnetic waves in rectangular wave guides", Philips Techn. Rev. **11**, 156-163, 1949 (No. 5), figs. 4 and 5.

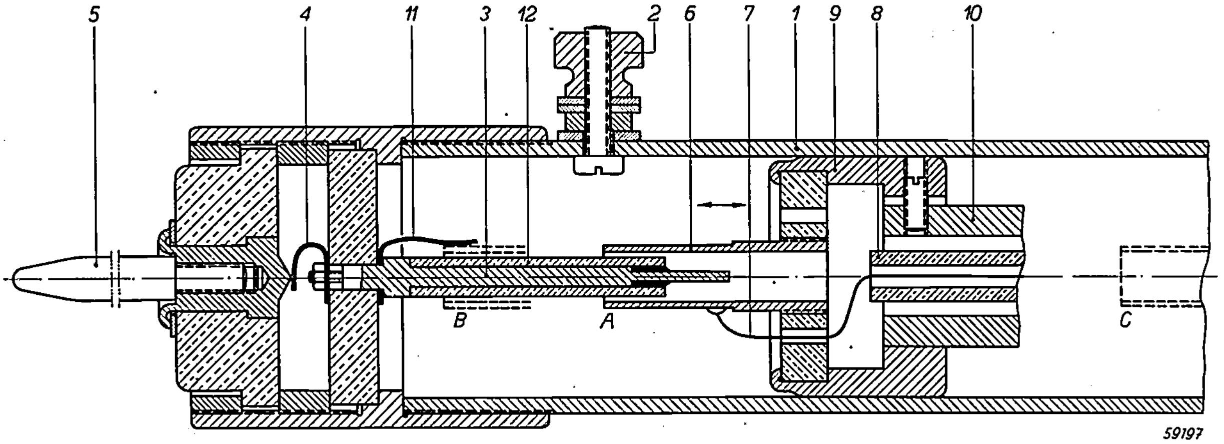


Fig. 10. Longitudinal cross section of the capacitive attenuator (slightly simplified; about 1.6 times the actual size). *1* = metal tube that has to be earthed by means of the screw *2*. *3* = fixed electrode connected via the leaf spring *4* to the contact pin *5*. *6* = tubular sliding electrode connected to the core *7* of the cable *8*. *6* is attached with intermediate insulation to the piston *9* in the tube *1*. The piston is attached to a rod *10* which is extended outside the tube *1* and which is marked with a calibrated scale (see fig. 8); acting upon this rod is a stop not shown in the illustration (a small steel ball pressed into a hole by a spring). The position *A* of the sliding electrode is that for 30 mV. The dotted lines at *B* indicate one extreme position (1 mV), in which the spring *11* short-circuits the attenuator. In the po.ition indicated at *C* the distance between the electrodes is so great that the exponential law applies. To get approximately the same law also in a position like that at *A* the ceramic tube *12* has been fitted in and the fixed electrode given a thin and a thick part.

For low sensitivities (voltage to be measured between 1 kV and approx. 1 V) the distance between the electrodes in the attenuator is such that the exponential law applies. In order that this law may also apply to a good approximation for higher sensitivities, the electrodes have been given a special shape: the sliding electrode is tubular and in the positions for high sensitivity it slides concentrically round the rod-shaped, fixed electrode. To approximate the right variation in capacitance the fixed electrode is enveloped by a small ceramic cylinder and is thin at one end (see *fig. 10* representing a cross section of the attenuator); this does not affect the exponential variation for low sensitivities (great distance between the electrodes).

As may be seen from fig. 8, the graduation of the scale in steps of $\sqrt{10}$ has indeed in this way been made practically linear in sensitivity.

In the position for the highest sensitivity (1 mV) a contact spring provides for the short-circuiting of the electrodes. (When measuring with this sensitivity, however, it is better to make a direct connection to the terminals of the amplifier, thereby avoiding the capacitance of the cable). In the other positions the input capacitance of the attenuator with respect to earth is only 2.8 pF.

Even the slighest displacement of the fixed electrodes with respect to the sliding one would upset the calibration. Therefore, in order to avoid derangement of this electrode in the event of something knocking up against the contact pin the electrical connection between the fixed electrode and the contact pin is brought about by means of a weak spring (fig. 10).

In the cable connecting the attenuator to the amplifier, and which is 0.85 m long, owing to the presence of self-inductance and capacitance there would be a tendency, at the highest frequencies, to a certain boosting of the voltage, in consequence of which the voltage at the amplifier end of the cable would be higher than that at the beginning. The effect has been neutralized by shunting a suitably chosen damping resistor across the input of the amplifier.

Summary. A description is given of a millivoltmeter for a wide frequency range (10^3-30×10^6 c/s) which gives full deflection at an input voltage of 1 mV and with the aid of an attenuator has been made suitable for measuring voltages from 1 mV up to 1000 V. Incorporated in the apparatus is an amplifier with six EF 42 valves and an amplification factor of 1500, to which a moving-coil meter is connected via a crystal detector. In the amplifier special measures have been taken to get an amplification that depends as little as possible upon frequency, fluctuations in mains voltage and ageing of the valves. For the purpose of calibration the amplifier is made to oscillate, thereby setting up across the input an alternating voltage of 1 mV (fixed by a constant threshold voltage), so that the meter should then give the full deflection; any deviations are corrected by readjusting the bias at the control grids. The instrument can withstand heavy overloads. The attenuator is connected to the amplifier with a cable and consists of a sliding capacitor of special construction with 12 calibrated positions; it is of such a construction that with the same displacement the sensitivity changes by the same factor in each step. The amplifier can serve also for other purposes for amplifying 500 times an alternating voltage smaller than 1 mV.

AN APPLICATION OF GEIGER COUNTER TUBES FOR SPECTROCHEMICAL ANALYSIS

by O. G. KOPPIUS *).

621.385.842:545.82

It has long been known that Geiger counter tubes, using the photoemissive effect of the cathode metal, can be applied to the measurement of very feeble visible or ultraviolet radiation. As a direct indicating instrument for measuring radiation intensities, the counter tube could be expected to offer special advantages in the detection of traces of chemical elements by their characteristic spectral emission lines. A simple apparatus based on this principle was designed and for several years successfully utilized for the detection of lead in the atmosphere of industrial areas.

The detection of traces of an element by means of its characteristic spectral lines emitted in an electric arc or spark is rather old. Kirchhoff and Bunsen used this principle first in the isolation and discovery of caesium and rubidium; Gerlach established the so-called "internal standard method" which enabled the principle to be used for precise quantitative analyses. In the last decade many industrial applications have been made, for example in the routine analysis of iron and steel for calcium, silicon, manganese, chromium, magnesium, and other elements. With modern equipment, the method is capable of high speed and precision [1]).

The application of spectrochemical analysis which will be described in this article concerns the detection of lead in air. Atmospheric contamination by lead may occur in several branches of the chemical industry, due to small leaks in plant installations. Because of the well known toxic effect of lead, a maximum permissible lead concentration in air has been established in several countries. To make sure that the lead concentration in a plant does not exceed the limit of safety, a rapid means of analyzing air samples for lead is desirable.

When the problem presented itself in one of the E. I. du Pont de Nemours plants, spectrochemical analysis, because of its specificity and sensitivity, was considered to offer the best prospects for meeting the rather exacting requirements. A photographic technique was developed for the purpose [2]). In the plant where the purity of the air was to be examined, an electric spark was run continuously between two copper electrodes. The spectrum of the spark was photographed with the aid of a small quartz spectrograph. Part of the spectrum, consisting principally of copper lines, with a certain number of other lines attributable among other things to the water vapor content of the air, is shown in *fig. 1a*. If the air contains traces of a lead compound this is decomposed in the hot spark, and lead lines will appear in the spectrum. The most sensitive line, i.e. the line which appears at the lowest lead concentration, is the line at about 2203 Å, i.e. well up in the ultraviolet. A photograph of the copper spark spectrum with this lead line is shown in fig. 1*b*. Inspection of the photographic plate for the presence of this line enabled a lead concentration as low as one part in 50 million (on a weight basis) to be detected. This is about 7 times lower than the maximum permissible concentration of 0.15 mg lead/m^3 air. The exposure time required

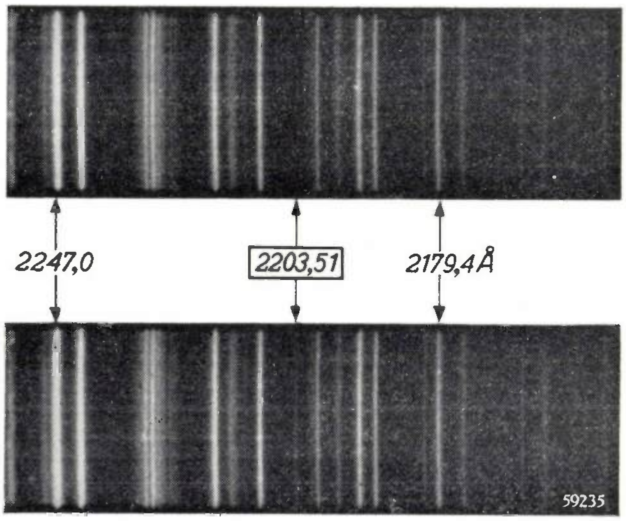


Fig. 1. *a*) Spectrum of the spark discharge in the absence of lead. Most lines are due to the copper of the electrodes, a few are caused by the water vapor content of the air.
b) Spectrum of the spark in lead contaminated air, showing the lead line at 2203 Å.

*) Philips Laboratories, Inc., Irvington on Hudson, N.Y., U.S.A.

[1]) Cf. for example R. Sawyer, Experimental Spectroscopy, Prentice Hall, New York 1944.

[2]) H. Aughey, J. Opt. Soc. Amer. **39**, 292-293, 1949 (No. 4).

was about one minute. An approximate quantitative analysis was possible by comparing visually the intensity of the lead line with that of adjacent copper lines. Calibration was accomplished by comparison with chemical analyses performed on samples taken simultaneously from a common air stream.

Although the photographic method worked quite satisfactorily and the superiority of spectrochemical analysis over previously used chemical methods was striking enough, still its application was hampered by the fact that the method was essentially discontinuous. In the operation of a plant, safety measures should be based, not on the detection of too high a lead concentration in one place, but rather on the early discovery of an increase in the lead concentration as a function of time, revealing the presence of a leak in plant installations before dangerously high lead concentrations in the atmosphere are attained. Detecting such a trend of the lead concentration to increase and locating the leak by the photographic method involved a number of successive samplings and exposures. In general, one to three hours was required before any resulting data could be utilized.

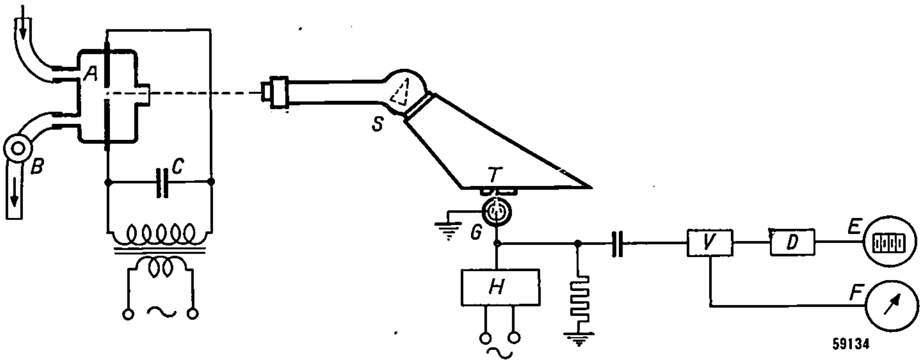


Fig. 2. Apparatus for the detection of lead. The spark *A* between copper electrodes, condensed by a capacitor *C*, excites the lead contained in the air pumped through the spark housing by the blower *B*. The radiation emitted by the spark is dispersed by the spectrograph *S*; the adjustable exit slit *T* isolates the lead line 2203 Å, which is measured by the Geiger counter tube *G*. *H* is the high voltage supply for this tube. The discharges triggered in the tube are amplified in *V* and counted by the scale-of-8 *D* and the mechanical register *E*. The deflection of the mA-meter *F* is proportional to the frequency of the discharges.

Considerable improvement was achieved by substituting a photoelectric Geiger counter tube for the photographic plate [3]) as a means for detecting the energy of the spectral line at 2203 Å. In fact, this line, if isolated by a spectrograph, can be measured by a radiation detector, such as the Geiger tube, and the information concerning other spectral lines which is given by the spectrograph can be dismissed (except for alignment and calibration purposes). The important advantage of the Geiger counter tube, provided with means for directly measuring the rate of arrival of the radiation quanta, lies in its being a continuously working, direct reading device. Such a device can serve the purpose of leak detection much better than any discontinuous method, as it will immediately reveal the trend of variation in the lead concentration. Thus, if the air to be analyzed is pumped to the spark gap by means of a long flexible hose with which the operator can scan pipes, valves etc. of an installation, a very effective method of leak hunting is obtained.

For reasons to be explained later, the Geiger counter method in the present case is not capable of a similar sensitivity and accuracy as the photographic method. However, for the restricted objective of locating areas of high lead concentration at possible leaks in plant installations, performance requirements are less severe and the Geiger counter method fully proved its ability to comply with them.

A few details of the apparatus used, which was constructed for E. I. du Pont de Nemours & Co. by Philips Laboratories at Irvington, N.Y., are described here [4]).

The complete set-up is shown in *figs 2* and *3*. A slit which can be moved across the spectrum is mounted in the focal plane of the quartz spec-

[3]) The application of Geiger counter tubes to spectrochemical analysis was initiated by O. S. Duffendack and W. E. Morris, J. Opt. Soc. Amer. 32, 8-24, 1942.

[4]) A short description was published earlier in: O. G. Koppius, J. Opt. Soc. Amer. 39, 294-297, 1949 (No. 4). — The author wishes to express his appreciation to Mr. H. Aughey of the E. I. du Pont de Nemours & Co. for his valuable assistance in this work.

trograph which disperses the radiation from the spark source. The position and width of the slit are adjusted so as to let only the 2203 Å lead line pass.

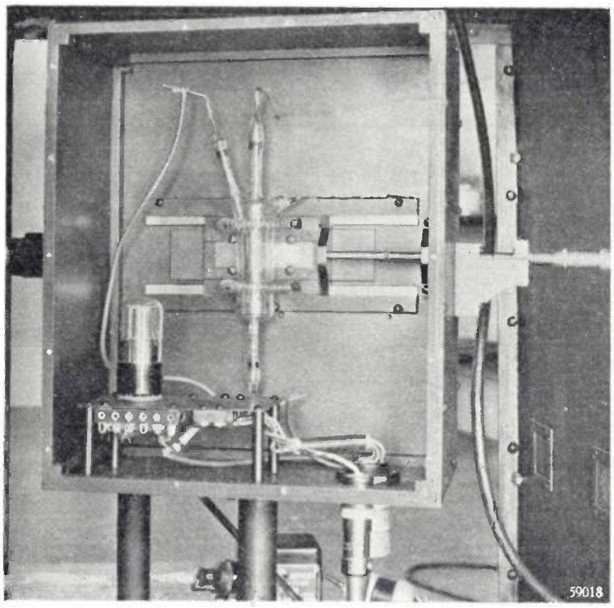


Fig. 3. Photograph of exit slit assembly with Geiger counter tube and preamplifier, at the back of the spectrograph. By the micrometer screw at the right the slit position in the spectrum may be adjusted.

The Geiger counter tube is placed behind the slit. Between the axial wire anode and the surrounding cylindrical cathode of this tube a voltage of about 1500 V is applied. Short discharge pulses are triggered in the tube by the photo-electrons liberated from the cathode metal by the ultraviolet radiation impinging on the inside wall. The radiation enters the cathode cylinder through a rectangular slot cut lengthwise in the wall as shown in *fig. 4*. The discharge pulses are amplified and fed to a mechanical counter or to a circuit measuring the rate of arrival of the quanta at the counter tube by the mean value of a current. The circuits used for this purpose are similar to those used in the Geiger counter X-ray spectrometer described earlier in this Review [5]).

The deflection of the rate meter milliammeter provides a direct check on the lead concentration at the place where the air is pumped off to the spark. It is possible to make a continuous record of this concentration if desired, or to make the instrument sound an alarm as soon as the lead content exceeds a given limit. The whole apparatus including the power supplies is mounted on wheels so that every part of the manufacturing area can be surveyed. The spectrograph and the counter tube with a pre-amplifier are mounted in an airtight box in which an inert atmosphere can be maintained, in case the air of the area to be examined should contain some inflammable vapor. The spark source is mounted in an air duct which traverses the airtight box. Air is forced through the duct by a small fan driven by an explosion-proof motor. To prevent a "flash-back", which may occur in case of an explosion, fire screens are mounted on the intake and outlet of the air duct.

The fact that the Geiger tube in this application serves as a photo-electron counter entails several peculiarities in its design not encountered in the more common counter tubes intended to respond to electrons, gamma-rays or X-rays. In X-ray applications, for instance, the counter tube must be designed to absorb the entering X-rays in the gas volume between the electrodes, the absorbed X-ray quanta setting electrons free (cf. the article quoted in[5])) ; the choice of the electrode metal is not important in this case. In the photo-electron counter, however, the gas filling must be chosen so as not to absorb any of the entering ultraviolet radiation quanta, because such an absorption would not in all cases give rise to a discharge pulse, and the cathode must be made of a metal giving a strong photoemissive effect. In our case a very pure (O.F.H.C.) copper cathode, cleaned by prolonged heating in hydrogen, was used. The energy necessary for liberating an electron from this metal amounts to $E = 4.42$ electronvolt; hence, according to Einstein's law, the maximum wavelength capable of liberating an electron, given by $\lambda = ch/E$($c =$ velocity of light, $h =$ Planck's constant) is about 2800 Å. This shows that the counter tube will be sensitive in the spectral region with which we are concerned. The anode is a tungsten wire 75 μ in diameter. The electrodes are mounted in an envelope of clear fused quartz and the tube is filled with argon at a

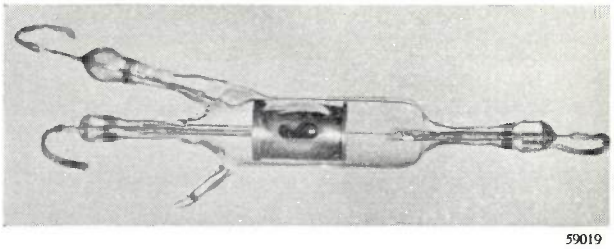


Fig. 4. Quartz Geiger counter tube used for detection of radiation at 2203 Å.

[5]) J. Bleeksma, G. Kloos and H. J. Di Giovanni, X-ray spectrometer with Geiger counter for measuring powder diffraction patterns, Philips Techn. Rev. **10**, 1-13, 1948 (No. 1).

pressure of 20 cm Hg. A small amount of alcohol vapor is added for rapidly quenching the discharge (cf. [5])). Neither of these gases absorbs the 2203 Å line to an appreciable extent. The counters have proved very reliable during their rated lifetime, so that during the past three years the instrument has been kept in almost continuous operation.

In the spectrograph (a Hilger E3), the rays of every wavelength converge onto their respective places in the focal plane in beams comprising an angle of about 10 to 30°, while the focal plane itself is situated obliquely to these beams; cf. *fig. 5.* As a consequence the radiation of the line 2203 Å, passing through the slit, is contained in a beam of a rather unfavorable configuration to be caught in the counter tube. A simple solution for thisdifficulty, avoiding the necessity of a rather artificial tube design, was found by mounting a mirror of stainless steel on the back of the slit, in the manner shown in fig. 5. When the angle of the mirror is adjusted to a suitable value, all the radiation passing through the slit is reflected into the hole of the counter cathode. The reflectivity of steel (and most other materials) for the ultraviolet radiation of 2203 Å is rather poor at normal incidence, but at near grazing angles of incidence as occur in our case, practically complete reflection of the radiation is obtained.

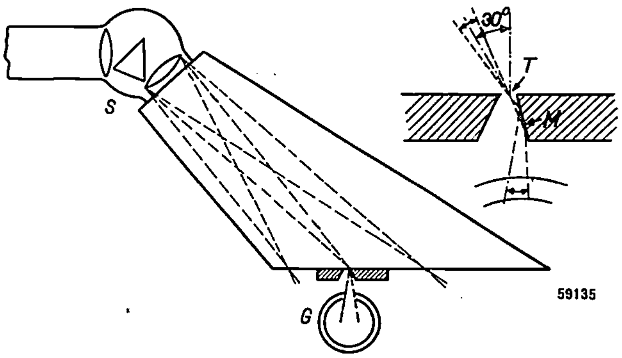


Fig. 5. Geometrical configuration of the beam of radiation of 2203 Å arriving at the exit slit of the spectrograph. By a mirror *M* of stainless steel mounted on the back of the slit, the entire beam is reflected into the hole of the counter tube cathode.

Once in several months it is necessary to scan the spectrum on either side of the lead line 2203 Å for readjustment of the slit. The adjacent copper lines are easily located, whereupon the slit is set on the lead line position by a simple interpolation. Of course, on moving the slit across the spectrum the configuration of the beam reflected by the mirror will change to some extent. The hole in the counter tube cathode wall is made large enough to receive the whole of the reflected beam for all positions of the slit within the portion of the spectrum which it may be desired to scan. The alignment of the spectrograph with the source, of course, must be checked more frequently, viz., once or twice during an eight hour period, as it is not possible for a simple portable instrument to maintain good optical alignment for a longer period under plant conditions. Every now and then the overall stability and sensitivity of the apparatus may be checked by measuring the relative intensities of the 2200 and 2195.8 Å copper lines.

A chart of the spark spectrum, as determined by setting the slit in subsequent positions about $^1/_2$ Å apart and noting the number of counts per second, is shown in *fig. 6a.* Fig. 6*b* gives a microphotometer trace of a photograph of the spectrum taken with the same spectrograph. The resolution of lines in the two cases is almost identical and more than adequate to resolve the 2203 Å lead line from nearby copper lines. A striking difference between the two spectra is the much lower background of the counter tube trace. This is due to the fact that the counter, because of the photoelectric threshold, is not affected by scattered radiation of wavelengths longer than 2800 Å arriving at the slit, whereas the photographic plate responds to this radiation.

The low background sensitivity of the Geiger counter tube is important, both for a practical and a fundamental reason.

In the photographic method, where the whole spectrum is recorded, the background contribution to the plate density can be recognized as such on inspection of the plate and appropriate corrections made. So the background does not prevent a very feeble line from being observed, and after corrections for the background have been applied a fairly accurate quantitative evaluation of the lead concentration from the relative line intensities is possible.

With the direct indicating Geiger counter method, however, no information is available during normal operation as to whether the measured number of counts per second is due to the lead line or to the background. The presence of the lead line must be derived and its intensity (L) evaluated by subtracting a constant number of counts (B), ascribed to the background and obtained as a result of a zero experiment, from the total number ($B+L$). In view of the statistical character of the arrival of radiation quanta at the counter tube, this total number contains a probable error $\sqrt{B+L}$, which in full is transmitted to the result, L, of the subtraction. The relative error $\sqrt{B+L}/L$ can be made small enough only by prolonged counting. Obviously the necessary counting time for a given accuracy will be shorter the smaller the background intensity *B*. This will be especially important when the counting rate (intensity *L*) is low.

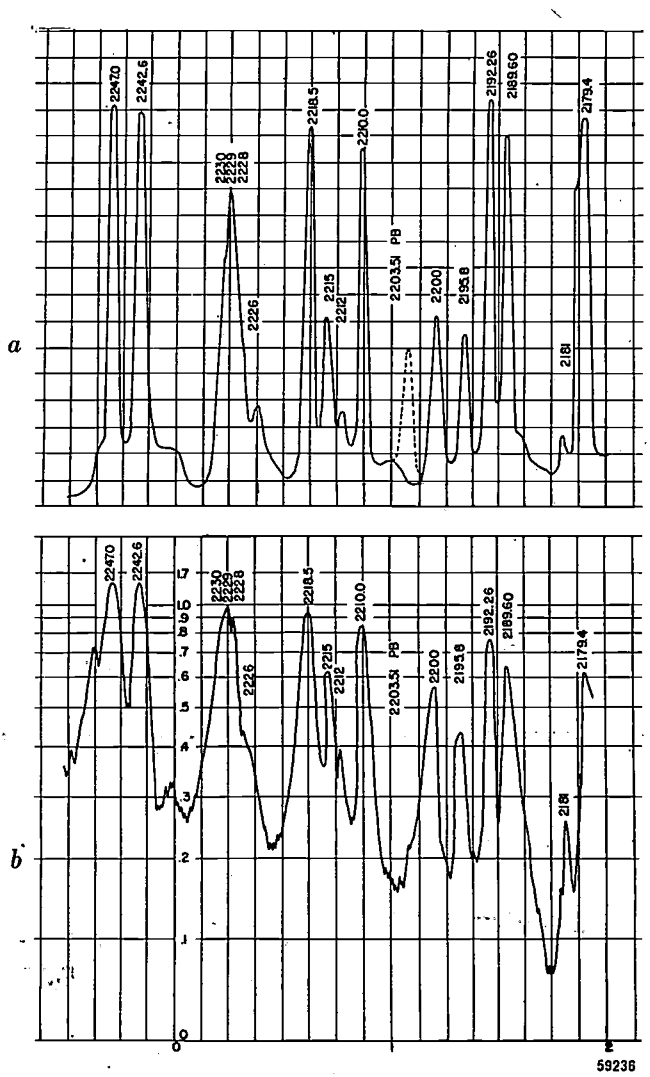


Fig. 6. *a*) Copper spectrum, determined by measuring the number of counts per second for a large number of positions of the exit slit (about $^1/_2$ Å apart). The position of the lead line 2203 Å is indicated by a dotted line.
b) Microphotometer trace of copper spectrum obtained photographically.

This is a practical consideration. A fundamental one is that, in reality, even the statistical mean value *B* is not a constant; the background is subject to continuous changes due to the erratic behaviour of the spark discharge. These changes have much more influence with the Geiger tube than with the photographic plate, owing to the shoiter averaging time. By the uncertainty of the value of *B* to be subtracted, a lower limit of the detectable lead concentration is imposed. So it may be said that the low background sensitivity of the counter tube is chiefly responsible for the comparatively high lead sensitivity of the apparatus. The lower limit of detectability, in a single measurement, is about 0.6 mg lead per m^3 air, or 1 part in about 2 million on a weight basis. Although this sensitivity cannot compete on equal terms with that of the photographic method, it is fully adequate for the purpose of leak hunting, where the discovery of local variations in lead concentration is all-important.

Several years' experience has shown the inherent soundness of the method for the detection of lead. However, it should be emphasized that in the construction of the apparatus there is little that points to its specificity for lead. The flexible hose and other tubing are made of "Saran" (a plastic) instead of rubber because certain lead compounds have a high affinity for rubber. The mirror behind the slit is specifically adjusted for the 2203 Å position of the slit. Apart from these minor details, there appears to be no fundamental reasons why the instrument could not be used for the detection of other elements, provided these elements show a sensitive emission line in a suitable spectral region and the slit is placed in the correct position. The analysis of dusts for elements such as arsenic, barium and beryllium are suggested examples.

Using a more constant source and a more elaborate technique, spectrochemical analysis with the Geiger counter is also feasible in cases where accurate quantitative data are required, as was shown in recent publications concerned with the analysis of the phosphor content of steel [6]). The background difficulties mentioned previously can be obviated by a method of "internal control", similar to the one used in the photographic procedure: by using two Geiger counter detectors it is possible to measure the ratio between the intensities of the lead line and an adjacent copper line, and thereby average out fluctuating background. Though, in some situations, also in our case accurate quantitative data may be desired, the method of internal control was discarded in the instrument described, as it would have unnecessarily complicated the instrument for the purpose of leak hunting; in a specific case it is easy enough to obtain quantitative information from a photographic recording of the spectrum.

[6]) R. Hanau and R. A. Wolfe, J. Opt. Soc. Amer. 38, 377-383, 1948 (No. 4); F. R. Bryan and G. A. Nahstoll, J. Opt. Soc. Amer. 38, 510-517, 1948 (No. 6).

Summary. A spark discharge is maintained between two copper electrodes in the air of a plant where atmospheric contamination with lead or lead compounds can occur through possible leaks in plant installations. If lead is present the

spark will contain the lead line at 2203 Å. By a small quartz spectrograph provided with a movable slit in its focal plane, this spectral line is isolated and its intensity is measured by a photoelectric Geiger counter tube with counting rate meter. The sensitivity of the apparatus in normal operation is limited to about 0.60 mg lead per m^3 air, by the fluctuations of the background intensity emitted by the spark. Although this is more than the maximum permissible concentration of lead in the air (0.15 mg per m^3), the instrument has proved very useful as a means for detecting and locating small leaks in pipes or valves, in whose vicinity high lead concentrations may occur. The superiority of the method as compared with the usual and more sensitive photographic procedure of spectrochemical anlysis is due to the fact that the direct indicating, continuous working, radiation meter is ideal for monitoring changes in concentrations.

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

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

1857: A. Bierman: A new type of betatron without an iron yoke (Nature London **163**, 649, 1949, April 23).

A new type of betatron is described in which the magnetic field is obtained by means of coils in air, through which the discharge current of a 6.5 μF condenser battery passes. There is only a small iron core, by means of which it is easy to fulfil the "flux condition". Saturation of the core effects contraction of the orbits towards the end of the acceleration period, until the electrons strike a tungsten target 0.1 mm thick. The betatron as a whole weighs no more than 50 kg, the iron core itself weighing less than 5 kg. The peak value of the magnetic induction is 0.4 Wb/m^2 (4,000 gauss), the radius of the orbit 8 cm; the energy of the accelerated electrons amounts to 9 MeV. Cf. Philips Techn. Rev. **11**, 65-78, 1949 (No. 3).

1858: J. A. Haringx: Het merkwaardig gedrag van op druk belaste schroefveren (Voordrachten Kon. Inst. Ingenieurs **1**, 298-313, 1949, No. **3**). (The remarkable behaviour of compression-loaded helical springs; in Dutch.)

A survey is given of the remarkable phenomena demonstrated by compression-loaded helical springs in respect of their elastic stability, lateral rigidity and natural frequencies for transverse vibrations. The existence of these phenomena was predicted on account of a theoretical calculation which was based upon the current simplification of replacing the helical spring by an elastic prismatic rod and a new interpretation of the latter's rigidity against shearing. At the end of the paper it is shown that the respective problems had to be treated first before helical springs could succesfully be applied as resilient elements for vibration-free mountings.

1859: J. L. Snoek: Time effects in ferromagnetism (Physica 's Grav. **15**, 244-252, 1949, No. 1/2).

In an attempt to give a short systematic survey of time effect in ferromagnetism a distinction is made between effects involving structural charges of the lattice (ionic time-effects) and effects affecting only the conditions of the 3*d*- and 4*s*-electrons (electronic time-effects). The ionic effects may be divided into time-effects involving plastic deformations, time-effects involving interstitial diffusion and effects of unknown origin (in ferrites). The electronic time-effects may be divided into eddy-current effects and ferromagnetic resonance effects.

1860*: C. J. Bouwkamp: On the transmission coefficient of a circular aperture (Phys. Rev. **75**, 1608, 1949, No. 10).

In connection with an article of Levine and Schwinger on the diffraction of a scalar plane wave by an aperture in an infinite plane screen, the writer briefly indicates the derivation of the correct expression for the transmission coefficient of a circular aperture.

1861: G. W. Rathenau: Enige nieuwe resultaten op het gebied van rekristallisatie (De Ingenieur **61**, Mk 57-Mk 63, 1949, No. 20). (Some new results on recrystallization; in Dutch.)

Some recent views on recrystallization. Review of recent literature on polygonization, recrystallization and grain growth. Some new results concerning the secondary recrystallization of nickel-iron alloys are discussed.

VOL. 11 No. 8, pp. 221-252 FEBRUARY 1950

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 HELIX AS RESONATOR FOR GENERATING ULTRA HIGH FREQUENCIES

by B. B. van IPEREN.

621.385.1.029.6:621.396.615.143

It is a known fact that constructional difficulties are encountered when trying to make a triode suitable for ultra high frequencies. In recent years various types of valves have therefore been developed on entirely new principles. Remarkable results have been obtained with the travelling wave tube, a tube in which electromagnetic waves travel along a helix. This tube, which is particularly used for amplifiying ultra-short waves, can of course also serve for generating such waves. An oscillator based upon this principle is described in this article and its working is explained.

Introduction

In order to avoid, or at least greatly reduce, the difficulties arising when using triodes for ultra high frequencies [1][2], in recent years various valves have been developed which in this case could be used to replace the triode.

An example of such a type of valve is the induction tube [3]). Here, instead of striking an anode forming part of the resonant circuit, the electrons impart their energy inductively to the output circuit and are then taken up by an anode which may be of any size. This eliminates one of the drawbacks of the triode, namely that with the small dimensions essential for high frequencies only little power can be dissipated. But the difficulties with the input circuit remain the same.

A considerable improvement in this respect is the velocity-modulation valve [4]). In this valve a beam of electrons of a constant velocity passes a high-frequency alternating field. The electrons which thereby undergo a certain change in velocity, depending upon the moment at which they pass the field, then arrive in a field-free space where this velocity modulation is converted into a density modulation, so that the current receives an alternating-current component. Also with this valve the energy is given off inductively.

The two grids required for the velocity modulation may be of a coarser mesh in the case of a velocity-modulation valve than the control grid of the triode or the induction tube. This reduces the constructional difficulties. Since, however, the electrons remain a finite time in the modulating field (and also in the inductor fields) this gives rise to effects which have an unfavourable influence upon the working of the valve.

In the first place this finite transit time results in the action of the high-frequency electric field upon the electrons being less effective: the velocity modulation or the induced alternating current is less than when the transit time is negligible. In the second place there is a transit-time damping, the modulating field giving off energy to the beam.

It is possible to avoid the effect of the transit-time damping by giving the electrodes such a

[1] C. J. Bakker, Some characteristics of receiving valves in short-wave reception, Philips Techn. Rev. 1, 171-177, 1936.

[2] M. J. O. Strutt and A. van der Ziel, The behaviour of amplifier valves at very high frequencies, Philips Techn. Rev. 3, 103-111, 1938. See also K. Rodenhuis, Two triodes for reception of decimetric waves, Philips Techn. Rev. 11, 79-89, 1949 (No. 3).

[3] This valve is described, i.a., by C. G. A. von Lindern and G. de Vries, Flat cavities as electrical resonators, Philips Techn. Rev. 8, 149-160, 1946.

[4] For a detailed description of the velocity-modulation valve see F. M. Penning, Velocity-modulation valves, Philips Techn. Rev. 8, 214-224, 1946.

shape and placing them at such a distance that the damping becomes nil or even negative [5]). But then (if one keeps to the simple configuration of two grids or a gap) the velocity modulation or alternating current induction aimed at is diminished, so that in practice this remedy is not very attractive.

From this it is not to be deduced that a negative transit time damping is never applied. There is a valve in which this is utilized. The electron beam is then made to pass only one modulating field. If the distance of the electrodes has been well chosen then the beam can give off energy to the field and if this energy is more than is needed for generating the high-frequency electric field required then the system can oscillate. The valve referred to, which bears the name of monotron [6]), works, however, only with a very high current, owing to the small absolute value of the negative transit-time damping. The small value of this damping is due to the fact that with the usual high-frequency fields the potential is a linear or at least monotone function of place. This will be reverted to later.

However, in the development of valves for very high frequencies a new principle is now being applied whereby use is made of an electric field the strength of which is a periodic function of place. The valves built according to this principle were at first considered for use as amplifiers [7])[8]) and were then given the name of "travelling wave tube". The simple construction of these valves makes them attractive also for generating ultra-short waves. In this article we shall particularly discuss the use of this valve as oscillator, but first of all a brief account will be given of its working as an amplifier.

Principle of the travelling wave tube used as amplifier

In a travelling wave tube it is ensured that the electro-magnetic wave that is to be amplified has a velocity, in a certain direction, which is only a fraction of the velocity in the free space. The wave thus retarded then has a reciprocal action with an electron beam.

There are various ways of retarding an electromagnetic wave, but theoretical investigations indicate that the application of a helix for this purpose offers the most favourable solution.

When an electromagnetic wave is introduced at one end of the helix it travels along the wire at a velocity approximating that of light. Viewed in the axial direction of the helix however the velocity is greatly reduced, in fact in the ratio of the pitch of the helix to its circumference. In this way it is possible to obtain along the axis of the helix a travelling electromagnetic wave with a velocity of, say, 1/10 of the velocity of light and with the electrical vector lying along thc direction of the axis. When an electron beam is sent along the axis a reciprocal action is set up between the beam and the electromagnetic wave. This situation is diagrammatically represented in *fig. 1.* Calculation shows that in a certain range of electron velocities the beam may impart energy to the wave and thus amplify it.

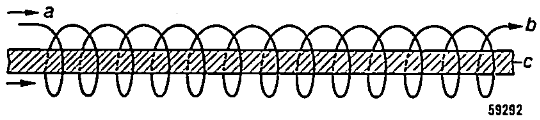


Fig. 1. Diagrammatic representation of the application of a helix with an electron current passing through it, used as amplifier. a = the arriving wave, b = the amplified wave, c = the electron current.

In broad terms this can be explained in the following way. The wave gives a velocity modulation to the beam and this is converted into a density modulation, so that the electron current in the beam receives an alternating-current component. The latter in turn yields energy to the wave and the amplified wave modulates the beam again, and so on. It appears that owing to this mechanism the amplitude of the wave increases exponentially with the distance travelled.

A closer investigation shows that the action of the valve is somewhat more complicated than this. The amplitude of a travelling wave as a function of the distance x and the time t can be represented by the formula

$$A(x, t) = A(0)\, \mathrm{e}^{\mathrm{j}(\omega t - \Gamma x)} \quad . \; . \; . \; . \; . \; . \; . \; . \quad (1)$$

where ω is the angular frequency and Γ the so-called constant of propagation. Now in the theory of the travelling wave tube it is assumed that all electrical quantities counting for this wave movement can also be represented in the manner indicated by formula (1). Calculating the reciprocal action between the electron beam and the field of the helix, we arrive at an equation of the fourth degree in Γ. Of the four roots of this equation two are real, one positive and one negative. This means that a wave travelling to the left and one travelling to the right can move along the tube with constant amplitude. The other two solutions for Γ are conjugate complex, which means to say that in the exponent of the e-power is a real term which is positive for one root and negative for the other.

[5]) These effects will be dealt with later.
[6]) The monotron has been fully discussed by J. J. Müller and R. Rostas in Helv. Phys. Acta 13, 435-450, 1940.
[7]) R. Kompfner, Wireless World 52, 369-372, 1946.
[8]) J. R. Pierce, Proc. Inst. Red. Engrs 35, 111-123, 1947.

These two solutions thus both correspond to a wave travelling to the right, one of which is amplified and the other attenuated. The amplification appears to reach the maximum when the velocity of the electrons is equal to the velocity which the wave would have along the axis of the helix if there were no beam.

An important property of the arrangement described here is the fact that the amplification is only very slightly dependent upon the frequency. Care must be taken, however, to ensure that only travelling waves do indeed occur; in other words, for all these frequencies there must not be any reflection at the end of the helix. If this condition is fulfilled then the band width — by which is to be understood the difference between the two frequencies at which the power amplification has dropped to one half — may amount to about one-third of the mean frequency.

The generation of oscillations

In the foregoing a brief description has been given of the manner in which the travelling-wave tube functions as amplifier. The usual way of turning an amplifier into an oscillator is to feed back part of the energy from the output side of the tube to the input side. With the tube in question however there is a much simpler method. By removing entirely the matched load that was essential when using the tube as an amplifier, then, owing to the reflection, standing waves are obtained at the ends of the helix, which then acts as a resonator.

This is quite comparable to other devices where both travelling and standing waves may occur, as for instance in a transmission line or a wave guide; there, too, a travelling wave is obtained if there is a matched load and a standing wave if there is no such matching. The electric field along the axis of the helix then likewise becomes a standing field, such that the amplitude varies along that axis roughly sinusoidally. As will be explained below, a beam of electrons with a suitably chosen velocity will then yield energy to that field. This can also be expressed by saying that in that case a negative transit-time damping arises, just as it does in a monotron. It appears, however, that with the sinusoidal field of the helix this negative transit-time damping may be much greater in absolute value than is the case with an approximately constant field, such as obtained with a monotron.

An oscillating "travelling wave tube" may also be regarded as a sort of velocity-modulating valve, but with a number of modulating fields and a number of inducing fields placed one behind the other, instead of one of each.

The question now arises as to whether the improvement due to the more favourable form of the field is not neutralized by the replacement of the high-impedance cavity resonators usually employed with velocity-modulation and similar valves by a helix, which is generally considered to be a much less effective resonator.

There is, however, the somewhat surprising fact that, as a calculation shows, the power required to generate a certain field in a part of a helix having a length of a half wavelength is practically equal to the power used in generating a field of the same size in a good cavity resonator. Thus the impedances of the circuits are approximately the same in both cases. This only holds, however, so long as there are so many waves on the helix that the radiation may be ignored.

Before going into these matters further, let us consider more closely the helix as a resonator.

Resonances of helices

The general problem of the field of a helix acting as the carrier of an electromagnetic wave is mathematically so difficult that no solution has yet been found for it.

For the case in which we are most interested, namely that of a helix which is long compared with its diameter, we may, however, to a good approximation say that the intensity of the current along the wire changes sinusoidally. Denoting the wavelength, measured along the wire of the helix, by λ_s and the total length of wire by L, we must have:

$$L = n\lambda_s, \qquad (n = \tfrac{1}{2}, 1, 1\tfrac{1}{2}, \dots).$$

Considering, however, the wavelength along the axis of the helix, λ_z, it is clear that it must equally hold that

$$l = n\lambda_z, \qquad (n = \tfrac{1}{2}, 1, 1\tfrac{1}{2}, \dots),$$

where l is the length of the helix (see *fig. 2*). Denoting the distance between two turns (the pitch of the helix) by h and the radius of a turn by R, we get the equation

$$\frac{l}{L} = \frac{\lambda_z}{\lambda_s} = \frac{h}{2\pi R} \doteq \sin\varphi\,,$$

where φ is the angle made by the wire with a plane perpendicular to the axis of the helix. The electric field along the axis of the helix is likewise sinusoidal with the wavelength λ_z.

We now have to put the question what resonance frequencies correspond to the various values of n, thus what wavelength λ_0 in the free space corresponds to each value of n (or λ_z and λ_s respectively). So far it has only been found possible to calculate this for the idealized case of an infinitely long cylinder along the surface of which currents are only possible at a given angle to the axis of the cylinder. One can imagine such a cylinder as being built up of a large number of helically wound, very thin wires lying parallel to each other on the periphery of the cylinder. For this case an exact solution can be given. Lenz has already done this for very small values of φ [9]). The solution is also possible for arbitrary values of φ [8]) [10]), it thereby having been found that for all values of φ occurring in practice the result is practically independent of φ, We can therefore confine our considerations to the case where φ is small.

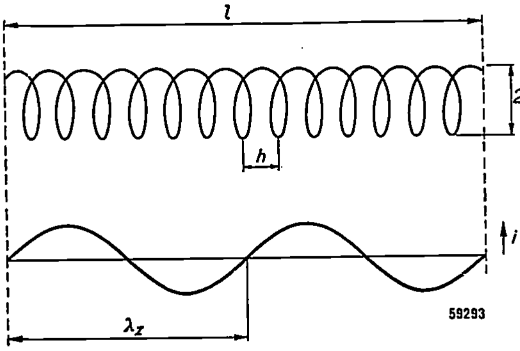


Fig. 2. In the case of a helix of which the length l is large compared with the diameter $2R$ the current flows along the wire sinusoidally. In the case drawn here $n = 2$. The wavelength along the axis is represented by λ_z.

The calculation then yields as result a relation between λ_s/λ_0 and $2\pi R/\lambda_z$. This relation is represented in *fig. 3* by the curve *I*. If λ_s/λ_0 were equal to 1 this would mean that a wave travels along the wire of the helix with the velocity of light. From the curve *I* it therefore appears that the velocity along the wire is indeed over a large area approximately the velocity of light [11]) and that a considerable change takes place when λ_z is large compared with R.

Although the above-mentioned theory cannot, of course, take full account of the actual behaviour of a helix, in the case that is of most interest to us, namely that of a long helix oscillating in one of its higher resonance frequencies, there appears to be little divergence from the theory. This may be seen from curve *II* in fig. 3, representing the result of measurements taken on a helix with $l = 155$ mm, $R = 6$ mm, wire thickness $d = 0.7$ mm and $h = 4$ mm. The numbers given at the measuring points indicate what harmonic the point represents, thus the number of half waves on the helix.

It is to be noted that the measurements give much greater deviations from this theory when the fundamental wave of the helix is measured (thus for $n = \frac{1}{2}$). In this connection Drude [12]) has carried out extensive measurements with helices of different dimensions. He found that the fundamental frequency depends but little upon d and h, but that it is particularly determined by the total length of the wire L and the ratio l/R. His results for a helix where $h/d = 2.4$ have been plotted in curve *III* for values of $l/(2R)$ varying from 6 to 0.05. The right-hand part of curve *III* thus relates to the range of short coils, for which the above-mentioned theory is no longer applicable.

For this range a theory has been worked out by Lenz [9]) and Szasz [13]), whilst later Hallén [14]) gave an improved theory for the fundamental wave of short coils (curve *IV*) which agrees rather well with Drude's measurements (curve *III*). Curve *V*, theoretically derived by Lenz and Szasz, relates to the various resonances of a very short coil of a certain shape (where $l/(2R) = 0.122$). Finally curve *IV* gives two measuring points by these authors for such a coil.

Contrary to what has been found in the case of long helices, the velocity of the wave in the case of short coils is always less than the velocity of light.

Transmission of energy from an electron beam to an electric alternating field

From what has been found in the foregoing it follows that in the case of a helix we have to do with an axially directed electric alternating field which can be represented by the formula

$$E = E_0 \cos\left(\frac{2\pi n}{l}\, z\right) \cos \omega t \quad . \; . \; . \; . \; . \quad (2)$$
$$(-l/2 < z < +l/2)$$

Travelling along the axis is a beam of electrons the initial velocity of which we shall denote by v_0. The question is now what power this beam will yield to the field.

In general such a problem can only be solved by very laborious graphical and numerical methods, but in the exceptional case where the power transmitted is small compared with the power with which the electrons reach the alter-

[9]) W. Lenz, Berechnung der Eigenschwingungen einlagiger Spulen, Ann. Phys. Leipzig, **43**, 749-797, 1914.

[10]) J. Chew and J. D. Jackson, Field theory of traveling wave tubes, Proc. Inst. Rad. Engrs **36**, 853-863, 1948.

[11]) This fact is closely related to the already mentioned property of the wide band width of travelling-wave amplifiers.

[12]) P. Drude, Ann. Phys. Leipzig **9**, 590-610, 1902.

[13]) O. Szasz, Mathematischer Beitrag zu der Abhandlung des Hrn. Lenz, (see [9])), Ann. Phys. Leipzig **43**, 789-809, 1914.

[14]) E. Hallén, Ueber die elektrischen Schwingungen in drahtförmigen Leitern, Upsala Universitets-årsschrift 1930.

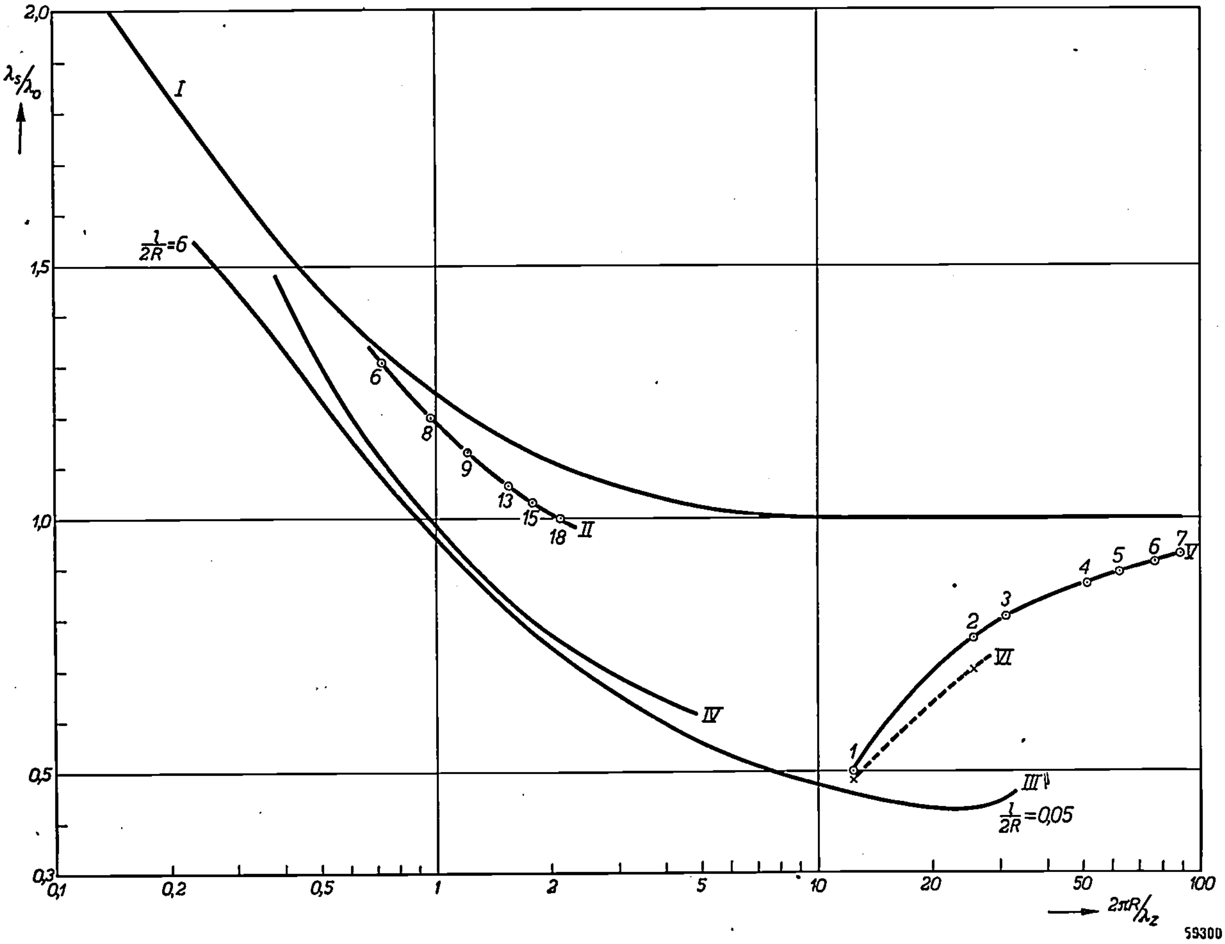


Fig. 3. The relation between the quantities λ_s/λ_0 and $2\pi R/\lambda_z$ (λ_0, λ_s and λ_z are the wavelengths respectively in the free space, along the wire and along the axis of the cylinder; R is the radius of the cylinder). *I* Theoretical curve according to Lenz for a helically conducting infinitely long cylinder. *II* Measurements with a helix of the length $l = 155$ mm and radius $R = 6$ mm, where the thickness of the wire was $d = 0.7$ mm and the spacing between the wires $h = 4$ mm. (The numbers at the measuring points denote the number of half waves on the helix.) *III* Measurements taken by Drude where only the fundamental wave was considered and with variations of $l/(2R)$ from 6 to 0.05. *IV* Theoretical curve according to Hallén for the fundamental wave ($n = \frac{1}{2}$). *V* Theoretical curve according to Lenz and Szasz for a short coil. *VI* Measurements taken by these authors with a coil with $l/(2R) = 0.122$.

nating field the solution can be given in a simple closed form, and for any arbitrary shape of the field along the path of the electrons [15]). One cannot, it is true, calculate in this way the yield that can be reached with a given system, but it is indeed possible to determine whether oscillations will arise with a given intensity of the electron current, and it can also be calculated at what electron velocity this will be the case.

The solution referred to is as follows. The electric field is represented by $E = E(z)\cos\omega t$, where $E(z)$ denotes the variation, as yet still arbitrary, of the electric field strength along the z-axis. It then appears that a passing electron receives from this field an amount of energy which to a first approximation is equal to:

$$\Delta_1 = e\cdot \}A(\xi)\cos\delta - B(\xi)\sin\delta\{ \quad . \; . \; . \; . \quad (3)$$

where $A(\xi) = \int_{-l_2}^{l_2} E(z)\cos\xi z\,\mathrm{d}z$ and $B(\xi) = \int_{-l_2}^{l_2} E(z)\sin\xi z\,\mathrm{d}z$.

Here $-l_2$ and l_2 are the limits of z beyond which $E(z) = 0$, δ is a parameter indicating at what instant the electron enters the field and $\xi = \omega/v_0$ (ω = angular frequency, v_0 = electron velocity), whilst e is the charge of the electron.

From formula (3) it can be determined what variation in velocity the electron undergoes in the field, thus what in the case of a velocity-modulation valve is called the velocity modulation. The total energy that the beam receives per second from the field is found by determining the mean value of Δ_1 over a cycle of the oscillation, thus over the area in which δ changes from 0 to 2π. This is found to be zero.

Considering the power transmission to a second approximation we find that the average is no longer zero. The formula then applying is:

$$\Delta_2 = -\frac{e^2\omega^3}{2m\,v_0}\,\}A(\xi)\,A'(\xi) + B(\xi)\,B'(\xi)\{ \quad . \; . \quad (4)$$

where $A'(\xi)$ and $B'(\xi)$ are derivatives with respect to ξ.

[15]) See B. B. van Iperen, On the generation of electromagnetic oscillations in a spiral by an axial electron current, Philips Res. Rep. 4, 20-30, 1949 (No. 1).

To give an example of the use of this method we shall apply it to the simple case of the monotron. There the electron beam passes two grids which are spaced a distance l apart and between which is an alternating voltage $V_\sim \cos \omega t$. Now $E(z) = V_\sim/l$. It is found that a beam with the intensity of current i_0 per second takes up an energy P:

$$P = \frac{2ei_0}{mv_0^2} V_\sim^2 \, \mathrm{f}(\tau), \quad . \quad . \quad . \quad . \quad . \qquad (5)$$

where τ is the number of cycles of the alternating voltage that an electron with the velocity v_0 requires to travel the distance between the grids, and $\mathrm{f}(\tau)$ is a function graphically represented in *fig. 4*.

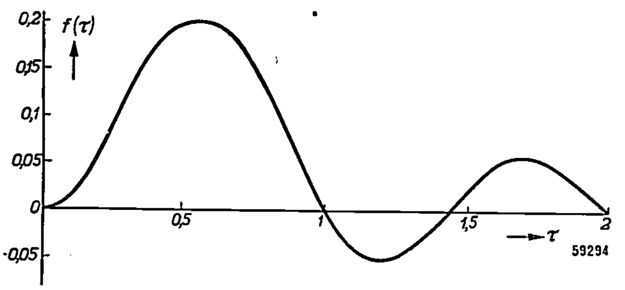


Fig. 4. The function $\mathrm{f}(\tau)$ to which is proportional the amount of power taken up by an electron beam when passing an electric alternating field between two grids. τ is the transit time, i.e. the number of cycles of the alternating voltage that an electron with a certain velocity needs to travel the distance between the grids.

If the impedance of the circuit coupling the grids is Z then the power dissipated therein is $V_\sim^2/Z$.

It is clear that oscillations can only occur if P is negative, and

$$|P| > V_\sim^2/Z \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (6)$$

Introducing the acceleration voltage of the beam V_0 (for which $eV_0 = \frac{1}{2}\, mv_0^2$ applies) formula (6) becomes

$$-\frac{i_0}{V_0}\, \mathrm{f}(\tau) > \frac{1}{Z}. \quad . \quad . \quad . \quad . \quad . \quad . \qquad (7)$$

The quantity $Z_0 = V_0/i_0$ may be termed the beam impedance of the valve. With the aid of this we can express the condition for the occurrence of oscillations as

$$-\frac{1}{Z_0}\, \mathrm{f}(\tau) > \frac{1}{Z}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (8)$$

Since for all values of τ the term $|\mathrm{f}(\tau)|$ is much smaller than 1, to comply with this condition Z_0 must be small. Consequently the current must have a high intensity.

From fig. 4 it may also be seen that for transit times τ between 0 and 1 the field always yields energy to the beam. Thus we have here a positive transit-time damping. For larger values of τ we arrive in a zone where the beam yields energy to the field. It is upon the effect of this negative transit time damping that the working of the monotron is based.

The fact that for small values of τ a positive transit-time damping will arise is also to be understood qualitatively, assuming for the sake of simplicity that the electric field variation with time is not just sinusoidal but follows a "square sine".

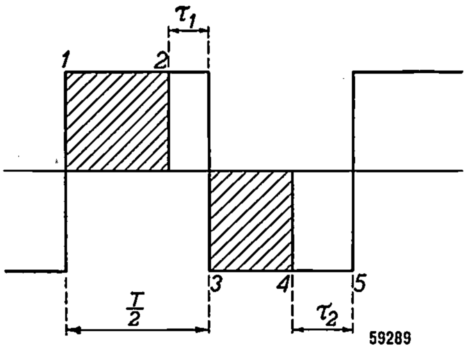


Fig. 5. A diagram for explaining qualitatively the variation of the function $\mathrm{f}(\tau)$ in fig. 4 for the case where τ is rather small. Here it is assumed that the field strength changes with time according to a "square sine". Since the transit time τ_1 of the accelerated electrons is smaller than that of the retarded electrons, τ_2, the number of electrons which are exclusively accelerated will be greater than the number of those which are only retarded. Thus on the average the beam takes up energy from the field.

All electrons entering the field at the instants *1* to *2* (see *fig. 5*) receive the same increase of energy, whilst all electrons arriving at the instants *3* to *4* undergo an equal reduction of energy. Since, however, the transit time τ_1 of the accelerated electrons is less than that of the retarded electrons, τ_2, the amounts of energy belonging to these two groups of electrons do not neutralize each other. The electrons arriving between *2* and *3* and remaining for a part of their transit time inside the retarding field are fewer in number than those which arrive between *4* and *5* and remain for a part of their time in the accelerating field. As a result, therefore, the beam takes up energy.

It is rather more difficult to understand that if τ is slightly greater than 1 the effect is the reverse, but this can be made plausible with the aid of *fig. 6*. We choose τ of such a value that the electron *1* reaching the field at the beginning of the accelerating needs just one cycle to pass the field. We can then choose a group of electrons entering between *1* and *2* in such a way that they leave the field in the area *1'-2'*. On the average these electrons take up energy from the field. In the area *2-3* (corresponding to *2'-3'*), however, there are

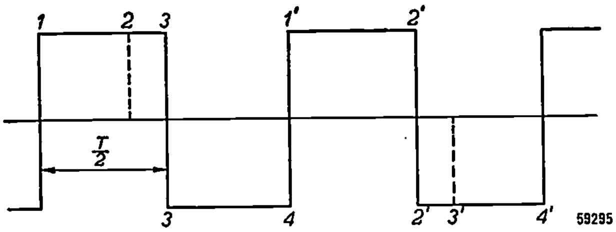


Fig. 6. With the aid of this diagram it is made plausible that for τ slightly greater than 1 on the average the electrons of the beam yield energy to the electromagnetic field.

electrons which yield energy to the field, whilst farther on all the electrons from the area *3-4* (corresponding to *3'-4'*) also yield energy. In the aggregate, therefore, there are more electrons yielding energy than there are dissipating energy.

Transmission of energy from an electron beam to the field of a helix

Having shown how in a simple case the exchange of energy between an electron beam and an electric alternating field can be studied, we shall now consider how matters stand in the case of an electric field changing sinusoidally along the path of the electrons, as occurs with an oscillating helix.

Let us first consider this case qualitatively before giving the result of the calculation.

The standing wave along the helix can always be imagined as being resolved into two travelling waves, one moving to the left and the other to the right. We assume that the electrons travelling to the right have a velocity v_0 which is approximately equal to the velocity of the two waves. Compared with the wave travelling to the right these electrons have a velocity which is small with respect to v_0, but the wave travelling to the left passes them with the velocity $2v_0$. Viewed from the electron, the latter wave causes the field strength to change very rapidly. Thus the effect of this upon the electron must be small, so that we can disregard the influence of this wave.

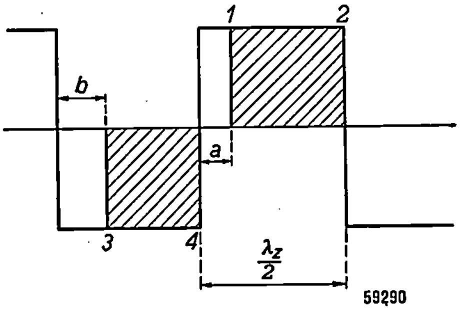


Fig. 7. The transfer of energy from the field of a coil to an electron beam. In this diagram the electric field along the axis of the helix is represented in a system of coordinates moving with the velocity of the wave. This field, which in the said system does not change with time, is actually sinusoidal but is represented here by a "square sine". With the aid of this diagram it is shown that on the average the electrons of the beam take up energy from the field when their velocity v_0 is slightly less than the velocity v_g of the wave on the helix. In that case the area *b* is greater than the area *a*.

To investigate the action of the other wave let us imagine a system of coordinates following the movement of this wave. The wave is represented by a sine curve which does not change with time. In order that the effect in question may be made clearer this sine is again replaced by a "square sine", as drawn in *fig. 7*.

Let us first assume that the velocity v_0 of the electrons is slightly less than the velocity v_g of the wave. In the beginning the electrons then travel along the wave to the left with the velocity $v_g - v_0$.

The point in the diagram where an electron begins depends upon the moment at which it enters the field of the helix. This starting point shifts to the left according to the moment of entry. The electrons entering in an interval of time corresponding to the area *1-2* are accelerated, so that their velocity to the left along the wave gets smaller and smaller. On the other hand the electrons beginning between *3* and *4* steadily increase in velocity to the left. The distance over which the electrons from the area *1-2* move to the left during their passage through the helix is therefore less than the distance covered by the electrons from the area *3-4* during their passage. The number of electrons dissipating energy during the whole of their passage is therefore greater than the number of electrons exclusively yielding energy to the wave, or, to put it in other words, the area *a* of the electrons passing over to the retarding field is smaller than the area *b* of the electrons passing over to the accelerating field. On the average, therefore, the electrons of the beam will take up energy from the field.

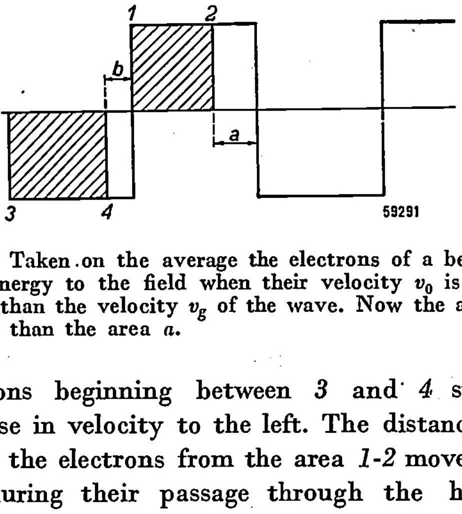


Fig. 8. Taken on the average the electrons of a beam will yield energy to the field when their velocity v_0 is a little higher than the velocity v_g of the wave. Now the area *b* is smaller than the area *a*.

Let us now suppose that the velocity v_0 of the electrons in the beam is slightly greater than v_g. In this case *fig. 8* applies. With respect to the wave the electrons now move to the right with the velocity $v_0 - v_g$. Those beginning between *1* and *2* are accelerated. We now find that the area a of the electrons passing over to the retarding field is larger than the area b of the electrons passing over to the accelerating field. Thus on the average the beam yields energy to the field, so that oscillations are generated.

From this reasoning it also follows that the desired effect can always be brought about to a sufficient degree, however small the electric field may be, by causing the electrons to move along with the wave for a sufficient length of time, thus by making the helix long enough.

The calculation of the exchange of energy for this particular case can again be carried out by the method already indicated. The alternating field arising along the axis of an oscillating helix is given by formula (2):

$$E(z) = E_0 \cos 2\pi nz/l \, .$$

In this case we find for the energy yielded per second by a beam with intensity of current i_0:

$$P = \frac{e\,i_0}{mv_0{}^3} \cdot \frac{E_0{}^2 l^3}{8} \, \mathrm{F}(2\pi n - \frac{\omega l}{v_0}) \quad . \; . \quad (9)$$

where n is the number of waves on the helix and $F(2\pi n - \omega l/v_0)$ is a function graphically represented in *fig. 9* [16]).

Since P is proportional to the third power of l, by increasing l the amount of energy yielded can always be made large. The maximum value of P is obtained when $2\pi n - \omega l/v_0$ is so chosen that $\mathrm{F}(2\pi n - \omega l/v_0)$ is a maximum [17]). This function appears to have a number of maxima. One of these, namely that for $2\pi n - \omega l/v_0 = 0.42 \times 2\pi$, is much larger than all the others. Thus this value is the most favourable for generating oscillations.

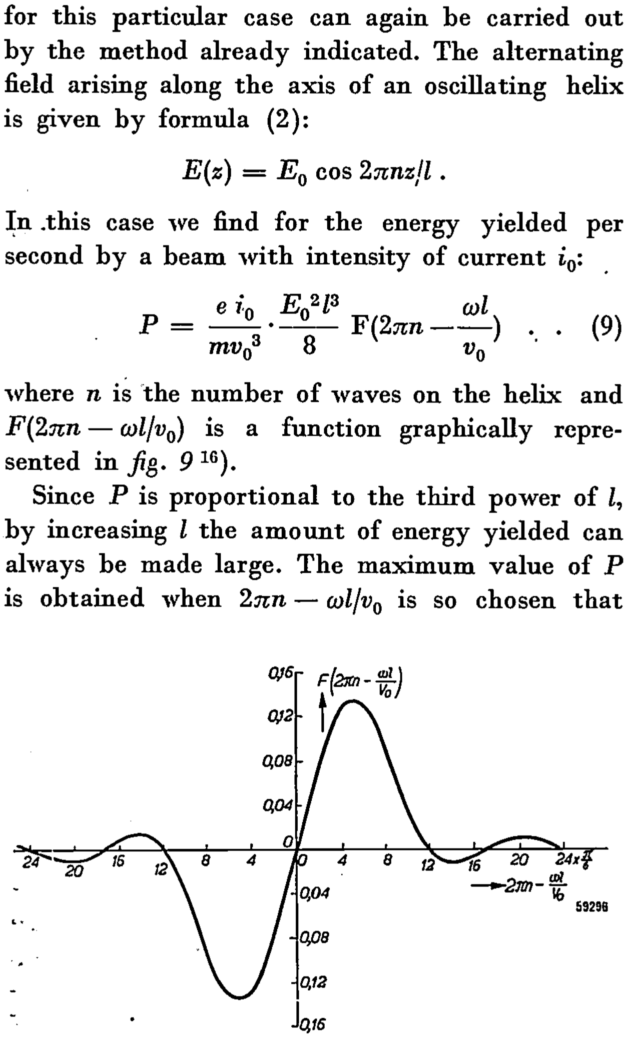


Fig. 9. Graph of the function $\mathrm{F}(\pi n - \omega l/v_0)$ to which is proportionate the power yielded by a passing electron beam to the sinusoidal electric field of a helix. The diagram shows that this function has several maxima, the greatest of which occurs at $2\pi n - \omega l/v_0 = 0.42 \times 2\pi$.

Considering that $\nu l = n\nu\lambda_z = nv_g$ (where $\nu = \omega/2\pi$), the condition for the maximum is:

$$n \frac{v_0 - v_g}{v_0} = 0.42, \; . \; . \; . \; . \; . \; . \quad (10)$$

from which it appears that v_g must indeed be smaller than v_0. For generating oscillations it is therefore necessary that the electrons travel slightly faster than the wave, and this difference in velocity has to be smaller the more waves there are on the helix.

Formula (10) can be reduced to a more practicable form by introducing the acceleration voltage V_0 and the wavelength λ_0, which then gives:

$$\frac{500\,l}{\lambda_0 \sqrt{V_0}} = n - 0.42 \quad (V_0 \text{ in volts}). \quad . \quad (11)$$

When we have a helix of given dimensions we can first determine from fig. 3 (curve *I*) the wavelength λ_0 corresponding to any value of n and then, with the aid of formula (11), calculate the optimum voltage V_0 required for generating these waves.

Formula (11) holds for the highest maximum of $\mathrm{F}(2\pi n - \omega l/v_0)$. A maximum about 10 times as small is found for $2\pi n - \omega l/v_0 = -1.16 \times 2\pi$. Oscillations can also be obtained for this value. Formula (11) then takes the form of:

$$\frac{500\,l}{\lambda_0 \sqrt{V_0}} = n + 1.16 \; . \; . \; . \; . \; . \quad (11a)$$

Calculation of the current intensity at which oscillations begin

In the foregoing section we found an expression (formula (9)) indicating what power a passing electron beam yields to the alternating field of a helix when the field has a given amplitude E_0. We shall now consider the case where $\mathrm{F}(2\pi n - \omega l/v_0)$ has reached its greatest maximum and is thus equal to 0.13. The power yielded is then $P = i_0 e/8mv_0{}^3 \cdot E_0{}^2 l^3$, which in a more practical form reads:

$$P = 26 \, \frac{l^3}{\lambda_0 V_0{}^{3/2}} \cdot i_0 E_0{}^2 \text{ watts} \; . \quad . \; . \; . \quad (12)$$

On the other hand we have to know what power is required to maintain the alternating field E_0. This is difficult to calculate exactly but a good approximation can be obtained in the following

[16]) This formula holds only to a good approximation if the number of waves on the helix is not too small say, at least 3.

[17]) This is not quite correct, because also the factor preceding F contains variable quantities. A correction for this, which is fairly small, is to be found in the article quoted in footnote [15]).

way. Let us first assume that the wire forming the helix has the thickness 0 and that through this wire an alternating current flows the amplitude of which changes sinusoidally along the wire with the wavelength λ_s. From the current distribution thus given one can also (with the aid of the continuity equation) find the distribution of the charge along the wire, and from that, according to the well-known solution of Maxwell's equations, with a given distribution of current density and charge density in space, the electromagnetic field in the surroundings of the helix can be calculated. Now all that we are interested in here is the field along the axis. The exact formula for this is rather complicated [18]), but for nearly all cases occurring in practice the following approximative formula suffices:

$$E_0 = 60\ I_0 \frac{\lambda_0}{R^{1/2}\,\lambda_z^{3/2}} \cdot \mathrm{e}^{-\frac{2\pi R}{\lambda_z}} \quad (\mathrm{V/cm}) \quad . \quad (13)$$

where I_0 is the amplitude of the current in the wire.

The next step is to calculate the power dissipated in the wire as a consequence of the alternating current I_0 flowing through it. Of course we must in the first place take into account the fact that owing to skin effect the current flows only along the surface of the wire. Actually there is also the fact that the current is not uniformly distributed over the circumference of the wire but flows for the greater part along the side of the wire facing the axis of the helix, but here no account has been taken of this effect. This having been disregarded, the calculated losses will be smaller than the actual losses.

By taking these calculated losses as being equal to the energy yield of the beam per second as given by formula (12) we find for the current intensity at which oscillations begin (for the case of a copper helix):

$$i_0 \doteq 3.8 \cdot 10^{-9}\, \frac{V_0^{3/2}}{n^3} \cdot \frac{RL}{d\,\lambda_0^{3/2}}\, \mathrm{e}^{\frac{4\pi Rn}{l}}, \quad . \ . \quad (14)$$

where d is the thickness and L the total length of the wire. In this formula i_0 and V_0 are expressed in amperes and volts, and L, l, d, λ_0 and R in centimetres.

Some experiments for checking the calculations

The easiest way to produce an oscillator working on the principle described in the foregoing is to take a vacuum tube of silica (e.g. 20 cm long and 15 mm thick) having a cathode system at one end and an anode at the other end, and to wind a copper wire round it. In order to avoid any divergence of the electron beam this tube has to be placed in the magnetic field of a coil.

But such an arrangement is not suitable for checking the validity of the formulae derived above, because owing to charges on the wall of the tube electrostatic fields are set up which affect the velocity of the electron beam. To overcome this difficulty we used a tube with a helix (of molybdenum wire with a cross section of 0.7 mm) placed inside it instead of around it. With the helix oscillating at a resonance frequency we then determined at what tube voltage the oscillations with the respective frequency were strongest. The corresponding value of n, the number of waves on the helix, was found in a simple manner by sliding along the helix a small incandescent lamp, which glowed strongest where there was a maximum field strength.

According to formulae (11) and (11a), $1/(\lambda_0\sqrt{V_0})$ when plotted as a function of n should give a straight line. The experiment was first carried out with the smallest intensity of current at which the tube oscillated. *Fig. 10* shows (curve *2*) that the points measured do indeed lie on a straight line. Measurements taken with a slightly stronger current (15 mA) appear to give a straight line slightly lower (curve *3*). Line *1* in fig. 10 represents the theoretical curve corresponding to the lower range of oscillation according to formula (11a). Since in that case oscillations could only be produced with a much stronger current, only one point of this curve could be measured. For the value corresponding to this point the test appeared to be in agreement with the theory.

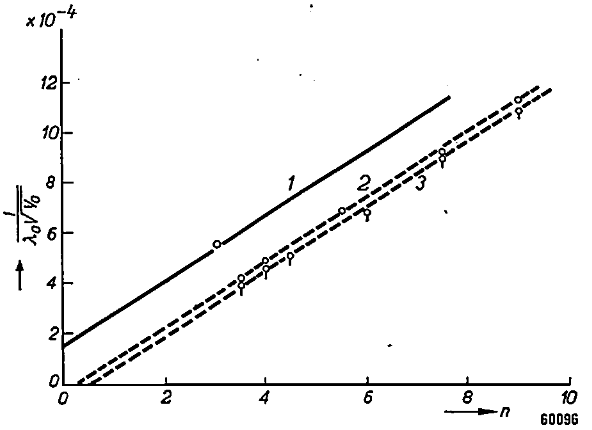


Fig. 10. The relation between $1/\lambda_0\sqrt{V_0})$ and n. (1) the theoretical curve for the lower range of oscillations according to formula (11a). (*2*) the measured points, which indeed appear to lie on a straight line. (*3*) measurements taken at a slightly stronger current (15 mA).

[18]) See E. Roubine, L'onde électr. 27, 203-208, 1947.

A second test was carried out to check the validity of formula (14) indicating at what intensity of current the oscillations begin. The results have been plotted in *fig. 11*. Curve *1* has been calculated from formula (14). Curves *2* and *3* are the results of two series of measurements. The fact that these do not coincide better is due to the value of i_0 being very sensitive to any small deviation of the direction of the tube with respect to the magnetic field in which it is placed.

The measured points lie considerably higher than would be expected from the theoretical curve. This is due in part to the effect, already mentioned, of the irregular distribution of current over the circumference of the wire. Moreover, since a certain fraction of the electrons always passes over from the beam to the helix, the temperature of the helix is higher than room temperature. As a consequence the specific resistance of the helix is greater than that reckoned with when calculating curve *1*.

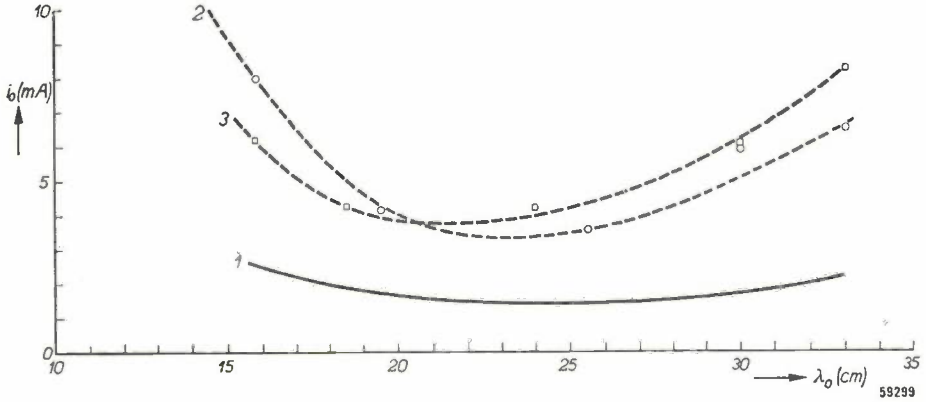


Fig. 11. The theory concerning the current intensity at which oscillations begin, checked by measurements. Curve *1* has been calculated from formula (14). Curves *2* and *3* have been plotted from two series of measurements.

Focusing of the electron beam by means of positive ions

It has already been pointed out that it is necessary to ensure that the electron beam does not diverge, and it has been said that this can be done with the aid of a magnetic field. The coils needed for this purpose, however, make the apparatus less easy to handle, though the tube itself is quite simple. Now it has been known for some time[19]) that the mutual repulsion of the electrons in a beam can be compensated or at least greatly reduced by introducing a certain quantity of some gas or other into the tube. The gas atoms are then ionized by the beam and owing to the presence of positive ions, which remain in the beam a fairly long time, the negative space charge of the electrons is neutralised. As a consequence the beam shows less tendency to diverge.

It has indeed been found possible, with the aid

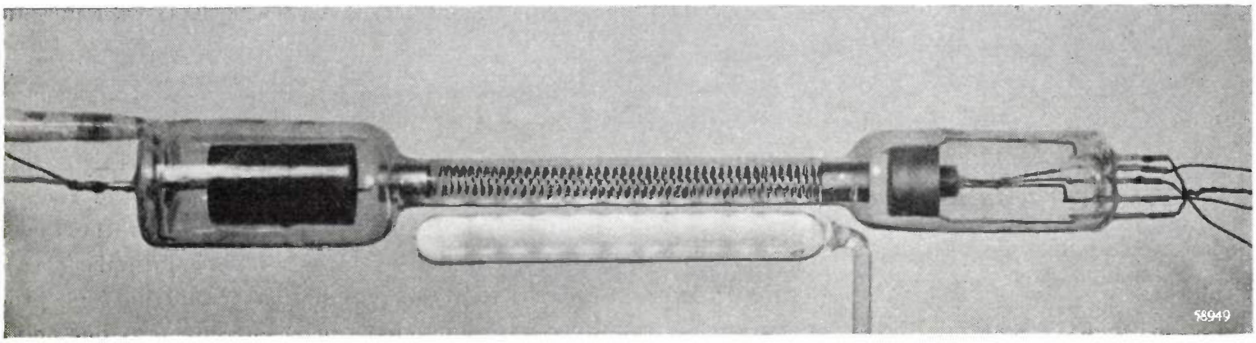


Fig. 12. Experimental model of a tube with gas-concentrated beam. The standing waves on the helix have been made visible with the aid of a neon tube. (Total length of the helix 12 cm; wavelength 20 cm.)

[19]) J. B. Johnson, Phys. Rev. 17, 420-421, 1921.

of a suitable electron gun and a rare gas of a certain pressure (for instance neon of 10^{-3} to $2 \cdot 10^{-3}$ mm Hg), to cause the tube to oscillate without employing focusing coils. The electron gun used gives a somewhat convergent beam, which beyond the focal point, if the tube were not filled with gas, would diverge again. However, owing to the presence of the gas ions the divergence is very much more gradual, so that the major part of the electron current is still able to reach the anode.

Fig. 12 is a photograph of an oscillating tube. It also shows how the standing waves on the helix can easily be made visible, by means of a small tube filled with neon (pressure about 10 mm Hg) placed alongside the helix; the neon glows strongly at the maxima of the electric field strength.

Summary. In the application of triodes for ultra high frequencies difficulties arise owing to the transit time of the electrons no longer being small compared with the cycle of the high-frequency alternating voltage. These difficulties are only partly overcome by using an induction tube, a velocity-modulation valve or a monotron. Matters are further improved by using a travelling-wave tube. This has already been studied in its application as amplifier. It also provides, however, a simple means of generating short waves. In this tube an electromagnetic wave travels along a coiled wire, whilst an electron beam is sent along the axis of the coil. After showing how this system can be used as an amplifier, the mechanism of the oscillation of the tube is discussed. The resonances of the coil play an important part in this mechanism. It is demonstrated that with a travelling wave tube it is a fairly simple matter to calculate under what conditions oscillations are possible and at what intensity of current they begin. Finally the results are given of some experiments carried out which confirm the accuracy of the theory given. For practical purposes it is of importance that the magnetic field for focusing the electron beam can be dispensed with when the tube is filled with a rare gas under a certain pressure.

HEATING BY MEANS OF HIGH-FREQUENCY FIELDS

II. CAPACITIVE HEATING

by M. STEL and E. C. WITSENBURG *).

621.364.16:679.562:674.24

The phenomenon of dielectric losses has been known for a long time, but its application as a source of heat (capacitive heating), on an industrial scale, dates back to only just before 1940. In the years following 1940 there has been a great advance in capacitive heating. This is connected with the rapid development in plastics, in the manufacture of which capacitive heating has in many cases led to a considerable saving and improvement of quality. Another important field of application is the wood-working industry, in particular the manufacture of plywood in flat and bent shapes.

Recently an article was published in this Journal dealing with induction heating [1]). By this is understood the heating of conductive materials by means of electric currents induced therein by a magnetic alternating field. We shall now deal with capacitive heating, the heating of non-conductive materials by means of the dielectric losses occurring therein when they are placed in an electric alternating field. Whereas these losses often constitute an undesired property in insulators, here they form the basis of the heating process.

Nature of the dielectric losses

Very briefly we shall recall here some of thc main features of dielectric losses.

As already discussed at length in this Journal [2]), dielectric losses are the result of after-effect phenomena. When the electrical field strength in a dielectric for instance is increased suddenly, but after that kept constant, then the dielectric displacement likewise makes a jump but instead of then remaining constant it continues to increase gradually until it approaches a final value. In simple cases this "after-effect" takes place according to an exponential function of time. The relaxation time (after-effect time) τ is in many cases in essence a diffusion time, shorter or longer according to whether charge-carriers or dipoles in the dielectric have more or less freedom of movement or are able to change their direction under the influence of the changing field strength.

If the field strength alternates sinusoidally with time, as is the case with the capacitor in *fig. 1a* connected to an alternating voltage, then the after-effect results in a phase shift δ between the field strength and the dielectric displacement. Accordingly the current flowing through the capacitor leads by an angle φ slightly less than 90° ($\varphi = 90° - \delta$). Corresponding to the phase shift δ are certain losses, viz. the dielectric losses, and that is why δ is also called the loss angle.

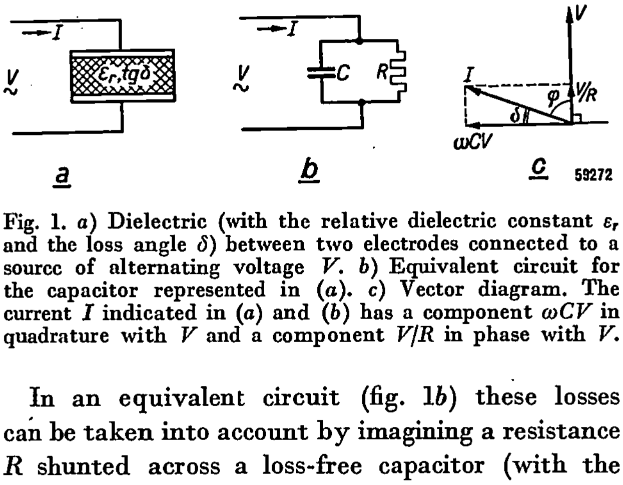


Fig. 1. *a*) Dielectric (with the relative dielectric constant ε_r and the loss angle δ) between two electrodes connected to a sourcc of alternating voltage V. *b*) Equivalent circuit for the capacitor represented in (*a*). *c*) Vector diagram. The current I indicated in (*a*) and (*b*) has a component ωCV in quadrature with V and a component V/R in phase with V.

In an equivalent circuit (fig. 1*b*) these losses can be taken into account by imagining a resistance R shunted across a loss-free capacitor (with the capacitance C of the capacitor represented in fig. 1*a*). The corresponding vector diagram in fig. 1c shows that, with V representing the R.M.S. value of the alternating voltage applied and ω the angular frequency, the power W converted into heat amounts to

$$W = \frac{V^2}{R} = \omega C V^2 \cdot \frac{1}{\omega CR} = \omega C V^2 \tan\delta\,. \quad (1)$$

For the simple cases mentioned above it can be calculated (see the article quoted in footnote [2])) that tan δ as a function of ω has a maximum at $\omega = 1/\tau$. Measurements taken on most dielectrics show, however, that in a wide frequency band tan δ is only little dependent on ω. This is due to the fact that the moving charge-carriers or dipoles

*) Philips Telecommunication Industry, formerly N.S.F., Hilversum (Netherlands).

[1]) E. C. Witsenburg, Heating by high-frequency fields, I. Induction heating, Philips Tech. Rev. 11, 165-175, 1949 (No. 6).

[2]) J. L. Snoek and F. K. du Pré, Several after-effect phenomena and related losses in alternating fields, Philips Techn. Rev. 8, 57-64, 1946.

do not all behave in the same way, which means to say that one has to do with not one but a series of after-effect times τ. It is not difficult to understand that as a consequence the curve representing $\tan\delta$ as a function of ω is much flatter than when there is only one after-effect time to be taken into account.

Something similar is the case with the dielectric constant, which, likewise due to the large number of after-effect times, in most materials changes but little with the frequency.

The heat generated by the dielectric losses will now be considered quantitatively.

Capacitive heating of a homogeneous medium

If the electric field in the capacitor of fig. 1*a* may be regarded as being homogeneous (thus ignoring boundary effects) then formula (1) may be written in the following form:

$$W = 0.556\cdot10^{-10}\cdot(\varepsilon_r \tan\delta)\cdot E^2 f\cdot\frac{M}{\varrho}\ \text{watts,} \quad (2)$$

where ε_r is the (relative) dielectric constant of the dielectric to be heated, E the field strength therein (R.M.S. value in V/m), f the frequency (in c/s), M the mass (in kg) and ϱ the density (in kg/m^3) of the dielectric.

Eq. (2) is derived from eq. (1) as follows: Denoting the surface area of the plates by A and the distance by s, the capacitance is [3]):

$$C = \frac{\varepsilon A}{s} = \frac{\varepsilon_0\varepsilon_r As}{s^2},$$

where ε = the absolute and ε_r the relative dielectric constant, ε_0 = the dielectric constant of the vacuum (in the Giorgi system = 8.855×10^{-12} F/m).

Considering that As = the volume of the dielectric = M/ϱ and that $V/s = E$, one easily arrives at formula (2).

In order to save time one will try to make the generated power W so great that the final temperature desired is reached very quickly. In that case, for a rough calculation, we may disregard the loss of heat during the short heating period, the more so since there is little conduction and radiation: little conduction because the material is a poor electrical and thus also a poor thermal conductor, and little radiation because with capacitive heating there are usually no higher temperatures than 200 °C.

Disregarding, therefore, the small loss of heat, we have:

$$W = Mc\frac{d\vartheta}{dt}\ \text{watts,} \quad . \ . \ . \ . \quad (3)$$

where c = the specific heat (in joules per kg and per °C), ϑ = the temperature (in °C) and t = the time (in seconds).

From (2) and (3) it follows that

$$\frac{d\vartheta}{dt} = 0.556\times10^{-6}\cdot\frac{\varepsilon_r\tan\delta}{\varrho c}E^2 f \quad . \ . \ . \quad (4)$$

As we have seen, for most materials ε_r and $\tan\delta$ show little dependency upon the frequency: $(\varepsilon_r\tan\delta)/\varrho c$ can then be regarded as a material constant (apart from temperature dependency). For a given material formula (4) gives the relation between the field strength, the frequency and the rate of increase of temperature.

If it is asked, for instance, what frequency is required to heat a piece of wood ($\varepsilon_r = 3$, $\tan\delta = 6 \times 10^{-2}$, $\varrho = 600$ kg/m^3, c = 1700 J/kg ×°C) to 90 °C in 1½ minutes, then with $E = 10^5$ V/m we find from (4): $f = 10^7$ c/s = 10 Mc/s.

As regards the field strength, 10^5 V/m = 100 V/mm is about the limit to which one can go without risk of breakdown. If vapours arise from the object during the heating process these may condense on the electrodes and thus increase the risk of breakdown, in which case it is advisable to choose a weaker field strength. In order to maintain the same rate of heating then, according to (4), a higher frequency is needed.

Capacitive heating of a non-homogeneous medium

In fig. 1*a* we started from a homogeneous dielectric filling the entire space between the plates of the capacitor. In practice, when gluing wood for instance, one often has to do with a charge that is not homogeneous. We shall therefore now consider two simple cases: a charge having two different homogeneous dielectrics, where the interface layer is parallel to the direction of the field strength, and the case where it is perpendicular thereto.

The first case is represented diagrammatically in *fig. 2a*; it occurs, inter alia, in the gluing together

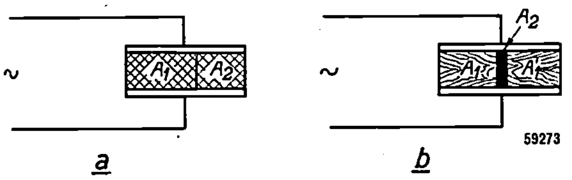


Fig. 2. *a*) Two dielectrics (A_1, A_2) connected in parallel. *b*) Example of parallel connection when gluing with capacitive heating. A_1, A_1' are blocks of wood with a layer of glue (A_2) between them. The glue is heated more quickly than the wood.

[3]) See for instance P. Cornelius, The rationalized Giorgi system with absolute volt and ampere as applied in electrical engineering, Philips Techn. Rev. 10, 79-86, 1948 (No. 3).

of two pieces of wood (fig. 2*b*). Electrically speaking we have here two capacitors connected in parallel. For each of the dielectrics formula (4) applies, with the same values of E and f but with different values for ε_r, $\tan\delta$, ϱ and c. Distinguishing the dielectrics with the indices 1 and 2, we have:

$$\frac{\mathrm{d}\vartheta_1}{\mathrm{d}t} : \frac{\mathrm{d}\vartheta_2}{\mathrm{d}t} = \frac{\varepsilon_1 \tan\delta_1}{\varepsilon_2 \tan\delta_2} \cdot \frac{\varrho_2\, c_2}{\varrho_1\, c_1} . \quad . \; . \; . \quad (5)$$

Somewhat less simple is the case of a stratified dielectric where the layers are perpendicular to the direction of the field strength, such as the layers of wood and glue in the manufacture of plywood. In *fig. 3* this case is illustrated for two layers, electrically connected in series.

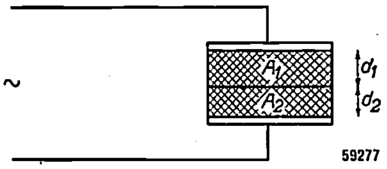


Fig. 3. Two dielectrics (A_1, A_2) in series.

As opposed to the case of fig. 2, the field strengths in the different materials are now unequal. In the layer with the dielectric constant ε_1, the loss angle δ_1 and the thickness d_1 the field strength E_1 is:

$$E_1 = \frac{\varepsilon_2}{\varepsilon_1\, d_2 + \varepsilon_2\, d_1}\, V , \; . \; . \; . \; . \; . \; . \quad (6a)$$

and in the other layer the field strength E_2 is:

$$E_2 = \frac{\varepsilon_1}{\varepsilon_1\, d_2 + \varepsilon_2\, d_1}\, V \; . \; . \; . \; . \; . \quad (6b)$$

For the ratio of the rates of heating we find from (4):

$$\frac{\mathrm{d}\vartheta_1}{\mathrm{d}t} : \frac{\mathrm{d}\vartheta_2}{\mathrm{d}t} = \frac{\varepsilon_2 \tan\delta_1}{\varepsilon_1 \tan\delta_2} \cdot \frac{\varrho_2\, c_2}{\varrho_1\, c_1} , \quad . \; . \; . \quad (7)$$

which differs from (5) only in the fact that here ε_1 and ε_2 are interchanged. It is this change which accounts for the somewhat surprising fact that when two materials are "connected in parallel" the ratio of the rates of heating may be entirely different from that in the case of "series connection" of the same materials. When, for instance, two pieces of wood are glued together with urea-formaldehyde glue in the manner according to fig. 2*b*, then the glue is heated much quicker than the wood (selective heating). With the "series connection" (fig. 3) on the other hand the heating of the glue is only a little quicker than that of the wood.

For gluing, selective heating is highly favourable, because it is mainly a question of the heating of the glue; the heat generated in the wood is useless. By choosing a certain position and shape of the electrodes one can often promote selective heating, as illustrated in *fig. 4*, where a layer of veneer has to be glued, over a large area, onto a thicker layer of wood. This is a situation that does not lend itself to ordinary "parallel connection". In order to get a better selective heating than is possible with "series connection" the electrodes here have been made in the form of parallel and alternately poled strips applied only on the side of the layer of veneer. As is the case with "parallel connection", here the lines of force run more or less parallel to the interface layer.

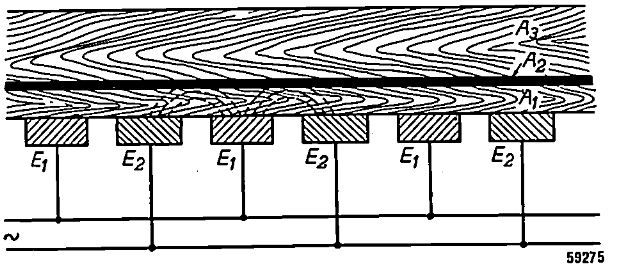


Fig. 4. A thin layer of veneer A_1 has to be glued onto a thick plank A_3 with a glue A_2. The electrodes E_1, E_2 are applied and connected in such a way that the lines of force (dotted arcs) in the glue are roughly parallel to the interface between the glue and the wood; the same effect (selective heating) is then obtained as with parallel connection.

Voltage distribution along the workpiece

In the foregoing it has been tacitly assumed that the potential is the same at all points of each of the electrodes. This, however, holds to a good approximation only so long as the wavelength corresponding to the working frequency is large with respect to the largest dimension of the electrodes. If the wavelength is comparable with that dimension (owing to the workpiece being large or the frequency being high) the field strength in the workpiece will vary from point to point, so that there will be an irregular distribution of heat.

With wood and similar materials $\tan\delta$ is so small that to a first approximation the wave phenomenon at the electrodes can be regarded as a standing wave. If V_0 is the potential difference (R.M.S. value) between the points of the electrodes where the input wires are connected, then the potential difference V_l between two other points of the electrodes at the distance l from the input points is:

$$V_l = \frac{V_0}{\cos\left(\frac{l}{\lambda}\,2\pi\right)}. \quad . \; . \; . \; . \; . \; . \; . \quad (8)$$

So long as $l \ll \lambda$, $V_l \approx V_0$, but when l becomes comparable to λ or $\gg \lambda$ then in certain areas V_l will be much greater than V_0.

If, for instance, a heat distribution is permitted where the extreme temperatures differ no more than 5% from the mean, thus where the lowest temperature is not more than 10% lower than the highest temperature, then the smallest field strength must not be less than 95% of the largest field strength. From (8) we then find for the maximum distance $l_{\max}$ between the edge of the dielectric and the input point:

$$l_{\max} = \frac{\text{arc cos } 0.95}{2\pi}\,\lambda = 0.05\,\lambda.$$

For λ we can write: $\lambda = v/f = c_l/(f\sqrt{\varepsilon_r})$, where v = velocity in the dielectric and c_l the velocity in vacuum $\approx 3 \times 10^8$ m/sec. Thus

$$l_{\max} \approx \frac{15 \cdot 10^6}{f\sqrt{\varepsilon_r}}\ \text{m}. \quad . \; . \; . \; . \; . \quad (9)$$

With, for instance, $f = 15 \cdot 10^6$ c/s and $\varepsilon_r = 3$ (wood) we find: $l_{\max} \approx 0.6$ m. Thus, with a temperature tolerance of ± 5%, this frequency is suitable for heating sheets of wood of a maximum length of 1.2 m (current input in the middle of the electrodes).

Generators for capacitive heating

As may have been seen from the article quoted in footnote [1]), in the case of induction heating the choice of the frequency is in the first place a question of depth of penetration and efficiency: the higher the frequency, the less is the depth of penetration and the greater the efficiency of the work coil; above the frequency $f_{\min}$, whereby the depth of penetration is about one-eight of the diameter of the workpiece, there is, however, hardly any further increase in efficiency. Only when an even smaller depth of penetration is favourable for the result of the heat treatment does it serve any purpose to choose a frequency higher than $f_{\min}$. Now the smaller the diameter of the workpiece (imagined to be cylindrical) the higher is this minimum frequency $f_{\min}$. Consequently, both in the case of small workpieces and in cases where a small depth of penetration is desired, one has to use high frequencies which in practice can only be obtained by employing a valve generator; in other cases lower frequencies suffice, and rotary generators — though of a special construction — may be used.

The main consideration in choosing the frequency for capacitive heating is the rate of working, which has to be as high as possible. Now, as is to be seen from the formula (4), the rate at which the temperature rises is proportional to the frequency, so that in the first place one will aim at getting the highest possible frequency. Consequently valve generators are used exclusively. However, in order to be able to employ normal types of valves and to keep the efficiency of the generators high, not too high a frequency should be chosen, this being usually in the order of 10 Mc/s. The material constants, the permissible field strength and the temperature desired are generally such that with frequencies of this order the heating time is very short and a considerable saving is obtained compared with the time that would be required for any other method of heating.

In a valve generator for capacitive heating the capacitor, formed by the electrodes and the workpiece, may be a part of the resonant circuit. It has to be taken into account, however, that the loss resistance R (fig. 1*b*) may be very much smaller or greater than the optimum load resistance of the generator valve. *Fig. 5* shows a system whereby the resistance R can be transformed to a favourable value by a suitable dimensioning of the capacitor C_1 and the inductor L_1.

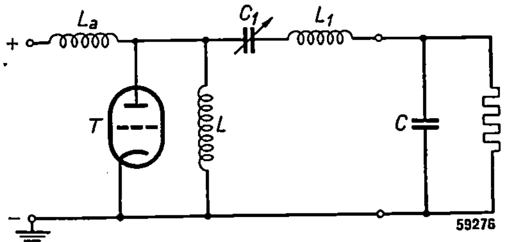


Fig. 5. Simplified circuit of a generator for capacitive heating. T = triode, C-R = equivalent circuit according to fig. 1*b*, L = inductance of the resonant circuit, C_1-L_1 = elements with which the resistance R is transformed to a value forming a favourable load for the triode, L_a = choke via which the anode is fed.

In this system there is apparently a resistance $R_{AC} = n^2R$ between the anode and the cathode of the generator valve, where the transformation ratio n is given by

$$n = \frac{L}{L + L_1} \cdot \frac{C + C_1}{C}.$$

With the aid of C_1 and/or L_1 it is therefore possible to give n any value greater or less than 1 which may be required to make R_{AC} equal the optimum load resistance of the valve.

Often only values of n greater than 1 are needed, in which case L_1 can be reduced to the self-inductance of the connecting wires and C_1 can be a variable capacitor.

Since the elements C_1 and L_1 bring about at the same time a transformation of the capacitance C, we have for the angular frequency generated :

$$\omega^2 = \frac{n}{LC}.$$

The maximum power generated in the workpiece with the Philips types of generators for capacitive heating lies between 0.3 and 50 kW. *Fig. 6* shows two generators, one for 2 kW and the other for 0.3 kW, in the special construction designed for the preheating of plastic pellets to be presently described. The generator illustrated in *fig. 7* has an output of 22 kW.

Applications of capacitive heating

The applications of capacitive heating, like those of induction heating, are still in course of development, but a number of fields have already been marked out where this method of heating offers striking advantages. Examples of such applications are the preheating of synthetic resins, the making of objects of bent plywood, the welding of certain thermoplastic materials, the twist-setting of rayon threads, etc.

We shall confine our considerations here to the first two applications (manufacture of plastics and the wood-working industry).

Capacitive heating in the manufacture of plastics

Phenoplasts (like "Philite") are thermo-setting substances, i.e. solids which when heated first become plastic but after a time harden and remain so upon cooling down, even if they should be heated again [4]). An article made of "Philite", for instance, may be manufactured in the following way. A weighed quantity of resins prepared from phenol and formaldehyde, mixed with fillers such as wood flour, mica and the like, is placed between the two halves of a steel mould. The mould is heated to about 160 °C, whilst at the same time the two halves are pressed together hydraulically. The moulding mass first becomes plastic, owing to the high temperature, and under the great pressure assumes the shape of the mould, after which it ultimately hardens.

The time that the mass has to be kept in the mould is mainly determined by the poor thermal conductivity of the mass. With a view to reducing

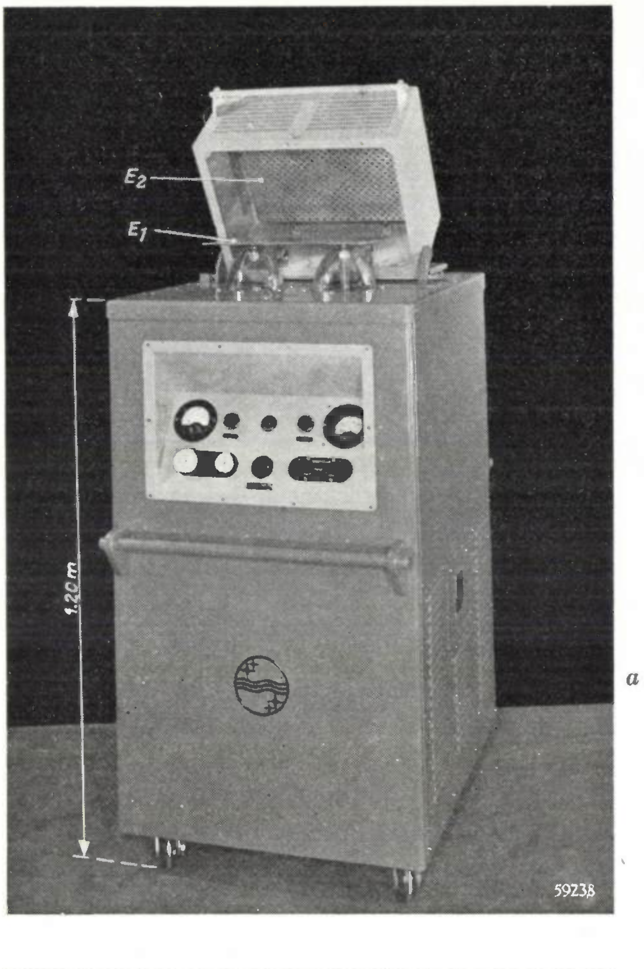


Fig. 6. *a*) Generator type SFG 136/21 for preheating plastic pellets, these placed on the plate E_1, which forms one of the electrodes. The other, earthed, electrode is the cap E_2, which is shown folded back. Frequency approx. 15 Mc/s. Maximum output 2 kW.
b) Generator for the same purpose for an output of 300 W, to be released shortly.

[4]) See for instance J. C. Derksen and M. Stel, Plastics and their application in the electrotechnical industry, Philips Techn. Rev. 11, 33-41, 1949 (No. 2).

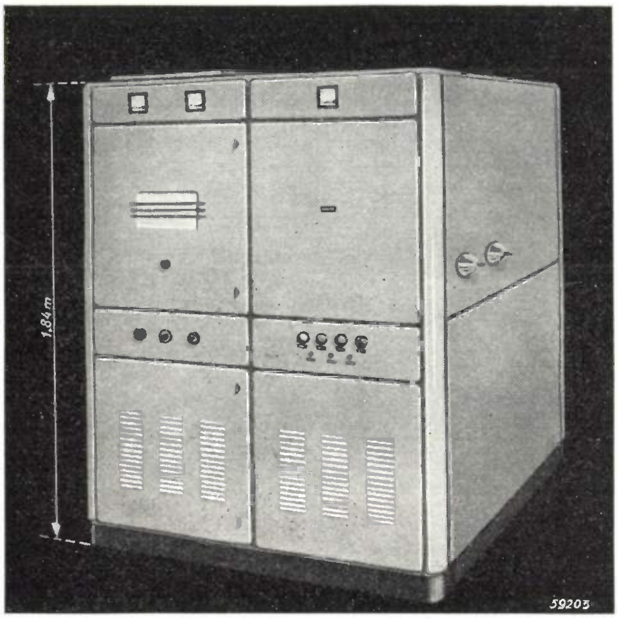


Fig. 7. Generator type SFG 134/01. Frequency approx. 2.5 to 8 Mc/s. Maximum output 22 kW.

the moulding time and thus increasing the output of a press per hour (or else reducing the number of presses required) it was very soon decided to preheat the mass before placing it in the mould. The idea was that if the mass could be preheated to a temperature just below that at which it begins to set, then only little additional heat would be needed in the mould. In this way much time would be gained in the moulding process, especially in the case of products with thick walls, which without preheating would take a long time to heat to the right temperature everywhere in the mould.

Now the conventional method of preheating, in an oven or furnace, is inefficient. When heat is applied externally the temperature on the outside of this poorly thermal-conducting moulding mass is much higher than that inside it, especially when it is desired to heat the mass quickly to a temperature approaching that of the mould. The danger then exists that after having become plastic the mass already begins to harden on the surface. To avoid this risk preheating is carried out very slowly and not beyond a rather low temperature, say 80 °C, but then, of course, it takes a long time and very little is gained in the pressing time. Moreover, in this way the mass does not reach that degree of plasticity that it does when it is preheated to a higher temperature and which has great advantages in the moulding process, since the more plastic the mass the less is the mould subject to wear and the lower is the pressure required (lighter presses suffice).

With capacitive heating the conditions are much more favourable, because the heat is generated inside the object itself. Owing to a certain amount of radiation the temperature on the outside of the object is always somewhat lower than that below the surface, but by means of thermal insulation or by keeping the electrodes themselves hot it is quite easy to ensure a highly uniform distribution of heat.

With the modern methods of manufacturing articles from "Philite" and similar plastics the mass is made in the form of pellets. A number of these pellets just sufficient to make the article that is being produced are placed in the mould after first being preheated. This preheating, to make the pellets plastic (see *fig. 8*), is done, for instance, with a generator like that illustrated in *fig. 9* or *10*. With the latter generator, for example, a charge of 100 grammes of artifical resin reaches a temperature of 110 °C in 1 minute. Once the generator has been running for a time the electrodes have become so heated by the charges already passed through that for the following charges very little heat is lost and thus a highly uniform distribution of heat is obtained. It is due to this that the pellets become plastic without showing any trace of the granulation that usually occurs when preheating in an oven. It is only with capacitive preheating that the aforementioned advantages of reduced

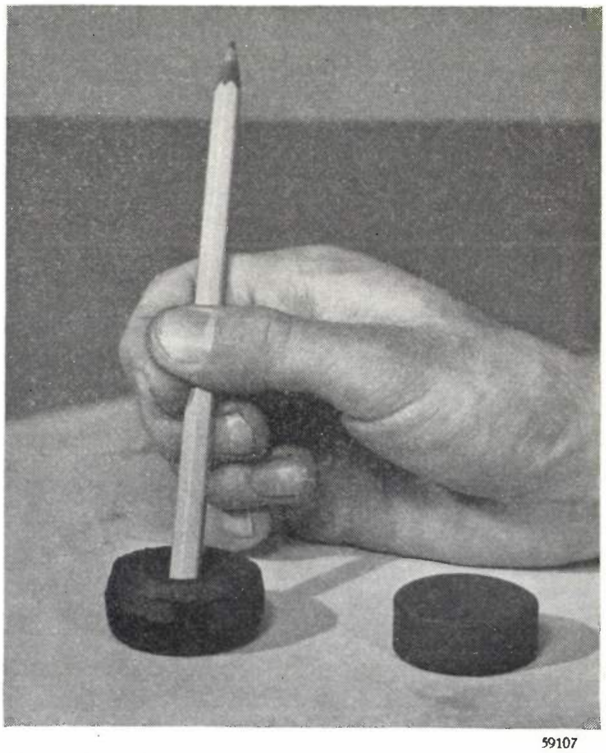


Fig. 8. Two thermosetting plastic pellets. The left-hand one has been made plastic by preheating and is ready for moulding.

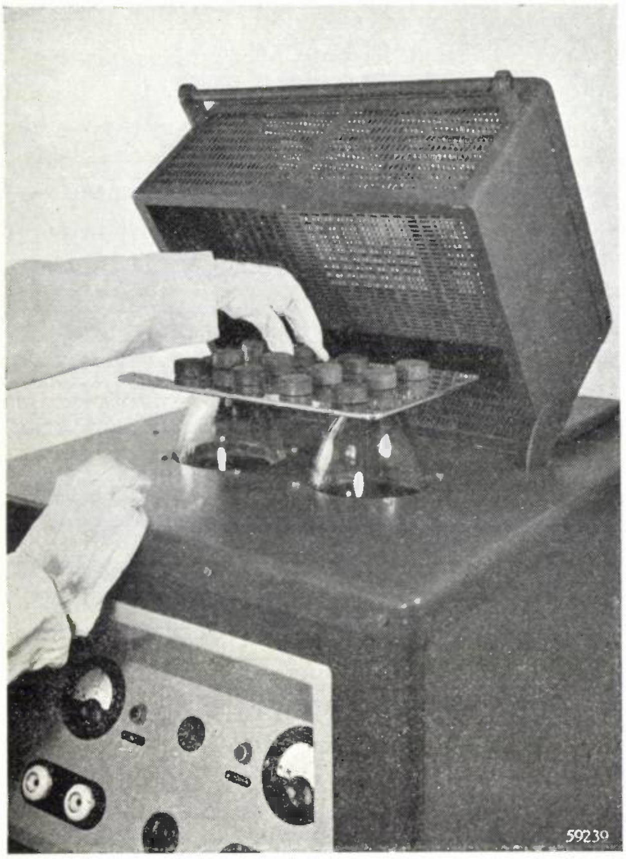


Fig. 9. Generator of the type illustrated in fig. 6*a*, with some preheated plastic pellets.

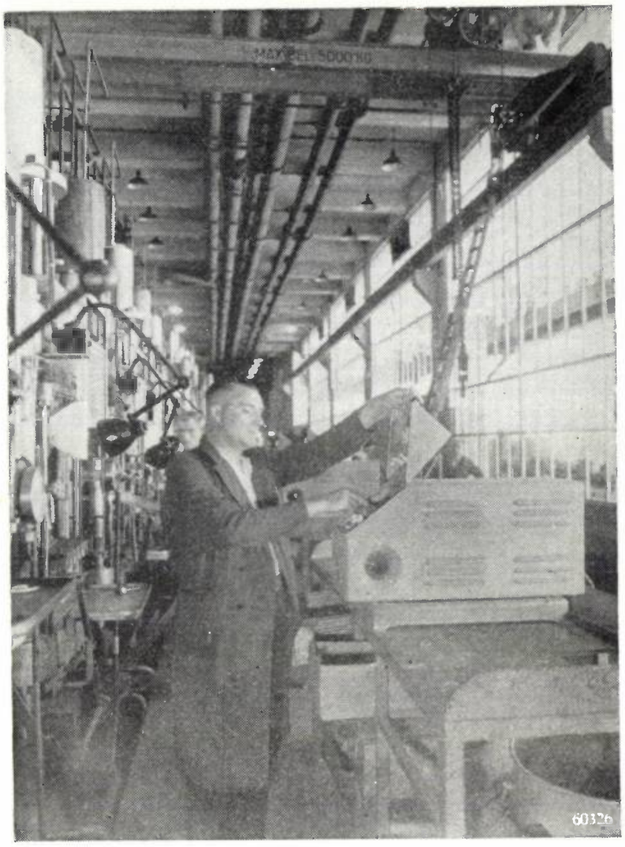


Fig. 10. Generators of the type depicted in fig. 6*b*, used in the "Philite" works at Eindhoven.

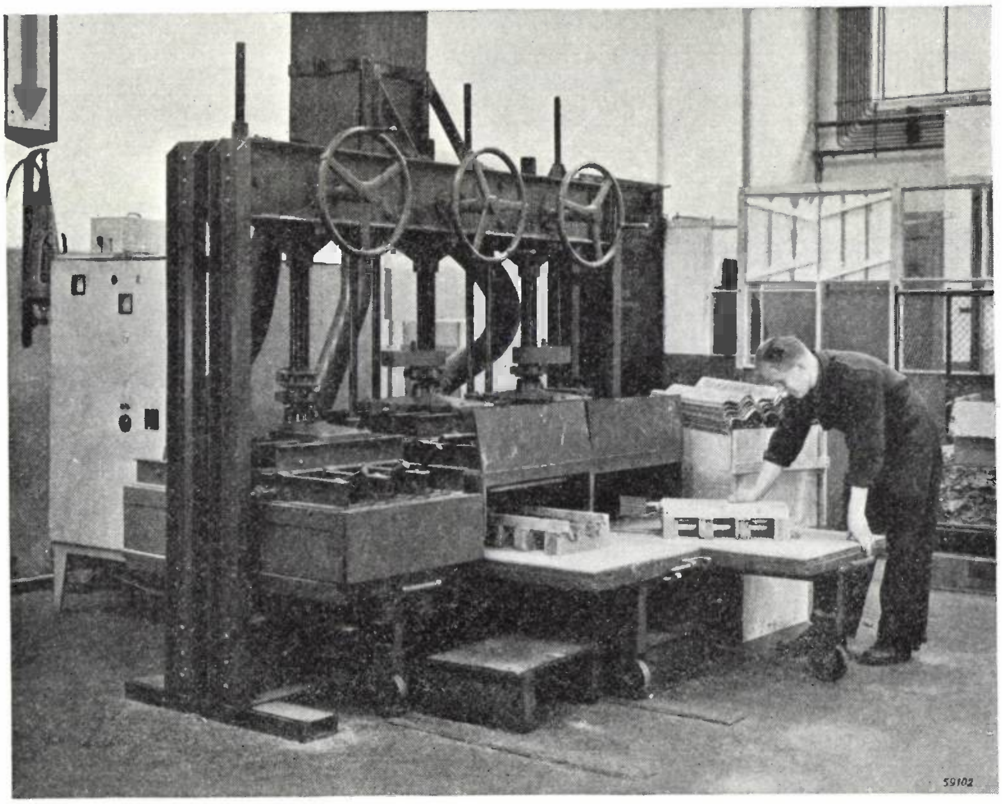


Fig. 11. Triple press for turning out corrugated sheets of plywood (seen on the right of the press). Two of the wooden moulds in which the sheets are pressed are shown. On the left in the background is the generator.

wear of the moulds and lower pressures are fully derived. Furthermore, if at the same time metal parts have to be moulded into the object the fact that the mass is made plastic throughout reduces the number of rejects; in a less plastic mass the metal parts are apt to get bent or be broken.

Compared with heating in an oven, capacitive heating has the further advantage that there is practically no inertia of the heat source, so that the rise in temperature can be stopped immediately, by switching off the generator (by hand or automatically). Thus it is easy to avoid heating to too high a temperature.

Consequently capacitive heating leads not only to the shorter moulding time mentioned but also, as follows from the foregoing, to a better control over all sorts of factors determining the quality of the product. Generally speaking it is therefore found that plastic objects produced by a manufacturing process where capacitive preheating is applied have better mechanical and chemical properties than those produced from a moulding mass preheated in some other way.

Capacitive heating in the wood-working industry

Plywood is made from a number of leaves of glued veneer stacked one upon the other. The glue used may be either urea or phenol formaldehyde with the addition of hardeners. Upon the stack of leaves being heated to a certain temperature (90 or 130 °C according to the kind of glue) and brought under pressure the hardeners bring about a polymerization of the glue which causes it to set hard. The plywood thus obtained is then ready for further processing (sawing, drilling, varnishing, etc).

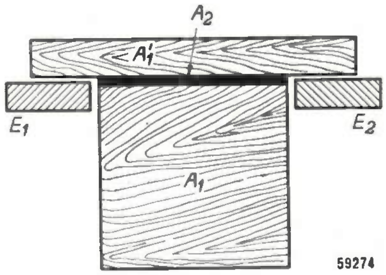


Fig. 12. A wooden block A_1 has to be glued onto a small plank A_1' with the glue A_2. The electrodes E_1, E_2 are applied in such a way as to obtain selective heating in favour of the glue.

By pressing in a shaped mould the plywood can be given all sorts of profiles, such as are often used for modern furniture, radio cabinets, etc.

The pressure needed in this process is so much lower than what has to be applied in the manufacture of plastics that no expensive steel moulds are required, simple wooden moulds answering quite well. This makes it possible for the capacitive heating to be carried out in a simple way during the moulding (in a period of time of the order of 60 seconds); all that is needed is two metal plates on the inside of the mould to act as electrodes.

Fig. 11 shows an installation for turning out corrugated sheets of wood as seen on the right of the press. This being a triple press, the generator can be run continuously.

In the manufacture of plywood the wood and the glue are connected in series. The material constants are such that no selective heating occurs. Selective heating can be obtained, however, in the gluing of larger workpieces, as we have seen in fig. 4. Another example of selective heating is the

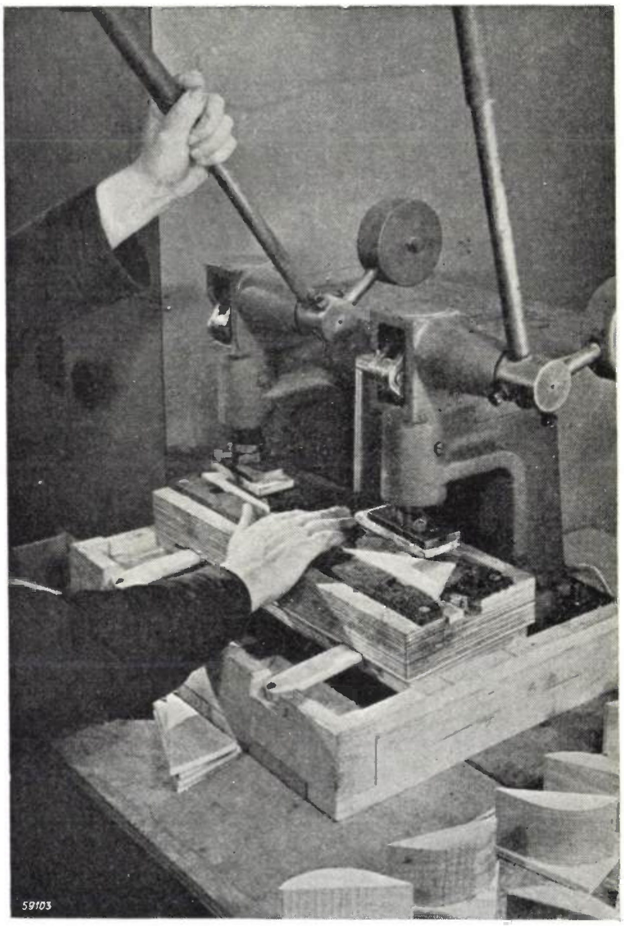


Fig. 13. Gluing by the method sketched in fig. 12. In the double hand press are to be seen the (triangular) planks; the blocks that are to be glued on are underneath the planks. The electrodes are out of sight.

gluing of a wooden block on a small plank (*fig. 12*). In order to direct as much of the energy as possible into the glue it has been purposely arranged to get, as it were, a parallel circuiting of the glue and the wood. Since the factor $(\varepsilon_r \tan \delta)/\varrho c$ of the glue is several times larger than that of the wood the heating is strongly selective in favour of the glue (cf. formula (5)). *Fig. 13* shows how this gluing is done with the aid of a simple hand press.

Summary. Capacitive heating is based upon the dielectric losses occurring in a dielectric situated in an alternating electric field. A formula is derived for the amount of heat generated per unit of time in a homogeneous dielectric. It appears that two simple cases of a non-homogeneous dielectric can be regarded as the parellel and series connection, respectively, of two homogeneous dielectrics; if the dielectrics are respectively glue and wood then with parallel connection an often desired selective heating may take place. After a brief consideration of the generators for capacitive heating, some important applications are discussed; manufacture of thermosetting plastics and of plywood, and the gluing of wood.

APPARATUS FOR PREPARATION OF METALS WITH AN EXACTLY KNOWN CONTENT OF IMPURITIES

by J. D. FAST.

621.365.56:66.046.516:669.77/.78

The presence of impurities in technical iron and steel determines for a large part the properties of these materials. It is difficult to study the influence of any particular element exactly because in nearly all cases several elements are present simultaneously. Only after thoroughly purifying the iron and then adding the desired element alone in an accurately weighed quantity is it possible to obtain data indicating the effect of that element.

Impurities in iron and steel

All normal technical steels and iron contain carbon, oxygen, nitrogen, sulphur and phosphorus. In special cases the last four constituents may, in combination with other elements, give the steel certain desired properties, but in most cases they are to be regarded as undesirable impurities which can never be entirely avoided in the large-scale production of steels. As a rule they have a particularly adverse influence upon the mechanical properties of the metal. The common grades of steel usually contain an amount of each of these four elements lying between 0.01 and 0.1 wt %. (In this paper all percentages are by weight.) In certain cases injurious effects may also be ascribed to carbon, which is not usually classed as an impurity and occurs in the various kinds of steel in greatly varying quantities (at least several hundredths per cent).

The harmful effect of these elements may become clear immediately after the shaping of the metal (by forging, rolling, drawing, etc) and the accompanying thermal treatment. Sometimes it also happens that the mechanical properties gradually change in the adverse sense subsequent to the last thermal or mechanical treatment, for instance while the metal is stored. One then speaks of ageing phenomena.

The question as to which impurity is actually responsible for a certain injurious effect that has come to light is often difficult to answer. The reason for this is that as a rule experiments are carried out with technical steels containing all the above impurities at the same time.

Probably the influence of sulphur is best known. In solid iron this element is practically insoluble but it combines with part of the iron, giving the chemical compound iron sulphide (FeS). In the liquid state iron and iron sulphide are homogeneously miscible in all proportions and form a eutectic with a comparatively low melting point (985 °C). When a sulphurous iron solidifies a metal is obtained in which the iron sulphide is located at the crystal boundaries. As a consequence in the mechanical treatment (e.g. forging) fractures occur along the crystal boundaries, both at low temperatures, because iron sulphide is brittle, and at high temperatures, owing to the presence of a thin layer of a liquid eutectic at the crystal boundaries. In order to counteract this injurious effect of sulphur an excess of manganese (with respect to the sulphur) is always added to the common technical steels, thereby binding the sulphur in the form of manganese sulphide. This compound has a higher melting point than iron and after solidification is found in the iron in the form of numerous small globules. Inclusions in such a form are comparatively harmless in metals.

Phosphorus cannot have the same effect as that described for sulphur.because it is highly soluble in solid iron even at room temperature (about 1%).

In order to study the effect of carbon, oxygen or nitrogen separately an apparatus has been designed which makes it possible to add exactly known quantities of these elements to molten pure iron. This apparatus will be described here.

The properties of the alloys prepared in this way will be discussed in a subsequent article, where attention will particularly be paid to the notched-bar impact strength of the material and to the ageing phenomena already mentioned.

The melting apparatus

In order to obtain a metal of known and reproducible composition the melting is done by induction heating. Here only a brief description of this heating method will be given; it has been dealt with at length in an article by E. C. Witsenburg recently published in this Journal [1]).

The energy is applied to the material to be melted by generating in the melting chamber a high-frequency alternating magnetic field. The heat developed in the metal is derived mainly from the eddy currents set up in it. In the case of iron,

[1]) Philips Techn. Rev. **11**, 165-175, 1949 (No. 6).

with which we are specially concerned here, below the Curie point the hysteresis losses also play a part. The eddy currents occur (when very high frequencies are applied) only in the outer layers of the material. The field is generated by placing around the melting chamber a coiled copper tube through which the high-frequency current flows and which is kept cool with running water. The high-frequency generator used has an output of 15 kW at a frequency of 330,000 cycles per second. The coil has 20 turns and a diameter of 13 cm (5").

One of the great advantages of such an induction oven is that the melting can be carried out in an apparatus made entirely of glass, the walls of which can be cooled with running water. Thus both the influence of the atmosphere and that of gases released from the walls is almost entirely precluded [2]). The apparatus can be evacuated up to a pressure of 10^{-8} atm; if desired it can then be filled with different pure gases. The crucibles are made of pure aluminium oxide. For some purposes crucibles of zirconium oxide or beryllium oxide are more suitable, but aluminium oxide is much cheaper.

Fig. 1 gives a diagrammatic representation (not drawn to scale) of the apparatus employed. On the extreme right is the cock (*1*) by means of which the melting chamber (*2*) can be connected to the high-vacuum pumps (not shown in the drawing). On the extreme left are cylinders containing nitrogen, hydrogen, oxygen, argon and carbon monoxide, which gases can be let into the melting chamber. Oxygen is admitted only in measured quantities, by using bulbs of known volumes (*3*). In the inlet pipe and also in the outlet pipe of the melting chamber are small valves, (*4*) and (*5*) respectively, with a ribbon of 18% Cr-8% Ni steel which can be heated by passing an electric current through it and which enables traces of oxygen or water vapour in hydrogen, nitrogen or argon to be detected by heat-tinting [3]).

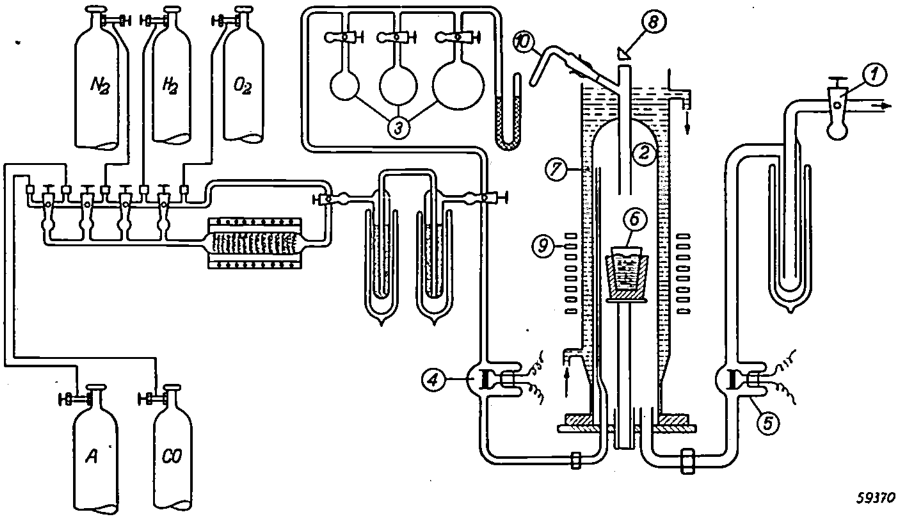


Fig. 1. Diagrammatic representation of the melting apparatus. *1* cock forming the connection with the high-vacuum pumps (not shown), *2* melting chamber, *3* bulbs with measured quantities of oxygen, *4* and *5* valves for detecting the presence of traces of water vapour and oxygen, *6* crucible, *7* cooling jacket, *8* prism for measuring the temperature of the melt with an optical pyrometer, *9* high-frequency coil, *10* rotatable tube for adding small quantities of solids to the melt. The diagram also shows a tube filled with copper turnings and placed in an oven, and two cooling vessels filled with the same material, serving to remove traces of oxygen and water vapour.

The crucible of Al_2O_3 (*6*) is contained in a larger crucible of clear quartz glass, the space between these being filled with coarse pieces of Al_2O_3 (fragments of discarded crucibles). The temperature of the molten metal can be measured with the aid of an optical pyrometer via a prism (*8*). A photograph of the whole set-up is given in *fig. 2*.

The preparation of metals of an exactly known composition

In order to produce iron with a definitely known content of impurities one has to start with the purest carbonyl iron. This is 99.9% iron and the

[2]) In this connection it may be interesting to point out that there is another method of achieving this object, namely by using concentrated solar energy by means of mirrors or lenses. See, e.g., F. Trombe, Utilization of solar energy, Research 1, 393-400, 1947 (No. 9).

[3]) G. W. Rathenau and H. de Wit, Metallurgia 40, 114, 1949,

impurities consist almost exclusively of carbon, oxygen and nitrogen. The nitrogen can be fully extracted by repeated melting in a good vacuum, whilst the same applies to the carbon if prior to the melting there was a large excess of oxygen in thc metal; if that is not already the case a known quantity of oxygen can be admitted to the metal from the bulbs (*3*). When all carbon has been driven out in the form of CO then the excess of oxygen is removed by reducing the liquid metal with flowing hydrogen for a long time. From the tinting of the Cr-Ni steel ribbon (*5*) it can be seen whether the oxygen has been completely removed. The hydrogen is removed in turn by melting in a good vacuum or in flowing argon.

The pure iron obtained in this way contains no more than 0.001% of carbon + oxygen + nitrogen, and the desired elements may then be added to it. Solids can be added to the melt by turning the branch tube (*10*) (fig. 1). Carbon, if desired as an impurity, can be added to the liquid iron in this way in the form of pure graphite. If the substances to be added evaporate rather easily, as for instance manganese, then the addition is made in pure argon instead of in vacuum. The additions blend homogeneously with the iron in a short time as a result of the circulation in the bath set up by the electromagnetic forces [4]).

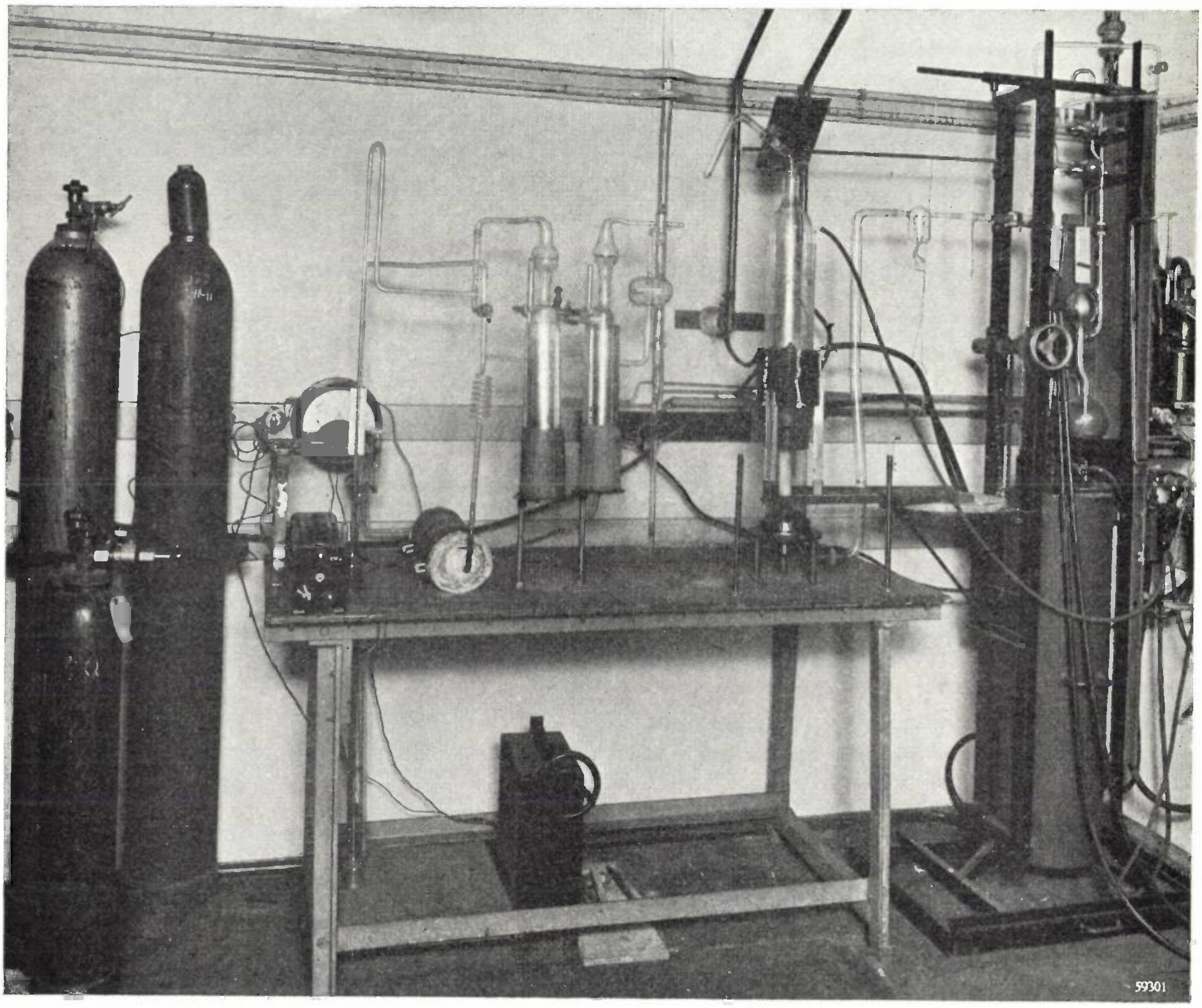


Fig. 2. Photograph of the set-up. On the extreme right are the high-vacuum pumps.

Oxygen and nitrogen are added in the gaseous state. The equilibrium pressure of oxygen over oxygen dissolved in liquid iron is negligibly low, so that the oxygen added is entirely absorbed by the iron. The equilibrium pressure of nitrogen over nitrogen in liquid iron, on the other hand, is high, so that only a relatively small part of the nitrogen added is absorbed. Moreover, as already discussed

[4]) See, for instance, Philips Techn. Rev. 1, 59, 1936.

in this Journal [5]), upon the metal solidifying it gives off a large part of the absorbed nitrogen. As a consequence the metal swells and after solidification shows porosity and large cavities.

In order to obtain nitrogen-containing iron without porosity one proceeds as follows. While the metal is liquid (temperature about 1600 °C) the nitrogen pressure is reduced from 1 to 0.2 atm, thereby reducing the equilibrium concentration in the metal from 0.046 to 0.021% of nitrogen. The high-frequency current is then switched off and the nitrogen pressure immediately raised again to 1 atm (before solidification begins). The metal then solidifies to a compact mass and after cooling down contains 0.022 to 0.030% of nitrogen. This effect has been found experimentally and can be understood, to a certain extent, from the solubility curve for nitrogen in iron. Smaller contents are obtained by choosing lower nitrogen pressures.

The procedure described is highly suitable for the preparation of comparatively small quantities of metal (300 to 400 grammes). When preparing the largest quantities that can be produced in the apparatus (about 2 kg) two difficulties arise. In the first place it takes too long to remove the last 0.01% of oxygen from the molten metal exclusively with the aid of hydrogen. For large melts, therefore, this last amount of oxygen is chemically bound by adding to the liquid metal the required quantity (determined by calculation) of pure zirconium, pure titanium or pure aluminium through the tube (*10*). The ZrO_2, TiO_2 or Al_2O_3 thereby formed remains in the iron for the greater part in a finely divided state. The influence of the oxygen is then much less than when it is bound in the form of FeO. The second difficulty lies in the fact that pure iron absorbs nitrogen so very slowly that it is almost impracticable to charge 2 kg of iron with nitrogen in the relatively deep crucibles used. On the other hand nitrogen is more rapidly absorbed in molten oxygenated iron and still more rapidly in molten carbon-containing iron. For experiments where an amount of material larger than, say, 1 kg is needed (for instance for measuring the impact strength as a function of temperature) the influence of nitrogen upon pure iron cannot, therefore, be studied directly but has to be deduced from the behaviour of metals containing oxygen, carbon, oxygen plus nitrogen and carbon plus nitrogen.

Experiments have been carried out not only with carbonyl iron but also with mild steel as basic material, in order to conform more to practice. The procedure followed with mild steel is exactly the same as that applied with carbonyl iron.

In conclusion it may be observed that with the apparatus described here it is also quite possible to prepare alloys with higher percentages of foreign elements, such as magnet steels for example.

Summary. In order to study the influence of carbon, oxygen, nitrogen and other substances upon the properties of iron a melting apparatus has been designed by means of which the impurities can be added to extremely pure iron in exactly known quantities. Melting is done by means of induction heating; the melting process can be followed throughout on account of the crucible being contained in a cooled glass jacket. The preparation of the purest possible iron is described. In the case of relatively large quantities (more than about 1 kg) it is found impracticable to study the influence of nitrogen separately. This influence can then only be deduced from the differences in the behaviour of iron containing oxygen or carbon with and without nitrogen as admixture.

[5] J. D. Fast, Philips Techn. Rev. 11, 101-110, 1949 (No. 4).

INSTABILITY OF SPRINGS

by J. A. HARINGX.

621.272.2:620.166

Little attention has so far been paid to the phenomena of instability that are apt to occur with resilient elements, such as helical springs, rubber rods and flat spiral springs, when these are subject to great distortions. It appears possible to explain this so-called buckling theoretically. Some remarkable properties which are brought to light can be confirmed by simple experiments.

Introduction

The occurrence of phenomena of instability in the case of constructional elements, such as rods, plates and thin-walled cylinders, may be assumed to be generally known. With such elements it is characteristic that so long as the load does not exceed a certain limit they retain their original shape. When that limit is reached, however, the element becomes suddenly distorted to a practically unlimited extent, so that it no longer answers its purpose. Here this phenomenon will be referred to as buckling.

For the purposes of the following explanation it will be useful to investigate the case of a metal rod of circular cross section with the aid of some simple formulae. When a straight rod is loaded axially at first it remains straight, but upon a certain compressive force being reached it suddenly deflects laterally. The critical compressive force causing this buckling depends upon the rigidity of the rod with respect to bending and is given, according to Euler's formula, by:

$$P_{\mathrm{cr}} = \pi^2 \frac{EI}{l^2}\,, \quad \ldots \ldots \ldots \quad (1)$$

where E is Young's modulus of elasticity of the material, l the length and I the moment of inertia of the cross section of the rod.

Although the changes in length of a metal rod are of course very small, for the sake of the analogy that is to be shown with the buckling phenomenon of helical springs a calculation will be made to determine in how far the rod is shortened before buckling takes place. The relation between the compression Δl and the compressive force P is given, as is known, by

$$\Delta l = \frac{Pl}{EF}\,, \quad \ldots \ldots \ldots \quad (2)$$

where F represents the cross-sectional area of the rod. The relative compression immediately prior to the buckling thus amounts to

$$\xi = \frac{\Delta l}{l} = \frac{P_{\mathrm{cr}}}{EF}\,. \quad \ldots \ldots \ldots \quad (3)$$

On account of equation (1), and with $F = \pi D^2/4$ and $I = \pi D^4/64$ respectively, for a rod of diameter D this formula becomes:

$$\xi = \frac{\pi^2}{16}\left(\frac{D}{l}\right)^2. \quad \ldots \ldots \quad (4)$$

The critical relative compression therefore appears to be dependent only upon the ratio l/D, which is to be construed as the slenderness of the rod, and not upon Young's modulus of the material.

The foregoing applies for a rod the ends of which are hinged or constrained parallel. For a rod with both ends clamped l has to be replaced in the formulae by half the rod length $l/2$, as may be seen from the following. *Fig. 1a* shows the deflection due to buckling in the case of a rod whose ends are constrained parallel. The rod is seen to assume the shape of a sine curve of half its wavelength. Fig. 1*b* shows a rod with the two ends clamped in the unstable state, the sine curve here consisting of a whole wave. Imagining this rod to be cut through the middle along the dotted line (s), it is seen that each half is in the same condition as the rod whose ends are constrained

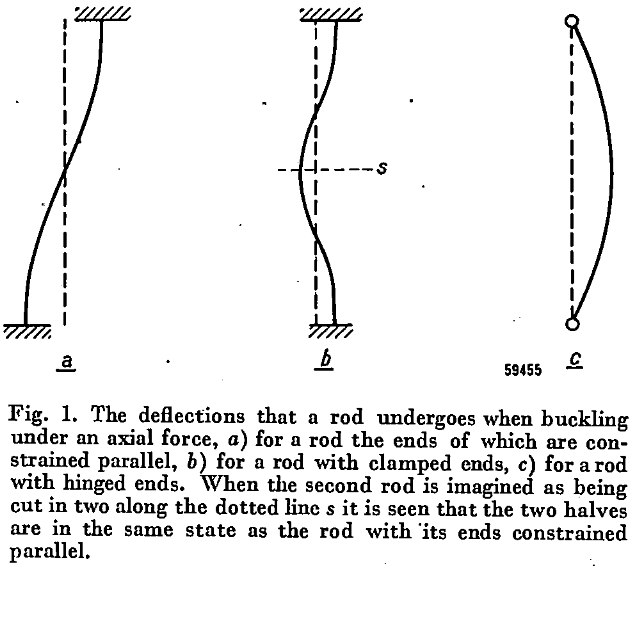


Fig. 1. The deflections that a rod undergoes when buckling under an axial force, *a*) for a rod the ends of which are constrained parallel, *b*) for a rod with clamped ends, *c*) for a rod with hinged ends. When the second rod is imagined as being cut in two along the dotted line *s* it is seen that the two halves are in the same state as the rod with its ends constrained parallel.

parallel. Thus here half the rod length determines the occurrence of buckling. The deflection due to buckling in the case of a rod with freely hinged ends is illustrated in fig. 1c, where the sine curve describes a half wavelength just as in fig. 1*a*.

It is to be noted that one must be careful when applying formula (4) because it holds only for elastic compressions of the rod (ξ smaller than about 0.001). In the event of the rod being subjected to such a load as to cause the yield point to be exceeded then considerable deviations arise.

It is obvious that under a compressive force helical springs may buckle just as well as straight rods. But in this respect, owing to their greater deformability, helical springs appear to possess some properties differing specifically from those of a straight metal rod. In what follows a description will be given of these remarkable properties as found by the author by calculation and subsequently confirmed experimentally. The buckling of elastic rubber rods is likewise briefly discussed. Further, attention will be paid to similar buckling phenomena shown by wound and unwound flat spiral springs, since it is also of importance to know under what conditions buckling takes place in these cases.

The behaviour of a helical spring under compression

Denoting the length of the spring in the unloaded state by l_0 and its diameter by D_0, in the case of a helical spring of circular wire section with its ends hinged or constrained parallel, the relative compression at which buckling takes place ($\xi = \Delta l/l_0$) amounts to:

$$\xi = 0.8125 \left\{ 1 \pm \sqrt{1-6.87\,(D_0/l_0)^2} \right\}. \quad . \quad (5)$$

This formula [1]) has been derived by introducing a simplifying assumption: the helical spring is imagined to be replaced by an elastic prismatic rod which has to consist of such a material that not only its rigidity with respect to bending but also that with respect to compression and shearing correspond to those of the helical spring. The fact is that, owing to the great deformability of the helical spring, its rigidities with respect to compression and shearing also play a part, while these rigidities in the case of a straight metal rod can safely be regarded as being infinitely large. From this it follows that the calculation for the helical spring differs essentially from that for a rod.

The introduction of the rigidity with respect to shearing involves a complication in regard to the transverse force responsible for the shearing. As the author has demonstrated, one has to take the transverse force in that cross section of the elastic rod (replacing the spring) which in the unloaded state is perpendicular to the central line of the rod. With the method of calculation hitherto commonly used, however, the transverse force has been taken which acts in the cross section perpendicular to the central line in the loaded state.

The formula given also holds for the case where the ends are clamped provided l_0 is replaced by half the spring length $l_0/2$. If the cross section of the wire is rectangular then only the values of the coefficients change.

It is to be noted that for $l_0/D_0 > 5$ the smallest value of ξ in formula (5) can approximately be written as

$$\xi \approx 2.80\,(D_0/l_0)^2\,, \quad . \quad . \quad . \quad . \quad . \quad (6)$$

which shows a decided agreement with formula (4). Apart from the indices only the coefficients differ.

Just as in the case of the metal rod, Young's modulus of the material has disappeared from the formulae and the critical relative compression depends only upon the ratio of the length to the diameter and upon the manner in which the ends are fixed. With the helical spring, however, this is much more striking than in the case of the rod, because it means that neither the diameter of the wire nor the pitch or number of coils has any effect upon the buckling phenomenon.

The relation, given in equation (5), between the critical relative compression ξ and the slenderness of the spring l_0/D_0 is graphically represented in *fig. 2.* The plotted curve is to be regarded as the line marking off the region of all combinations of spring slenderness and relative compression leading to instability. It shows that there is a certain critical slenderness ($l_0/D_0 = 2.62$ in the case where the

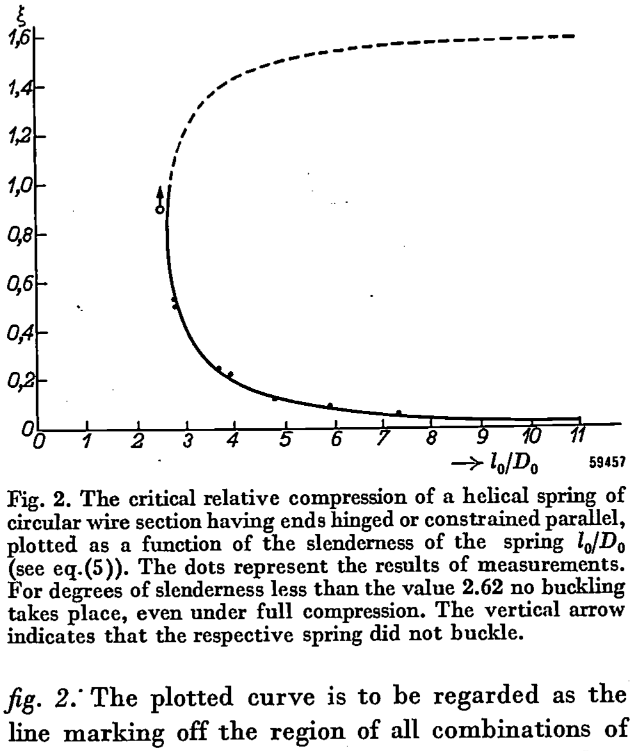


Fig. 2. The critical relative compression of a helical spring of circular wire section having ends hinged or constrained parallel, plotted as a function of the slenderness of the spring l_0/D_0 (see eq.(5)). The dots represent the results of measurements. For degrees of slenderness less than the value 2.62 no buckling takes place, even under full compression. The vertical arrow indicates that the respective spring did not buckle.

[1]) J. A. Haringx, On highly compressible helical springs and rubber rods, and their application for vibration-free mountings, Part I, sections 2 and 3, Philips Research Reports 3, 406-414, 1948 (No. 6).

ends of the spring are hinged or constrained parallel, and 5.24 if the ends are clamped) below which the spring does not buckle and above which it does. As a result of the steepness of the curve near this critical value, it so happens that any small variation of slenderness in the neighbourhood of 2.62 or 5.24 gives rise to very great differences in the behaviour of the spring. If this value of the slenderness is only slightly exceeded then there is a decided buckling of the spring when it is compressed to about half its original length. On the other hand there is no question at all of instability if the slenderness of the spring is slightly less than the critical value. This difference is demonstrated in *figures 3a* and *3b* by the two helical springs clamped at the ends. The right-hand spring with slenderness $l_0/D_0 = 5.3$ definitely buckles, whereas the left-hand one with slenderness $l_0/D_0 = 5.0$ remains straight and does not buckle.

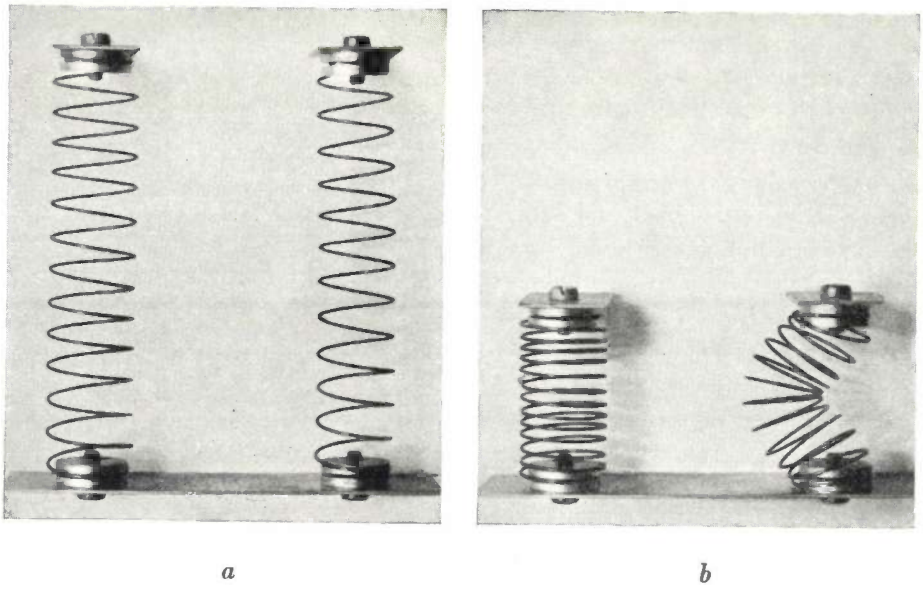


Fig. 3. *a*) Two non-loaded helical springs with clamped ends and differing slightly in size. The slenderness of the left-hand spring is 5.0, thus a little below the critical value, which for springs clamped at both ends amounts to 5.24; the slenderness of the right-hand spring is 5.3, thus a little above the critical value. *b*) The same springs under a relative compression of 60 %. The left-hand one remains straight, the right-hand one buckles.

The results of the calculation have been further checked experimentally by measuring the relative compression at which buckling takes place in the case of a number of helical springs of different degrees of slenderness ($D = 18$ mm, wire diameter $= 0.5$ mm and 2 mm, pitch $= 5$ mm). The results of these measurements are indicated in fig. 2 by dots. The arrow indicates that the spring of the corresponding slenderness did not buckle. As is to be seen, there is very satisfactory agreement with the calculated curve.

Under certain circumstances, however, there may be not inconsiderable deviations. In the first place the "clamping" of the ends of the spring is usually no fixture at all, but only a pressing of the flat-wound end coils against a flat surface. The elasticity of the end coils then plays a part, with the result that the spring buckles when the compression is apparently too small.

In the second place the length of the spring is understood to be the "free" length, that is to say, the length of that section of the spring taking part in the deflection. Now it may happen, especially if the pitch of the spring is comparatively small with respect to the thickness of the wire, that in the compressed state the last free coils make contact with the end coils, in which case buckling takes place at an apparently too large compression. It is possible to avoid these and other effects to a certain extent by giving the planes of the end coils in the unloaded state an oblique position. For further details reference is made to the calculation of the effect of the flat-wound end coils upon the lateral rigidity of helical compression springs [2]).

Behaviour of the helical spring at a compression greater than the original length

Normally a (cylindrical) helical spring can never be compressed beyond the point where all contiguous coils make contact. For this reason only the full-drawn part of the curve in fig. 2, corresponding to compressions Δl less than the original length of the spring l_0 $(\xi < 1)$, is of practical importance.

[2]) See footnote[1]), Part II, sections 11 and 12, Philips Research Reports 4, 57-68, 1949 (No. 1).

Theoretically, however, the region $\xi > 1$ is very interesting, because, according to the dotted part of the curve , a spring that has once become unstable would become stable again if it were compressed far enough. To show that this is indeed the case we need a spring compressed farther than its original length, a condition which at first sight appears to be difficult of realization but which can in fact be brought about by turning a normal helical spring inside out.

To do this we take a helical spring with a small wire diameter, for instance 1/50th of the coil diameter. Beginning at one end of the spring, we draw the second coil over the first one, then the third coil over the first and second ones, and so on, thus reversing the positions of each coil's neighbours. Without our help they cannot return to their original position, so that they are of necessity pressed against each other. Consequently the spring is already under an initial load and subject to a compression Δl exceeding the original length of the spring l_0 by the product of the number of coils and the diameter of the wire. The relative compression is then indeed greater than 1. When such a spring that has been turned inside out is "stretched" (corresponding to a further compression of the original spring) then ξ is still further increased. Now this stretching is accompanied by a very remarkable phenomenon. The coils first slide along each other and tend to take up a parallel position; the spring assumes the shape of a flat ribbon, as shown by the middle part of the spring in *fig. 4a*. When, however, ξ exceeds a certain value the wire of the spring suddenly takes the shape of a helix (fig. 4*b*). Upon the spring being made "shorter" again, with more or less the same critical value of ξ the coils tumble over and the wire of the spring resumes the shape of fig. *4a*, thus proving that the spring does indeed show a second transition from the stable to the unstable state in the region where $\xi > 1$.

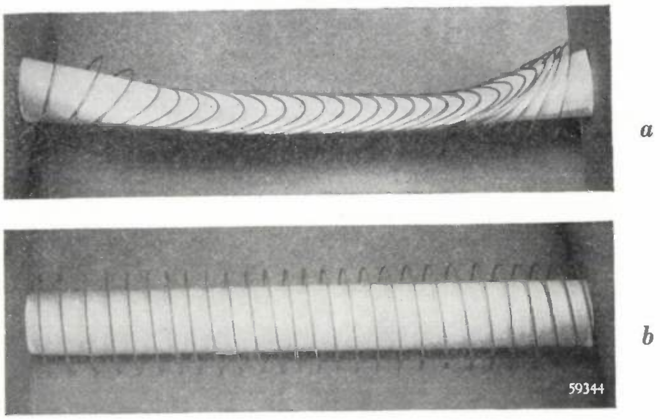


Fig. 4. *a*) A helical spring with clamped ends at a compression greater than the original length of the spring but not yet so great as to cause the second stability transition point to be passed. In order to give a better insight into the three-dimensional shape of the spring wire the spring has been passed over a thin paper cylinder, which does not constitute any hindrance for the free distortion of the spring. *b*) A helical spring with clamped ends at a compression greater than the original length of the spring, being so great that the second stability transition point is passed and the spring has again assumed the shape of a helix.

A further calculation [3]) has shown that the part of the curve in fig. 2 for $\xi > 1$ can only hold for

[3]) J. A. Haringx, Elastic stability of helical springs at a compression larger than original length, Applied Scientific Research, **A1**, 417-434, 1949.

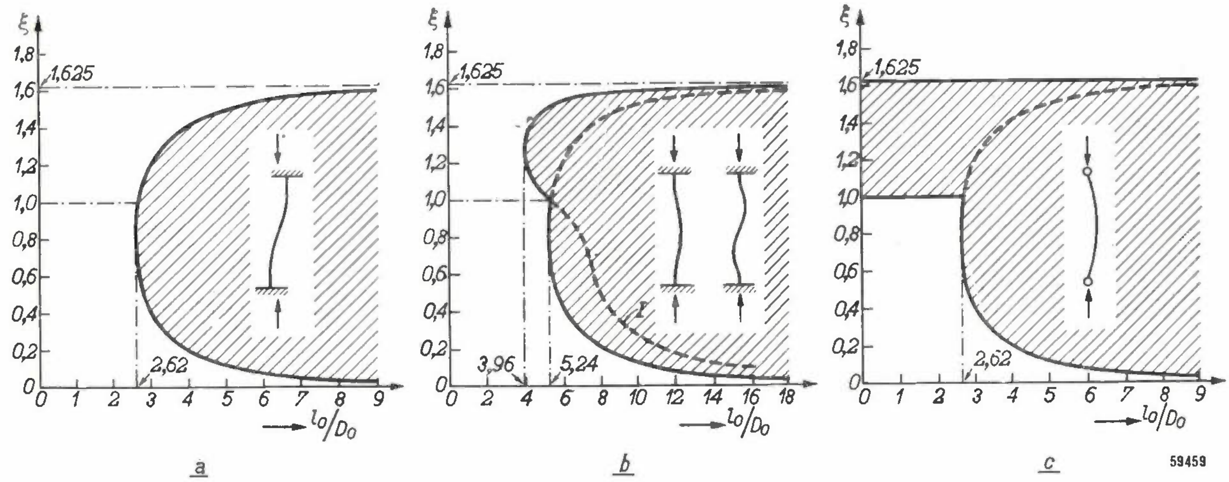


Fig. 5. The relation between the relative compression at which buckling takes place and the slenderness of the spring, for different methods of clamping the ends. *a*) Spring with its ends constrained parallel; this is the same curve as given in fig. 2; *b*) spring with clamped ends; here asymmetrical deflection takes place at values of $\xi > 1$; the curve relating to this deflection is that denoted by *I*; *c*) spring with hinged ends; the whole of the region between $\xi = 1$ and $\xi = 1.625$ corresponds to unstable states.

helical springs with the ends constrained parallel. The region of the combinations of spring slenderness and relative compression leading in this case to instability is represented by the hatched part of *fig. 5a.* The drawing in the open space of the diagram represents a possible deflection of the central line of the helical spring and gives an idea of the manner in which the ends of the spring are fixed. In the case of springs with clamped ends — taking into account a factor 2 for the horizontal scale — the upper part of the curve in fig. 2 assumes a different shape; see fig. 5*b*. The abrupt change at $\xi = 1$ from one curve to the other is related to the variable preference for the symmetrical or the asymmetrical deflection of the spring. Springs with hinged ends are always unstable for all values of ξ between 1 and 1.625, as indicated in fig. 5*c*.

Qualitatively the experiments confirm the results of the new calculation for the various methods of fixing the spring ends, but quantitatively there are certain deviations. It must be borne in mind, however, that in the region of $\xi > 1$ the coils of the spring tumble over through angles of almost 90° (see fig. 4*a*), so that the rotations can no longer be regarded as being infinitely small, as was assumed in the calculation.

Behaviour of a rubber rod under compression

Although, for the sake of simplification, the buckling of helical springs has been calculated with the spring imagined as being replaced by a (fictitious) elastic prismatic rod, the calculation for an elastic rod actually realizable is much more complicated. This is due, inter alia, to the fact that the relation between the compression of the rod and the axial load is no longer linear. A calculation for a rubber rod of circular cross section (original length l_0 and original diameter D_0) having the ends hinged or constained parallel yields the result [4]):

$$\xi = \frac{1}{1 + 1.62\ (l_0/D_0)^2} \quad . \ . \ . \ . \ . \quad (7)$$

Here again the critical relative compression $\xi = \Delta l/l_0$ depends only upon the slenderness l_0/D_0 of the rod. Where l_0/D_0 is greater than 5 formula (7) is practically identical with formula (4), as was to be expected.

The curve in *fig. 6* is a graphical representation of equation (7), whilst the dots plotted have been taken from measurements carried out by Kosten [5]). The agreement between the calculation and the experiment can be said to be very satisfactory. Whereas, however, according to the calculation every rod, however short, must buckle, experimentally a certain critical slenderness is found below which there is no possibility of buckling ($l_0/D_0 \approx 0.6$; see the measuring point in fig. 6 with vertical arrow indicating that the respective rod did not buckle). This behaviour bears a strong resemblance to that of helical springs, except that in the latter case the calculation provided a direct explanation for this behaviour. It must not be forgotten, however, that the elastic properties of highly compressed rubber can only approximately be expressed in formulae. For compressions greater than 50 % these formulae are certainly no longer reliable, and it is for this reason that the calculated curve for values of ξ greater than 0.5 has been drawn with a dash line. Moreover, under compression rubber rods usually lose the prismatic shape owing to their ends being fixed, and a thin layer of the material at the ends has to be regarded as being absolutely incompressible. Although this effect too can be partly taken into account, it still remains a source of deviations.

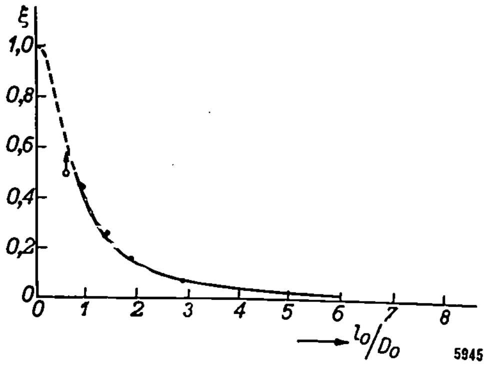


Fig. 6. The critical relative compression of a rubber rod of circular cross section and ends hinged or constrained parallel plotted as a function of l_0/D_0. The dots indicate the results of measurements by Kosten [5]). The arrow denotes that the respective rod did not buckle.

Phenomena of instability of a flat spiral spring

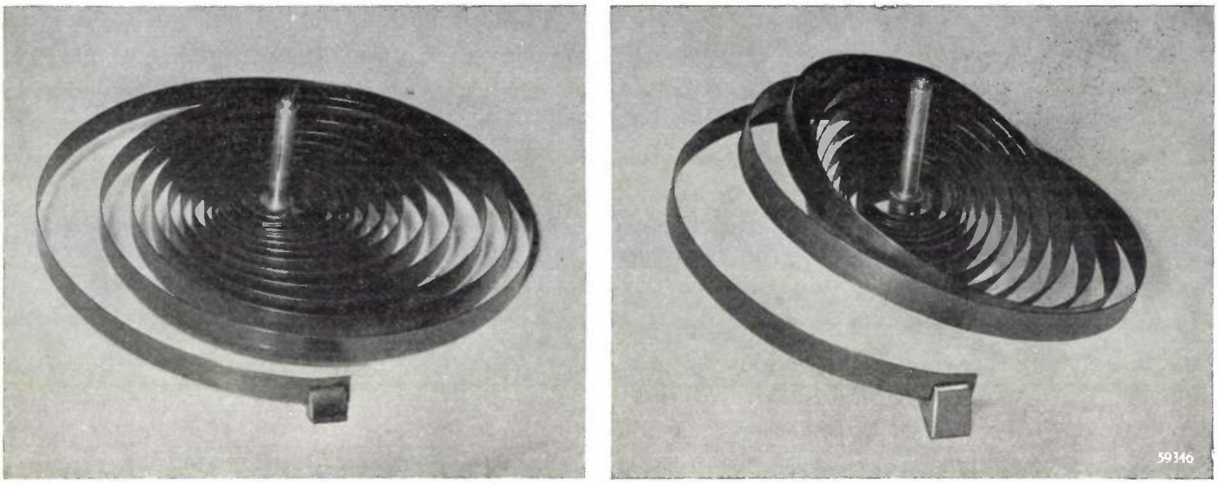
Apart from an axial compressive force, also a torque or a combination of these two loadings may cause a helical spring to buckle [6]). Since such cases, however, seldom occur in practice we shall not pay any further attention to them in this paper. Nevertheless this fact is mentioned because a relationship exists between the helical spring subjected to torsion and a flat spiral spring which has been wound up. This is readily understood

[4]) See footnote [1]), Part III, sections 1 and 2, Philips Research Reports 4, 206-216, 1949 (No. 3).

[5]) C. W. Kosten, On the elastic properties of vulcanized rubber (in Dutch), Thesis Delft 1942, p. 66.

[6]) See footnote[1]), Part I, sections 8 and 9, Philips Research Reports 3, 435-449, 1948 (No. 6).

when the length of the helical spring is imagined as being reduced to nil. It will therefore not be surprising that flat spiral springs, like those used in timepieces, are apt to buckle when being wound up or unwound, just as is the case with helical springs subjected to torsion.



a *b*

Fig. 7. *a*) A flat spiral spring wound up just far enough to avoid buckling. *b*) The same spring after instability has set in.

When one of the ends of such a spiral spring is clamped and the other end is turned about the axis of the spiral then at first the coils continue to lie in the plane of the spiral as shown in *fig. 7a*, but as soon as the spring has been wound up a critical number of turns, the coils suddenly tilt in the manner indicated in fig. 7*b*. This critical number of turns has been calculated for spiral springs wound up and unwound [7]). In the case of spiral springs having a large number of coils the calculation leads to the following result. The number of coils that the spring has to be wound up or unwound before buckling takes place does not appear to depend upon Young's modulus of the material nor upon the number of coils and the dimensions of the spiral. The only determining factor is the ratio of the sides *a* and *b* of the (rectangular) wire section. The relation between this ratio and the critical number of coils N is represented in *fig. 8*. The side *a* of the wire section is directed radially and the side *b* is perpendicular to the plane of the spiral.

When b/a is greater than 10, as will nearly always be the case in practice, the influence of this ratio b/a disappears almost entirely (N varies only from 2.84 to 3.08). It may therefore be said that for most spiral springs the number of turns required to cause buckling is practically the same. This number of turns is approximately three.

As a rule the number of coils of a spiral spring will be rather small and the question is whether the results found also hold for such a spring. For spiral springs with a small number of coils the calculation is much more complicated, but it can nevertheless be carried out for a (fictitious) spiral spring having circular coils of equal diameter. It appears that a distinction has to be made between the number of turns the spring has to be wound up and the number of turns it has to be unwound to cause buckling. The differences with respect to the results for spiral springs having a large number of coils are, however, small.

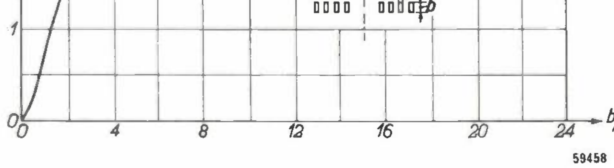


Fig. 8. The relation between the number of turns N that a flat spiral spring with (infinitely) large number of coils has to be wound up or unwound to cause it to buckle, and the ratio of the sides b/a of the rectangular wire section.

[7]) J. A. Haringx, Elastic stability of flat spiral springs, Applied Scientific Research, A2, 9-30, 1949.

To verify the results of the calculations tests were carried out with some spiral springs with different numbers of coils and different wire sections to determine at what number of turns they buckle. The results are given in the table below, where a comparison between the third and fourth columns shows that there is satisfactory agreement with the calculation.

The number of turns made in winding up the spring has been taken as positive, so that the negative sign on the last line in the table means that the spring was unwound. For judging the effect of the number of coils it is indicated in the last column what number of turns would be critical for spiral springs with (infinitely) large number of coils. The difference does indeed appear to be small.

It is to be noted that during the experiments the clamped ends of the springs were so arranged that successive coils could not touch each other and the spring remained flat until the moment that it suddenly buckled. If these steps for a careful adjustment of the position of the wire ends are not taken the spring will jump out of its plane at a 10 to 20 % smaller number of turns.

Number of coils	Ratio of sides of the wire section b/a	Number of turns required to cause buckling		
		Measured	Calculated	Calculated for infinite number of coils (fig. 8)
3.1	10	2.5	2.52	2.84
4.2	8.6	2.6	2.55	2.78
4.5	10	2.6	2.62	2.84
13.5	19	2.75	2.89	3.00
13.5	19	–2.9	–3.10	–3.00

Buckling can be avoided by placing the spiral spring in a box or between two parallel discs, but then there is inevitable friction between the spring and the wall or between the coils themselves, with the result that the magnitude of the torque of the spring is not reproducible. This objection can usually be accepted, but it is quite possible that in certain cases a frictionless construction is to be preferred. For instance, friction must certainly be avoided if the spiral spring is to be used as a resilient element in a controlling or measuring apparatus which has to be rotatable over large angles. When only one spiral spring is used the angle of rotation is then limited to the maximum of three turns mentioned above.

Summary. Some properties of helical springs, rubber rods and flat spiral springs (as used in timepieces) are discussed in their relation to buckling. The analogous phenomenon in the case of a metal rod is dealt with as an introduction to the discussion of the behaviour of a helical spring. The relative compression at which the spring of circular wire section buckles is found to depend only upon the ratio of the length to the diameter of the spring (its so-called slenderness) and upon the manner in which its ends are fixed. There is a minimum slenderness below which the spring does not buckle even in the fully compressed state. A second transition from the stable to the unstable state occurs at a compression greater than the original length of the spring. Such a compression can be brought about experimentally. Also in the case of the rubber rod the buckling is governed by the ratio of the length to the diameter and the manner in which the ends are fixed. In the case of a flat spiral spring the angle over which this has to be wound up or unwound to cause buckling depends only to a very small degree upon the dimensions of the spring; it appears to be impossible to design a spiral spring which can be wound up more than three turns without buckling.

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

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

R 112: G. W. Rathenau and J. F. H. Custers: Secondary recrystallization of face-centred Ni-Fe alloys (Philips Res. Rep. 4, 241-260, 1949, No. 4).

The primary texture of severely rolled Ni-Fe alloys is the cubic orientation. At high temperatures abnormal grain growth may occur: secondary recrystallization. The orientations of secondaries are different from the cubic orientation. Several new orientations were found. Alike orientations are encountered, whether the secondaries grow in a normal way or under quite different conditions (1-3 below). Normal secondary recrystallization is supposed to be the grain growth of primary crystals with a high temperature of primary recrystallization. One is led to this view by the following experimental evidence: (1) Thoroughly annealed strips were locally deformed by a pin prick and reheated. Large crystals grew frequently from the pinhole into the cubic matrix. Pin pricks made in secondary crystals did not give rise to grain growth outside the area of local deformation. (2) By intermediate local annealing a soft region was inserted in a cold-rolled strip. On recrystallization large secondaries grew from the boundary of the soft region into the cubic matrix. (3) Besides rolling, a deformation of another kind, such as cutting, hammering, was locally applied. Recrystallization produces secondaries which grow from the boundary of the locally deformed area into the cubic matrix. On cutting rolled strips, the direction of cutting and the orientation of the resulting secondaries proved to be related. (4) Rapid heating in the temperature region of primary recrystallization favours the nucleation of secondaries. (5) Large secondaries were observed only within a restricted region of rolling deformation, the extension of which depends on the rate of heating.

R 113: J. A. Haringx: On highly compressible helical springs and rubber rods, and their application for vibration-free mountings, IV (Philips. Res. Rep. 4, 261-290, 1949 No. 4).

In this paper (see Abstracts Nos **R 94**, **R 101** and **R 109**) the different types of vibration-free mounting are treated, though in the one-dimensional case only. Of the well-known elementary systems with relative or absolute damping the latter is superior, since the relative damping unfavourably affects the forced vibrations at frequencies exceeding $\sqrt{2}$ times the resonant frequency. Absolute damping can unfortunately not be realized, but it can be imitated by applying an auxiliary mass in addition to the damping element. Usually the application of such a damped dynamic vibration absorber is limited by the demand that the auxiliary mass shall be small in relation to the main mass; with vibration-free mounting of instruments, however, there is in general no objection to the auxiliary mass being so large as to be able to take full advantage of the damping properties of this system in the region of the resonances. For optimum results some constructional conditions in respect of the spring rigidities and the coefficient of damping have to be satisfied. To determine these, the forced vibrations are examined by means of frequency characteristics, whilst for the judgement of the decay of the free vibrations a fictitious logarithmic decrement is introduced. Cf. Philips Techn. Rev. **9**, 16-23, 85-90, 1947.

R 114: E. Labin: Théorie de la synchronisation par contrôle de phase (Philips Res. Rep. 4, 291-315, 1949, No. 4). (Theory of synchronization by phase control; in French.)

An auto-oscillator, the frequency of which is controllable by means of D.C. voltage, can be synchronized with a pilot by taking the control voltage from a mixing stage in which the phases of the the two oscillators are compared. This method is first discussed when both oscillations are sinusoidal, and then extended to the case where one of the two oscillations consists of pulses, a condition that has been realized in the so-called I.G.O. system (impulse-governed oscillator). The salient properties of this method of synchronization as revealed by experiment are explained.

R 115: A. Guinier and J. Tennevin: Comparison of the perfection of the crystals of primary and secondary recrystallization (Philips Res. Rep. 4, 316-318, 1949, No. 4).

In connection with experiments on secondary recrystallization of Ni-Fe alloys (see Abstract No. **R 112**) the degree of perfection of primary and secondary crystals has been investigated, using a method previously described by the authors. Secondary and primary crystals proved to be equally perfect, whereas in the case of the growth of large crystals of Al, preceding primary crystallization, as described by Chevigny, considerable imperfections were found. Considering their limited sensitivity the experiments do not contradict the opinions expressed by Rathenau and Custers.

VOL. 11 No. 9, pp. 253-280 MARCH 1950

Philips Technical Review

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

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

A PRESERVATION RECTIFIER WITH ELECTRONICALLY STABILIZED CHARGING VOLTAGE

by E. CASSEE.

621.314.6.076.7: 621.316.261

In almost every field of electrical technology there are nowadays installations in which electronic control is applied. The present article explains how with such a control a rectifier can be given a characteristic highly suitable for preserving or maintaining the charge of an accumulator battery.

Preservation charge

In the course of time electricity supplies have been so improved by developments as to form a source of energy quite reliable for industrial and domestic purposes notwithstanding the breakdowns that are still apt to occur at times. Nevertheless there are a number of cases where even a very short interruption of the current supply may have unpleasant and even fatal consequences; we have only to think of telephonic, telegraphic and telex communications and the lighting of theatres, traffic tunnels, operating theatres, etc. Wherever this is the case an emergency supply system, consisting for instance of a battery of accumulators, is therefore installed.

An article which appeared earlier in this journal [1]) dealt at length with what was termed the "hygienics" of the battery, i.e. its maintenance. There it was shown that a single battery kept under what is called a preservation charge has great advantages compared with the former system generally employed for telephone and similar plants where two batteries are charged and discharged alternately. As the first of these advantages is to be mentioned the much longer life of the battery, and further the saving in initial cost, space and upkeep when instead of two batteries only one is used. Other advantages lie in the higher efficiency of preservation charging and the absence of any formation of gases and noxious vapours.

Preservation charging means that the battery is kept fully charged by feeding it continuously with a low current just sufficient to compensate the internal losses. Thus the battery is permanently connected in parallel with the source of the charging current (and, for instance in the case of a telephone exchange, also with the direct-current-consuming network, in which case the battery is said to be kept in a "floating" condition). This charging-current source, which, as an example, may be a rectifier fed from the A.C. mains, has to answer very special requirements. It has to feed the battery continuously with a small current just sufficient to make up for the internal losses of the battery, this being achieved by maintaining a voltage of 2.1 to 2.2 V per cell [2]). In the case of a telephone exchange the charging current has also to supply the current required for the working of the plant (this current is not usually supplied by the battery, except in a special case to which we shall revert later, and of course during a failure of the A.C. mains).

The manner in which a rectifier possessing these properties — "a preservation rectifier" — can be devised has already been mentioned in the article referred to in footnote [1]) and also in an article of an earlier date [3]). In the simplest case the principle is as represented in *fig. 1*: a set of

[1]) H. A. W. Klinkhamer, Emergency supply systems with accumulator batteries, Philips Techn. Rev. 9, 231-238, 1947.

[2]) Here we are speaking of lead accumulators. Our arguments apply, mutatis mutandis, in broad lines also for nickel-iron cells.

[3]) H. A. W. Klinkhamer, A rectifier for small telephone exchanges, Philips Techn. Rev. 6, 39-45, 1941.

valves (for instance connected for single-phase full-wave rectification) is fed from the series-connected secondary windings of two transformers, one of which (T_2) is connected direct to the A.C. mains while the other (T_1) is connected to the mains via a capacitor. It is essential that the iron core of the latter transformer should be strongly saturated. For the working of this system reference may be made to the articles [1] and [3] already quoted. In the latter article it was stated that, with a certain form of application of the system, in a wide range of the currents supplied by the rectifier (0.3 to 3 A) the charging voltage (i.e. the battery voltage) varies by only a few per cent (from 63 to 60 V). Furthermore the charging voltage is very little affected by variations in the mains voltage, a fluctuation of 10% causing a change in the charging voltage of at most 3%.

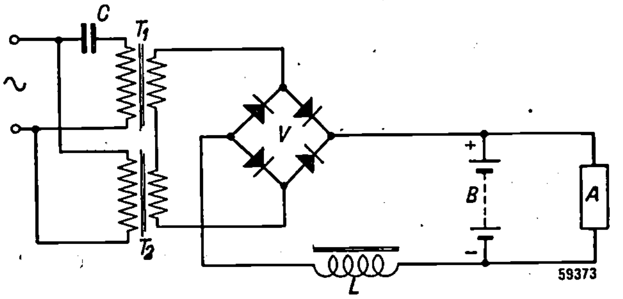


Fig. 1. Circuit of a preservation rectifier previously described (see footnotes [1] and [3]). T_1 = transformer with strongly saturated core, T_2 = ordinary transformer, C = capacitor in series with the primary of T_1, V = single phase, full wave rectifying circuit, L = smoothing choke, B = battery, A = plant.

In three respects, however, the rectifiers working according to this system have less favourable properties. In the first place there is the fact — already mentioned in the article quoted in footnote [3]) — that the charging voltage is to a rather considerable extent dependent upon the frequency of the A.C. mains, a deviation of 1% from the rated frequency causing a change of about 1.6% in the charging voltage. Now since the war frequency variations up to about 4% often occur during a number of hours per day, at least in the networks in Western and Central Europe. The resultant variations in the charging voltage, together with those due to mains voltage fluctuations and to fluctuations in the consumption of the direct current, may be greater than is permissible for the proper preservation of the battery.

The second drawback is that the charging voltage is dependent upon the voltage loss in the valves, which is not constant in the course of time. In selenium rectifiers, for instance, the voltage loss increases with the current passed through, whilst, moreover, at a certain current intensity it becomes gradually greater, especially during the first period after the valve has been taken into use. It is of course possible to compensate changes in the voltage loss occurring with a given strength of current by connecting the valve to a different transformer tapping, but this is only a makeshift solution. In gas-filled rectifying valves the voltage loss (arc voltage) varies only slightly with the current but it nevertheless sometimes shows irregular fluctuations which cannot be compensated at all.

The third drawback of rectifiers according to the principle of fig. 1 is not of a fundamental nature but concerns the manufacture. The characteristic of these rectifiers representing the charging voltage as a function of the direct current is highly dependent upon the shape of the B-H curve of the iron in the transformer T_1, particularly in the region of strong saturation. Now the prescriptions generally applied for the testing of transformer sheets say nothing about this and transformer sheets from different batches show considerable differences, which of course give rise to difficulties in manufacture.

The increasing demand for preservation rectifiers has therefore made it necessary to find a system that is free of these drawbacks.

Preservation rectifier with electronic control

Stabilization of the charging voltage

The solution has been found in a system where the battery voltage itself performs a controlling action by regulating the D.C. magnetization of a transductor. A transductor is a choke whose impedance for alternating current can be varied by means of an auxiliary direct current (the control current) flowing through a separate winding on the core of the A.C. coil. With a transductor it is possible to arrange for a small D.C. power to control a much greater A.C. power or, if the alternating current is rectified, a much greater D.C. power.

The circuit is shown in *fig. 2*, where *Td* represents the transductor. When, for example, the D.C. output I_0 of the rectifier increases then the charging voltage tends to drop. However, by making the control current I_a of the transductor dependent upon the charging voltage — by means of the circuit denoted by F — the drop of this voltage can be made to increase the control current to such an extent that the impedance of the A.C. coil — and thus the voltage drop in that coil — decreases far enough to compensate practically the drop of the charging voltage.

For this purpose the control current cannot be taken from the battery direct, in the first place because the permissible fluctuations in the battery voltage are so small that they would have very little effect in the transductor, and in the second place because this effect would be just in the wrong direction: for instance, when the battery voltage increases the control current would likewise increase, and, as already shown, this would result in a further increase of the battery voltage. That is why the battery voltage fluctations have to be amplified (in F, fig. 2) before they control the transductor, this amplification being accompanied by the necessary inversion of sign.

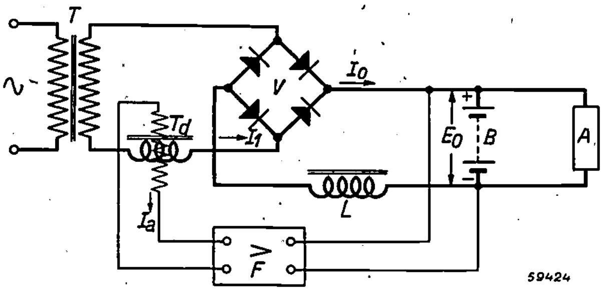


Fig. 2. Circuit of a preservation rectifier with a transductor (Td), the control current of which, I_a, is supplied by an amplifier F controlled by the charging voltage E_0. T = supply transformer. A, B, L and V as in fig. 1.

This amplification can be brought about in various ways. The simplest method is with electronic valves, for instance in the manner indicated in *fig. 3*, where the charging voltage E_0 is compared with a reference voltage E_1 that can be varied with the aid of a potentiometer R_5 and kept constant by means of a voltage reference tube St, fed from a small auxiliary rectifier. Variations in the potential difference $E_0 - E_1$ are amplified by the valves P_1 and P_2; in this case the anode current of P_2 is at the same time the control current I_a of the transductor. It can easily be seen that a change in the charging voltage changes the control current in the right sense.

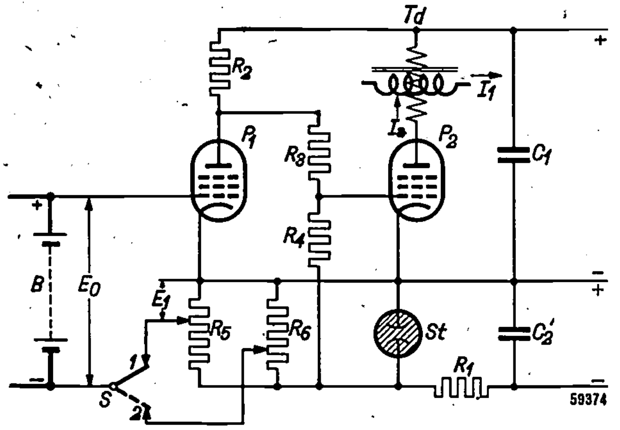


Fig. 3. Circuit of a part of the amplifier F (fig. 2). The control-grid voltage of the pentode P_1 (EF 41) is the difference between the battery voltage E_0 and a constant reference voltage E_1 adjustable with the potentiometer R_5. R_1 = current-limiting resistor of the voltage reference tube St (85 A1) which keeps E_1 constant. R_2, R_3 and R_4 are resistors for the intervalve coupling between P_1 and P_2 (EL 60), the anode current of which, I_a, controls the transductor Td (fig. 2). C_1, C_2 = smoothing capacitors of two small auxiliary rectifiers (not drawn). With the switch S one can change over to a higher reference voltage (for rapid charging) adjustable with a potentiometer R_6.

With this system any change of the charging voltage is counteracted, no matter whether it arises from a variation of the voltage loss in the valves, from fluctuations of the mains voltage or from mains frequency variations. In how far the charging voltage can be kept constant in a practical case may be judged from *fig. 4*, representing the charging voltage as a function of the D.C. output current at different mains voltages. These characteristics have been plotted from measurements taken with a preservation rectifier type 3067 (*fig. 5*) for 64.5 V (30 lead cells), maximum 15 A, designed according to the system of fig. 3. With a frequency deviation of 4% or a mains voltage variation of 10% the change in the charging voltage is less than 0.5%. Thus not only the dependency upon the frequency (the main point at issue) but also the dependency upon the mains voltage is appreciably reduced in comparison with the former system. With the frequency and mains voltage fluctuations just given and with currents between 10% and 100% of the nominal rating, the total changes of the charging voltage remain between +1 V and —1 V.

In the article quoted in footnote [1]) it was described how very little dependency upon mains voltage fluctuations can be obtained also with preservation rectifiers according to the system of fig. 1. There use was made of the remarkable feature of these rectifiers that the charging voltage varies in the opposite sense to the mains voltage. By connecting such a rectifier in series with an ordinary rectifier, the two voltages of which act in the same direction, compensation can be brought about to a high degrec.

As to the bends seen in the characteristics at small and at large currents, these will be referred to again farther on.

Regarding the question of the transformer sheet, with the transductor no difficulties arise of the nature of those encountered with the transformer T_1 (fig. 1). This is due in part to the manner in which the new rectifier works in bringing about the smallest possible difference between the charging voltage and the given, constant, reference voltage.

It is no exaggeration to say that the functioning of this rectifier stands or falls with the constancy of the reference voltage, thus with the properties of the voltage reference tube (*St*, fig. 3). A tube has therefore been used (type 85 A1) which shows only very small deviations from the rated voltage (85 V) either during its lifetime or with varying currents or temperature fluctuations; these deviations are never more than 1 V and mostly not even more than a few tenths of a volt. This tube, which works with a glow discharge in neon (average current 4 mA), has a carefully prepared cathode of molybdenum. A layer of the same metal is precipitated by atomization on the inner wall of the glass bulb and prevents the release of impurities from the glass. For further particulars reference may be made to another article published in this Journal [4]).

The potentiometer R_5 (fig. 3) allows of some regulation of the charging voltage, for instance between 2.10 V and 2.30 V per cell in the flat part of the characteristic.

Current limiting

The sharp bend of the characteristic at a high value of the current (fig. 4) is of great importance, for it means that the current I_0 supplied by the rectifier is limited. When the working of a plant requires from time to time a current higher than this threshold value then the difference has to be made up by the battery; thus the rectifier itself need not be calculated for these higher currents. It has to be explained, however, how this limitation of current comes about — it occurs also with preservation rectifiers of the old type, but in a different way — since the arrangement according to fig. 3 is not sufficient.

This effect is obtained by arranging for the grid voltage of the valve P_1 to consist, from a certain strength of current onwards, not of $E_0 - E_1$ (fig. 3) but of $E_2 - E_1$, where E_2 is a direct voltage proportional to the current I_0 supplied by the rectifier; the transition from $E_0 - E_1$ to $E_2 - E_1$ is brought about when E_2 becomes greater than E_0.

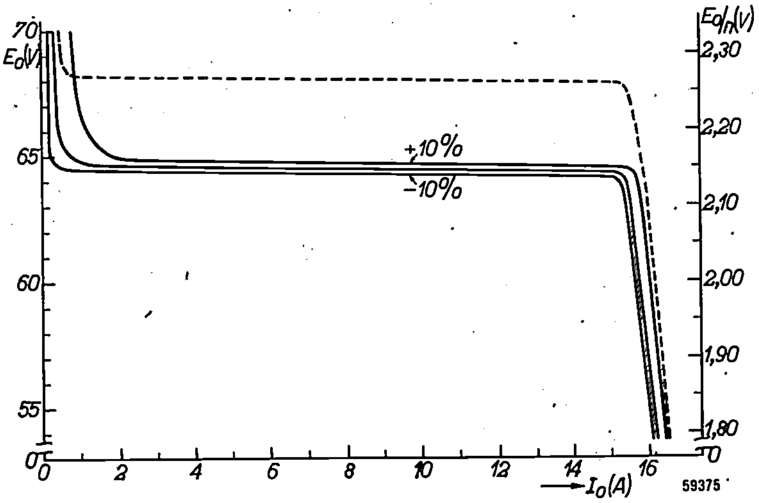


Fig. 4. Charging voltage E_0 as a function of the direct current I_0 of the preservation rectifier type 3067 for 64.5 V (30 lead cells), max. 15 A. Fully drawn lines: preservation charging currents respectively with 10% too low, with normal and with 10% too high mains voltage. Dotted line: rapid charging. On the right-hand scale the voltage per cell E_0/n (n = number of cells in series).

Fig. 6a (disregarding the derivation of the voltage E_2 proportional to I_0) shows how this process takes place: so long as $E_2 < E_0$ the diode D_1 does not pass any current and, as explained above, the grid voltage of the valve P_1 is formed by $E_0 - E_1$, but when E_2 becomes greater than E_0 current flows through the diode and the resistor R_7, and the grid voltage becomes equal to $E_2 - E_1$. Thus as I_0 increases this grid voltage now changes in the positive direction and this — as may easily be seen from fig. 3 — reduces the control current. As a consequence the voltage drop in the A.C. coil of the transductor is greatly increased and this results in a rapid drop of the charging voltage.

It now remains to be seen how the voltage E_2 proportional to I_0 is obtained. The alternating

[4]) T. Jurriaanse, A voltage-stabilizing tube for very constant voltage, Philips Techn. Rev. 8, 272-277, 1946.

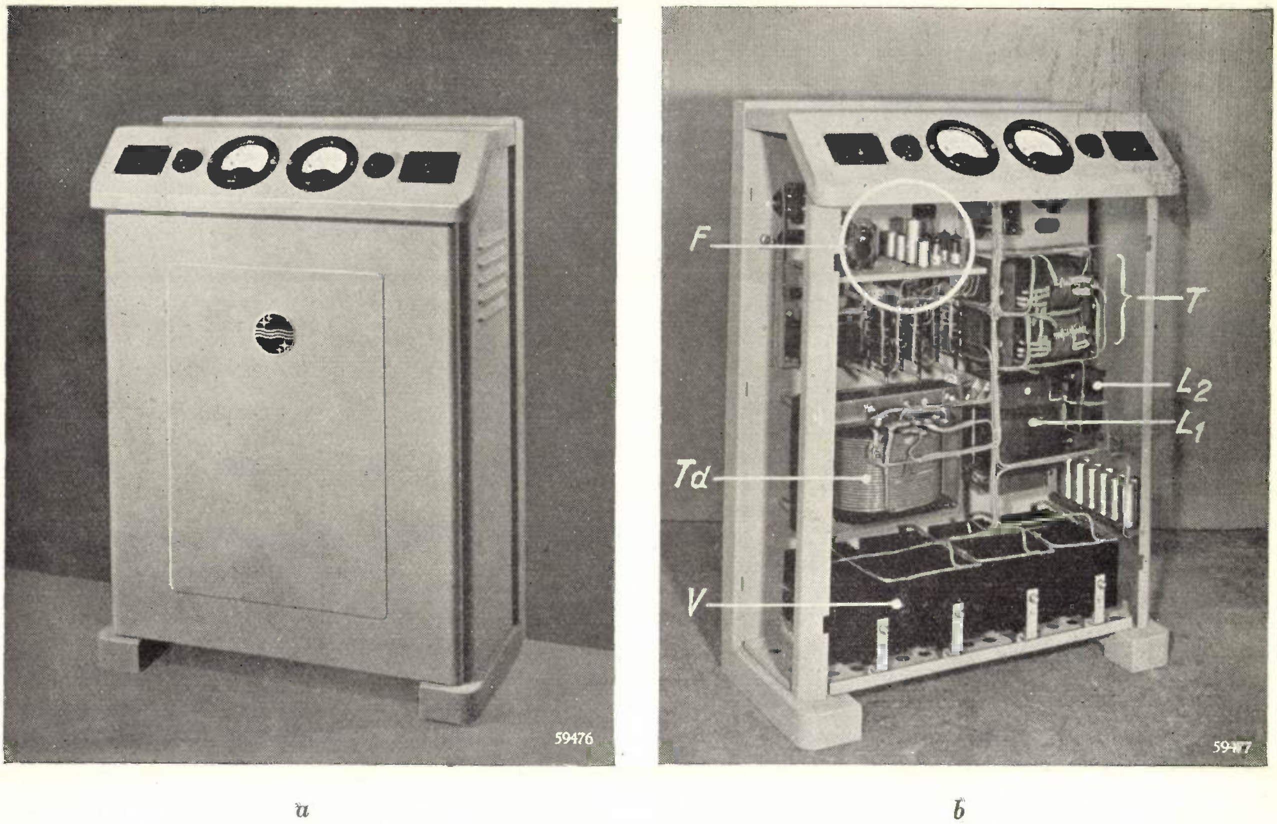


Fig. 5. Electronically controlled preservation rectifier type 3067 for 64.5 V, max. 15 A. *a*) closed, *b*) open. L_1, L_2, two smoothing chokes. Other lettering the same as in fig. 2.

current I_1 taken up by the rectifier is proportional to I_0 and flows through the primary coil of a small auxiliary transformer (T_1, fig. 6*b*), inducing in the secondary winding a voltage which is proportional to I_1 and thus also proportional to I_0. Rectification of this voltage, with the aid of a diode D_2, and smoothing yields the desired D.C. voltage E_2. By making this voltage adjustable, for instance with the potentiometer R_8, one can adjust, if required, the value to which I_0 is limited.

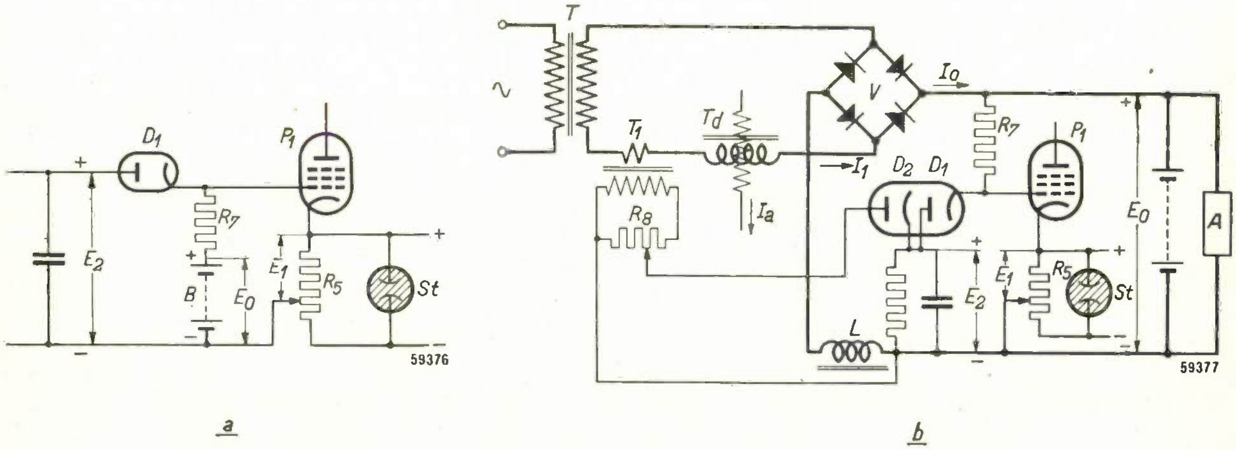


Fig. 6. Circuit for limiting the current. *a*) When the voltage E_2 exceeds E_0 current flows through the diode D_1 and the resistor R_7, and the control-grid voltage of P_1 (fig. 3) is then no longer $E_0 - E_1$ but $E_2 - E_1$. The other letters are the same as in fig. 3. *b*) The voltage E_2 is obtained by rectification (diode D_2) and smoothing of the alternating voltage (adjustable with a potentiometer R_8) of the secondary of the transformer T_1. Flowing through the primary of T_1 is an alternating current I_1 proportional to the direct current I_0. Thus E_2 is also proportional to I_0. The diodes D_1 and D_2 are contained in one bulb (EB 41). The other letters are the same as in the foregoing diagrams.

The working point of the rectifier is shifted to the bend in the characteristic (fig. 4) at low current — this bend also occurs with preservation rectifiers of the old type — when the battery voltage rises (for instance, due to there being no load for a long time) so far that the valve P_2 is cut off and thus the control current drops to about zero. The transductor then acts as an ordinary choke (with constant, high impedance) and the rectifier as an ordinary rectifier. The current at which this bend in the characteristic occurs can be made smaller by using a transductor designed for a greater self-inductance (without control current), but this would necessarily be heavier and more expensive. With the normal design of transductor this bend occurs at a current which is so small that usually the self-discharge of the battery, together with the basic consumption nearly always present (relay coils, etc), is already sufficient to prevent the rectifier from coming into action in the steep part of the characteristic. Should this load be insufficient — in which case undesired "gasing" of the battery might arise — then a small, additional, basic load can be applied, for instance in the form of a resistor shunted across the battery and the rectifier, or else a resistor or a choke shunted across the A.C. terminals of the rectifying circuit.

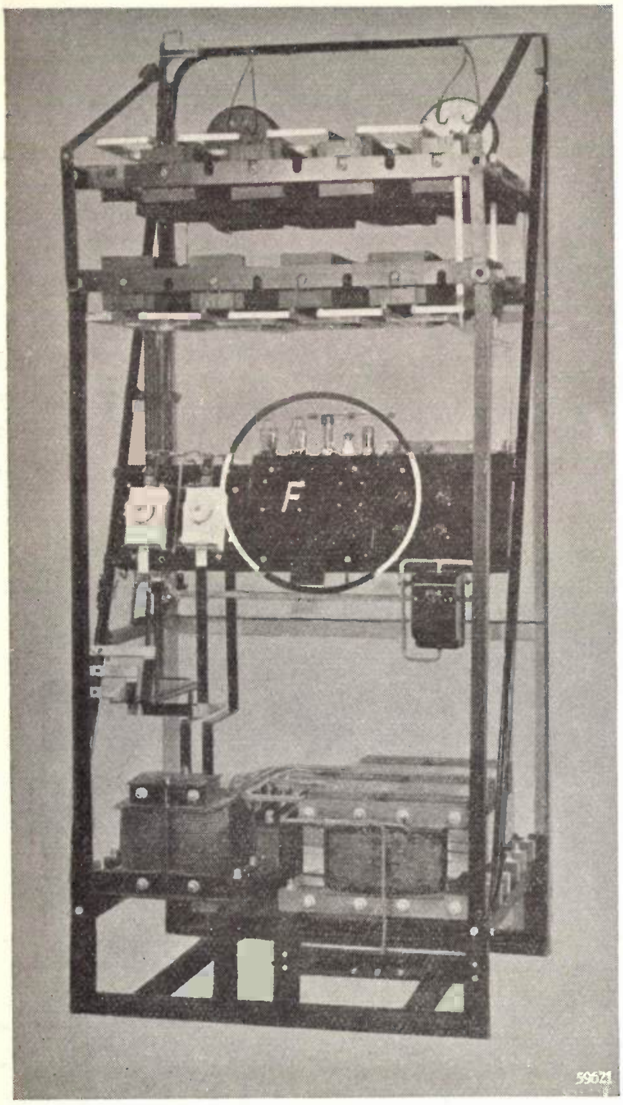


Fig. 7. Back view of a preservation rectifier for 64.5 V, max. 105 A, destined for large telegraph stations. The rectifier is of the three-phase construction but otherwise it functions in the same way as type 3067 (fig. 5). The amplifier is again denoted by *F*.

Charging of the battery after interruption of the mains

When the battery has performed its duty as a reserve source of energy during interruption of the mains supply then upon the latter being restored it has to be rapidly recharged so as to be ready for another possible emergency. At the beginning of this charging the cells have a voltage of 2.0 V or less and the charging current is therefore of the maximum value (e.g. 15 A in the case of the rectifier 3067, the characteristics of which are represented in fig. 4). With a normal characteristic, the horizontal part of which lies at, say, 2.15 V per cell, the charging current would drop to a small value (about 1 A in the case of the type 3067) already when that voltage is reached and the charging would therefore take a long time, so that the full capacity of the battery would not become available for a considerable time. To reduce this charging time provision has been made for rapid charging, simply by increasing temporarily the reference voltage E_1, thus raising the charging voltage to a higher level corresponding, for instance, to 2.3 V per cell. The charging current is then maintained at its maximum value until this higher voltage is reached and the battery is therefore charged more rapidly. The change-over to rapid charging is brought about by turning a switch (*S*, fig. 3) to position *2*, thereby replacing the potentiometer R_5 by the potentiometer R_6, by means of which the charging voltage can be adjusted between, say, 2.2 V and 2.4 V per cell (dotted curve in fig. 4), this being desired so as to permit of the battery being kept connected to the plant during the rapid charging. In telephone exchanges, for instance, for the proper functioning of certain relays the supply voltage must not be allowed to exceed a given value; for other plants there is no such restriction and more rapid charging is therefore possible.

Auxiliary transductor for high powers

An output valve like that of the type EL 60 (P_2 in fig. 3) with a maximum permissible anode dissipation of 25 W is just capable of supplying the control current for the transductor of a preservation rectifier for 64.5 V, 15 A. For rectifiers of a higher power a larger output valve could be chosen, but for rectifiers having a power of, say, 64.5 V, 105 A (*fig.* 7) we have followed another way, namely a sort of cascade circuit of two transductors. Here

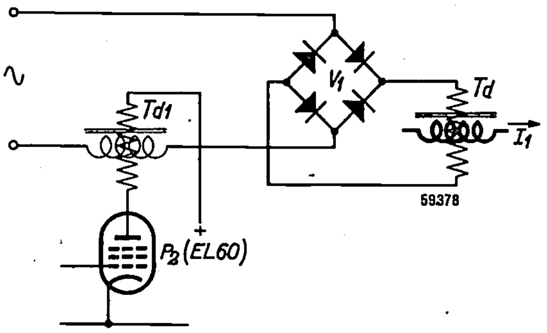


Fig. 8. Cascade circuit of an auxiliary transductor Td_1 and the main transductor *Td*. V_1 = auxiliary rectifier. In this manner a main transductor of high power can be controlled with a relatively small output valve (EL 60).

the valve EL 60 controls a small auxiliary transductor (Td_1, *fig. 8*). This again controls an auxiliary rectifier (V_1), which in turn supplies the control current for the main transductor (*Td*). For very high powers the cascade circuit can be extended, if necessary, by more auxiliary transductors.

Summary. The fact is recalled that for the power supply of telephone exchanges, for instance, a single battery with preservation charge is better than two batteries alternately charged and discharged, as used to be the case. Preservation charging means that the battery is permanently connected to a charger and that its voltage is maintained between 2.1 and 2.2 V per cell (for lead cells). As source of the charging current use can be made of a rectifier with a constant charging voltage. Here a system is described for such a preservation rectifier where the charging voltage is compared with a voltage which is to be regarded as being constant and which is stabilized by a voltage reference tube with narrow tolerance. The difference of the two voltages is amplified by electronic valves and controls the D.C. magnetization of a transductor (a choke the impedance of which is variable by means of D.C. saturation) connected in series with the A.C. side of the rectifier. Between wide limits of the D.C. output the charging voltage is very little affected by the voltage drop in the valves (thus also by the direct current), by the mains voltage or by the frequency. In this and other respects there is a great improvement compared with preservation rectifiers of an older system. An addition to the controlling mechanism makes it possible to limit the direct current to a certain adjustable value. By means of a switch a higher charging voltage can be chosen for rapidly recharging a battery after it has been wholly or partly discharged in the event of a prolonged interruption of the mains supply. In the controlling mechanism a pentode EL 60 (with a maximum permissible anode dissipation of 25 W) is used as output valve; this supplies direct the control current for the transductor of a rectifier for, say, 64.5 V, 15 A. In the case of higher powers, instead of choosing a larger output valve it is often more advantageous to arrange for the output valve to control an auxiliary transductor which in turn controls the main transductor (if necessary applying more than one auxiliary transductor).

THE THEORY OF SAMPLING INSPECTION PLANS

by H. C. HAMAKER.

620.113.2:658.562:31

To gain an insight into the various factors playing a part in the establishing of sampling inspection procedures some study of the mathematical foundations of the problem of sampling is needed. This leads to a systematic comparison of the different sampling-inspection systems available and to a choice amongst them.

The evaluation of the operating characteristic

In the previous article [1]) some fundamental aspects of sampling inspection have been considered and the three most common inspection procedures, single, double and sequential sampling, have been briefly discussed. It was found that the inspection performance of a sampling plan was duly expressed by means of the operating characteristic, a curve representing the probalitity of acceptance P of an inspection lot as a function of the percentage defective contained in it. The operating characteristics for three single sampling plans reproduced in *fig. 1* may serve to illustrate some of the fundamental properties of these curves.

As is to be expected, the operating characteristic shifts to higher percentages when the acceptance number c_0, that is the maximum number of rejects permitted in the sample, is increased. Furthermore the characteristic is steepest at the point where the percentage defective in the lot is about equal to the maximum permissible percentage of rejects, $c_0/n_0 \times 100\%$ (where n_0 is the sample size).

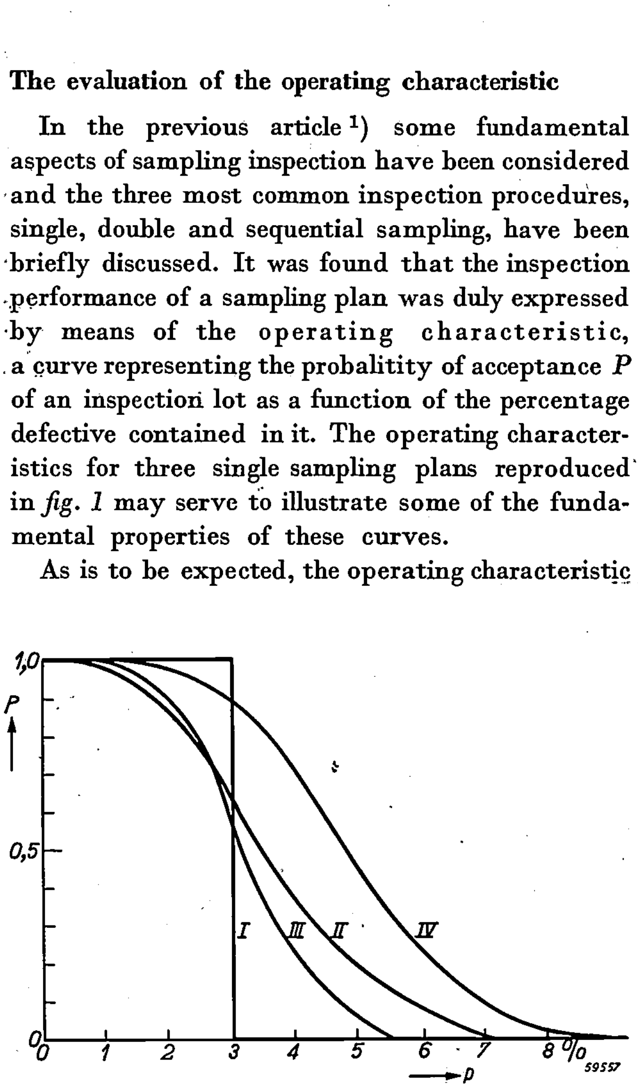


Fig. 1. Some operating characteristics for single sampling plans. *I* the characteristic when the sample size is equal to the size of the inspection lot; *II* the characteristic for a plan with a sample size $n_0 = 200$ and acceptance number $c_0 = 6$; *III* the same but for $n_0 = 400$, $c_0 = 12$. In these two cases the ratio c_0/n_0 has been altered but the sample size has been increased; as a consequence curve *III* is steeper than curve *II*. *IV* the characteristic for $n_0 = 200$, $c_0 = 9$. Compared with *II* the characteristic has shifted towards the right, owing to the increase in c_0, the sample size being the same in both cases.

Lastly, if we keep the ratio c_0/n_0 constant while increasing the sample size, the curve remains at about the same place but becomes steeper.

In the limit, when the sample comprises the entire inspection lot, the operating characteristics will consist of a horizontal branch at $P = 1$ for percentages of defects less than the percentage of rejects permitted, and a second horizontal branch at $P = 0$ for percentages of defects greater than that value, these two branches being joined by a vertical. This is the operating characteristic corresponding to an ideal inspection in which all defects are detected and removed. In actual sampling, however, we inspect only a fraction of the lot and it is clearly of some interest to calculate the operating characteristic under such circumstances. To this end we shall have to resort to the calculus of probability.

Let us consider an inspection lot consisting of N pieces, M of which are defective; if from this lot we draw a random sample of size n_0, what will be the probability of finding m rejects in it? The correct formula expressing this probability is too cumbersome for numerical evaluation but fortunately an approximation can be used.

In most practical cases the sample is but a small fraction of the lot, so that we have

$$n_0 \ll N,$$

which automatically implies

$$m \ll M.$$

Moreover, in factory applications the percentage defective $p = M/N$ in the inspection lots is usually

[1]) H. C. Hamaker, Lot inspection by sampling, Philips Techn. Rev. **11**, 176-182, 1949 (No. 6), hereinafter referred to as I.

very small, of the order of a few percent; in other words we have

$$p \ll 1.$$

These conditions being satisfied, the probability of finding m defects in a sample of n_0 items is given by Poisson's formula:

$$\Pi(m; n_0, p) = \frac{e^{-n_0 p}(n_0 p)^m}{m!}. \quad . \quad . \quad . \quad . \quad (1)$$

According to this expression the probability depends on m and on the product $n_0 p$, but not on n_0 and p separately.

In single sampling, however, we are not so much interested in the probability of finding exactly m rejects in the sample as in that of finding c_0 or fewer rejects, which is obtained by summation as:

$$P(c_0; n_0 p) = \sum_{m=0}^{c_0} \frac{e^{-n_0 p}(n_0 p)^m}{m!}. \quad . \quad . \quad . \quad . \quad (2)$$

This is also the mathematical expression for the operating characteristic of a single sampling plan.

The probability of acceptance for double sampling, deducible in a similar way, leads to a somewhat more complicated formula, viz.:

$$P(c_1; c_2; c_3; n_1 p; n_2 p) = P(c_1; n_1 p) + $$
$$+ \sum_{k=c_1+1}^{c_2} \Pi(k; n_1 p)\, P(c_3 - k; n_2 p). \quad . \quad . \quad . \quad (3)$$

Here the three criteria, c_1, c_2, and c_3, are so defined that an inspection lot is accepted when the number of rejects in the first sample is c_1 or less, and rejected when this number is greater than c_2, while acceptance after a second sample requires that the total number of rejects in both samples together shall be equal to or less than c_3 (see I page 179). On the right-hand side of (3) the first term measures the probability of accepting a lot after the first sample, and the second term measures the probability of accepting the lot after a second sample, no decision being reached in the first.

We are interested not only in the operating characteristic, but also in the average number of observations, or the average sample size, required per lot inspected. For single sampling this number is constant and equal to n_0, but for double sampling the situation is less simple. Sometimes one sample will suffice and sometimes two samples will be needed; and even if all the lots inspected contain exactly the same percentage defectives it still depends on the vagaries of chance whether one or two samples have to be drawn. All we can do, therefore, is to define an average sample size, and this is given by

$$\bar{n} = n_1 + n_2 \sum_{k=c_1+1}^{c_2} \Pi(k; n_1 p), \quad . \quad . \quad . \quad (4)$$

i.e. by the size of the first sample plus the product of the size of the second sample and the probability that a second sample has to be taken. As the formula shows, this average sample size depends on the percentage defective in the inspection lot. The fate of decidedly good or decidedly bad lots is usually determined by the first sample, and for such lots the average sample size is almost equal to n_1; but for lots of doubtful quality a second sample is often needed and the average sample size is consequently larger, so that as a function of the percentage defective it passes through a maximum.

The Poisson functions Π and P have been extensively tabulated [2]), and from these tables the operating characteristics for single sampling can at once be taken, while those for double sampling may easily be calculated.

For sequential sampling, too, formulae for the operating characteristic and for the average sample size have been derived, but as these are of a more complex character we shall not consider them in detail. They may be found in the writings of Barnard and of Wald [3]). Suffice it to say that, as with double sampling, the average sample size goes through a maximum as a function of the percentage of defects in the lot.

The fundamental parameters

In the concluding section of the previous article it was emphasized that the practical choice of a sampling plan is in the main influenced by economic factors which it is impossible to evaluate in a precise manner; and from this it was concluded that an accurate knowledge of the operating characteristic is not of essential importance, specification of the curve by a suitable set of parameters being all that is practically required. Now two important features of the operating characteristics strike the eye, namely the place where the curve shows its steepest descent, and the degree of its steepnes (see e.g. the characteristics of fig. 1). We shall use

[2]) E. C. Molina, Poisson's exponential binomial limit, D. van Nostrand, New York 1947. Less extensive tables in: E. L. Grant, Statistical quality control, McGraw Hill, New York 1946, and in T. C. Fry, Probability and its engineering uses, D. van Nostrand, New York 1928.

[3]) A. Wald, Ann. Math. Stat. 16, 117-187, 1945 and J. Am. Stat. Ass. 40, 277-306, 1945. G. A. Barnard, Supp. J. Roy. Stat. Soc. 8, 1-27, 1946.

these two properties as a basis for the definition of two parameters which will play a fundamental part in our further investigation. To this end we shall locate the operating characteristic by its "point of control", p_0, defined as the percentage defectives corresponding to a probability of acceptance of $\frac{1}{2}$:

$$P(p_0) = \tfrac{1}{2}, \quad . \ . \ . \ . \ . \ . \ . \ . \qquad (5)$$

and we shall specify the steepness of the curve by its relative slope, h_0, in this point, defined by

$$h_0 = -\left\{\frac{p}{P}\frac{\mathrm{d}P}{\mathrm{d}p}\right\}_{p=p_0} = -\left\{\frac{\mathrm{d}\log P}{\mathrm{d}\log p}\right\}_{p=p_0}, \ . \ . \qquad (6)$$

a negative sign being added to render h_0 an essentially positive quantity. Here the relative slope has been chosen instead of the absolute slope $(\mathrm{d}P/\mathrm{d}p)_{p=p_0}$ for a very special reason. As already pointed out, Poisson probabilities depend on the product np, and not on n and p separately. Consequently, by increasing or diminishing the sample size of a single sampling plan, or by altering the sample sizes of a double sampling plan both in the same ratio, while keeping the acceptance number c_0 or the criteria c_1, c_2 and c_3 at fixed values, the operating characteristic undergoes no other change than such as is equivalent to a change in the scale along the p-axis. Such a transformation does not alter the value of h_0, as is easy to see, so that this parameter is independent of the absolute size of the sample, or samples. This is a considerable advantage, as will appear farther on.

It also follows from the argument just given that any sampling plan can always be adjusted to some pre-assigned value of the point of control simply by changing the sample size or sizes. This is an adjustment that can always easily be effected, so that from now on we may consider all single sampling plans with the same acceptance number, or all double sampling plans with the same criteria and the same ratio of the sample sizes, as belonging to one class. Such a class may be represented by a single operating characteristic, which in its main features may be specified by one single parameter, namely its relative slope, h_0. All the operating characteristics of the sampling plans belonging to one class may be made to coincide in a simple and convenient way by plotting the probability of acceptance P not against the percentage defective p but against the ratio p/p_0.

Single sampling plans

For single sampling plans two simple empirical relations have been found to exist between the sample size n_0 and the acceptance number c_0 on the one hand, and the fundamental parameters p_0 and h_0 on the other, namely

$$n_0p_0 = c_0 + 0{,}67\ , \ . \ . \ . \ . \ . \ . \qquad (7)$$

and

$$\frac{\pi}{2}h_0^2 = n_0p_0 + 0{,}06 = c_0 + 0{,}73 \ . \ . \ . \qquad (8)$$

The validity of the first of these equations was noted as early as 1923 by Campbell[4]), who calculated accurate values of n_0p_0, some of which are reproduced in the second column of *table I*; comparison with the first column shows the high accuracy with which (7) is fulfilled.

For high values of c_0 equation (7) gives $n_0p_0 \approx c_0$ or $p_0 \approx c_0/n_0$, a result that might have been expected: inspection lots with a percentage defectives equal to the percentage of rejects permitted in the sample should roughly have an equal chance of being accepted or rejected.

For large values of c_0, or of n_0p_0, Campbell also deduced the approximate relation

$$\frac{\pi}{2}h_0^2 = n_0p_0 \ . \ . \ . \ . \ . \ . \ . \qquad (8a)$$

The empirical correction term $+0.06$ in (8) has the effect of rendering this equation valid down to the lowest values of c_0, as may be seen on comparing the last two columns in table I; the degree of approximation is highly satisfactory.

Table I. Values of n_0p_0 and h_0 for single sampling plans with a given acceptance number.

c_0	n_0p_0	h_0 from (8)	h_0 calculated from (2)
0	0.693	0.682	0.693
2	2.674	1.319	1.319
4	4.671	1.736	1.735
6	6.670	2.070	2.069
8	8.669	2.357	2.356
10	10.668	2.613	2.613
15	15.668	3.164	3.164
20	20.668	3.633	3.632
25	25.667	4.047	4.047
30	30.667	4.423	4.422

It goes without saying that with the aid of (7) and (8) the parameters p_0 and h_0 may be computed when n_0 and c_0 have been given, and vice versa.

Here, however, a slight difficulty arises owing to the fact that by their very nature n_0 and c_0 are positive integers, so that p_0 and h_0 cannot assume all the conceivable sets of values. This inconvenience may, however, be removed by the

[4]) G. A. Campbell, Bell System Technical Journal 2, 95-112, 1923.

artificial introduction of single sampling plans with broken values of c_0 and n_0. For example, a plan with $c_0 = 3.63$ is defined as follows. We set up a lottery between the numbers 3 and 4 such that the 4 has the probability of 0.63 of being drawn and the 3 the complementary probability of 0.37. Simultaneously with the taking of a sample we also draw a lot from this lottery, using the figure obtained as our acceptance number. Let P_3 and P_4 denote the probabilities of acceptance of an inspection lot when the acceptance numbers are 3 and 4 respectively, then the probability of acceptance for $c_0 = 3.63$ will be

$$P_{3\cdot 63} = 0.37\ P_3 + 0.63\ P_4.$$

Our definition is apparently equivalent to a linear interpolation between the operating characteristics corresponding to successive acceptance numbers.

Broken values of n_0 may be introduced in similar fashion. At first sight it may seem rather unnecessary and artificial to extend the definition of single sampling plans in the way we have just explained, and it is certainly not our opinion that plans with broken values of n_0 or c_0 should be of practical application. But the extended definition is of theoretical importance; since the operating characteristics for successive acceptance numbers form a regular sequence, we may conclude that equations (7) and (8), which were derived for integer values of n_0 and c_0, will hold good for broken values as well, so that by applying these equations we can from now on construct a single sampling plan corresponding to any set of positive values of the two parameters p_0 and h_0. This is of advantage when it is desired to compare different sampling systems, as we shall see in the next section.

Double sampling

Arguments set forth in the previous article suggested that a double sampling plan must be more efficient than a single sampling plan, in the sense that it may achieve the same inspection performance at the cost of a smaller number of observations. In what manner can we measure the gain thereby obtained?

To answer this question let us start by considering as a concrete example the double sampling plan with sample sizes $n_1 = 100$, $n_2 = 200$ and with the criteria $c_1 = 2$, $c_2 = 6$ and $c_3 = 10$. *Fig. 2* gives the operating characteristic of this plan as resulting from equation (3); precise calculation yields the following values for the fundamental parameters:

$$p_0 = 0.0379\ ; \quad h_0 = 2.422.$$

Now it is natural to expect that the operating characteristics of two sampling plans having the same point of control and the same relative slope will be largely coincident, so that we are logically led to compare the double sampling plan under consideration with a single sampling plan with the same p_0 and h_0. According to (7) and (8) this requires

$$n_0 = 241.7 \quad \text{and} \quad c_0 = 8.49,$$

and this single plan corresponds to the charateristic indicated by the dots in fig. 2. The coincidence of the two characteristics appears to be almost complete, so much so in fact that for all practical purposes the inspection performances of the two plans may be considered as identical; we shall express this by calling them "equivalent".

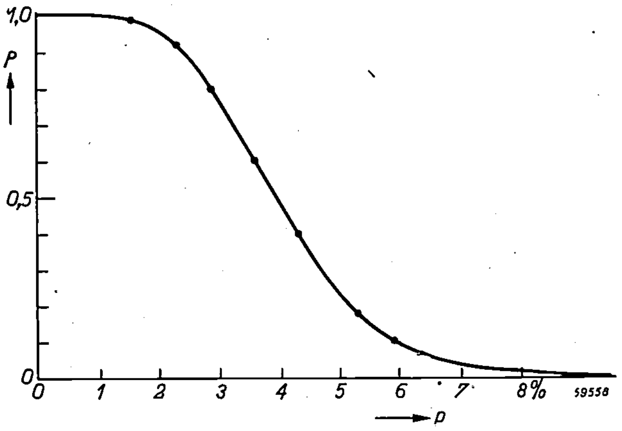


Fig. 2. Operating characteristic of a double sampling plan with $n_1 = 100$, $n_2 = 200$, $c_2 = 2$, $c_2 = 6$, $c_3 = 10$. The dots represent the characteristic of the equivalent single sampling plan, the curve that of the double plan.

Next, by calculating according to (4) the average sample size of the double sampling plan, we obtain curve *I* in *figure 3*, while the sample size of the "equivalent single sampling plan" is given by curve *II*. The advantage of double sampling is clearly demonstrated by these curves, but the choice of the sample sizes and hence of the scale on the left is still arbitrary. We shall arrive at a measure of the efficiency independent of this choice if we divide the average sample size of the double

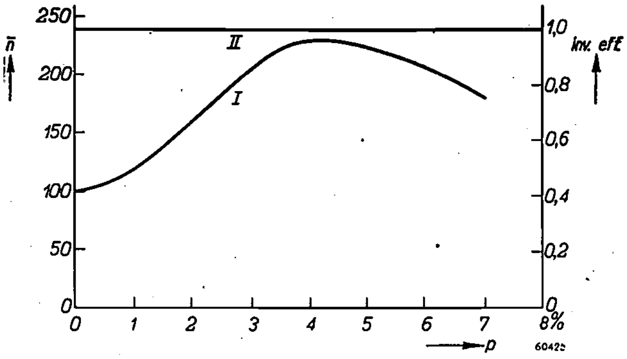


Fig. 3. The average sample size as a function of the percentage defective in the lot. Curve *I* gives the average sample size of the double sampling plan of fig. 2, which depends on the percentage defectives; curve *II* represents the (constant) sample size of the equivalent single sampling plan. For very good or very bad lots the average sample size of the double sampling plan is considerably smaller than the sample size of the single plan; the average amount saved is about 30%. Interpreted in terms of the scale on the right, curve *I* represents the inverse efficiency of the double sampling plan.

sampling plan by the sample size of the single sampling plan. The result is still represented by curve I in fig. 3, provided we use the scale on the right for its interpretation. The curve thus obtained measures the average number of observations required by the double sampling plan in terms of the number required by an equivalent single sampling plan, and may suitably be called the "inverse efficiency characteristic" [5]). We may read at a glance from this characteristic that for good batches the number of observations needed in double sampling is only about 40% of that needed in single sampling, whereas for lots of doubtful quality, near the top of the curve, the amount saved is relatively small. Further towards the right, beyond the range of the figure, the curve continues to descend until for very bad lots the inverse efficiency again reaches a value of 40%.

The reader will easily discern the general principle lying at the root of the example we have just discussed. By the extended definition of the previous section there exists one and only one single sampling plan for every set of positive values of the parameters p_0 and h_0. Since, moreover, of all sampling methods single sampling is by far the simplest, it is natural to adopt single sampling plans as a general standard of reference; the efficiency of any single sampling plan is then by definition equal to unity, while other sampling plans may be judged on the basis of their inverse efficiency, that is by comparing them with their single sampling equivalents. As a general principle this method of comparison will, of course, only be acceptable if the coincidence of operating characteristics is equally satisfactory as it was in fig. 2. The fact that this is generally so will become evident from the further cases we shall now proceed to consider.

As already pointed out, it is of some advantage to use a "class characteristic" obtained by plotting the ratio p/p_0 along the abscissa instead of p itself. The data needed for specifying such a class characteristic of a double sampling plan will then be four in number, viz. the ratio of the two sample sizes and the three criteria. Henceforth these will be written in the following notation:

$$D(n_2/n_1\,;\ c_1, c_2, c_3);$$

the symbol

$$D(2;2,6,10)$$

for example signifying all double sampling plans for which the second sample is twice as large as the first and for which the criteria are 2, 6 and 10 respectively. The plan considered in figs 2 and 3 belongs to this class.

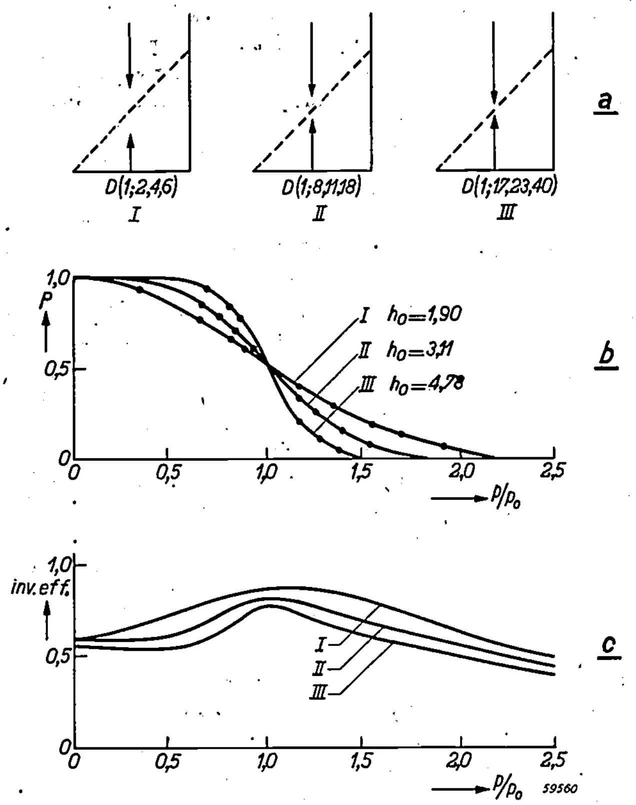


Fig. 4. a) Random walk diagrams corresponding to three double sampling plans with approximately the same configuration; b) The operating characteristics for these three plans, or the probability of acceptance as a function of the percentage defectives. The figure illustrates the high degree of concordance of the operating characteristics of double sampling plans with those of equivalent single sampling plans, the latter being indicated by the dots. c) Efficiency characteristics of the three plans under consideration, or the inverse efficiency as a function of the percentage defectives.

In *figs 4 a* to *c* three double sampling plans are compared for which the screens in the random walk diagram possess approximately the same geometrical configuration (see fig. 4a), but for which the criteria differ in magnitude. The coincidence of their operating characteristics with those of the equivalent single sampling plans is again highly satisfactory, as is illustrated in fig. 4b.

From the similarity of the configuration of the random walk diagrams we should intuitively expect the three plans to be approximately equal in efficiency, a conclusion corroborated by the efficiency characteristics in fig. 4c; only, the steeper the operating characteristic the sharper is the peak in the inverse efficiency curve. To remove this difference, which is more apparent than real, it may be of some practical advantage to plot the

[5]) The qualification "inverse" has been added because of the properties of these quantities; their inverse efficiency increases as the number of observations saved becomes less.

inverse efficiency not against p/p_0, as done in fig. 4c, but against the corresponding probability of acceptance P. The equivalent efficiency of the three plans under consideration is then brought to the fore much more conspiciously, as shown by *fig.* 5. Moreover, in practice sampling plans are mostly used under such circumstances that about 95% of the lots submitted for inspection are accepted and only 5% rejected; in other words they are used in the neighbourhood of a probability of acceptance $P = 0.95$. A comparison of different sampling plans at this point can be most conveneintly performed with the aid of fig. 5. For example, we read at once from this figure that the saving in the number of observations attainable by double sampling is under practical conditions in the order of 25 to 30%.

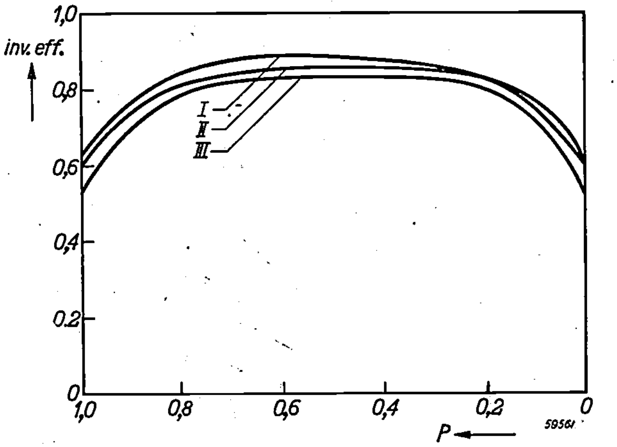


Fig. 5. Efficiency characteristics of the three plans of fig. 4; the inverse efficiency is now plotted as a function of the probability of acceptance P.

According to equation (8) we have for single sampling approximately

$$p_0 n_0 \approx \sqrt{h_0}\,.$$

If, therefore, we wish to find a double sampling plan which, for a prescribed p_0, requires on the average the same number of observations as a given single sampling plan, the relative slope h_0 of this double plan may be chosen about $1/\sqrt{0.70} \approx 1.2$ times as high as for the single plan. This principle may sometimes be of service in changing over from one plan to another.

The most general double sampling plan requires five data for its specification, the two sample sizes and the three criteria, while in all existing sampling tables [6]) only four parameters are employed. This is because they use only those plans for which $c_2 = c_3$, a principle originally due to Dodge and Romig (see note [6])). From the random walk diagram in *fig. 6a* we may infer that this simplification implies that lots with fairly high percentages defectives will often be considered as doubtful after the first sample, while they will almost certainly be rejected after the second sample; hence the Dodge and Romig principle must be expected to lead to sampling plans with a relatively poor efficiency on the side of bad lots. This conclusion is affirmed by the efficiency characteristic *II* in fig. 6*b*. If, in addition to making $c_2 = c_3$, we increase c_1 so as to bring the opening in the first screen entirely above the line connecting the origin with the dividing point in the second screen (as in fig. 6*a* case *III*), this relative inefficiency is considerably enhanced. As shown by curve *III* in fig. 6*b*, the inverse efficiency of such a plan is greater than unity except for a small range of exceptionally good inspection lots, so that the number of observations needed is on the average greater than with an equivalent single sampling plan. Though there seems to be little sense in using plans of this kind they are nevertheless sometimes encountered in literature.

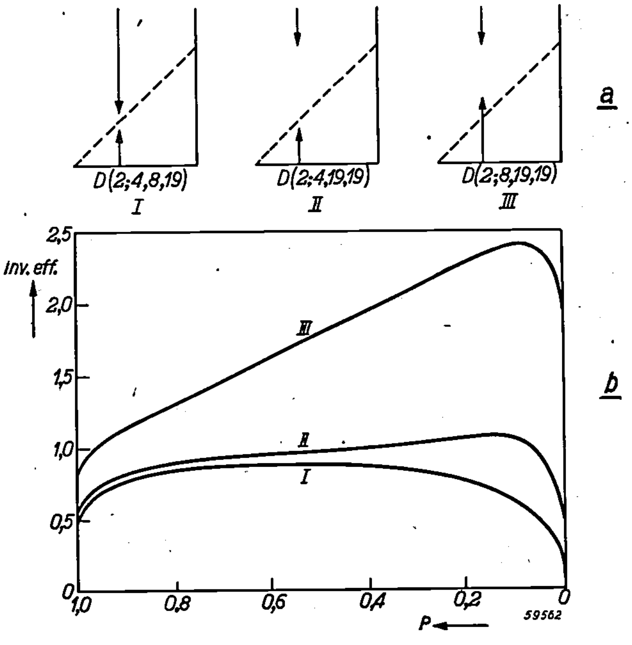


Fig. 6. Efficiency characteristics (*b*) of three double sampling plans corresponding to the three random walk diagrams (*a*). The pronounced influence of an incorrect choice of the criteria is clearly brought out, especially on the side of bad lots.

Case *I* in figs 6*a* and 6*b* on the other hand represents a plan where the Dodge and Romig principle has been abandoned and the criteria have been adjusted in the best possible manner.

[6]) H. F. Dodge and H. G. Romig, Single sampling and double sampling inspection tables, Bell System Technical Journal **20**, 1-61, 1941. Sampling inspection tables; single and double sampling, Wiley, New York 1944. Sampling inspection, Statistical Research Group Columbia University, Mc Graw Hill, New York 1948. See also E. L. Grant's book cited under reference [2]).

The total number of double sampling plans conceivable is of course unlimited, but from among them only a limited number are needed in practice, and it will by now have become clear that the efficiency characteristic may be helpful in guiding our choice and in weighing the advantages and disadvantages of different plans against one another.

A systematic study revealed that it is advantageous to make the second sample twice as large as the first, and that there is little point in altering this ratio. Furthermore, if for the sake of simplicity we adhere to the Dodge and Romig principle $c_2 = c_3$, it turns out that the achievement of a reasonable efficiency requires that c_3 must be 5 to 7 times as large as c_1. Also it was found convenient to classify double sampling plans according to their relative slope h_0; for the higher the value of h_0 the steeper is the operating characteristic and the larger the average sample size. On the basis of these principles we finally arrived at a set of double sampling plans as listed in *table II.*

Table II. Data for construction of a double sampling plan when p_0 and h_0 have been prescribed.

Plan	h_0	n_1p_0
$D(2; 0, 1, 1)$	0.93	0.84
$D(2; 0, 2, 2)$	1.20	1.07
$D(2; 0, 3, 3)$	1.40	1.35
$D(2; 0, 4, 4)$	1.63	1.64
$D(2; 1, 5, 5)$	1.69	2.19
$D(2; 1, 6, 6)$	1.90	2.44
$D(2; 1, 7, 7)$	2.08	2.71
$D(2; 2, 10, 10)$	2.41	3.78
$D(2; 2, 12, 12)$	2.74	4.35
$D(2; 3, 15, 15)$	2.97	5.40
$D(2; 4, 20, 20)$	3.39	7.02
$D(2; 5, 25, 25)$	3.82	8.66
$D(2; 6, 30, 30)$	4.28	10.31

By way of illustration let us suppose that we wish to replace the single sampling plan $n_0 = 180$, $c_0 = 4$ by a double plan. From equations (7) and (8) we have $p_0 = 0.026$ and $h_0 = 1.74$ and from the second column in table II we see that the double plan D (2;1,5,5), with $h_0 = 1.69$, furnishes the best approximation. For this plan we have $n_1p_0 = 2.19$, so that to fix p_0 at 0.026 we must take $n_1 = 2.19/0.026 = 84$, which in practice may be rounded off to 85. The double plan sought is now fully specified.

It would of course be possible to extend table II so that the successive values of h_0 lie closer together, but for practical purposes the set given is amply sufficient.

Sequential sampling

As explained in I, the random walk diagram of a sequential sampling plan is bounded by two parallel straight lines, one for acceptance and the other for rejection. Such a set of boundaries is completely specified by three parameters, for which we may choose, for instance, the angle α and the lengths of the two segments cut off from the ordinate axis (see *fig. 7*). Since we have found that in its main features an operating characteristic is fixed by two parameters only, p_0 and h_0, we are led to expect that for every set of values of p_0 and h_0 there exists a one-parametric set of corresponding sequential sampling plans. This is indeed the case.

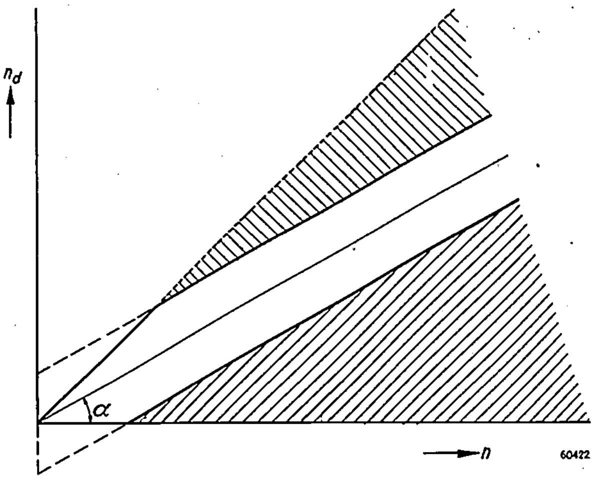


Fig. 7. Random walk diagram of sequential sampling plans. Only those sequential plans were found to be really efficient for which the line segments cut off from the vertical axis on both sides of the origin are approximately equal.

By calculating the efficiency characteristics on the same basis as above it was found, however, that from such a set only those sequential plans are reasonably efficient for which the two decisive lines in the random walk diagram lie nearly symmetrical with respect to the origin. Furthermore, in the case of symmetry the equations specifying the boundary lines become extremely simple, namely

$$n_d = \pm h_0 + np_0 , \quad \ldots \ldots \ldots \quad (9)$$

where n_d denotes the number of rejects observed after inspecting n items. We go on inspecting as long as n_d remains within these limits.

From eq. (9) the construction of a sequential sampling plan with prescribed values of p_0 and h_0 becomes extremely simple. It must be added, however, that sequential sampling is frequently found to be impracticable, because one has to check for acceptance or rejection after every unit inspected (which might be remedied by an automatic recording apparatus), while it may sometimes also happen that one has to inspect a very large number of items before a decision is reached.

Often it will be more convenient to steer a middle course by taking a limited sequence of concrete samples, and deciding between acceptance, rejection or a further sample only when the inspection of the sample at hand has been completed. Sampling plans of this type have been developed by the Statistical Research Group, Columbia University (see ref. [6])) and by J. H. Enters [7]). The method of application adopted by Enters is particularly simple, because he only varies the sample size, which in our terminology means that he uses only one value of h_0. It is possible to derive plans of the same simple form as the one he uses but with different values of the relative slope [8]), and these may be set out in just such a simple table as table II. However, our researches in this direction have not yet been completed.

Other sets of parameters

So far p_0 and h_0 have figured in our arguments as the two fundamental parameters, a practice that led to the simple equations (7) and (8) and to the successful introduction of the concept of inverse efficiency. But it does not necessarily follow that this set of parameters is also adapted to practical requirements. In all existing sampling tables which have been developed for practical application different parameters have been used. Some of these will now be considered in more detail.

The aim of sampling inspection is as a rule to protect the consumer against the delivery of bad lots occasionally produced. But at the same time it will be necessary to adjust the sampling plan in such a manner that the number of good lots perchance returned as unsatisfactory is not unduly large. We therefore have to take into account the requirements of both the consumer and the producer, and one set of parameters has been specially designed to achieve this purpose as directly as possible. Two points are chosen on the operating characteristics, one near the top and the other in the bottom part of the curve as shown in *fig. 8*.

The probabilities of acceptance for these two points are set at definite values, say at $P_1 = 0.90$ and $P_2 = 0.10$, and the corresponding percentages defectives, $p_{0.90}$ and $p_{0.10}$, are used as the two parameters by which the operating characteristic is fixed. The producer can then be guaranteed that of good lots with $p_{0.90}$ percent defectives not more than 10% are on the average rejected, while the consumer can rest assured that of lots with $p_{0.10}$ percent defectives no more than 10% are accepted. For this reason $p_{0.90}$ and $p_{0.10}$ have been called the 10 to 1 producers' and consumers' risk points; they have been employed by Wald and Barnard in their theoretical treatment of sequential sampling plans (see ref [3])), while Enters [9]) has constructed a nomogram from which the single sampling plan corresponding to given risk points can be read in a sample manner. This last nomogram may be further simplified and generalized. We then arrive at a simple graphical method by which the risk points corresponding to any set of values of p_0 and h_0 can be deduced in a very simple way, and vice versa. This method is described in the appendix to this paper.

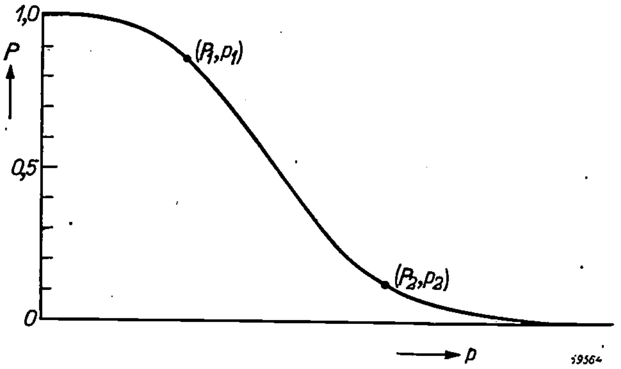


Fig. 8. Illustration of a set of parameters intended to take into account the requirements of producer and consumer. Two points on the operating characteristic are chosen for which the probabilities of acceptance have certain preassigned values, P_1 and P_2.

In the "Army Service Forces" tables developed in the United States during the second world war, and with minor alterations subsequently taken over by the Statistical Research Group, Columbia University (see the tables mentioned under [6]) and Grant's book cited in reference [2])), the sample size has been adopted as one fundamental parameter together with one point on the operating characteristic, for which the 5% risk point of the producer was chosen ($p_{0.95}$). The reason for this was that in practice the sample size is often found to be the main limiting factor, so that it should more emphatically be taken into consideration. Again we may pass over from this set of parameters to p_0-h_0 and vice versa with the aid of a simple set of curves. For details we refer to the appendix.

[7]) J. H. Enters, T. Eff. Doc. 18, 262-266, 1948 (No. 11).
[8]) J. H. Enters and H. C. Hamaker, not yet published.
[9]) J. H. Enters, The choice of the sample size in single sampling (in Dutch), Statistica 1, 228-234, 1948.

Yet another parameter, already introduced in the earliest tables of Dodge and Romig and since then frequently employed, is the so-called "average outgoing quality limit", usually abbreviated as the AOQL. The argument leading to this concept runs as follows.

Suppose, firstly, that all inspection lots contain the same percentage defectives p and, secondly, that every lot rejected undergoes a 100% inspection and is then passed on to the consumer free of defectives.

If P denotes the probability of acceptance and p the percentage defectives in the lots then, under the said circumstances, a fraction P of the lots submitted for inspection is on the average accepted and a fraction (1—P) rejected. Hence the average percentage defectives in the lots delivered is

$$p \cdot P + 0 \cdot (1 - P) = p \cdot P,$$

which is equal in value to the shaded area in *fig. 9*. We see at a glance from this figure that as p is made to increase, the average outgoing quality passes through a maximum, which is actually reached when the diagonal AC runs parallel to the tangent to the operating characteristic in point B. This maximum, called the "average outgoing quality limit", or AOQL, will be denoted by the symbol $\overline{p}_{\max}$. The consumer can always be guaranteed that even under the most unfavourable circumstances, which never obtain in practice, the average percentage defectives in the lots he receives can never exceed the value of $\overline{p}_{\max}$ corresponding to the sampling plan adopted.

As in the previous cases, a simple method for finding $\overline{p}_{\max}$ for a given plan has been set out in the appendix.

The choice of a set of parameters

The principal sets of parameters to be found in literature have now been discussed. Which of these will prove most convenient in practice depends largely on personal taste and on the circumstances under which a sampling plan is operated. One aspect of the problem should, however, always be borne in mind, namely that by disposing of two parameters we have to serve three masters, the consumer, the producer and the inspector. It may be true that as a rule the inspector belongs to some department of either the producer or the consumer, but even so he will have only a limited personnel at his disposal, so that as far as the sample size goes he has an independent voice in the matter.

The producer is for obvious reasons always inclined to push the producers' risk point to the highest value possible, while the consumer is likewise inclined to maintain his risk point at the lowest level he can get accepted, so that if we adopt these risk points for our two parameters we are always likely to be driven to too steep an operating characteristic requiring an average sample size unacceptable to the inspector.

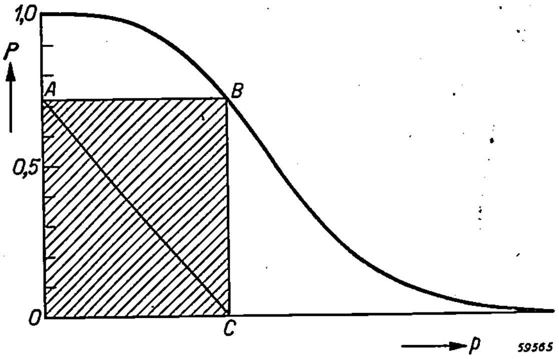


Fig. 9. Illustration of the concept of the "average quality limit".

If, on the other hand, we pitch, as in some of Dodge and Romig's tables, on the average outgoing quality limit and the consumer's risk point, we are using two parameters both intended to protect the consumer and we consequently run the risk of troubles with producer and/or inspector; or if, as in the tables of the Statistical Research Group, Columbia University, the leading part is allotted to sample size and producers' risk point, too little attention may be paid to the interests of the consumer.

Considered in this light the use of the point of control seems to offer some definite advantages. When producer and consumer have specified their respective risk points at say 2 and 5% the point of control should obviously be chosen about midway between them, say at 3.5%. By fixing p_0 in this fashion we are in a way simultaneously discounting the demands of both producer and consumer, while in the final adjustment of the slopc h_0 we still possess a certain degree of freedom to satisfy the demands of the inspector. On these grounds the point of control p_0 and the sample size n_0 seem to constitute a particularly attractive set of parameters, the more so since the simple relation

$$n_0 p_0 = c_0 + 0.67$$

yields at once the acceptance number c_0 of the corresponding single sampling plan, while a double sampling plan can easily be found with the aid of table II.

In a third article to appear in a subsequent issue of this journal we hope to explain how these principles have been employed in setting up a sampling table now in use on an extensive scale in the Dutch factories of our concern.

Appendix: The transition from one set of parameters to another

The transition from the 1-to-10 risk point to the fundamental parameters p_0 and h_0

As found above, sampling plans with the same relative slope h_0 possess operating characteristics that may be made to coincide by a simple adjustment of the sample size. It follows at once that the ratio $p_{0\cdot10}/p_{0\cdot90}$ is a function of h_0 alone, and the same holds true for the ratio $p_{0\cdot10}/p_0$ [10]). By plotting these two ratios together as functions of h_0 we obtain a graph by means of which the transition from one set of parameters to the other can be readily achieved. If, for example, we have

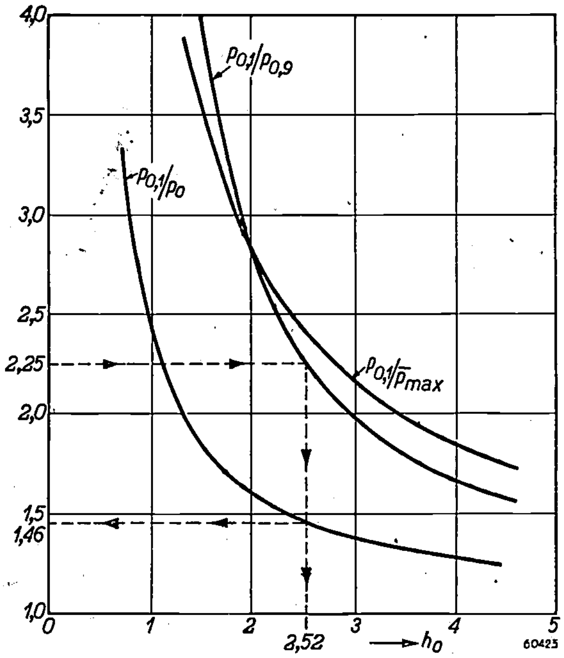


Fig. 10. Graph for the transition from one set of parameters to another. The broken lines with arrow points show how p_0 and h_0 can be found when $p_{0\cdot10}$ and $p_{0\cdot90}$ are known.

$$p_{0\cdot10} = 4.5\%, \quad p_{0\cdot90} = 2\%\,,$$

and consequently

$$p_{0\cdot10}/p_{0\cdot90} = 2.25\,.$$

we start from the corresponding point on the ordinate in *fig. 10* and travel, as indicated by arrow points, first to the right until the curve $p_{0\cdot10}/p_{0\cdot90}$ is intersected and then downwards to the points of intersection with the $p_{0\cdot10}/p_0$ curve and with the horizontal axis. On the horizontal axis we read

$$h_0 = 2.52,$$

and on the vertical axis

$$p_{0\cdot10}/p_0 = 1.46,$$

whence

$$p_0 = 4.5/1.46 = 3.08\%.$$

Thereby p_0 and h_0 have been obtained. By substituting these values in (7) and (8) we find

$$c_0 = 9.25\,;\ n_0 = 322,$$

which are rounded off to

$$c_0 = 9 \text{ and } n_0 = 320.$$

If a double sampling plan should be preferred table II provides D (2; 2, 10, 10) with $n_1 = 3.78/0.0308 \approx 123$ as the most suitable approximation.

The transition from n_0 and $p_{0\cdot95}$ to p_0 and h_0

For single sampling the product n_0p_0, the ratio $p_{0\cdot95}/p_0$ and consequently the product $n_0p_{0\cdot95}$ depend on h_0 alone. A graph giving the two products n_0p_0 and $n_0p_{0\cdot95}$ as a function of h_0 enables us to achieve our present purpose in a simple manner. Suppose we are looking for a sampling plan with

$$p_{0\cdot95} = 2\% = 0.02 \text{ and } n_0 = 130\,,$$

or

$$n_0p_{0\cdot95} = 2.6.$$

Then, starting from the corresponding point on the vertical axis, we travel in just the same way through *fig. 11* as we did in fig. 10 (see the arrow points). We thereby obtain

$$h_0 = 1.90 \text{ and } n_0p_0 = 5.60\ ;\ \text{hence } p_0 = 5.60/130 = 0.043, \text{ and } c_0 = 5.60 - 0.67 = 4.93 \approx 5,$$

which fixes the single plan required. In deriving a double sampling plan on the same basis an approximation has to be accepted, because the average sample size now depends on the percentage defectives in the inspection lot and, consequently, on the circumstances under which the double sampling plan is operated. As mentioned earlier, however, double sampling roughly saves about 30% in the number of observations. A double plan requiring on the average about 130 observations per lot corresponds to a single sampling plan with a sample

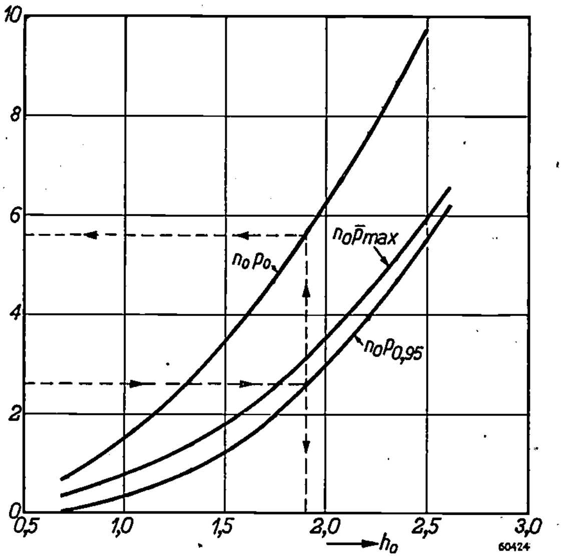


Fig. 11. As fig. 10. The arrow points indicate how h_0 and p_0 can be found for given n_0 and $p_{0\cdot95}$.

[10]) In this connection it would be more consistent to write $p_{0\cdot50}$ instead of p_0.

size of (100/70) × 130 = 185. Proceeding as before with $p_0 = 2\% = 0.02$ and $n_0 = 185$ we find $h_0 = 2.16$ and $p_0 = 4\%$, and table II indicates D (2; 1, 7, 7) with $n_1 = 2.71/0.04 = 68$ as the most suitable choice.

Determination of $\overline{p}_{\max}$ for a given sampling plan

In analogy with the foregoing we may conclude at once that ratios such as $p_{0\cdot 10}/\overline{p}_{\max}$ or $p_0/\overline{p}_{\max}$ depend on h_0 alone, the same being true for the product $n_0\overline{p}_{\max}$ in single sampling. This product and the first of the two ratios mentioned have also been plotted in figs 11 and 10 respectively, and, with the aid of the resultant curves, $\overline{p}_{\max}$ can be found for a given sampling plan.

For example, if it is asked to find $\overline{p}_{\max}$ for a single sampling plan with $n_0 = 130$, $c_0 = 8$, we have from (8) $h_0 = 2.36$ and then from fig. 11 $n_0\overline{p}_{\max} = 5.23$, so that we must have $\overline{p}_{\max} = 5.23/130 = 0.04$, which is the same value as that given in the tables of Dodge and Romig[6]). Proceeding in the same manner for the sampling plan $n_0 = 670$, $c_0 = 3$ we arrive at $\overline{p}_{\max} = 0.0025$ as against 0.0029 according to Dodge and Romig, a difference of no practical consequence.

Conversely we read from Dodge and Romig's tables that the double sampling plan $n_1 = 150$, $n_2 = 350$, $c_1 = 1$, $c_2 = c_3 = 9$ corresponds to $\overline{p}_{\max} = 0.011$ and $p_{0\cdot 10} = 0.03$. Consequently $p_{0\cdot 10}/\overline{p}_{\max} = 2.7$ and with the aid of fig. 10 we find $h_0 = 2.08$, $p_{0\cdot 10}/p_0 = 1.58$, and finally $p_0 = 0.03/1.58 = 0.019$.

This provides the simplest method of finding p_0 and h_0 for the sampling plans given in Dodge and Romig's tables.

Summary. First, making use of Poisson's probability formulae, the expressions needed for a numerical theory of sampling plans are derived, and next two fundamental parameters are introduced for describing the main features of an operating characteristic, namely the "point of control" and the "relative slope". Between these two parameters on the one hand and the sample size and acceptance number for single sampling plans on the other, two simple relations, (7) and (8), are found to exist, as is illustrated in table I. The operating characteristics of double sampling plans and single sampling plans possessing the same values of the fundamental parameters are shown to coincide to a very high degree; such plans will give the same inspection performances in practice and may be called "equivalent". By comparing the average sample size of a double plan with the sample size of an equivalent single sampling plan one arrives at the concept of "inverse efficiency", which measures the relative number of observations that can be saved by resorting to the double sampling principle. Next the dependence of the inverse efficiency on the position of the screens in the random walk diagram is discussed and on the basis of these arguments double sampling plans are selected for practical use. These are set out in a table, from which a double plan with prescribed values of the fundamental parameters can easily be constructed.

A similar reasoning is then applied to sequential sampling and the resultant conclusions are briefly discussed.

Finally, various other parameters that have already made their appereance in literature are considered, such as the producers' and consumers' risk points, and the average outgoing quality limit. This leads to a discussion as to which of these parameters is to be considered most suitable in practice, and arguments are produced tending to recommend the point of control and the sample size as a particularly practical set.

In the appendix simple graphical methods are described which allow of a rapid transition from one set of parameters to another.

CONDUCTION PROCESSES IN THE OXIDE-COATED CATHODE

by R. LOOSJES and H. J. VINK.

537.311.33: 621.315.592.4: 621.3.032.216

From the discovery of the thermionic emission of the alkaline earth oxides the modern oxide-coated cathode has been developed which is widely used in the field of electronics, for instance in radio valves, in cathode ray tubes and in gas-discharge lamps. An essential part of the emission process of this cathode is the electronic conduction of the oxide layer. Some new experiments and theoretical considerations concerning the nature of this conduction have made it possible to understand the behaviour of the oxide-coated cathode under various conditions better than has previously been the case.

Introduction

In 1903 Wehnelt discovered that a glowing strip of platinum covered with a small quantity of calcium oxide emits an appreciable quantity of electrons at temperatures where the thermionic emission of the metal itself is negligible. From this discovery in the course of time the modern oxide-coated cathode, as used, for instance, in most radio valves, has been developed, which consists of a porous sintered layer of mixed crystals of barium oxide and strontium oxide, 10-100 μ thick, on a metal core. The so-called indirectly heated cathode consists of a nickel tube coated on the outside with the oxide and surrounding a heating element (tungsten wire with an insulating coating). By means of this heating element the tube is heated to 1000-1100 °K and when the cathode is brought to a negative potential with respect to a collector electrode (anode) a strong thermionic emission takes place from the oxide coating.

At the temperatures mentioned a good oxide-coated cathode gives a thermionic emission (saturation current) of the order of 1 A/cm^2. The electrons emitted from the coating are restored from the metal core. From this it follows that the oxide coating itself must be conductive, and the object of the present article is to study this conduction process more closely, in the light of new experimental investigations.

Preparation of the oxide coating

As already remarked, with indirectly heated cathodes a tube of nickel is used as core for the oxide coating. For reasons which will be made clear later, the metal used is not pure nickel but an alloy obtained by adding small quantities of a readily oxidizable metal (Mg, Si, Mn, Al, Ti) to the nickel. First the metal core is coated with a thin layer of carbonates of alkaline earth metals (barium, strontium). These carbonates can be obtained, for example, by precipitation from an aqueous solution of the corresponding nitrates or hydroxides with ammonium carbonate. After being washed and dried the substance is ground for a number of hours in a ball mill with a solution of what is known as a binder (e.g. nitrocellulose dissolved in a volatile organic solvent). The resulting milkwhite suspension (size of particles 2-10 μ) is applied to the metal core, for instance, by means of a spray-gun. The solvent evaporates very quickly, leaving on the metal a porous coating of fine carbonate crystals cemented together and to the core by the binder.

The cathode is then mounted in the tube and heated in vacuo to 1100-1400 °K, whereupon first the binder decomposes and after that the carbonates decompose into oxides and carbon dioxide. The released carbon dioxide oxidizes the carbon of the binder to carbon monoxide. The gaseous products are pumped away, leaving a coating of pure oxides.

During the heating process the oxide coating sinters somewhat and adheres well enough to the core. The dissociation of the carbonate crystals requires great care. It must not be done too quickly nor at too high a temperature. If this is not properly attended to it may happen that the oxide coating sinters until it is no longer porous. Experience has taught that this is detrimental for the emission, for only a porous layer (porosity 50-60%) gives good emission. For the oxide coating mixed crystals of BaO and SrO. (1 mol BaO to 1 mol SrO) are used because this material has proved to possess the best thermionic emission properties.

Activation of the cathode

The layer of earth alkaline oxides thus formed on a metal core is not capable of yielding the desired electron emission directly. To bring this about the cathode must first be "activated", and this is generally done in the following way. The cathode is heated to 1100-1250 °K, while a voltage of some tens of volts positive with respect to the cathode is applied to the anode. Immediately after applying the voltage the emission is still very small, but it increases at first gradually and later on more rapidly. Ultimately the temperature of the cathode and the anode voltage have to be reduced in order to avoid very high emission currents, which appear to have a detrimental effect. After a few minutes a stationary state is reached and a cathode is obtained from which a continuous emission current of 10^3-10^4 A/m^2 (0.1-1 A/cm^2) at a temperature of 1000-1100 °K and a pulse emission current (10^{-4} sec) of 2×10^4 to 10^5 A/m^2 can be drawn.

Although drawing current is mostly necessary for good activation, it is sometimes possible to activate the cathode without applying a voltage, i.e. without drawing current. This is particularly the case when the core contains the above-mentioned reducing materials (such as Si and Mg). Full activation can then be reached merely by heating the cathode in vacuo to a high temperature (1100-1200 °K).

This shows that a partial reduction of the oxide is essential for activating the cathode. It has been found that as a result of this reduction a certain quantity of free barium is formed in the coating of BaO-SrO mixed crystals. In the method of activation by drawing current described above this barium is released by electrolysis of the oxide coating, whereby Ba^{2+}-ions move in the direction of the metal core and O^{2-}-ions in the opposite direction. Finally the oxygen ions emerge from the layer in the form of free O_2 and the Ba^{2+}-ions are likewise neutralized, so that ultimately a quantity of free barium is left in the coating.

The quantities of barium in question are very small (in the order of 0.01%), but this free barium is nevertheless of essential importance for the emission. The fact is that as soon as the amount of barium is reduced, for instance by a momentary increase of temperature causing the barium to evaporate, or by chemical conversion (heating in a gaseous atmosphere containing traces of O_2, Cl_2 or of H_2O), the emission is greatly diminished. By introducing new free barium, for instance by evaporation, the emission is increased again.

The thermonic emission of a well activated oxide-coated cathode is, of course, strongly dependent on the temperature. In the case of the thermionic emission from metals (such as tungsten) Richardson's formula applies:

$$J_s = A\, T^2 \exp\,(-e\varphi/kT) \quad . \; . \; . \quad (1)$$

or [1]):

$$\log J_s - 2 \log T = \log A - 0.434\; (e\varphi/k)\, \frac{1}{T}, \quad (2)$$

where A represents a universal constant ($= 1.2 \times 10^6$ A/m^2 °K^2) and φ is the work function amounting, for instance, to 4.5 V for tungsten ($e\varphi$ is the energy required to liberate an electron from the metal in vacuum). If, in analogy with this, one plots for an oxide-coated cathode $\log J_s - 2 \log T$ against $1/T$, a straight line

$$\log J_s - 2 \log T = \log A' - B/T. \; . \; . \quad (3)$$

is found. If B is then replaced by

$$0.434\, e\, \varphi'/k\,, \; . \; . \; . \; . \; . \; . \; . \; . \quad (4)$$

we find the quantity φ', which in this case is also called the work function. For a well activated oxide-coated cathode φ' lies between 0.9 and 1.1 V.

When emission takes place electrons leave the oxide coating. In the stationary state this loss of electrons is compensated by electrons leaving the metal core and entering the oxide coating. Obviously a continuous thermionic emission is only possible if the electrons are able to pass through the coating into vacuum, i.e., if the oxide coating is an electronic conductor.

The oxide coating as semi-conductor

At first sight it may seem surprising that a layer of (Ba, Sr) O should be capable of conducting electrons, since it consists of transparent oxide crystals, and it is a known fact that such non-metallic substances generally behave as perfect insulators.

One would be inclined to ascribe this to electrolytic conduction of the oxides, and if the temperature is high enough the phenomenon of electrolysis indeed occurs; we have already seen from the foregoing that electrolysis plays a part in the process of activation. But the ionic conduction is far too small to account for the emission currents observed, and particularly so at the relatively low temperatures (1000-1100 °K) at which the cathode is used; it has been estimated that at such temperatures with a fully activated cathode the ionic current forms only 0.001% of the total current through the oxide coating, so that ionic conduction need not be considered.

It is known, however, that only absolutely pure crystals of stoichiometric composition are good insulators; impure crystals, contaminated in some

[1]) By "log" is to be understood the logarithm with the base 10.

way or other, for instance with foreign atoms or an excess of one of the constituent atoms, show conduction and are therefore called electronic semi-conductors.

The conductivity γ of an electronic semi-conductor depends strongly on temperature. A semi-conductor containing only one kind of impurity can be described by a formula of the type:

$$\gamma = K \exp\left(-E/kT\right), \quad . \; . \; . \quad (5)$$

where E represents a constant with the dimension of energy and K is likewise a constant which depends, however, on the amount of the impurity.

Fig. 1. The logarithm of the conductivity γ as a function of $1/T$ for a semi-conductor to which applies:

$$\gamma = K_1 \exp\left(-E_1/kT\right) + K_2 \exp\left(-E_2/kT\right)$$

(fully drawn line). The two dotted lines can be described by the separate terms.

When for such a material log γ is plotted as a function of $1/T$ a straight line is obtained from the slope of which the energy constant E can be deduced. It has been found experimentally that E is also slightly dependent on the impurity concentration, the value of E decreasing with increasing impurity concentration. Sometimes the semi-conductor contains more than one impurity, and when for instance there are two kinds of impurity γ can be described by a formula of the type:

$$\gamma = K_1 \exp\left(-E_1/kT\right) + K_2 \exp\left(-E_2/kT\right) . \; . \quad (6)$$

When in such a case log γ is again plotted against $1/T$ a curve is obtained as represented in *fig. 1.* From the slope of the two straight parts one can deduce E_1 and E_2.

Theory of electronic semi-conductors

Let us now consider the conduction and the thermionic emission of electronic semi-conductors more closely. According to the modern atomic theory (the quantum theory) electrons in the periodic field of the crystal lattice cannot have all possible energies. On the contrary, the possible energy levels are restricted to so-called allowed energy bands, whilst the intermediate values are prohibited [2]). In a crystal some of the allowed energy levels are always occupied by electrons while others are not. The lowest allowed bands are "filled", while the higher bands are "empty". Sometimes there is a half-filled band between the full and the empty ones, and this is the case with metals, the conductivity of which is due to this fact, since in this case the electrons in the half-filled band can be raised to a higher level in that band at the cost of very little energy taken from an external field, which means that an electric current flows through the metal.

If an entirely filled band is followed by an entirely empty band then the substance is an insulator, because the electrons cannot then be raised to a higher energy level, i.e. in the empty band, at the cost of little energy. This is only possible when there are partly filled levels between the filled and empty bands, as is the case with the above-mentioned crystals containing foreign atoms or, in the case of the oxide coating, with an excess of barium. These additional energy levels are usually at a small distance E_0 (in the order of 1 to 2 eV) below the empty band (*fig. 2*). Now it is possible that under the influence of the thermal agitation electrons are raised from the impurity levels to the empty band and become conducting electrons. Putting the number of impurity levels occupied by electrons (at $T = 0$ °K and per unit volume) as N, then according to the theory the number of electrons per unit volume (n) in the empty band at the temperature T is given by

$$n = 2\, N^{\frac{1}{2}} \cdot \left(\frac{2\pi mkT}{h^2}\right)^{3/4} \exp\left(-\frac{E_0}{2kT}\right), \quad . \quad (7)$$

where m is the electron mass (9×10^{-31} kg) and h is Planck's constant (6.6×10^{-34} W sec^2).

For the conductivity we find:

$$\gamma = b\, n\, e\,, \quad . \; . \; . \; . \; . \; . \; . \quad (8)$$

where b, the mobility of the conducting electrons, may be regarded as a temperature-independent

[2]) See for instance E. J. W. Verwey, Electronic conductivity of non-metallic materials, Philips Techn. Rev. 9, 46-53, 1947

constant. When $\log \gamma$ is plotted as a function of $1/T$ in a certain, fairly narrow, temperature range then one obtains a straight line, since the term with $T^{3/4}$ has little effect compared with the exponential function. In this range γ can therefore be represented by the function

$$\gamma = K \exp\left(-E/kT\right), \quad . \; . \; . \quad (9)$$

where E differs but little from $E_0/2$.

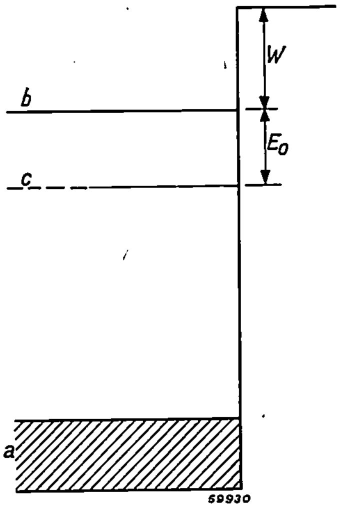


Fig. 2. Energy bands and additional levels of an electronic semi-conductor with an excess of metal atoms. *a* is the full band. E_0 is the distance between the narrow band of impurity levels (*c*) and the lowest level of the empty band (*b*). In the theory the factor $E_0/2$ plays an important part. *W* is the distance from the lowest level of the empty band to the zero level (vacuum).

A semi-conductor can also emit electrons. The foregoing theory is able to account for this and can describe the emission properties of the oxide-coated cathode. For a thermionic emission not only the energy separating the electrons in the impurity levels from the lower boundary of the empty energy band has to be overcome, but also the energy difference W between this lowest empty energy level and the "zero level" (corresponding to an electron with zero velocity in vacuo). For the thermionic emission of the semi-conductor one finds theoretically the formula

$$J_s = cne\left(\frac{kT}{m}\right)^{\frac{1}{2}} \exp\left(-W/kT\right), \quad . \; . \quad (10)$$

where c represents a numerical constant, viz. $(2/\pi^3)^{1/4}$. From the formulae (10) and (7) it follows that $e\varphi'$ (see (4)) differs little from $\frac{1}{2}E_0 + W$.

In what follows it will be investigated whether the conduction of the oxide coating can be explained from what is known about electronic semi-conductors. The results of the experimental investigation to be described will show that this is not fully the case.

Experimental investigation of conduction

Many investigations have been carried out regarding the conductance of the oxide coating, but hitherto it had not been possible to derive a coherent picture from the data obtained. What had been definitely established, however, was that almost exclusively electronic conduction takes place in the oxide-coated cathode.

For the investigations of this conduction carried out in the Philips Laboratories at Eindhoven a cathode was used as employed in cathode ray tubes. This cathode consists of a hollow cylinder of nickel (6 mm × 8 mm^2) closed at one end and containing a heating element. The coating of Ba-Sr carbonate, 50 μ thick, covers the flat outer face of the closed end. Two of these cathodes, coated as smoothly as possible but otherwise quite normal, were pressed together with the carbonate coatings facing each other (*fig. 3*) and held in that position by springs. Thus between the two metal cylinders there was a carbonate coating 2 × 50 = 100 μ thick, which could be homogeneously heated by means of the two filaments. The temperature was measured with a chromel-alumel thermocouple.

The cathodes pressed together in this way were mounted in a glass bulb and this connected to a high-vacuum pump. Of course both the bulb and the cathodes were subjected to a thorough degassing process, so that ultimately, after dissociation of the carbonates and degassing of the whole unit, the bulb could be sealed with a vacuum of 10^{-5} mm Hg. In order to improve and maintain the vacuum, just before sealing the bulb a quantity of barium was evaporated as "getter", care being taken that this barium did not reach the cathode.

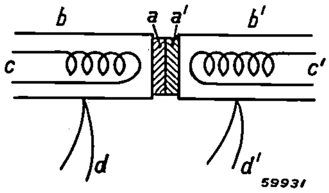


Fig. 3. Arrangement employed for measuring the conductance. *aa'* the two oxide coatings, *bb'* nickel tubes, *cc'* filaments, *dd'* thermocouples.

The core of the cathode consisted of ordinary "cathode nickel" containing Mg and Si, so that the oxide coating could be thermally activated. By carrying this out at relatively low temperatures it was possible to prolong the activation process over a number of hours, so that it could be interrupted at any moment to allow of measurements being taken in successive stages of the process. The conductance of the oxide coating approximately 100 μ thick was measured in various stages of activation as a function of temperature. These measurements were made with alternating current in a bridge circuit ("Philoscope" [3])), the voltage across the unknown resistance being about 1 V. At the same stages of activation current-voltage characteristics of the coating were recorded with the aid of short current pulses, the peak value of the voltage amounting to about 10 V.

Discussion of the results of the measurements

We shall first deal with the results obtained by measuring the conductance at low voltage: in *fig. 4* a graph is reproduced representing for a given oxide coating the logarithm of the resistance R ($-\log R$ is equal to $\log \gamma$ except for an additional constant) as a function of $1/T$ for various successive states of activation. It is clearly seen that as the activation increases, i.e. with increasing quantity of free barium in the coating, the conductance likewise increases and becomes less dependent on temperature. Further, in *fig. 5* another similar graph has been plotted, which shows that this curve can be divided into three distinct parts: part *I* (600-800 °K), only slightly dependent on temperature, part *II* (800-1000 °K), more strongly dependent on temperature, and part *III* (> 1000 °K), again less dependent on temperature. Disregarding part *III*, one would be inclined to explain the conduction mechanism as being that of a semi-conductor in which two kinds of foreign atoms occur, for instance two kinds of barium atoms placed differently in the lattice, so that the conductivity could be represented by formula (6). The fact that the curves become less steep as the activation increases agrees with what has already been said about the effect of the number of impurity atoms on E. When the quantities E_1 and E_2 are calculated from the curves in accordance with formula (6) we find the following values for the successive stages of activation:

Stage	1	2	3	4	5	6
E_1	0.22	0.14	0.12	0.11	0.10	0.09 eV
E_2	0.98	1.14	1.24	0.96	0.96	0.94 eV
$e\varphi'$	1.07	—	1.10	0.90	1.02	0.87 eV

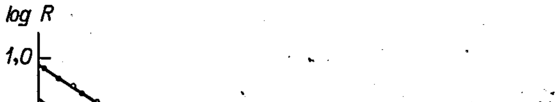


Fig. 4. The logarithm of the resistance R ($-\log R$ is equal to $\log \gamma$ except for an additional constant) as a function of $1/T$ for an oxide-coated cathode in successive stages of activation (from bottom to top).

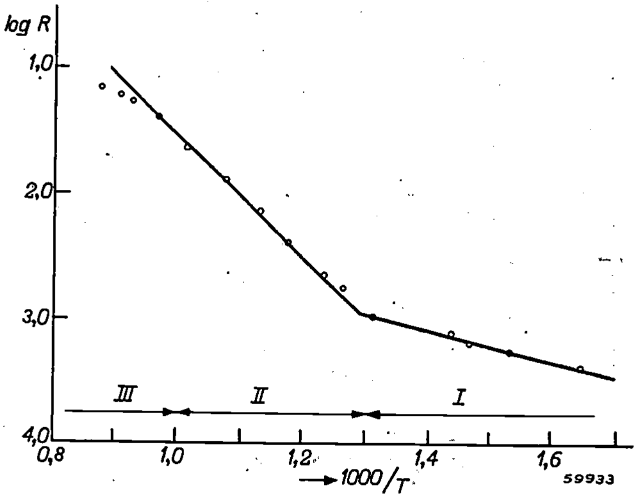


Fig. 5. The logarithm of the resistance R as a function of $1/T$ for an oxide-coated cathode. This curve clearly shows that there are three temperature ranges *I*, *II* and *III* in which the behaviour of the coating is different. The straight lines correpond to thc dotted lines in fig. 1.

It appears that for all stages of activation E_2 is equal, within the limits of the accuracy of the measurements, to the product $e\varphi'$ of e and the work function at the respective stage. Now with indisputably real semi-conductors E has always been found to be much smaller than $e\varphi'$, so that the assumption of two kinds of impurity levels is improbable. Neither can the occurrence in part *III* of the $\log \gamma$-$1/T$ characteristic be explained by the semi-conductor theory.

Something similar is encountered when studying the current-voltage characteristics, an example of which is given in *fig. 6*, from which it may be seen

[3]) See Philips Techn. Rev. 2, 270-275, 1937.

that with the J-V characteristics the same temperature ranges can be distinguished as those which played a part in the log γ-$1/T$ curves. Up to 700-800 °K the characteristics are straight, whilst between 800 and 1000 °K they curve towards the V-axis. Around 900 °K the curvature begins to diminish and above 1000 °K the characteristics are again straight. Now it is possible to explain a curved J-V characteristic by electronic conduction. For this it has to be assumed that rectifying contacts (barrier layers) exist at the boundary between the metal and the oxide coating. It is unlikely, however, that electrons of one kind of impurity levels (E_1) whose influence is preponderant at low temperatures ($<$ 800 °K) do not exhibit this effect, whilst those of another kind (E_2) whose influence predominates at higher temperatures ($>$ 800 °K) do show it. Whereas, therefore, there is nothing against ascribing that part of the conductance represented by the first term (the term with E_1) of formula (6) to real electronic conduction, for the behaviour represented by the second term (the term with E_2) some other explanation has to be found.

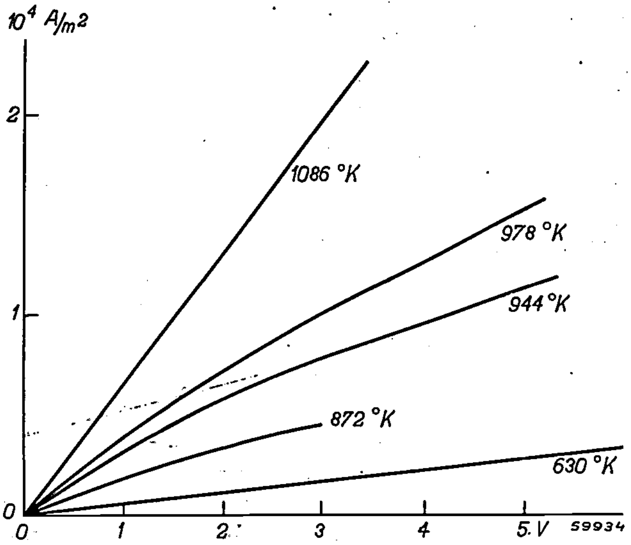


Fig. 6. J-V characteristics of an oxide-coated cathode at various temperatures. Between 800 and 1000 °K the characteristics are curved towards the V-axis.

A new supposition regarding the causes of conductance

Attention has already been drawn in the foregoing to the fact that a good oxide-coated cathode has to be highly porous and that a closely sintered, non-porous cathode has poor emission. It is therefore surprising that hitherto so little attention has been paid to this question of porosity in trying to account for the phenomena in the oxide coating. It is our belief that the conduction by means of electrons present in the pores plays an important part in the phenomena observed. It is not illogical to suppose the existence of an electron gas of a certain density in the pores. For a long time it has in fact been accepted that owing to their emitting power the outermost grains of the oxide coating set up a cloud of electrons in a layer immediately adjacent to the surface of the cathode. If the outermost grains can do that, then why not the grains inside the coating too? The pores in the coating would then be filled with an electron gas coming from the electron clouds of the surrounding grains. As early as in 1918 von Laue [4]) investigated theoretically the electron density in a space bounded by walls emitting electrons. He proved that at low temperatures the density varied with temperature in the same way as the emission. At higher temperatures, and thus with increasing emission, owing to the mutual repulsion of the electrons, the density of the electron gas in the centre of the cavity diminishes as compared with the density close to the emitting surface. As a matter of fact the density in the centre of the cavity increases only in direct proportion to the temperature, whilst that close to the wall always increases according to the same exponential law as the emission itself. We shall revert to this presently, but here we can already find support in von Laue's results for our conceptions regarding the part played by the pores in the conduction process.

The pores referred to above should not be imagined as being cavities separated from each other, because then the electrons would always have to pass through a semi-conducting layer to move from one cavity to the next. The cavities are to be regarded more as a coherent sequence, so that winding channels are formed in the material, leading from the metal core to the free cathode surface. We shall now see what effect this has on the conduction.

In these channels the electrons move with a mean free path l, which is of the order of magnitude of the cross section of the channel, i.e. of the diameter of the grain. When a field F is applied, during the time t taken by the electrons to travel their mean free path l these electrons (with charge e and mass m) undergo an acceleration

$$a = \frac{eF}{m}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

[4]) M. von Laue, Jahrb. Radioaktivität und Elektronik 15, 205-256, 1918; Ann. Physik 58, 695-711, 1919.

During this time the average increase in velocity in the direction of the field is

$$\overline{\Delta v} = \frac{eF}{2m}\, t\,.$$

We now distinguish between two cases:

1) The average velocity $\bar{v}$ is great compared with $\overline{\Delta v}$. In this case $t = l/\bar{v}$ and therefore

$$\overline{\Delta v} = \frac{e}{2}\frac{F}{m} \cdot \frac{l}{\bar{v}} \quad \ldots\ldots\ldots \quad (12)$$

Since the acceleration is destroyed after the electrons have traversed the mean free path, $\overline{\Delta v}$ represents at the same time the average velocity of the electron in the direction of the field, so that for the current density we find:

$$J = n\, e\, \overline{\Delta v} = n\, \frac{e^2 F}{2m} \cdot \frac{l}{\bar{v}}\,, \quad \ldots\ldots \quad (13)$$

where n represents the number of electrons per unit volume. A similar formula was derived long ago by Drude to explain the electronic conduction of metals. Since J is proportional to the field strength F, and this in turn is proportional to the applied voltage V, this formula yields a rectilinear J-V characteristic (Ohm's law).

2) The average velocity $\bar{v}$ is small with respect to $\overline{\Delta v}$. This is the case if l or F is sufficiently large. Then one can write for t: $t = l/\overline{\Delta v}$ and thus

$$\overline{\Delta v} = \sqrt{\frac{eF}{2m}\, l}$$

and

$$J = ne \sqrt{\frac{eF}{2m}\, l}\,. \quad \ldots\ldots \quad (14)$$

Since F is proportional to the potential difference applied, Ohm's law no longer holds in this case and we get a parabolic J-V characteristic, curved towards the V-axis.

The possiblility of this being the case with the oxide-coated cathode may be seen from the following consideration. With a potential difference of 10 V across a coating 100 μ thick thc increase of the kinetic energy per mean free path (of the order of magnitude of the grain, thus 2-10 μ) is 0.2-1 eV. This is very much greater than the average thermal kinetic energy ($^3/_2\, kT$), which at 800 °K amounts to 0.1 eV. The fact that at very high temperatures (> 1000 °K) the characteristic again straightens out can be explained by the theory of von Laue, according to which the electrons are driven out of the centre of the cavities (*fig.* 7) owing to the space charge. As a result the electrons move mainly in a thin layer along the wall of the pores, because an electron leaving the wall does not travel across the cavity but is reflected by the space charge back to the wall on the same side. Thus the free path is greatly reduced and we then again have the case of formula (11), the characteristic becoming straight again. Since, moreover, the average density in the layer referred to is less than the boundary density, which is proportional to the saturation current, it also follows that with increasing temperature the conduction in the pores within this temperature range does not increase so much as it does in the temperature range where the space charge is still negligible and where, therefore, the conduction increases with the emission.

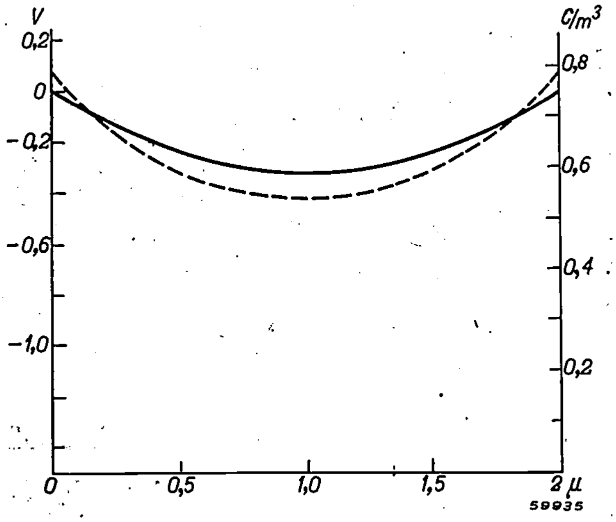


Fig. 7. Potential V in volts (fully drawn line) and charge density in coulombs/m³ (dotted line) between two thermionic plates at a distance of 2μ (T = 952 °K, $J_s = 3.8 \times 10^4$ A/m²).

In order to ascertain whether the explanation given for the curved J-V characteristic is correct we calculated J as a function of V according to formula (14) for a cathode which at 1000 °K should have an emission of 8×10^4 A/m² with $\varphi' = 1.1$ V. The electron density is taken for a simple case (two infinite, parallel, flat plates at a short distance from each other). For a distance of 2 μ we found $J = 9.0 \times 10^3 \sqrt{V}$ A/m² and for a distance of 5 μ $J = 8.5 \times 10^3 \sqrt{V}$ A/m². These current densities are of the same order as those found experimentally.

Summarizing it may therefore be said that in the porous oxide coating the conduction by the electron gas between the grains is of essential importance. Without taking this conduction into account it is not possible to give a satisfactory explanation of all the phenomena found experimentally. The conduction by the electron gas and the electronic conduction of the crystals of the oxide coating are to be regarded as two parallel processes which together determine the conduction of the coating. At low temperatures conduction through the crystals predominates and at high temperatures the conduction by the electron gas.

Strong support for the theory of conduction by the electron gas is found in the fact that not only can the phenomena be predicted qualitatively but that they can also be calculated quantitatively. It is interesting to note that the only parameters in this theory are the size of grains, the porosity and the magnitude of the thermionic emission, values which were known from literature.

Summary. A description is given of the methods of preparing and activating an oxide-coated cathode, from which it appears that the carbonate coating originally applied has to be subjected to such a heating process as to result in a loosely sintered, porous oxide coating. The electrons emitted from the coating have to be restored by conduction through the oxide coating. It is shown to be probable that the conduction of the oxide coating is due to two mechanisms acting in parallel: the electronic conduction of the grains, predominating at low temperatures (< 800 °K), and the conduction through the electron gas in the pores between the grains, which is preponderant at high temperatures (> 800 °K). With the aid of the conduction through the pores it can also be seen why the *J-V* characteristic for high voltages curves towards the *V*-axis. The fact that at the highest temperatures (> 1000 °K) the conduction lags behind the emission, as also the fact that in this temperature range the curvature of the *J-V* characteristic disappears, can be explained by taking into account the effect of the space charge upon the electron density and upon the field in the pores.

BOOK REVIEW

Fundamentals of Radio-valve Technique, by J. Deketh, 535 pages, 384 illustrations. — Philips' Technical Library (Book I of the series of books on electronic valves) — Published by N.V. Philips' Gloeilampenfabrieken, Technical and Scientific Literature Department, Eindhoven, Netherlands, 1949.

In every field of science or technical engineering there is a need not only of a detailed handbook for the specialist but also of a more concise source of information serving as a general guide for the student and at the same time as a book of reference for others indirectly concerned with the subject matter. Without being written in such a popular language as to make it unnecessarily superficial, such a book must nevertheless be comprehensible and, moreover, bring forward all the essential points in a logical manner.

For the author of the book discussed here on the fundamentals of radio valve technique this difficult task was perhaps made somewhat easier by reason of his working in surroundings where he was continually in contact with various classes of people engaged in this particular field of practical science. In those very same surroundings it appeared that a book such as this is appreciated not only by young technicians but also by those who are mainly interested in chemistry and physics and who need to know something about electronic systems in connection with their own particular work.

This book deals exclusively with radio receiving valves and their application in radio receivers, thus not with transmitting valves, cathode ray tubes, and so on.

The first three chapters are devoted to the fundamental principles upon which the working of electronic valves is based. Chapters IV and V go deeper into the matter of the thermionic emission. The next three chapters are concerned with the technology of the valves, a subject which has been gone into at some length, and rightly so, because also in the application of radio valves this technology influences the possibilities and limitations.

Following upon a review of the various functions of the electronic valve in a radio receiver and of the types of valves that have been developed for reception purposes, in chapters XII to XVI the author deals with the characteristic properties of these valves, whilst in the next seven chapters it is explained how these properties can be utilized in the various functions, such as low-frequency amplification, high- and intermediate-frequency amplification, detection, etc. Thus a complete insight is given into the working of a radio receiver, without, however, dealing in detail with other components like capacitors, inductors, loudspeakers, etc.; this limitation has undoubtedly contributed towards the sound treatment of the subject matter.

The advantage that the author had from his close contact with a large factory making radio valves and receivers is manifest from the space that has been reserved in this book to the treatment of phenomena of interferences such as noise, hum, microphony, secondary emission and the ageing of valves. These are points which certainly have to be taken into account in the designing and developing of receivers, and the reader will find, in a somewhat condensed form, a great deal of information about these phenomena which will enable him to evaluate for himself the extent of all sorts of effects.

to be regarded as two parallel processes which together determine the conduction of the coating. At low temperatures conduction through the crystals predominates and at high temperatures the conduction by the electron gas.

Strong support for the theory of conduction by the electron gas is found in the fact that not only can the phenomena be predicted qualitatively but that they can also be calculated quantitatively. It is interesting to note that the only parameters in this theory are the size of grains, the porosity and the magnitude of the thermionic emission, values which were known from literature.

Summary. A description is given of the methods of preparing and activating an oxide-coated cathode, from which it appears that the carbonate coating originally applied has to be subjected to such a heating process as to result in a loosely sintered, porous oxide coating. The electrons emitted from the coating have to be restored by conduction through the oxide coating. It is shown to be probable that the conduction of the oxide coating is due to two mechanisms acting in parallel: the electronic conduction of the grains, predominating at low temperatures (< 800 °K), and the conduction through the electron gas in the pores between the grains, which is preponderant at high temperatures (> 800 °K). With the aid of the conduction through the pores it can also be seen why the *J-V* characteristic for high voltages curves towards the *V*-axis. The fact that at the highest temperatures (> 1000 °K) the conduction lags behind the emission, as also the fact that in this temperature range the curvature of the *J-V* characteristic disappears, can be explained by taking into account the effect of the space charge upon the electron density and upon the field in the pores.

BOOK REVIEW

Fundamentals of Radio-valve Technique, by J. Deketh, 535 pages, 384 illustrations. — Philips' Technical Library (Book I of the series of books on electronic valves) — Published by N.V. Philips' Gloeilampenfabrieken, Technical and Scientific Literature Department, Eindhoven, Netherlands, 1949.

In every field of science or technical engineering there is a need not only of a detailed handbook for the specialist but also of a more concise source of information serving as a general guide for the student and at the same time as a book of reference for others indirectly concerned with the subject matter. Without being written in such a popular language as to make it unnecessarily superficial, such a book must nevertheless be comprehensible and, moreover, bring forward all the essential points in a logical manner.

For the author of the book discussed here on the fundamentals of radio valve technique this difficult task was perhaps made somewhat easier by reason of his working in surroundings where he was continually in contact with various classes of people engaged in this particular field of practical science. In those very same surroundings it appeared that a book such as this is appreciated not only by young technicians but also by those who are mainly interested in chemistry and physics and who need to know something about electronic systems in connection with their own particular work.

This book deals exclusively with radio receiving valves and their application in radio receivers, thus not with transmitting valves, cathode ray tubes, and so on.

The first three chapters are devoted to the fundamental principles upon which the working of electronic valves is based. Chapters IV and V go deeper into the matter of the thermionic emission. The next three chapters are concerned with the technology of the valves, a subject which has been gone into at some length, and rightly so, because also in the application of radio valves this technology influences the possibilities and limitations.

Following upon a review of the various functions of the electronic valve in a radio receiver and of the types of valves that have been developed for reception purposes, in chapters XII to XVI the author deals with the characteristic properties of these valves, whilst in the next seven chapters it is explained how these properties can be utilized in the various functions, such as low-frequency amplification, high- and intermediate-frequency amplification, detection, etc. Thus a complete insight is given into the working of a radio receiver, without, however, dealing in detail with other components like capacitors, inductors, loudspeakers, etc.; this limitation has undoubtedly contributed towards the sound treatment of the subject matter.

The advantage that the author had from his close contact with a large factory making radio valves and receivers is manifest from the space that has been reserved in this book to the treatment of phenomena of interferences such as noise, hum, microphony, secondary emission and the ageing of valves. These are points which certainly have to be taken into account in the designing and developing of receivers, and the reader will find, in a somewhat condensed form, a great deal of information about these phenomena which will enable him to evaluate for himself the extent of all sorts of effects.

An appendix of about 50 pages gives definitions, formulae and tables of great value to the technician in his daily work.

Finally, an extensive bibliography is given at the end of the book, whilst also in the text, where necessary, references are made to books dealing with a particular detail at greater length.

Naturally there is not much scope within the confines of such a book to give the full derivation of all sorts of formulae. As a rule indispensable formulae, like those for emission, anode current, noise voltage, etc., are given direct. Where formulae have been obtained by derivation the reader is assumed to have a knowledge of algebra, goniometry and the principles of differential calculation. Further, of course, a reasonable electrotechnical grounding is required.

H. van Suchtelen.

The books of Philips Technical Library are distributed in:

Norway, Sweden, Holland, Indonesia, Dutch West Indies: Meulenhoff & Co N.V., Beulingstraat 2-4, Amsterdam
Denmark: Jul. Gjellerups Boghandel, Sølvgade 87, København K.
Finland: Akateeminen Kirjakauppa, 2 Keskuskatu, Helsinki
Germany (Trizone): Buch- und Zeitschriften Union, Harvestehuder Weg 5, Hamburg 13
Switzerland: A. Francke A.G., Bubenbergplatz 6, Bern
S. A. Payot, Lucerne
Belgium: N.V. Alg. en Techn. Boekhandel, v.h. P. H. Brans, Prins Leopoldstr. 28 - Borgerhout, Antwerp
N.V. Standaard-Boekhandel, Huidevetterstraat 57-59, Antwerp
Luxemburg: Librairie Paul Bruck, 50 Grande Rue, Luxemburg.
France: Maison Dunod, 42 Rue Bonaparte, Paris VI
Great Britain and Eire: Cleaver Hume Press Ltd., 42 A South Audley street, London W. 1.
Spain: Editorial Pueyo, Tetuan 5, Madrid
Portugal: Livraria Bertrand, Rua Garrett 73, Lisbon
Italy: Librairie Internationale Corticelli Via S Tecla 5, Milann
Austria: Buchhandlung Minerva, Mölkerbastei 4, Vienna I
Yugoslavia: Jugoslovenska Knjiga, Marsala Tita 32, Belgrade
Greece: "Eleftheroudakis", Constitution Square, Athens
Turkey: Librairie Hachette, 469 Istikalât, Caddesi, Beyoğlu, Istanbul
Egypt: Lehnert & Landrock Succ., 44 Sherif Pasha str., Cairo
Palestine: Pales Press Ltd., Allenby Road 119, Tel Aviv
U.S.A. and Canada: Elsevier Book Comp. Inc., 215 Fourth Ave., New York 3
Argentine: Editorial Sud Americana S.A., Alsina 500, Buenos Aires
Brasil: Livraria Editoria Kosmos, Rua do Rosario 135-137, Rio de Janeiro
Uruguay: Libreria International S.R.L., Calle Uruguay 1331, Montevideo
Venezuela: C.A. Philips Venezolana, Apartado 1167, Caracas
Australia: Philips Electrical Industries of Australia (Pty), Ltd., "Philips House", 69-73 Clarence Street, Sydney
New-Zealand: Philips Electrical Industries of New-Zealand Ltd., G.P.O. Box 1673, Wellington G. 1.
South-Africa: Technical Books & Careers, 37 High Court Buildings, 17 Joubert Street, Johannesburg
India: Philips Electrical Co., (India) Ltd., "Philips House", 2 Heysham Road, Calcutta 20
Pakistan: Philips Company of Pakistan Ltd., Bunder Road, P.O. Box 301, Karachi 3
Syria and Lebanon: Philips Liban-Syrie S.A., P.O.B. 670, Beyrouth

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

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

1862: P. C. van der Willigen: Contact arc welding (Sheet Metal Industries **26**, 393-398, 402, 1949, Febr. No. 262).

A survey is given of the new method of arc-welding with contact electrodes. Constant arc length, independency of physical conditions of the welder, ease of starting of the arc, ease of welding and more reliable welds are the chief advantages. Other features are a special form of penetration and greater welding speed. Special application such as contact-arc spotwelding and under-water welding are briefly dealt with. See Philips Techn. Rev. 8, 161-166 and 304-308, 1946.

1863: J. L. Snoek: The Weiss-Heisenberg theory of ferromagnetism and a new rule concerning magnetostriction and magnetoresistance (Nature London **163**,837,1949, May 28).

The Heisenberg model of ferromagnetism, in contradiction to the band theory, leads one to expect that the exact filling up of lattice points with an equal number of electrons will make itself felt in some properties of a ferromagnetic alloy. From available data and from special experiments it is shown that under these conditions the magnetostriction is zero and the change in the electric resistance due to the applications of a magnetic field passes through a maximum.

1864: J. H. van Santen and F. de Boer: High indices of refraction of barium titanate and other heteropolar compounds (Nature London **163**, 957, 1949, June 18).

It is shown that the polarizability of the oxygen ions in $BaTiO_3$ and similar compounds as derived from a corrected Lorentz-Lorenz formula (see these abstracts, No. **1833**) is mainly due to a low value of the characteristic frequency ν_e occurring in the expression $c/(\nu_e^2—\nu^2)$.

An estimation of ν_e on the basis of lattice data is given. At the same time a qualitative explanation is given of the high refractive index of compounds containing titanium ions surrounded by six oxygen ions.

1865: H. Rinia: The Schmidt optical system (Bull. Schweiz, Elektrotechn. Verein **40**, 580-585, 1949 No. 17).

In this article the principle of the Schmidt optical system is outlined. It is shown that the conventional form gives fifth order coma for low magnifications. Means are indicated to compensate this coma. The cause and magnitude of the lateral spherical aberration are discussed and a new method to compensate this aberration is shown.

1866*: J. D. Fast: The influence of oxygen, nitrogen and carbon on the impact strength of iron and steel (reprint No. 7, Int. Foundry Congress 1949, Amsterdam).

Description of method of preparation of reproducible samples of pure metals charged with known quantities of oxygen, nitrogen and carbon.

1867: G. H. Jonker and J. H. van Santen: Properties of barium titanate in connection with its crystal structure (Science **109**, 632-635, 1949, June 24).

The behaviour of the permittivity of barium titanate and related compounds as a function of temperature is extensively discussed. (See these abstracts, No. **R92** and Philips tech. Rev. **11**, 176-186, 1949, No. 6.)

1868: W. Hoogenstraaten and F. A. Kröger: The intensity-dependence of the efficiency of fluorescence of willemite phosphors (Physica **15**, 541-556, 1949, No. 5/6).

The efficiency of the fluorescence of zinc silicate activated with a high concentration of manganese is dependent on the temperature and on the exciting intensity when excited by short-wave ultra violet. The efficiency is constant over a wide range of low exciting intensities, increases at intermediate exciting intensities and will probably tend to a constant value again at high exciting intensities. Similar effects are observed with Zn_2SiO_4-Mn_2SiO_4 in which small quantities of Fe_2SiO_4 have been incorporated. This behaviour can be explained on the basis of the theory of hole-migration for bimolecular phosphors of Schön-Klasens, if the roles of positive holes and excited electrons in that theory are interchanged. Values of the activation energies for migration from fluorescence centres and killer centres are computed, and the theoretical temperature-dependence curves calculated from these data for different intensities are shown to be in accordance with the observed curves.

1869: F. A. Kröger and W. Hoogenstraaten: Temperature quenching and decay of fluorescence in zinc-beryllium silicate activated with manganese (Physica **15**, 557-588, 1949, No. 5/6).

Zinc silicate and zinc-beryllium silicate activated by manganese contain various types of fluorescence centres which differ in the wavelength of the emitted radiation, the probability of the fluorescence transition, and the vibrational interaction with the surroundings, which regulates the dissipation of energy of excitation. The centres are assumed to be manganese ions with different numbers of the other manganese ions at neighbouring sites (clusters). Quenching of fluorescence is due to radiation-less processes starting from the excited state of the centres (characteristic quenching). If particular conditions are fulfilled, however, quenching may also occur from (higher) excited states of a different type. In this case energy of excitation is transferred to centres especially suited for the dissipation of energy (quencher centres). This type of quenching is found in zinc silicates at medium manganese concentrations, and in zinc- and zinc-beryllium silicates containing foreign quenchers (iron). It is assumed that quencher centres are fluorescence centres with a low quenching point.

VOL. 11 No. 10, pp. 281-312 APRIL 1950

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 "EXPRESSOR" SYSTEM FOR TRANSMISSION OF MUSIC

by R. VERMEULEN and W. K. WESTMIJZE.

534.86:621.395.665.1

In the transmission — either by radio or via a system of sound-recording — of the musical performance of a large orchestra the following difficulty is encountered. Whereas in the concert hall the differences in sound level may amount to about 80 db, the systems of sound transmission referred to, if equipped for a wide frequency band, cannot cope with any greater differences than, say, 50 db without introducing distortion or background noise arising in the reproduction. "Compression" is therefore commonly applied, which means that the loud passages are less amplified than the soft ones. Provided this compression takes place in the right way, for normal reproduction the result is reasonable, but dynamically it is of course far inferior to the original music. To improve matters in this respect systems of "expansion" have been devised with the object of compensating the compression in the reproduction. The present authors maintain that manual compression by a skilled expert is better than automatic compression. The "Expressor" system ensures that the degree of expansion is unambiguously determined by the degree of compression.

It is a well-known phenomenon that the reproduction of music is accompanied by more or less disturbing noises. Owing to the improvements that have gradually been introduced in the technique of reproduction the imperfections still remaining (among which is the presence of foreign noises) strike the eye — or rather the ear — more and more, with the result that the demands made of reproduction in this respect are becoming more and more severe. The case can be likened to the blacking out of a window: shutting off the large openings makes small cracks all the more noticeable. An example may illustrate what is meant.

One of the most outstanding recent improvements in the field of the transmission of sound [1]) is stereophonic reproduction, whereby, according to the system developed by Philips [2]), the sound from, say, two loudspeakers, each connected to a microphone or a pick-up, is caused to merge into one single "sound picture". The listener has the impression that this sound picture is somewhere between the two loudspeakers. This does not apply however to the background noise (insofar as this comes from the transmitting apparatus), since the noise contributions from the two loudspeakers are independent of each other and consequently the ear localizes them in the direction of each of the loudspeakers. Thus the noise is heard as coming from directions different from that of the music, so that it becomes easier to pay no attention to those noises. It might therefore be expected that with stereophonic reproduction the background noise would be less troublesome than in the case of ordinary reproduction. In conformity with what has been said in the previous paragraph, however, the reverse is often the case: the illusion of hearing original music, to which the stereophonic effect so strongly contributes, is repeatedly broken on account of the background noise reminding one that the music heard is only a reproduction.

As a matter of fact much depends upon the ability and the willingness of the listener to concentrate upon the music and thus not to listen to the

[1]) By transmission is to be understood here not only transmission from one place to another (via telephone lines or by radio) but also the production of a sound record, as for instance a gramophone record, a Philips-Miller tape, etc. for subsequent playing.

[2]) K. de Boer, Stereophonic sound reproduction, Philips Techn. Rev. 5, 107-114, 1940; The formation of stereophonic images, Philips Techn. Rev. 8, 51-56, 1946.

background noise. There are not many listeners prepared to make any effort in this respect, as is evidenced by the position usually occupied by the tone control of radio-sets in the homes of radio listeners. These obviously prefer the loss of the high-pitched notes in the music to the disturbing effect of the background noise. The reason for this lies mainly in the common aversion to the hissing that forms one of the principal components of the background noise. Naturally, when the tone control is turned back the high notes are for the greater part lost. Although technical experts disagree on the question if and in how far the public can appreciate the faithful reproduction of the entire audible frequency range, the suppression of noise at the cost of the high notes of the music is generally considered to be an unacceptable makeshift.

The noise problem occurs in many fields of communication technique, but not everywhere to the same extent. In the case of telegraphy and telephony for instance it is a matter of transmitting the signals economically in such a way that they can be recognized with sufficient certainty in spite of interference. For the transmission of music, however, much more is demanded: the interferences must be at least unobtrusive, if not imperceptible. Lowering the level of the interferences below the auditory threshold is so difficult, in view of the enormous sensitivity of the ear, that it is to be doubted whether any success will ever be attained in that direction. Fortunately, however, there is another property of the ear that tends to act as a counterbalance, and that is the difficulty of distinguishing a weak sound in the presence of loud ones. Thanks to this property of the ear it suffices to attenuate the noise to such a level that it becomes masked, either by the music itself or by other sounds always present both in a concert hall (or studio) and in the auditorium where the music may be reproduced. These noises, apart from those penetrating into the hall from the outside, come from the musicians and the audience itself. This "hall noise", as we may call it, should, objectively speaking, be classified under the disturbing sounds, but, as a "natural" noise it is more acceptable to the listener than the noise of the reproduction system which it helps to mask. In fact it even contributes towards creating the desired atmosphere in the hall where the music is being reproduced.

Level ranges

What is the range of intensity levels possible between the hall noise and the loudest music on the one hand and the range between the noise of the transmission apparatus itself and the maximum signal it can handle on the other hand?

The highest peaks occurring in the sound produced by a large orchestra reach close to the "feeling level" of the human ear. This is borne out by measurements by American investigators [3]), who found that an orchestra of 75 musicians may produce in the peaks a power of about 70 W, which in the average concert hall corresponds to a level of about 110 db above the usual zero level of 10^{-12}W/m^2.

Measurements of the intensity of audience noise taken in cinemas [4]) show that the sounds penetrating from the outside into an empty hall reach an intensity of 25 db above the zero level, while in halls where an audience is present the level of the noise varies between 38 and 44 db (average 42 db), dropping to 32 db during particularly thrilling parts of the film. The latter figure will not be far off the level to be reckoned with in a concert hall. Passages of music below this level will be drowned in the hall noise. Thus the difference between the highest and the lowest sound level in a concert hall — the "range" — amounts to about 80 db.

In the transmission of sound one has to take into account, on the one hand, the danger of non-linear distortion which threatens to arise at the high peaks in the intensity of the sound, and on the other hand noises inherent in the transmitting system. In the case of the Philips-Miller tape [5]), to which the following considerations will be mainly confined, a difference of at most 50 db can be reckoned with between the noise level and the level at which this non-linear distortion becomes noticeable. (The former also depends on the width of the frequency spectrum, which is taken here as being 8000 c/s.)

As a rule the intensity range afforded by the transmitting system is not wide enough to cope with the intensity range of the original music; in our case it is 30 db too short.

Levels determining the intensity range in transmitting systems

Before discussing the measures that have to be taken to meet this discrepancy we have to go further into the factors determining the intensity range of a system with optical scanning, like that employed for sound films. Let us first consider the lower limit, which is determined by the background noise.

[3]) L. J. Sivian, H. K. Dunn and S. D. White, Absolute amplitudes and spectra of certain musical instruments and orchestras, J. Acoust. Soc. Amer. 2, 330-371, 1931.

[4]) W. A. Mueller, J. Soc. Mot. Pict. Engrs. 35, 48-53, 1940.

[5]) R. Vermeulen, The Philips-Miller system of sound recording, Philips Techn. Rev. 1, 107-114, 1936.

Between the original and the reproduced sound there are the following links: microphone - amplifier - sound recorder - sound tape - light ray - photocell - amplifier - loudspeaker, each of which contributes its share in the total noise. These sources of sound are incoherent, so that the powers of each noise have to be added and not the amplitudes. The contribution of a source with an amplitude say half of that of another source is thus only 1 db. Therefore, the strongest source of noise practically determines the total noise level.

When modern microphones are used in a suitable circuit and the amplifiers are of a carefully designed construction these elements are certainly not the main source of the background noise, and neither are the loudspeakers.

The photocell calls for rather more attention. The noise of the photocell results from the fact that the electrical signal is the integrated effect of many electrical particles (electrons) and as such this noise is proportional to the square root of the number of particles passing through the cell per second, and is thus proportional to the square root of the average value of the photocell current. This average current is limited at the lower end by the fact that it must always be at least equal to the peaks with which the current is modulated by the sound track of the film. There are in fact systems for counteracting noise (commonly termed "noiseless") whereby the average photocell current is varied so as not to be greater at any time than what is required for the modulation, the photocell noise thus being minimized. Any further improvement could only be obtained if it were possible to increase the illumination of the track. If this illumination could be doubled, for instance by using a lamp with greater brightness or by employing a better optical system, the signal amplitude at the output of the photocell would likewise be doubled but the noise would only increase by a factor $\sqrt{2}$, so that the intensity range between the maximum signal and noise would become greater by a factor $\sqrt{2}$, i.e. 3 db. With the apparatus commonly employed however any improvement of this nature is hardly to be expected, nor is it urgently needed.

The light, too, has a "granular" structure (photons), so that the beam of light striking the photocell forms in itself a source of noise. Since, however, on an average about 200 photons are required to release one electron from the usual photo-cathode, the structure of the electron current flowing through the photocell is much coarser than that of the beam of light, and the noise contribution of the latter may be ignored.

Finally there is the noise of the sound track. This is stronger than that of any of the sources summed up in the foregoing, even in the case of a freshly cut Philips-Miller tape, although, in comparison with other systems for optical scanning, its inherent noise is very low. This low noise level of the Philips-Miller tape results from the fact that the cutter entirely removes the opaque coating from the sound track (see *fig. 1a*), while outside the track the amount of light passing through is quite negligible (in contrast with photographic films, where the grained structure of the emulsion gives rise to noise; see for instance figs. 8 and 12 in the article quoted in footnote [5]). The main source of the low tape noise lies in the narrow wedge-shaped edges of the track (A in fig. 1b), where some fluctuation is possible in the density. To this are to be added the irregularities in the cutting process itself. When it is borne in mind that the edges of the track are formed by the cutter tearing away the material it is really surprising that it is possible to get the excellent definition of the edges of the track that is needed for this purpose.

More serious is the noise that is to be ascribed to quite commonplace causes but which often exceeds all other contributions, namely the noise due to particles of dust and scratches or indentations on the surface of the film and to impurities in the material of the tape. Specks of dust and impurities intercept, and scratches scatter light, all this resulting in a kind of noise that could best be called "sputtering".

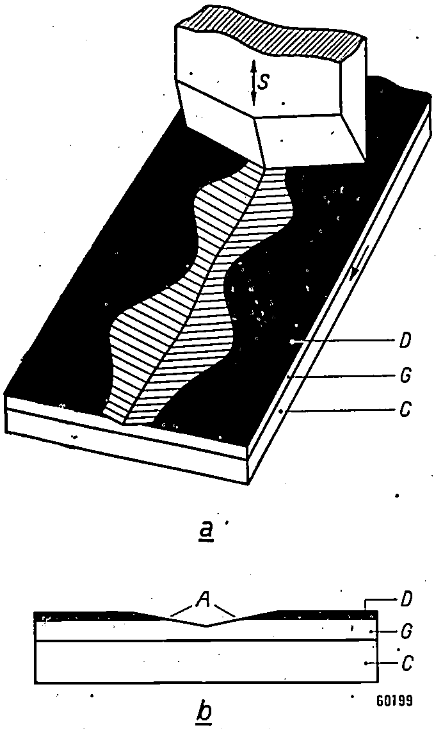


Fig. 1. a) The recording of sound on a Philips-Miller tape. This tape consists of a strip of celluloid C, on which a layer of transparent gelatine G and a thin opaque coating D are applied. The tape moves along underneath the cutter S in the direction indicated by the arrow, the cutter moving up and down, thereby cutting a sound track in the opaque coating. b) Enlarged cross section of a Philips-Miller tape; A = edges of the sound track; the other letters have the same meaning as in (a).

Enlarging the intensity range by raising the upper limit is rendered difficult by a number of factors, some of which are of a fundamental nature while others are of a more practical nature. The Philips-Miller recorder could be made to yield larger amplitudes without distortion by giving it a lower resonant frequency. (To avoid a rise in noise owing to the track then necessarily being wider, the above-mentioned "noiseless" system could be applied.) This greater amplitude, however, can only be obtained at the cost of the high-frequency response [6]).

We would mention here another method, likewise not without its drawbacks, which would in theory enable the intensity range to be enlarged. When the sound spectrum is imagined as being divided into a number of frequency

[6]) With the Philips-Miller recorder developed for stereophonic recording (two tracks on one tape) the opposite course has been followed: in order to get a flatter frequency characteristic a narrower track is used. The two stereophonic tracks together take no more room than the former single track.

bands it is found that the large amplitudes do not usually occur in all bands. Given this distribution (which is known for most musical instruments and for speech in different languages), the recording installation can be given such a frequency characteristic that the bands in which normally only small peaks occur are amplified more than the others, so as to make it equally probable for the peaks in each band to reach the limits of the sound track. To obtain a reproduction with a flat overall characteristic, this difference in amplification must be compensated by giving the reproducing amplifiers the inverted frequency response of the recording apparatus. This means that the frequency bands with the small signals which were given extra emphasis in recording are now correspondingly attenuated.

The result is then a similar attenuation of the noise contribution of those bands. Obviously this measure can only be successful if in certain bands there is only a weak "signal" (weak components of the music or speech) — in relation to the maximum amplitudes — and much noise. In favourable circumstances (orchestral music) the intensity range can in this way be widened by 10 db [7]), but for speech, at least in strongly sibilant languages, this method does not hold. Another drawback is that owing to its special frequency characteristic the reproducing apparatus is not suitable for those forms of recordings where this method has not been applied, and vice versa.

Compression and expansion

The discrepancy between the intensity range in orchestral music (80 db) and the intensity range of the transmitting apparatus (50 db) leads to the following difficulty. When it is so arranged that the noise level of the apparatus coincides with the level of the hall noise, i.e. 30 db, (respectively C_1 and A in *fig. 2*), all peaks in the music ascending higher than 80 db above the zero level are overmodulated. When the limit for the maximum permissible level of the apparatus is made to coincide with the highest sound level (respectively D_2 and B in fig. 2) no distortion will occur, but the noise level (C_2) becomes so high as to drown all music not reaching 60 db above zero level.

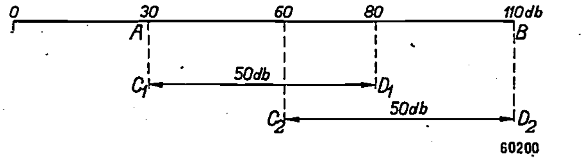


Fig. 2. A = level of the hall noise (30 db), B = level of the highest peaks in orchestral music (110 db) above the zero level 0 (10^{-12} W/m²). If the 50 db range of an installation for sound transmission is laid along C_1D_1 then distortion arises at all peaks higher than 80 db. If the range is laid along C_2D_2 then the noise level (60 db) will be high.

The remedy usually followed to overcome this difficulty consists in reducing the intensity range of the music to 50 db by what is known as "compression" — for instance by means of a volume control operated either by hand or automatically — thus attenuating the loud passages compared with the soft ones [8]). These reduced differences in level are familiar to everyone who listens to radio and gramophone music.

Musicians quite rightly object to their work being treated, or rather ill-treated, in this way. Although "compressed" music may be acceptable provided the compression is carried out by a skilled expert, different means have been sought to restore the original contrasts in the reproduction, so long as there is no alternative to compression. These means are all based upon the application of expansion, restoring the original relations in the music in some way or other by giving more emphasis to the high levels, thereby compensating the compression. The expandor used for this purpose may be operated automatically, for instance by causing it to respond to the amplitude of the signal fed to it [9]).

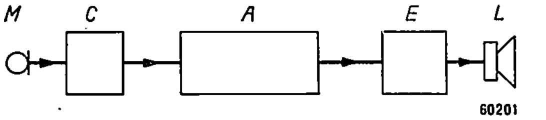


Fig. 3. Diagrammatic represention of an installation for sound transmission. M microphone, C compressor, A complex of amplifiers and any recording and reproduction apparatus that may be used, E expandor, L loudspeaker.

The chain between microphone and loudspeaker is then as represented in *fig. 3*, where A is a complex of amplifiers and possibly a recording and scanning device. An important point, as we shall presently see, is the fact that the main sources of noise are contained in A, that is between the compressor (C) and the expandor (E).

Shape of the compression characteristic

For the moment we shall defer a consideration of the question whether compression can be carried out better by hand or automatically, and assume for the time being that it is done automatically, according to a fixed relation between the degree of attenuation C, introduced by the compressor, and the intensity I_1 of the original music. It will be taken

[7]) J. C. Steinberg, The stereophonic sound-film system; pre- and post-equalization of Compandor systems, J. Soc. Mot. Pict. Engrs. **37**, 366-379, 1941, in particular p. 374.

[8]) See, e.g., R. Vermeulen, The relationship between fortissimo and pianissimo, Philips Techn. Rev. **2**, 266-269, 1937.

[9]) See, e.g. V. Cohen Henriquez, Compression and expansion in sound transmission, Philips Techn. Rev. **3**, 204-210, 1938.

for granted that when I_1 rises from 30 to 110 db the attenuation must increase from 0 to 30 db [10]). Thus in *fig. 4a* the initial point A and the final point B of the compression characteristic are fixed.

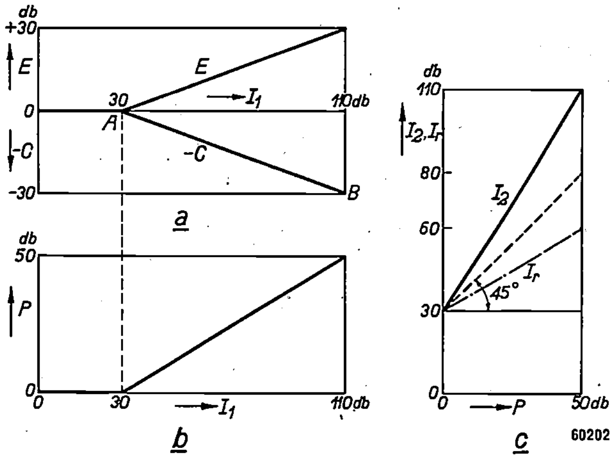


Fig. 4. *a*) The compression C and the expansion E plotted as functions of the intensity I_1 of the sound in the concert hall. It is assumed that the compression increases linearly according to AB when I_1 is increased from 30 to 110 db. To restore the original dynamic contrast it is necessary that $E—C = 0$. *b*) The amplitude P of a sound record made with the compression characteristic according to (*a*) plotted as a function of I_1. *c*) Plotted as functions of P: the intensity I_2 of the reproduction (dotted line for the case where expansion is not applied, full line for the case where expansion is applied), and the intensity I_r (dot-dash line) of the noise level (with expansion).

The shape of the characteristic between A and B is to a certain extent arbitrary. The simplest characteristic is the straight line AB (fig. 4*a*). The amplitude P of the recorded signal [11]) then assumes a shape, as a function of I_1, as represented in fig. 4*b*, whereby the range of 80 db in I_1 is reduced to 50 db in P. If no expansion is applied in the reproduction then the intensity I_2 of the reproduced music, as a function of P, follows the dotted line in fig. 4*c*, with the same range (50 db) as P. By applying expansion the level of the reproduced sound I_2 can be given a range equal to that of I_1 (full line in fig. 4*c*); the expansion characteristic $E = f(I_1)$, fig. 4*a*, must then be the inverted image of the compression characteristic (E-C = 0 for every value of I_1).

As already pointed out, the main sources of noise are located between the compressor and the expandor (in A, fig. 3). The noise level I_r of the reproduction therefore follows the movements of the expandor (see the dot-dash line in fig. 4*c*) and as these have to follow the movements of the compressor (in order to maintain the relation between the two characteristics, $E—C = 0$) the noise in the reproduction therefore varies with the intensity of the music. This is most unpleasant, for, in spite of the masking effect of the correspondingly varying intensity of the music, a varying noise attracts attention much more than a constant noise, from which one can detach one's mind even if it is rather strong. Hence the question arises whether the compression characteristic represented in fig. 4*a* is the most favourable, since according to that characteristic compression is already applied (and noise is thereby emphasized and varied) as soon as I_1 rises above the level of 30 db, where there is no danger of overloading whatever. It is therefore obvious that it would be more favourable if the compression were made to start at a higher level of I_1, in fact at such a level that the noise is already masked by the music, for instance at 70 db, as indicated in *fig. 5a*. The corresponding variations of the characteristics $P = f(I_1)$ and $I_2 = f(P)$ (figs 5*b* and *c*) need no explanation. From fig. 5*c* it is seen that the noise level is now constant over a large range of P.

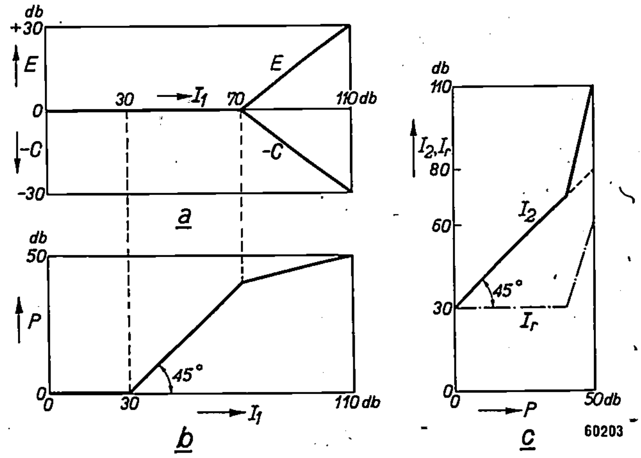


Fig. 5. The same as fig. 4 but with a compression characteristic starting at $I_1 = 70$ db.

Is it possible to go still farther in this direction? Strictly speaking, there is no need for compression so long as there is no danger of overmodulation, i.e. before P reaches 50 db, corresponding to $I_1 = 80$ db above zero level. The characteristics then assume the shape given in *figs 6a, b* and *c*. A difficulty then arises, however, in that, in the vertical part of the characteristic in fig. 6*c*, I_2 is then no longer a single-valued function of P. In other words,

[10]) This figure of 30 db agrees with what H. Fletcher gives for variable-area sound-film records (J. Soc. Mot. Pict. Engrs. **37**, 331-352, 1941, in particular p. 338). Some of the ideas developed here have been taken from this article by Fletcher.

[11]) Where no record is made, as in the simultaneous reproduction of music elsewhere, P may be taken as representing the signal at the input of the expandor.

contrary to the cases of figs 4 and 5, the expandor cannot of itself deduce from the amplitude of the input signal P what the expansion has to be. A solution for this problem will be indicated presently.

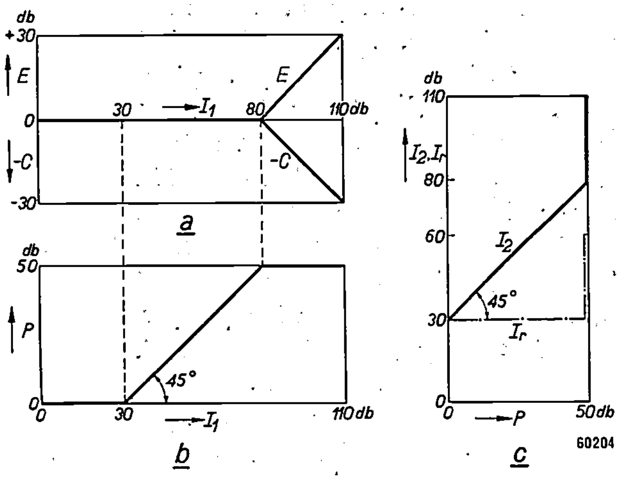


Fig. 6. The same as fig. 4 but with compression starting at $I_1 = 80$ db, thus at the level at which, without compression, distortion would arise. At higher values of I_1 there is no longer any unambiguous relation between I_2 and I_r on the one hand and I_1 on the other hand.

Automatic or hand-regulated compression?

It is impossible for the relation between the compression and the level of the music to be so simple as represented in the foregoing. A relation between momentary values, such as has been roughly assumed, can certainly be obtained but this would lead to strange and undesirable effects, because it would in essence mean a strong non-linear distortion.

A compression characteristic relating to momentary values can be realized, for instance, by feeding the microphone voltage to the control grid of a pentode having a characteristic of the shape $i_a = c_1 v_g{}^n$ (i_a = anode current, v_g = control-grid voltage, c_1 a constant, n a number between 0 and 1). The non-linear distortion then obviously occurring can be eliminated, at least in theory, by using for the expansion a valve having a characteristic of the shape $i_a = c_2 v_g{}^{1/n}$ (where c_2 is again constant). It is then essential, however, that the factors n in the two exponents have exactly the same value. Another requirement, just as difficult to meet, is that no phase shifts may occur between the components of the signal on its way from the compressor to the expandor. In practice such a system would result in strong distortion being heard not only in the highest levels (as would be the case without any compression) but almost continuously.

To avoid this distortion the variations in the attenuator can be made to take place so gradually as to allow the system always to be quasi-stationary with respect to the sound waves, but the compressor would then be unable to fulfil its task properly. This task is on the one hand to avoid overloading and on the other hand to avoid too high a noise level. If, however, the compression has to lag somewhat behind the variations in intensity of the music, then upon the sudden starting of a fortissimo passage there is bound to be a clipping, and an abrupt pause of the music will be followed by noise dying away. In the latter case, it is true, owing to reverberation the sound in the concert hall will die away gradually and thus more or less mask the noise, but, nevertheless, variations in noise level are likely to be noticed. The clipping at the sudden rise of the level is even more annoying.

Various methods have been suggested for avoiding this distortion arising when an automatic compressor is used. In the first place the compression can be made to start at a slightly lower level than where overmodulation occurs, for instance as indicated in fig. 5*d*, thus leaving some margin for the first peaks of a sudden fortissimo; this margin, which would have to be about 6 db, could only be obtained, however, at the cost of the intensity range of the recording, because normally it cannot be used.

Other suggestions are based upon the idea of bringing the automatic compressor into action just before the first peak of the fortissimo reaches the recording apparatus. This idea can be realized by placing between the orchestra and the microphone an auxiliary microphone for picking up the sound earlier and operating the compressor before the same sound has reached the main microphone. The objection against this solution, however, is that it sets intolerable limitations to the choice of the microphone positions. Moreover, the interval of time that can be obtained in this way (or alternatively, for instance, with a delaying network between the compressor and the recording apparatus) is at most of the order of a few milliseconds. If clipping of the peaks is to be effectively avoided the compressor would have to act so rapidly that no quasi-stationary state could possibly be reached and, therefore, distortion at low levels would replace the clipping.

A third method, frequently applied with automatic compressors, consists in causing the compressor to respond as quickly as is possible (limited for other reasons) to an increase of the sound level but to lag behind a decrease in sound level. The first peaks of a fortissimo passage then still cause clipping, but this is only of short duration, owing to the now almost instantaneous response of the compressor. Any delay in the

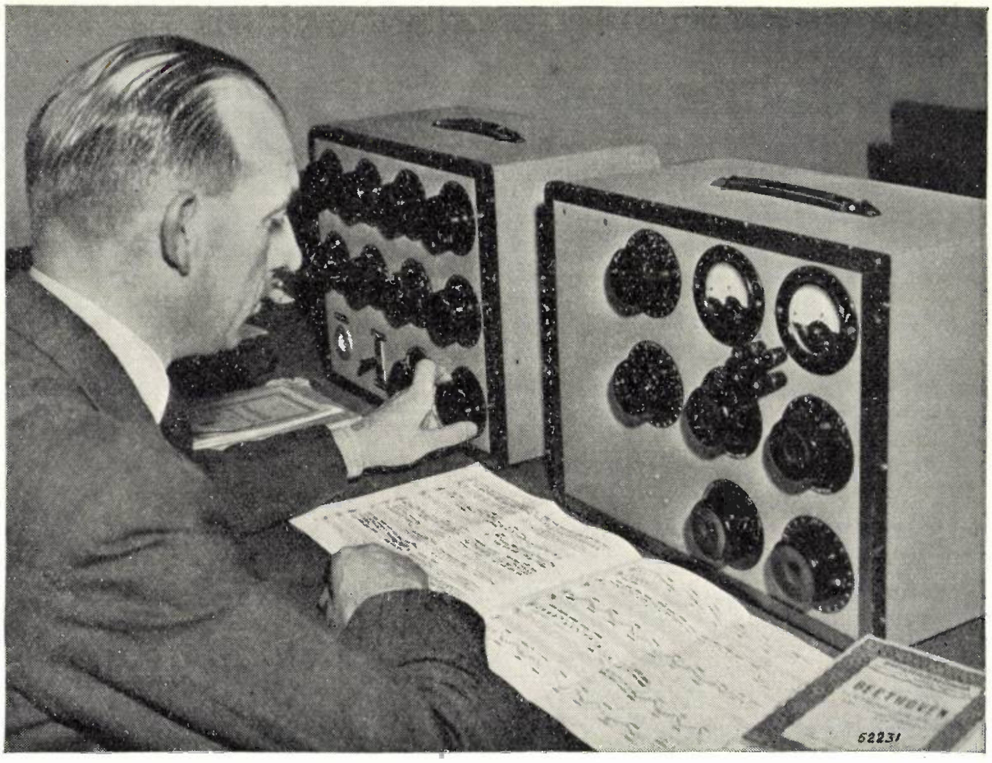


Fig. 7. The "music pilot" at work. He regulates the compression guided by the music score in front of him. By means of the other controls seen in the photograph the signals from the microphones set up in the hall can be mixed in the right proportions.

compensation by the expandor also lasts too short to be noticeable. The time constant of the delay at a sudden drop of the level has to be so chosen that the "tailing on" of noise is the least troublesome. In this way a compromise can be reached where an untrained ear scarcely notices the momentary distortion or the varying noise.

In our opinion, however, such an automatic system avoids the issue and it is by far preferable for the compressor to be operated by hand, but then by a capable hand! No mechanism has the power of prediction that is indispensable if it is to take action in good time and in the right way before a peak arises. Only one trained in music has the foreknowledge — especially when he has the score in front of him — which, together with a good dose of routine, is necessary to be able to reduce the amplification gradually and in good time before a fortissimo passage starts, and to increase the amplification in the right way upon the termination of a loud passage. To reach the ideal solution, sense and an ear for music are indispensable [12]).

In the majority of cases compression is in fact regulated by hand (see *fig.* 7). When listening to radio, gramophone or film music anyone can judge how satisfactory the result can be, provided the operator is skilled in his job.

The "Expressor" system

Once it has been decided that the compressor is to be operated by hand there is no longer any question of a certain fixed relation existing between the degree of compression and the level of the original music as was supposed to exist in figs. 4*a*, 5*a* and 6*a*; the personal views and the skill of whoever may be operating the compressor play a part. To put it in other words, the amplitude P of the sound record is no longer a single-valued function of the sound level I_1. If expansion is to be applied then a difficulty arises similar to that encountered in the case of fig. 6*c*, where I_2 is no longer everywhere a single-valued function of P — likewise arising from the fact that I_1 is not everywhere a single-valued function of P (see fig. 6*b*). This difficulty lay in the fact that as soon as these fixed relations are broken the expandor can no longer deduce from the signal P what level I_1 actually had (and thus what the level I_2 should be). A solution of this difficulty is the following: the position of the compressor at any moment can be communicated to the expandor via a separate

[12]) The experiments which will be mentioned presently were made possible by the devoted and capable cooperation of Mr. M. J. C. van der Meulen (Electro-Acoustical Dept., Eindhoven).

transmission channel and the expandor can be made to compensate the compression continuously. In more technical terms, through the intermediary of a pilot signal the compressor has to control the expandor in such a way that $E—C$ is always equal to zero.

We have given this system the name of "Expressor", which is made up from parts of the words *ex*pandor and com*pressor* and at the same time signifies the greater possibilities of expression which can be obtained by this system in the reproduced music.

When listening to the monitoring loudspeaker one will not notice in the music any influence whatever of the adjustment of the control of the "Expressor", apart from the fact that turning the knob in one direction results in an increasing distortion, while the noise level steadily increases in the other direction. The readings of the meters on the control panel indicate how the level of the output signal of the compressor is varied and how it should be adjusted.

There are two points of view from which one can consider the operation of the "Expressor". The simplest way is to disregard the expansion and to pilot the signal in the usual way between noise and distortion. In the case of reproduction without expansion the result will have the same dynamic character as ordinary radio and gramophone music; with expansion the original dynamic nuances are fully restored. Operated in this way, however, the inherent possibilities of the "Expressor" are not utilized to the full. This is only the case when the noise level is continuously kept as low as possible, that is to say, when the sound track is continuously modulated close up to the limit of distortion. This imposes a heavier task upon the "music pilot" operating the compressor, for he has to be on the alert all the time and can scarcely do without the music score in front of him, or even a rehearsal, if he is to take action in good time for every crescendo. The reward for such painstaking attention is a reproduction which only exceptionally reminds one of the technical element involved. Particularly striking is the absence of noise during the "rests" and the pianissimo passages. Only in the fortissimo passages, if one concentrates on it sufficiently, can any foreign noise be detected. This is just the opposite of what one has become accustomed to in the reproduction of music, and in our opinion it means an appreciable improvement.

For safety's sake an automatic auxiliary compressor might be added to come into action only when there is a risk of overloading. This seems to be a wise precaution, but we have not adopted this measure because it was feared that the "music pilot", relying on this automatic compressor, would not be induced to concentrate to the full on his exacting work. It has been found that if the pilot takes his work seriously there is no need for any such precaution.

Technical execution

It now remains to discuss the method chosen for controlling the expandor by the position of the compressor.

The music pilot operates a potentiometer controlling the degree of compression. The expandor consists of another potentiometer which has to respond to every change in the compression to a like degree in the opposite sense. (Of course for stereophonic transmission two compression and two expansion potentiometers are needed.) The problem of effecting this coupling (in space and time!) between the compression and the expansion potentiometers is possible of solution in various ways. Here only a brief description will be given of an impulse system which was applied for the experiments carried out in the Concert Hall in Amsterdam in 1947 [13]).

With stereophonic transmission the compression

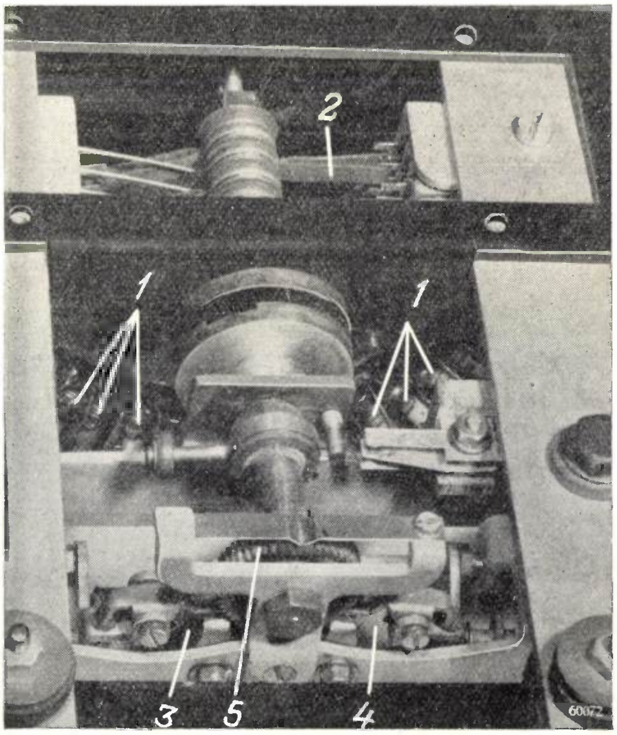


Fig. 8. Mechanism of the expandor. At *1* are to be seen some of the resistors which form a potentiometer and are connected to a series of contacts traversed by a contact brush *2*. This is turned in one direction or the other by the action of the armature of the electromagnet *3* or *4* causing the ratchet wheel *5* to turn one or more steps.

[13]) R. Vermeulen, Duplication of concerts, Philips Techn. Rev. **10**, 169-177, 1948 (No. 6).

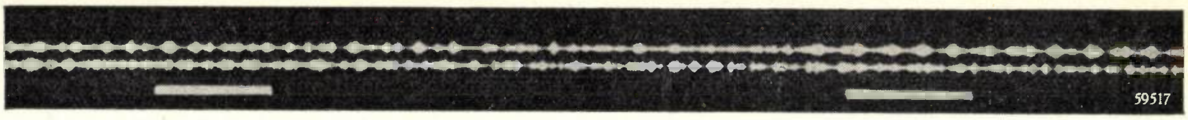


Fig. 9. Illustration (twice the actual size) of a strip of Philips-Miller tape with two stereophonic sound tracks. Underneath these tracks are marks (pilot track) made as the result of the compression potentiometer having been moved two steps in a certain direction. When scanned during the play-back these marks cause the expansion potentiometer to be rotated the same number of steps in the opposite direction. Above the sound tracks similar marks can be made for turning the expandor potentiometer in the other direction.

has to be adjusted in the two channels in exactly the same way by means of one single knob (*G* in fig. 1 in the article just quoted), so as to avoid any displacement of the stereophonic sound picture. Consequently the potentiometers used have to be of a very high quality. The most reliable potentiometers are variable in steps with wire-wound resistors. If the steps are no more than 1 to 2 db the ear does not perceive the discontinuities in the adjustment,

The compression potentiometers are fitted with an auxiliary contact which at every step produces an impulse that is transmitted via the pilot channel to the expandors, where it activates an electromagnetically operated ratchet gear (*fig. 8*) which adjusts the expansion potentiometers in the desired

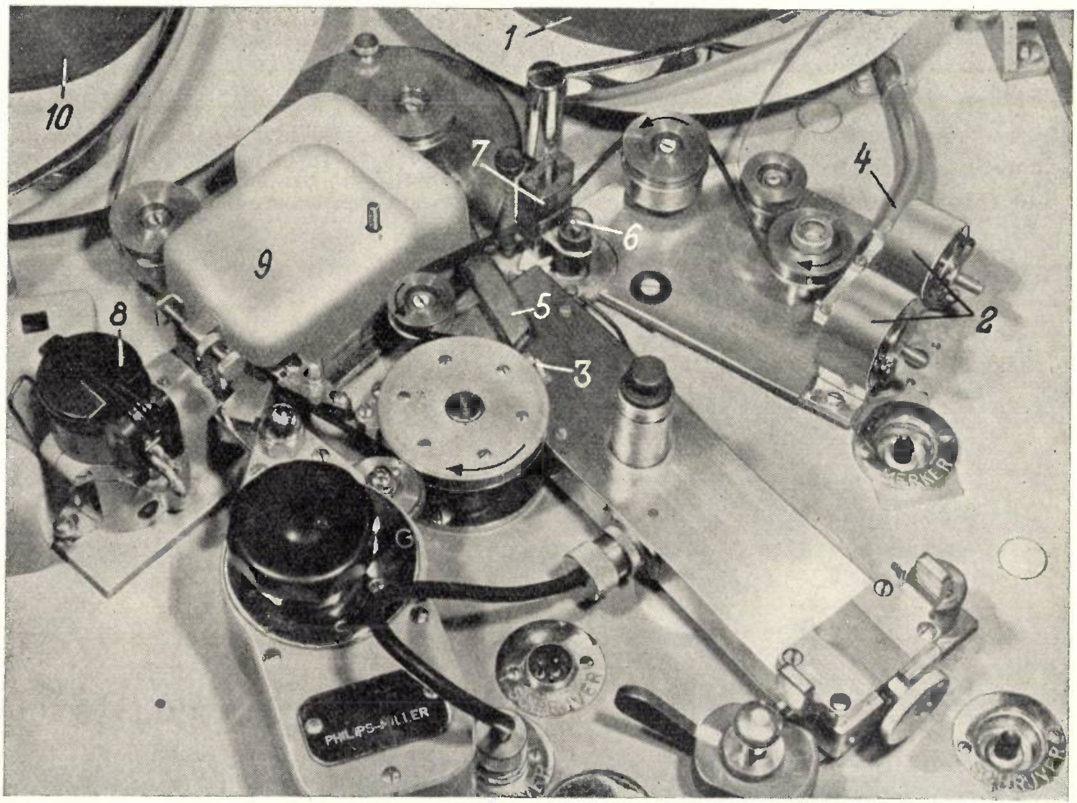


Fig. 10. The principal parts of a Philips-Miller machine with "Expressor" system. This machine is suitable both for recording and for play-back. For recording, *1* is a roll of blank Philips-Miller tape, passing along in front of the markers *2*, which cut the pilot track indicating how many steps and in which direction the compressor has been moved. It then passes by the cutters, only one of which (*3*) is to be seen here; the other, which should be on top of *3*, has been removed so as not to obscure the first one. At *4* and at *5* are suction tubes for the swarf (the material cut out of the tape). The tape is rewound at *10*. If so desired, the music can be played back while the recording is still proceeding. The tape is then passed — as shown in the picture — along the optical scanning devices, one for the pilot tracks and the other for the sound tracks. The exciter-lamp of the former scanning device is to be seen at *6*; the photocells (selenium cells) are at *7*. The exciter-lamp for the sound-track is at *8* and the two photocells are at *9*. For the expandor to be brought into action at exactly the right moments the length of the tape between *2* and *3* must be equal to that between *6-7* and *8-9*. When a previously made tape record is to be played then of course the tape has only to pass along *6-7* and *8-9*.

direction. Naturally the impulses corresponding to one direction of rotation of the compression knob have to be distinguished, or kept separate, from those for the other direction of rotation.

This can be done, in the case of simultaneous reproduction, for instance by modulating each of the two series of impulses on a carrier of its own and transmitting these carriers to the expandor via a telephone line or by radio. For recording on Philips-Miller tape the space at the sides of the sound-track offer room for two "pilot tracks", each of them containing a series of impulses (*fig. 9*). In *fig. 10* the cutting units ("markers") for these two pilot tracks are to be seen at *2*, and their optical scanners at *6-7*.

By choosing impulses for the pilot signal its exact amplitude becomes of no consequence and spurious voltages in the pilot channel will not readily give rise to errors. The only requirement that has to be fulfilled is that the contrast between the impulses and spurious voltages is sufficient and that successive impulses do not overlap.

When the steps of the compression and expansion potentiometers are logarithmic it is not necessary for these potentiometers to be in corresponding positions, but this is nevertheless desirable because otherwise the range covered by the two together is reduced. To be able to bring them easily into corresponding positions the expandor potentiometers are provided with a stop in the extreme positions; all one need do, therefore, is to turn the compressor knob from one extreme position to the other just before the transmission of the music begins, to obtain the desired correspondence in position of the potentiometers.

Summary. In systems for sound transmission with a wide frequency band the difference between the level at which overloading starts and the level of background noise is less than the difference that may occur in orchestral music between the loudest passages and the noise in the concert hall, so that one is faced with the necessity of applying compression in order to avoid both overloading and high background noise. In the reproduction of the music the original dynamic differences can be regained by means of expansion. — It is claimed that adjustment of the compression by a capable hand has great advantages over automatic control. With non-automatically adjusted compression however there is no unambiguous relation between the intensity of the original music and that of the input signal of the expandor, and consequently special measures have to be taken to ensure that the expandor always exactly compensates the compression. In the "Expressor" system a pilot signal causes the potentiometer of the expandor to follow continuously the movements of the compressor. This pilot signal is transmitted via a separate channel and may consist, for instance, of impulses. If Philips-Miller tape is used as a sound record the pilot signal can be registered on the same tape in two pilot tracks (one for each direction of rotation of the compressor).

PHANTOM TESTS WITH X-RAYS

by G. C. E. BURGER.

778.33:621.386.1:
616-073.75:771.534.5

The medical practitioner wishes a lung radiograph to be capable of revealing details which are hardly perceptible. If he does not see a certain suspected detail does this then mean that that detail is non-existent, or is it that the X-ray picture is unable to reveal it? For a study of the general problem of perceptibility of details in X-ray technique good use can be made of specially constructed phantoms.

In the formation of an X-ray image there are a large number of factors playing a part. Besides the absorption and the scattering of the X-rays in the object there are the properties of the photographic film or, in fluoroscopy, those of the fluorescent screen and, furthermore, also the properties of the optical system (camera) if such is used.

The quality of the image can be judged according to its two main characteristics, definition and contrast of details. Contrast is defined by the ratio of the brightnesses of the detail and the background. A small detail is more readily observed when it stands out clearly against the background, that is when the contrast is great, than when such is not the case. Definition of details depends upon the X-ray tube used (dimensions of the focus) and, to a large extent, upon the fineness of grain of the photographic film or the fluorescent screen. Contrast is mainly determined by the physical properties of the object itself, by the voltage on the X-ray tube and by the gradation of the film.

In this paper attention will be specially directed to medical radiology and in particular to its application in the X-raying of the lungs. In practice different methods are followed for radiographing the lungs, the most important of which are:

1) contact radiography, whereby the X-rays, after passing through the object, impinge directly on the film, which may or may not be provided with one or more intensifying screens;
2) fluoroscopy, whereby the X-rays impinge on a fluorescent screen and the image thus formed is viewed visually;
3) fluorography; a camera is set up behind the fluorescent screen and the image is photographically recorded on a film on a reduced scale.

With all three methods it is possible to obtain a certain amount of magnification by placing the object at some distance in front of the screen or film.

It is difficult to make a quantitative comparison between these methods and between the various ways in which one particular method may be carried out, owing to the previously mentioned large number of factors playing a part in the formation of the X-ray image. The introduction of the phantom in X-ray technique is a great improvement in objective testing. In this article it will be explained how a comparison between the various methods can be arranged on a quantitative basis with the aid of phantoms.

Description of a phantom. Contrast-detail diagram

In its simplest form a phantom consists of a plate of some kind of material the thickness of which varies in a known manner. By radiographing this plate the value of the radiographic method used can to a certain extent be determined from the differences in brightness of the image. Tests with such a phantom have been carried out by Bronkhorst [1]) and Luft [2]), among others.

It is not possible, however, to determine with such a phantom any relation between the visibility on the one hand and the contrast and definition of detail on the other, because the phantom does not possess any small details. Phantoms having small details have already been commented in this journal [3])[4]). Here a description will be given of the phantom that has been used for the investigation dealt with in this paper and which in principle resembles the phantom that was used for the article quoted in footnote [4]).

The phantom consists of a number of plates of

[1]) W. Bronkhorst, Kontrast und Schärfe im Röntgenbild, published by Thieme, Leipzig, 1927.

[2]) F. Luft, Experimentelle Beiträge zur Detailerkennbarkeit und Detaildarstellbarkeit bei verschiedener Aufnahmetechnik, Fortschr. Röntgenstrahlen 51, 412-417, 1935.

[3]) B. van Dijk, Several problems of X-ray fluoroscopy, Philips Techn. Rev. 4, 114-117, 1939.

[4]) G. C. E. Burger, B. Combée and J. H. van der Tuuk, X-ray fluoroscopy with enlarged image, Philips Techn. Rev. 8, 321-329, 1946.

"Philite", one of which serves as object plate (*fig. 1*) and has 225 holes drilled in it. The holes in each horizontal row are of the same diameter but vary in depth between 8 mm and 1 mm in 15 steps of 0.5 mm; those in each vertical column are equal in depth but vary in diameter, likewise between 8 mm and 1 mm in steps of 0.5 mm. These holes serve as details, the visibility of which is investigated in the X-ray image. The other "Philite" plates can be inserted in or taken out of the phantom as desired for imitating the thickness of the thorax, the aim being to collect data about the methods applied in practice for pulmonary diagnostics. A constructional diagram of the phantom is given in *fig. 2*. To make it as near as possible comparable to an actual radiograph of the lungs the phantom can be placed in an apparatus like that depicted in *fig. 3*, with which the movements resulting from the heart beat of the patient can be imitated; the phantom is moved by means of a pendulum and the rate of movement can be adjusted by varying the maximum deflection. A small lead plate that can be attached to the phantom makes it possible to measure exactly the motional unsharpness and thus the exposure time.

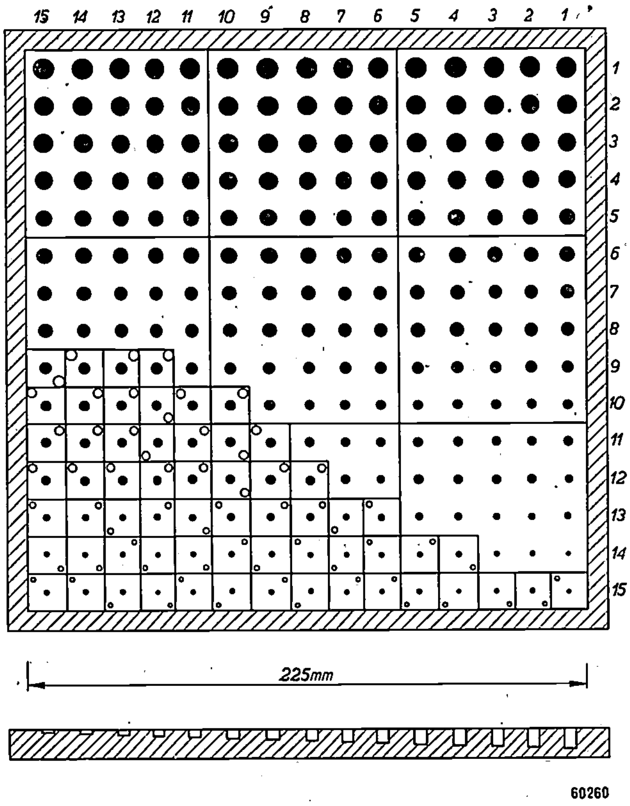


Fig. 1. Object plate for contrast-detail diagram. The holes in each vertical column are of the same depth but vary in diameter; those in a horizontal row have the same diameter but vary in depth. Near the small holes additional holes have been drilled which are indicated by circles.

"Philite" has been chosen as the material for the phantom because it most closely resembles the lung tissue as regards its properties determining the attenuation of X-radiation. The following *table* gives a comparison between some of these properties for "Philite" and for water:

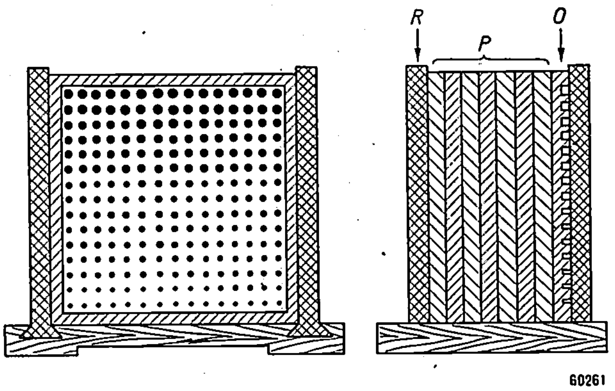


Fig. 2. Constructional diagram of the phantom. It comprises the object plate *O* and a number of plates *P* without details; all plates are of "Philite". The frame *R* can be moved to take a varying number of plates.

	"Philite"	Water
Average atomic number	6.9	6.0
Density	1.3 g/cm³	1.08 g/cm³
Ratio of transmission at 100 kV and at 50 kV	1.34	1.33
Mass absorption coefficient	2.3 λ^3 cm²/g	2.6 λ^3 cm²/g
Mass scatter coefficient	0.18 cm²/g	0.19 cm²/g

Thus it is seen that "Philite" does indeed correspond well to the human tissue, which for 70% consists of water. Moreover, it is an easily workable material.

A fluorogram of the phantom produced on a 45 mm film with an apparatus for mass chest radiography is reproduced in *fig. 4*. From this it is fairly simple to judge the effect of contrast and size of detail upon the visibility of the details, both separately and together.

Roughly speaking there are two areas to be distinguished in this radiograph. In the first area all the holes are visible, whereas in the other it is impossible to discern any detail at all. There is no sharply defined transition between these two areas. In order to determine the boundary as well as possible, in the part of the object plate where the holes are small in diameter and shallow in depth additional holes have been drilled (indicated in fig. 1 by circles); these have exactly the same dimensions as the corresponding holes situated in the centre of the respective squares marked off on the phantom with lead oxide. The extra hole may be in any one of four possible positions in its square, and in order to decide whether a certain hole is visible or not all the observer has to do is to indicate the position in which he thinks the extra holc is situated. In this way the border line can be determined fairly sharply.

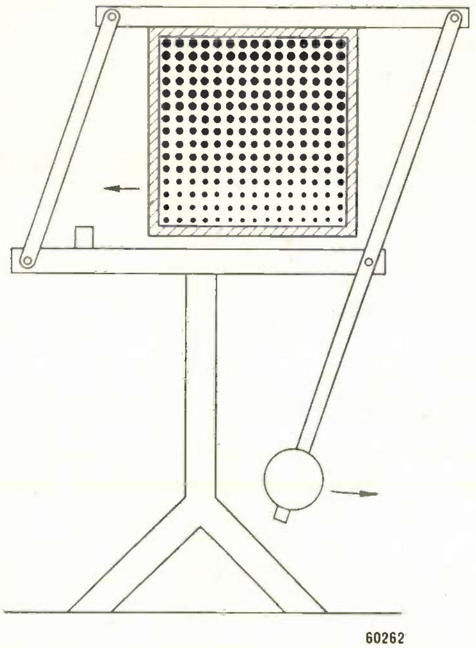


Fig. 3. Apparatus for determining the effect of movement of the phantom upon the sharpness and contrast of the X-ray image. The phantom is suspended in a frame made in the form of a hinged parallelogram. By means of a pendulum coupled to the frame an almost purely translational movement is imparted to the phantom. A device not shown in the drawing ensures that the motion is practically uniform, whilst the speed of the movement can easily be adjusted and recorded.

In order to construct a contrast-detail diagram (*fig.* 5) the diameters *d* of the holes are plotted along the ordinate and their depths *h* along the abscissa. The dimensions of the holes that are just visible in the radiograph correspond to a curve in the diagram, the visibility curve. And since the diagram represents, as it were, the phantom itself, the visibility curve corresponds to the border line on the phantom. This line has roughly the form of a hyperbola, which means to say that small details must have great contrast and details with a small contrast must have large dimensions if they are to be perceived. When we attempt to define the resolving power of the X-ray method employed as the reciprocal value of the volume of the smallest visible object (in the form of a cylinder) then it becomes apparent that this factor largely depends upon the shape of the cylinder to be observed; in other words, the curve of observation is not a curve of constant volume (cf. the dotted curve in fig. 5). A thin and relatively long cylinder is more easily perceived than a short and thick cylinder, as is to be seen from *fig. 6a*, where the reciprocal of the detail has been determined and plotted against the diameter of the holes for various points along the curve of observation, and from fig. *6b*, where the reciprocal volume has been plotted against the depth of thc holes. What is of particular interest for prac-

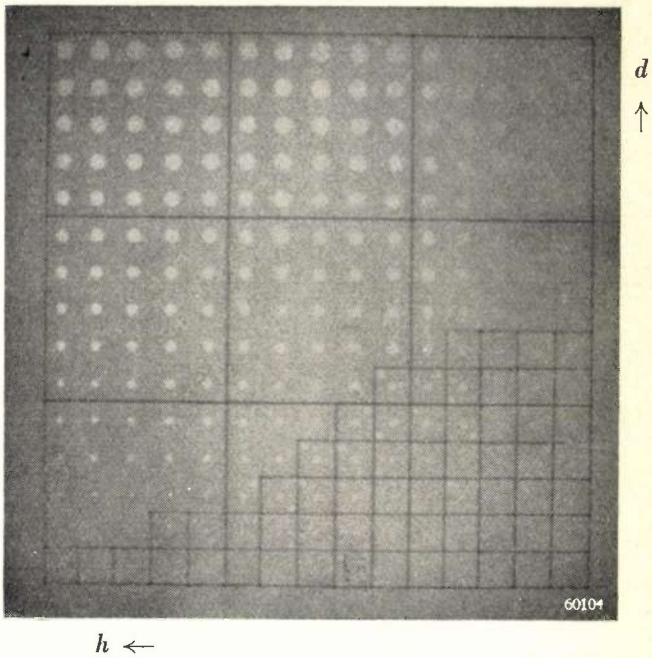


Fig. 4. Fluorogram of the contrast-detail phantom on 45 mm film. (The reproduction is the reflected image of fig. 1.)

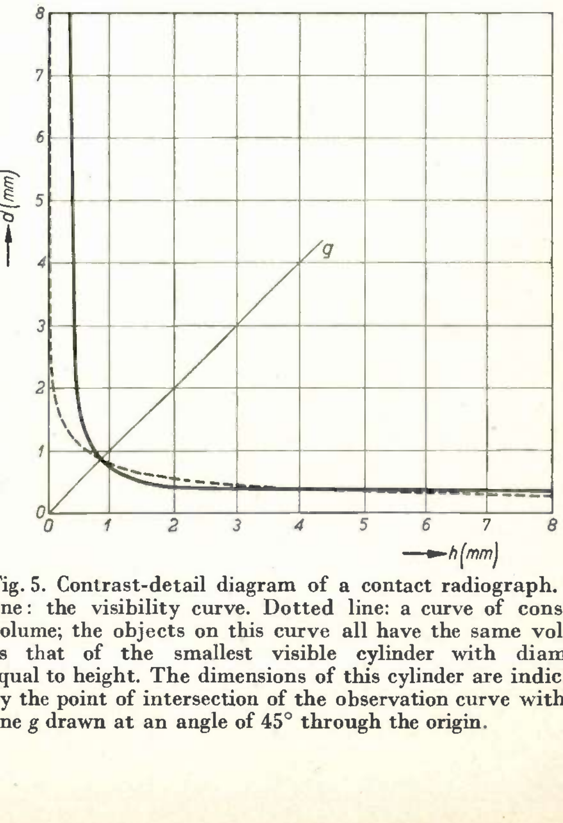


Fig. 5. Contrast-detail diagram of a contact radiograph. Full line: the visibility curve. Dotted line: a curve of constant volume; the objects on this curve all have the same volume as that of the smallest visible cylinder with diameter equal to height. The dimensions of this cylinder are indicated by the point of intersection of the observation curve with the line *g* drawn at an angle of 45° through the origin.

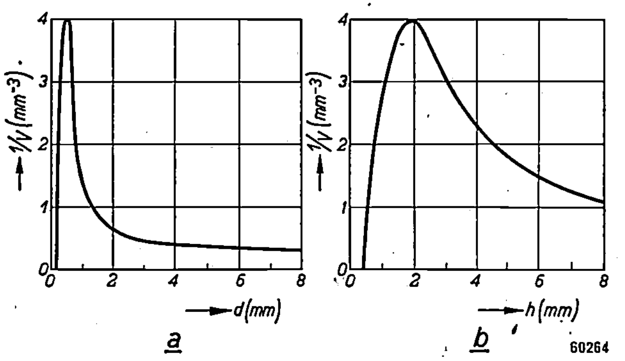


Fig. 6. *a*) Reciprocal volume V of the holes just visible according to the visibility curve of fig. 5, plotted as a function of the diameter of the holes. *b*) ditto, plotted as a function of the depth of the holes.

tical purposes is a cylinder of which the diameter and height are equal, since this most closely approaches a sphere. (A recently formed nodule in the lung is also roughly spherical.) The smallest still visible cylinder of this shape is given in the contrast-detail diagram by the point of intersection of the visibility curve with the straight line (g) drawn at an angle of 45 degrees through the origin (fig. 5). In a fluorogram like that in fig. 4 the size of this cylinder is about 2 mm.

Apperception diagram and the phantom used for it

It appears that the time taken to judge an image of the phantom described in the foregoing depends largely upon the quality of the X-ray image. The better the X-ray method the better is the curve and the less time it moreover takes to observe it. The latter factor does not find expression in the contrast-detail diagram.

Furthermore, any judgment of an X-ray method according to what is revealed by the contrast-detail diagram does not fully answer practical requirements, inasmuch as the phantom test discloses whether details in the image can be observed at places where one knows they should really be, whereas in actual practice it is not known where the details sought may occur.

For this reason, following upon some tests that were carried out in regard to the influence of an enlargement of the X-ray image (see below), a so-called "apperception phantom" has been constructed which makes it possible to ascertain how long it takes to judge an X-ray image in its details and how many errors are made in that judgment.

The object plate of the apperception phantom (*fig. 7*) also has a number of holes, but not arranged in regular rows. The plate is divided into three parts, each marked off into twenty squares with lead oxide. In each square is a number of holes varying from two to six, all in different positions. The holes in each part of twenty squares are of the same dimensions, viz. diameter 2 mm and depth 5 mm, 3 mm and 3 mm, and 5 mm and 2 mm respectively for the three parts. These dimensions have been so chosen that the detail falls well within the limit of visibility as was determined from the contrast-detail diagram for an apparatus for fluorography. The observer counts the number of holes he can see in a square in a radiograph or fluoroscopic image of the phantom, and as quickly as possible. If he fails to give the right number this counts as an error. The time taken to count the holes in twenty squares of one of the three parts (which part depends upon the wishes of the observer) is noted down[5]). In a diagram constructed by plotted the time along the abscissa and the number of errors along the ordinate a complete observation corresponds to one point. Different X-ray methods will give different points, and in this way an "apperception diagram" is obtained, an example of which is reproduced in *fig. 8*.

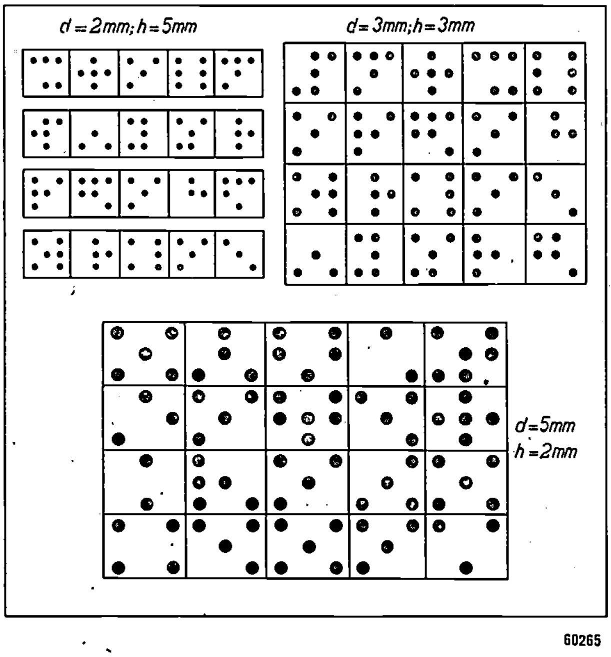


Fig. 7. Object plate for the apperception diagram. This is divided into three parts. The holes in each part have the same dimensions, respectively 2 mm diameter and 5 mm in depth, 3 and 3 mm, and 5 and 2 mm. The holes are drilled in groups, with two to six holes in each square marked off with lead oxide.

[5]) A similar method for judging the perceptibility of details has also been applied for formulating illumination standards according to Weston, as described by A. A. Kruithof and A. M. Kruithof in their article: Basic principles for the formulation of illumination standards, Philips Techn. Rev. **10**, 214-220, 1949 (No. 7).

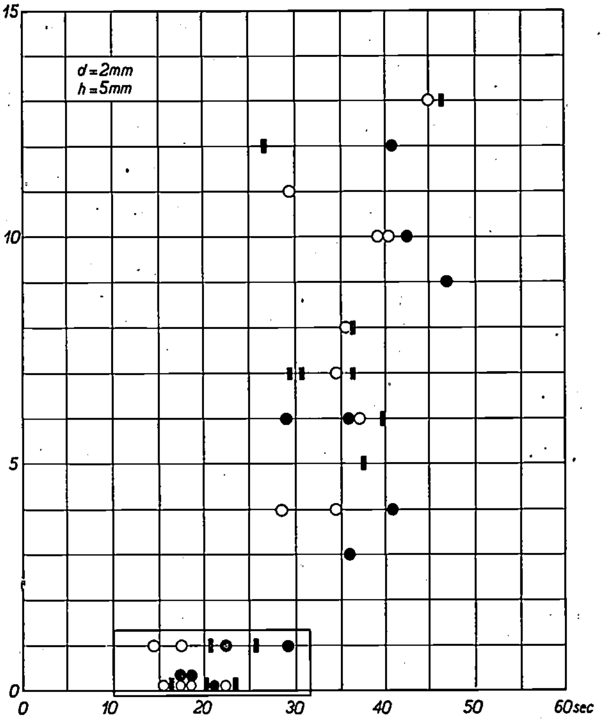


Fig. 8. Example of a apperception diagram. The observation time is plotted in seconds along the abscissa and the number of errors made by the observer along the ordinate. Each complete observation furnishes a point in the diagram. The points indicated by different signs refer to different films and different ways of observing the image of the apperception phantom produced by the method of fluorography (through a magnifying glass or by means of projection). The points indicated by the same sign refer to observations with films of different density. The frame at the bottom encompasses all the points obtained when judging enlarged radiographs.

Applications of the diagrams

a) With the aid of contrast-detail diagrams and apperception diagrams it can now be decided whether certain X-ray methods merit preference over others.

Owing to the aforementioned dependency of the resolving power upon the shape of the object an X-ray method is not to be judged solely, for instance, from the smallest visible sphere. Judging according to the shape of the whole of the contrast-detail curve is far more complete. The closer this curve follows the coordinate axes the better does the respective X-ray method answer the purpose. In eleven Scandinavian institutes the writer plotted contrast-detail diagrams with the aid of the twelve apparatus used there for fluorography. The phantom was placed immediately in front of the screen (except in one case, No. 11, where an enlarged image was produced); both mirror and lens cameras were used. In *fig. 9* twelve contrast-detail diagrams are represented for radiographs with an exposure time corresponding to a chest thickness of about 18 cm. From these diagrams it is to be concluded that there is a considerable difference in performance between the apparatus employed. Although such a difference is already fairly well indicated by the dimensions of the smallest visible cylinder with equal depth and diameter (see the points of intersection of the curves with the line *g*), the shape of the curves obviously furnishes a much better distinction.

b) The suitability of physicians and technicians for radioscopic work depends upon their ability to discern small details. This can be judged by getting them to plot a contrast-detail diagram of an image of the phantom on the fluoroscopic screen. It will be seen that there are striking differences between the results obtained by the various observers (see *fig. 10*). In this way it is possible to select the best observers.

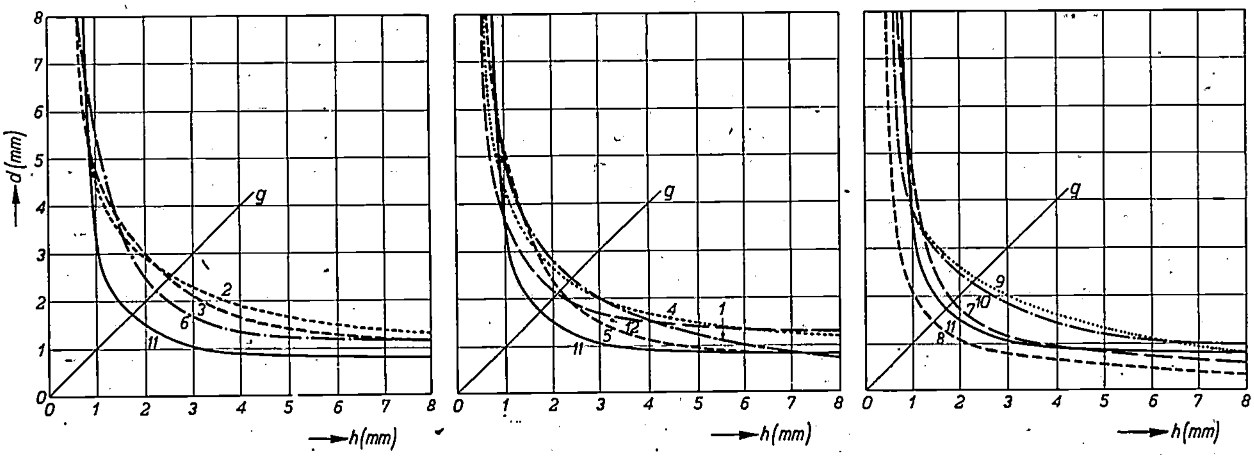


Fig. 9. Twelve contrast-detail diagrams recorded for comparing the properties of the X-ray installations used in eleven Scandinavian institutes. The dimensions of the smallest visible cylinder with diameter and height equal (intersection of the curves with the line *g*) already give an idea of the differences between the twelve methods; the shape of the observation curves as a whole makes these differences more pronounced. (To facilitate comparison the curve for method No. 11 is included in all three diagrams.)

c) In order to judge what improvement can be reached by X-raying with an enlarged image it is necessary to use, in addition to the contrast-detail diagram, also the apperception diagram, because this diagram reveals best the difference between an enlarged image and a normal one. (It is in fact the investigation into this improvement that led to the designing of the apperception phantom.)

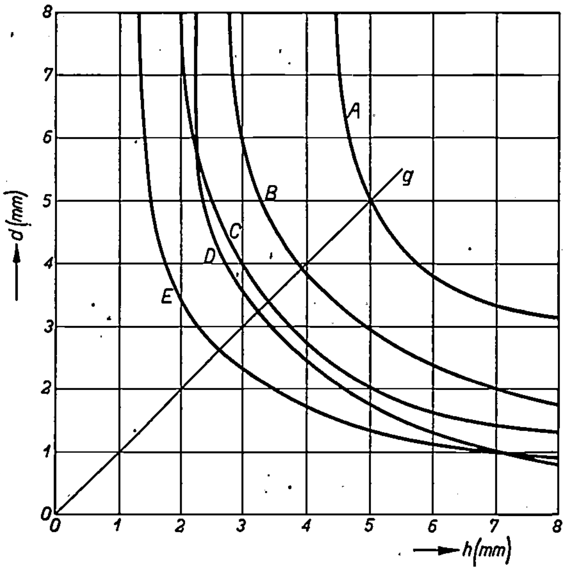


Fig. 10. Contrast-detail diagrams made by five different observers *A-E* from their examination of a fluoroscopic image of the contrast-detail phantom. The results of the five observers differ considerably.

An enlarged image is obtained when the object is placed at a greater distance from the photographic film or screen than is normally the case (usually the object is placed as close as possible to the film or screen). The advantages derived from enlargement of the image are as follows (see footnote [4])):

1) The blurring due to the granular structure of the film or screen (intensifying screen) is of less importance.
2) Contrast is greater, because a larger portion of the scattered X-rays does not reach the screen.

However, there are also disadvantages attaching to an enlargement, viz:

1) Geometrical blurring is increased.
2) When screening it is easier to discover any abnormalities if the doctor can move the patient a little to and fro, but this cannot be done so well when the patient is some distance away from the screen.

In order to investigate the effect of an enlargement of the image a number of radiographs of both phantoms were made, both normal and enlarged 2.8 times, by the X-ray-camera method. The fluoroscopic image was photographed on a 45 mm film under different exposure conditions (different densities). The radiographs were examined in three different ways: under a magnifying glass with a power of 1.8, after projection on a white, fine-grained screen to a picture of 15 cm × 15 cm, and after projection to a picture of 35 cm × 35 cm. Contrast-detail diagrams and apperception diagrams were made of the pictures obtained. Both kinds of diagrams clearly show that the image quality can be considerably improved by enlargement, as may be seen from fig. 8 and *fig. 11*. The group of points within the frame at the bottom of fig. 8 relate to enlarged radiographs; they belong without exception to a small number of errors and a relatively short observation time. In fig. 11 the apperception curve relating to an enlarged image likewise shows a more satisfactory trend than that relating to the normal image. The improvement obtained by enlargement is expressed in figures in the *table* below. It appears that the relatively small improvement of 0.5 mm in the size of the smallest visible detail is accompanied by greater ease of apperception, as a result of which the time taken for the observation with the apperception phantom is reduced to about half and hardly any errors are made.

Image	Smallest visible detail	$d=5$ mm, $h=2$ mm		$d=2$ mm, $h=5$ mm	
		observ. time	errors	observ. time	errors
Normal	2.0 mm	49 sec	5.5	44 sec	7
Enlarged	1.5 mm	28 sec	1	19 sec	1

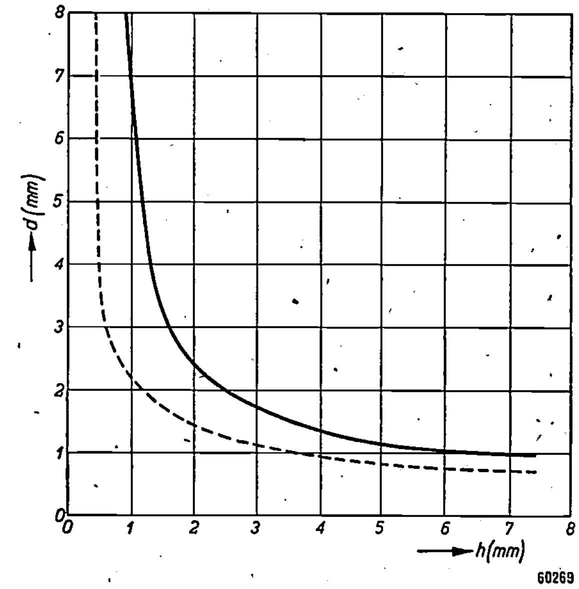


Fig. 11. Full line: contrast-detail diagram for screening with normal image. Dotted line: ditto with enlarged image.

The blurring of the screen, which sets a limit to the visibility of the details and is the main factor leading to the introduction of X-ray technique with enlarged image, can be determined by means of a so-called star [6]). This star (see *fg. 12)* consists of a number of copper strips arranged as sectors in a circle. The space between the strips is filled with paper. The circle is divided into four quadrants, each with different thicknesses of the copper sectors, viz: 0.5, 1.0, 2.0 and 3.0 mm respectively. The radius of the circle is 2.5 cm. Four circular wires indicate where the distance between the copper strips is about 0.25, 0.50, 0.75 and 1.00 mm. Thus it is possible to determine with a fair degree of accuracy the smallest distance separating the sectors in the X-ray image. In order to determine the screen blurring the fluoroscopic image of the star is photographed together with a number of contact-radiographs of the star fixed onto the side of the screen facing the camera (see *fig. 13*). The difference in sharpness between the picture of the fluoroscopic image and that of the contact images (fig. 12) has to be ascribed entirely to the effect of the fluorescent screen (care has to be taken to ensure that the brightness and contrast of the screen image and the contact images are practically the same).

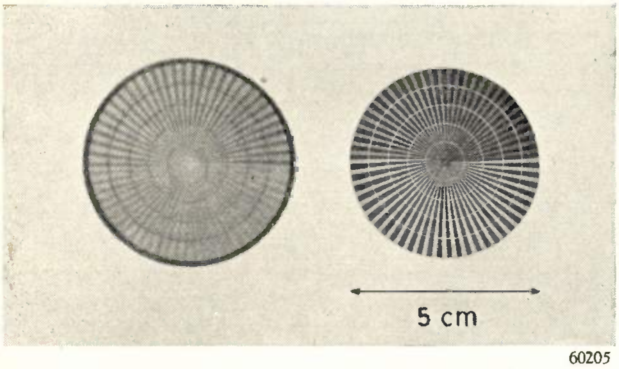


Fig. 12. Determination of the blurring of the fluorescent screen with the aid of a star consisting of a number of copper strips arranged as sectors in a circle. *a)* Photographic recording of the fluoroscopic image of the star. *b)* Contact radiograph of the star.

d) Finally mention has to be made of the possibility to investigate separately with the contrast-detail diagram the effect of the movement of the object upon the properties of the X-ray image. It has already been pointed out that such a movement cannot be avoided when radiographing the lungs. The contrast-detail phantom is placed in the moving apparatus shown in fig. 3. Contact radiographs of the moving phantom with exposures of 0.05 sec and 0.25 sec give contrast-detail diagrams as illustrated in *fig. 14*. The explanation of the various curves is given in the subscript. A small movement is of little effect; although the object is less sharply defined it can still be observed, because the image is somewhat protracted by the movement. When, however, the motional unsharpness is greater than 2 mm visibility is noticeably reduced.

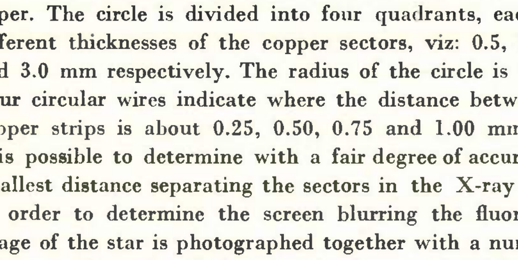


Fig. 13. Sketch of the set-up for determining the blurring of the fluorescent screen. R = X-ray tube, O = star, S = screen, F = contact radiographs of the star, fixed onto the screen, C = photographic camera.

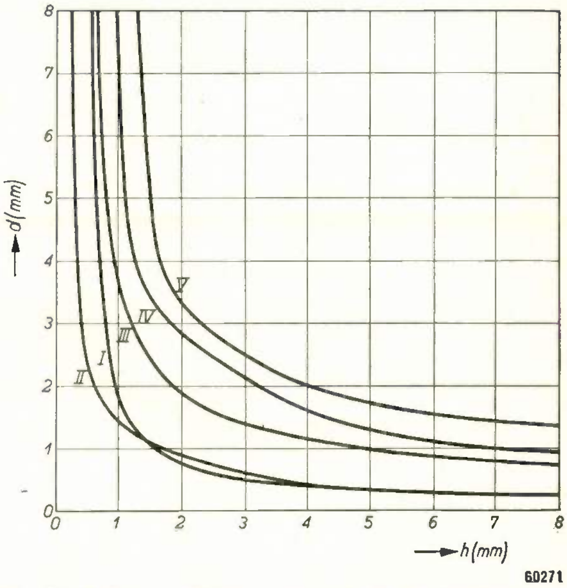


Fig. 14. Influence of the movement of the phantom upon the contrast-detail diagram. *I* motional unsharpness nil, *II* motional unsharpness 1 mm. Not much difference is to be seen between *I* and *II*. *III*, *IV* and *V* motional unsharpness respectively 2.5, 7 and 12 mm. The visibility curves show a less and less satisfactory trend.

Regular check of the quality of radiographs

In X-ray examination of the lungs it often happens that several radiographs have to be made of the same patient at more or less long intervals. To be able to compare these photographs and determine exactly how a certain affection of the lungs is progressing it is of the greatest importance to have data available regarding the sharpness and contrast properties of each picture. It is impracticable to collect these data with the aid of the contrast-detail phantom or the apperception phantom; not only are they too large to be photographed together with the patient, but it takes far too much time to scan the phantom images and construct the diagrams.

[6]) For another method of investigating this factor see H. A. Klasens, The blurring of X-ray images, Philips Techn. Rev. 9, 364-369, 1947/1948 (No. 12).

To form an idea of the properties of the thousands of lung radiographs that are made every year by the Medical Department of Philips Works at Eindhoven, a small phantom was designed which can easily be photographed together with the patient on either side of the neck and which enables the doctor to check regularly the quality of all his radiographs. This phantom (see *fig. 15*) consists of a set of aluminium "steps" cut into two parts. The first part has five steps of 4 to 8 mm thickness, the second part five steps of 9 to 13 mm. "Philite"' cannot be used for this; owing to the small dimensions of the phantom "Philite" has too little total absorption for X-rays. For such small objects aluminium has been chosen as being a more suitable material. This, however, does not correspond so well to the physical qualities of the thorax, and consequently preliminary tests were carried out to compare an aluminium phantom with one of "Philite"; the former has been calibrated, as it were, with the aid of the latter. To represent the "details" five small spheres of "Philite" varying in diameter between 1 and 4 mm have been affixed to each step. Furthermore, to check the quality of the rays one part of the phantom is fitted with a step of copper (0.2 to 0.6 mm thick), whilst the other part has a strip of lead for checking the fogging of the unexposed film.

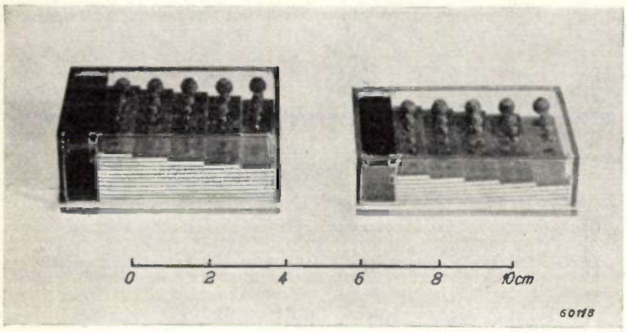


Fig. 15. Phantom for regular checking of the quality of radiographs. It is in two parts, each formed by five aluminium steps on which "Philite" spheres of varying diameter are affixed. At the left-hand side of the part on the left is a copper step and a strip of lead on the part on the right, for checking respectively the ray quality and the fogging of the film.

With the aid of this phantom it is quite easy to judge a radiograph of the lungs. In *fig. 16* a normal lung radiograph is reproduced; the image of the phantom is to be seen above the patient's shoulders. The part of the photograph to be examined is compared with the aluminium step showing the same density, and the smallest detail of that step that can still be perceived is noted. It appears that in the case of contact radiographs on an average four and sometimes all five spheres can be observed on the aluminium step corresponding in density to the free field of the lung (between the ribs). This agrees well with the dimension of about 1 mm of the smallest visible cylinder on a contact radiograph as determined by means of the contrast-detail diagram.

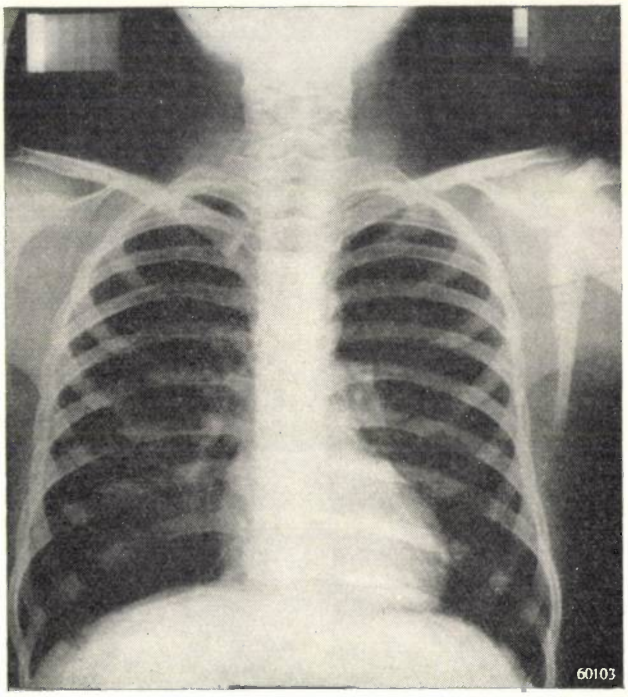


Fig. 16. Lung photograph, showing the check phantom above the patient's shoulders.

Summary. Following an introduction concerning the factors affecting an X-ray image, the construction is discussed of some phantoms enabling the merits of various radiologic methods to be compared. Some applications of these phantoms are dealt with: comparison of the properties of X-ray apparatus in different institutions, selection of doctors and others for X-ray screening, investigation into the influence of enlargement of the X-ray image, and the effect of movement of the object upon the X-ray image. Finally a phantom of smaller dimensions is described which serves for the regular checking of lung radiographs, which is particularly of importance when several radiographs have to be made of the same patient.

LIGHT SOURCES FOR LINE SPECTRA

by W. ELENBAAS and J. RIEMENS.

621.327.3/.4:535.338.3

This article contains data appertaining to a series of discharge lamps which have been developed to serve as monochromatic light sources for optical experiments.

Introduction

In physcial experiments or tests where the light plays a part, and also in chemistry (for example in polarimetry), the need is often felt for a strong monochromatic light source, or for a light source which emits a number of monochromatic rays of a known wavelength. Sometimes one can manage with a "monochromator", an instrument in which a narrow part of the spectrum of a light source with a continuous spectrum (incandescent lamp or carbon arc) is separated with the aid of prisms and slits. In many cases, however, such a light source is not sufficiently monochromatic. One can then have recourse to the colouring of flames with metals, present in salts, as already applied by Kirchhoff and Bunsen. An example of this is the well known sodium flame used in polarimetry and for interference experiments. These flames, however, have the drawback that they are of low luminosity, whilst moreover the constancy of the radiation leaves much to be desired. A complicated apparatus then has to be devised for feeding the flame regularly with air containing a constant proportion of salt particles.

Shortly after the discovery made by Kirchhoff and Bunsen investigators in the field of spectroscopy began to make use of the fact that gases and vapours can be made to give light by means of an electric discharge. Plücker got his glass-blower Geissler to make some discharge tubes suitable for this purpose, which since then have been named after their maker. A Geissler tube (*fig. 1*) is in essence a discharge tube with cold electrodes. The luminous part, called the positive column, is situated in a constricted part, so that the density of the current and the brightness are increased. Originally Geissler tubes were usually fed with the pulsating current from an induction coil. Provided care is taken to ensure adequate current limitation they can also be fed with alternating or direct current of a high voltage. But there is always the drawback that the capillary and the electrodes become hot only under a relatively small load, as a result of which the gas becomes contaminated and the tube is rendered useless.

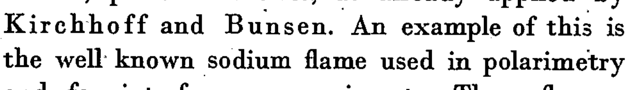


Fig. 1. Geissler tube for use in spectroscopy.

The development of modern gas-discharge lamps with oxide-coated cathode has created new possibilities, including such uses as referred to above, because an oxide-coated cathode can emit a strong thermionic current without being overloaded. It is now easy to make small discharge tubes containing some gas or other or a metallic vapour or a mixture of both in a very pure state and permitting a high current intensity, so that a light source is obtained capable of emitting considerable energy in one single spectral line or in a few lines.

Construction of the lamps

The discharge tubes at present being made by Philips for the above-mentioned purpose are either of normal glass or of quartz as needed. The dimensions of the discharge tube proper vary from case to case. The length of the luminous part of the discharge amounts to 3 to 4 cm.

At each end of the tube is a leading-in wire with an oxide-coated cathode attached, which is not separately heated but is brought to the required high temperature by the discharge and maintained at that level. In some cases, that is to say with some tubes filled with metallic vapour, the extremities of the discharge tube are metallized, with the object of keeping the temperature in the tube high enough to bring about a sufficiently high vapour pressure of the metal (e.g. cadmium). In *fig. 2* some forms of discharge tubes are shown diagrammatically in cross section.

In order to protect the discharge tube proper and to prevent oxidation of the leading-in wires, which sometimes get rather hot, the tube is built into an envelope, which is likewise made either of normal glass or of quartz as required. The space between the actual discharge tube and the envelope is evacuated or filled with nitrogen. The envelope is fitted with a screw base so that the lamp can be inserted in a normal lamp-holder. A number of these lamps are shown in the photogaph reproduced in *fig. 3*. The distance between the light centre and the lamp base is the same in all lamps, so that there is no need to readjust the height when a lamp has to be replaced.

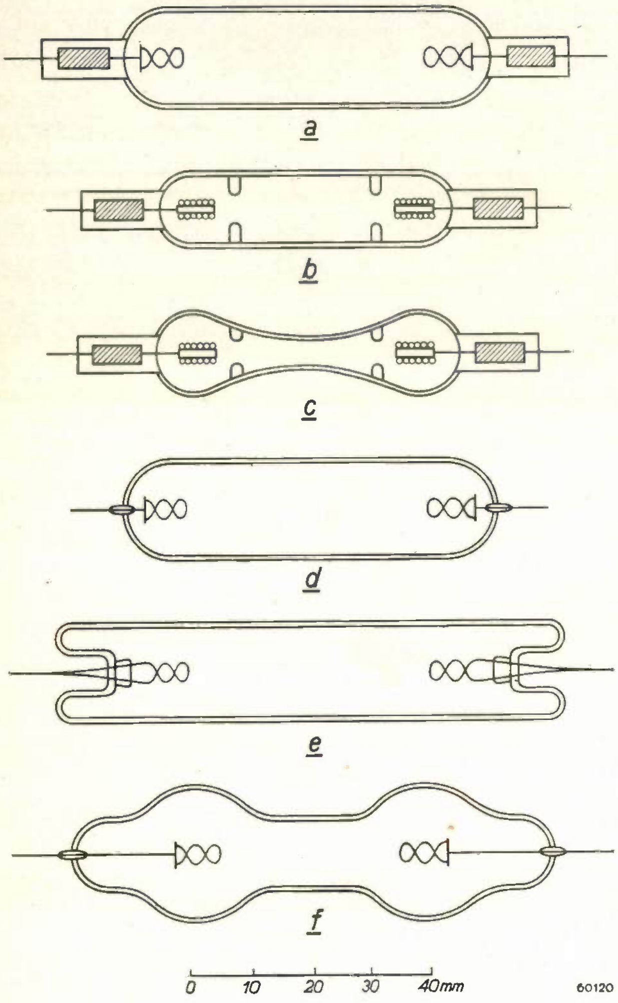


Fig. 2. Construction of the light sources for line spectra.
a) Mercury (low pressure, quartz),
b) mercury (high pressure) and cadmium (both in quartz),
c) zinc (quartz),
d) sodium, rubidium, cesium (glass),
e) mercury (low pressure), Ne, A, Kr, Xe (glass),
f) helium (glass).
In b and *c* (discharge tubes with high pressure) the oxide-coated cathode consists of a pin surrounded by coiled wire; in the other tubes the oxide-coated cathode is a twisted loop of coiled wire. In the quartz tubes *a*, *b* and *c* the inlet leads are foils of molybdenum. The tube *d* is lined with an alkali-vapour-resistant layer. The exceptional shape of *f* is necessary on account of the greater heat developed in helium.

Supply

Although some of these lamps could be connected to a 220 V A.C. supply, with of course suitable current limitation, for easy ignition it is better to supply them with a higher voltage. A transformer having an open voltage of 470 V at a primary voltage of 220 V is therefore supplied together with the lamp. The transformer is an autotransformer of the low power factor type. This means that provision is made for a considerable spread of the lines of force, which has the same effect as if a choke were connected in series with a normal transformer. The current is thereby automatically limited. Consequently the working current is practically equal to the short-circuit current, which amounts to about 0.9 A. Thanks to the high open voltage the tube ignites without the oxide-coated cathodes having to be pre-heated. The circuit is diagrammatically represented in *fig. 4*.

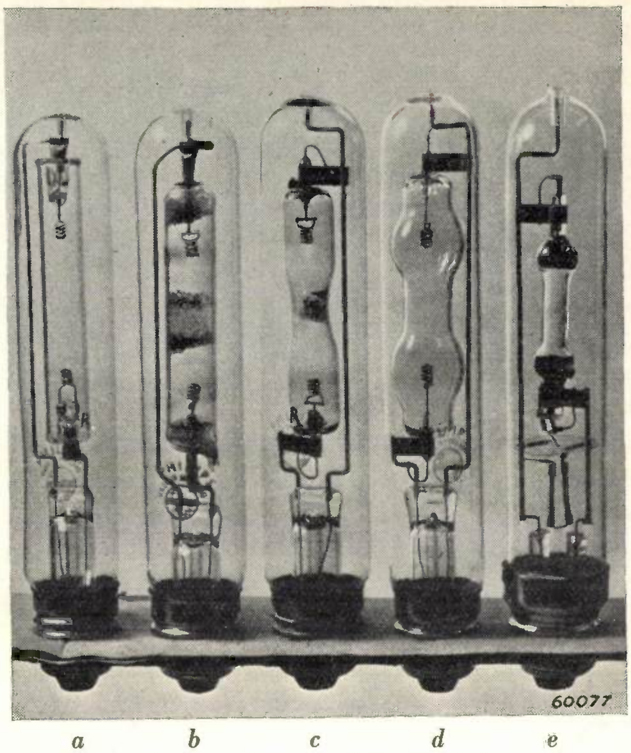


Fig. 3. Some lamps for line spectra.
a) Mercury (low pressure, glass),
b) sodium,
c) rubidium,
d) helium,
e) zinc.
The metal condensed on the glass wall of the lamps *b* and *c* is vaporized by the heat developed when the lamp is burning.

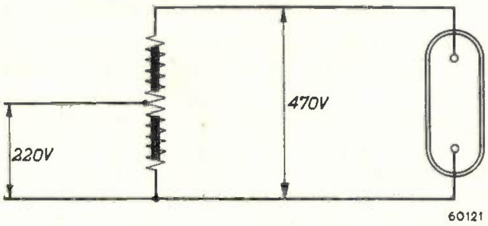


Fig. 4. Circuit diagram of the supply apparatus (autotransformer with low power factor) with the lamp connected to it.

The spectra

The lamps are filled either with one of the rare gases He, Ne, A, Kr, Xe, or with argon to which a metal has been added, this metal being vaporized by the heat of the discharge. Suitable metals are the alkali metals Na, Rb, Cs and the bivalent metals Zn, Cd, Hg. Mercury lamps can be made for low as well as for high pressures, in the latter case the amount of mercury being such that the metal is entirely vaporized at the working temperature. *Fig. 5* is a reproduction of photographs of the visible spectrum of the various lamps, whilst in *fig. 6* the ultra-violet spectra are shown.

If it is desired to separate a part of the spectrum,

Type No.	Gas or vapour	Wattage
93123E	Hg (low pressure)	15
93136E	Hg (high pressure)	90
93162E	Cd	25
103137E	Zn	25
93145E	Hg, Cd Zn	90
93098E	He	45
93099E	Ne	25
93100E	A	15
93101E	Kr	15
93102E	Xe	10
93122E	Na	15
93104E	Rb	15
93105E	Cs	10

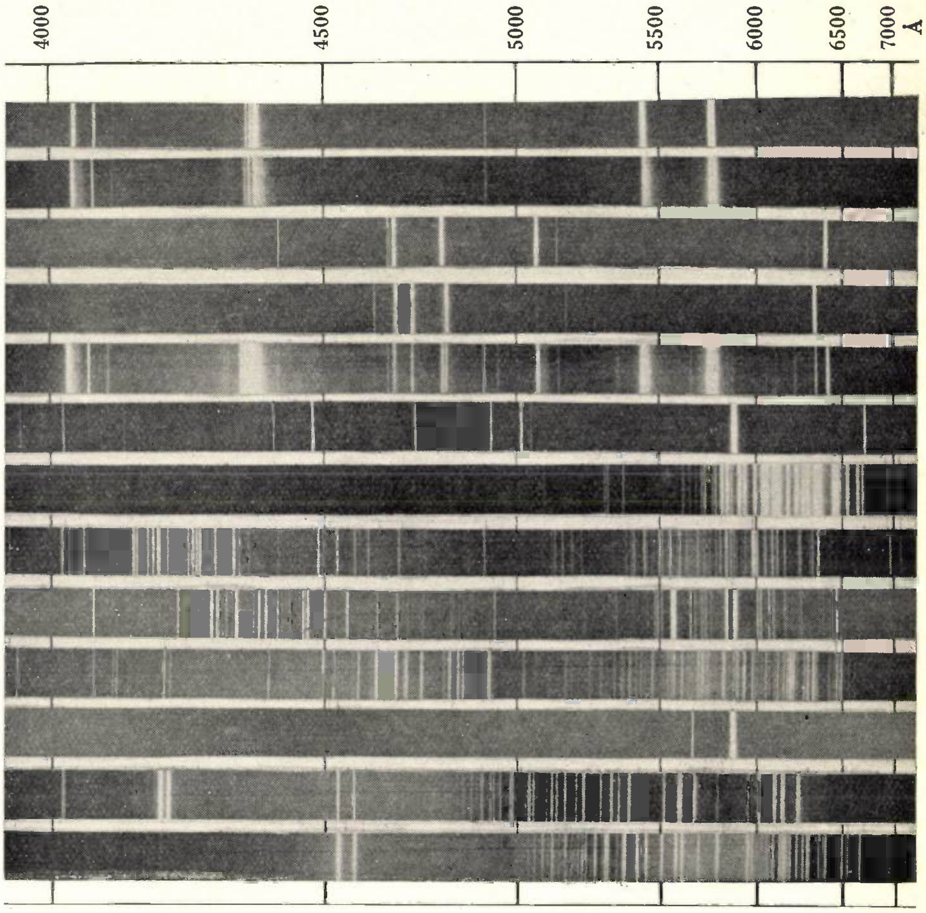


Fig. 5. Spectra of the various types of lamps that can be used for the visible spectrum. The envelope of all these lamps is of glass. The type number, the gas or vapour filling and the wattage of the lamps are indicated on the left.

Type No.	Vapour	Wattage
93109E	Hg (low pressure)	15
93110E	Hg (high pressure)	90
93107E	Cd	25
93106E	Zn	25
93146E	Hg, Cd Zn	90

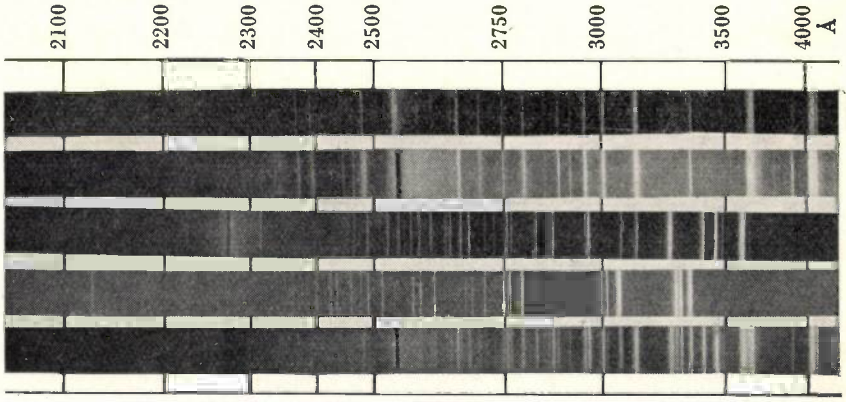


Fig. 6. The same as fig. 5 for the ultra-violet spectrum. All these lamps have an envelope made of quartz.

Table I. Combinations of lamp types and filters with which certain spectral lines can be isolated. If it is necessary to eliminate the infra-red radiation then the filters indicated in the last column have also to be added. The filters A, B, C, D and E are liquid filters of the composition given in table II (to be used in an optical glass cell 20 mm thick). The others, with the exception of those shown as Zeiss filters, are Schott filters. (The data for this table have been taken mainly from: H. Ewest and K. Larché, Handbuch der Lichttechnik, Springer Berlin, 1938, p. 192.)

Isolated wavelength in Å	Type No. of lamp	Filters	Filters for absorption of infra-red
3076	93106E	UG 5 (5 mm) + A + B (10 mm)	
3126/32	93110E	UG 5 (5 mm) + C	E
3261	93107E	UG 2 (2 mm) + A + C	
3341	93110E	UG 2 (2 mm) + A + D	
3650/63	93110E	UG 2 (2 mm) + BG 12 (4 mm)	E + BG 19 (2 mm)
4047	93123E 93136E	UG 3 (9 mm) + GG 4 (1.5 mm)	E + BG 19 (2 mm)
4358	93123E 93136E	BG 12 (4 mm) + GG 3 (4 mm) or Zeiss C	E + BG 19 (2 mm)
4555/93	93105E	GG 2 (2 mm) + BG 12 (2 mm)	E + BG 19 (2 mm)
4678/4800	93162E	GG 5 (1 mm) + BG 12 (2 mm)	E + BG 19 (2 mm)
5086	93162E	GG 11 (2 mm) + VG 3 (2 mm)	E + BG 19 (2 mm)
5461	93123E 93136E	BG 11 (20 mm) + OG 1 (1 mm) + BG 18 (3 mm) or Zeiss B	E + BG 19 (2 mm)
5770/91	93123E 93136E	OG 3 (1 mm) + VG 3 (1 mm) + BG 18 (1 mm) or Zeiss A	E + BG 19 (2 mm)
5890/96	93122E	OG 3 (1 mm) + VG 3 (1 mm) or OG 2 (2 mm)	E + BG 19 (2 mm)
6362	103137E	RG 1 (2 mm)	BG 19 (2 mm)
6438	93162E	RG 1 (2 mm)	BG 19 (2 mm)

Table II. Liquid filters used for isolating certain spectral lines according to table I.

No.	Composition	Weight per litre water
A	Nickel-cobalt sulphate $NiSO_4$. 1 H_2O $CoSO_4$. 1 H_2O	 303 g 86.5 g
B	Picric acid	31.6 mg
C	Potassium chromate K_2CrO_4	150 mg
D	Nitric acid HNO_3	12.6 g (0.2 n)
E	Copper sulphate $CuSO_4$. 5 H_2O	28.5 g

filters can be used. In favourable cases it can be so arranged that only light of one wavelength is emitted. If under certain circumstances it is not possible to achieve this with filters then a monochromator will have to be placed behind the lamp. In *table I* a number of lines are indicated which can be isolated with a suitable combination of lamp and filters; *table II* gives the composition of the liquid filters occurring in the combinations.

Summary. A description is given of the construction and supply apparatus of a number of light sources for line spectra, as also a summary of the various fillings (rare gases and metallic vapours) and the spectra emitted. A table shows how certain rays can be isolated from the various spectra with the aid of certain filters, which are further described

INVESTIGATIONS INTO THE IMPACT STRENGTH OF IRON AND STEEL

by J. D. FAST.

620.178.746.22:669.12:
669.141.24

For judging the strength of steels the so-called notched-bar impact test is often applied in addition to the usual tensile test. This impact test makes heavier demands upon the resistance of the material to brittle fracture than the tensile test, which mainly gives indications of the resistance to plastic deformation. This article deals not only with matters of common knowledge such as the execution and significance of the notched-bar impact test and the elementary concepts of stress and fracture, but also with original measurements by means of which the impact value of pure iron is determined as a function of temperature, and the effect thereon of oxygen, carbon and nitrogen is investigated.

The notched-bar impact test

In exceptional circumstances, steels which are shown to be ductile by the normal tensile test (great deformation preceding fracturing), when in actual use break in a "brittle" manner (no previous deformation worth mentioning). In the past ten years or so the collapse of some welded bridges and several welded ships has caused considerable consternation. The fractures found in these cases appeared to be of the brittle type, notwithstanding the fact that under tensile test the materials had behaved reasonably well. These and other experiences had long ago led to the view that the question whether fracture is tough or brittle is determined not only by the properties of the material but also by the manner in which the material is loaded during testing or in use.

The important part that the manner of loading plays in tough or brittle fracture is evident from bending tests with notched bars. Long ago it was found that such tests make heavier demands upon the resistance of materials against rupture than the classical tensile test. The notched bars were bent in a vice in such a way that the notch opened while the material was being deformed. It was occasionally found that, using two kinds of steel which behaved in practically the same way during tensile testing, one showed tough fiacturing in the bending test, whereas the other showed brittle fracturing. The test could be made more severe still by giving the clamped bars a blow with a hammer over the notch so as to cause them to bend more quickly. Based upon these findings machines have been constructed which enable us to test the metal in such a way that a satisfactory result of the test (notched bar impact test) offers a better guarantee against accidents than a satisfactory result of the normal tensile test.

Various types of such machines are in use in different countries and the specifications for the dimensions of the bars to be tested also differ widely. In our investigations we used bars of 10 mm × 10 mm × 55 mm with a notch 2 mm wide and 5 mm deep (see *fig. 1*), as recommended by the International Federation of the National Standardizing Associations, and tested them in a Charpy machine. In this test the bar rests on two horizontal supports and against two vertical supports 40 mm apart and is struck in the centre of the plane opposite the notched side by a pendulum hammer H released from a certain angle α corresponding to the height h (see *fig. 2*). The hammer is attached to an arm S. The shape and distribution of the mass of the hammer are so chosen that when the blow is struck practically no shock is transmitted to the pivot of the arm [1]). Upon the blow being struck the notched bar breaks or is torn open and bent so far that it can be forced out

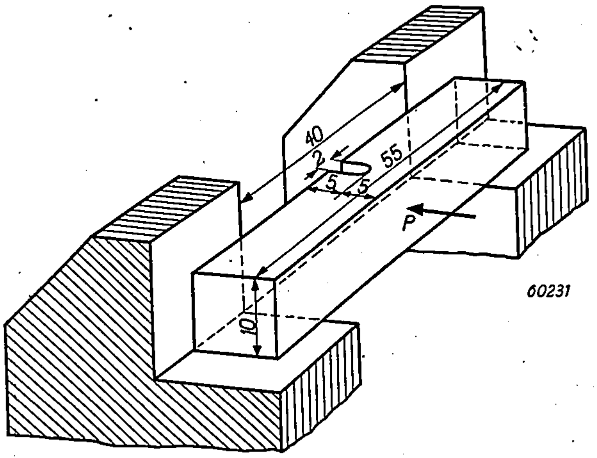


Fig. 1. Sketch representing a notched bar for the impact test. The dimensions given are in mm. The bar is struck by the hammer at the point P.

[1]) For this purpose the hammer has to strike the bar with its centre of percussion not far removed from its centre of gravity.

through the opening of 40 mm between the supports. Under the action of its remaining energy the hammer then rises on the other side and reaches, say, an angle β corresponding to the height h'. The work A performed upon the bar is given by the equation

$$A = G\,(h - h') = Gl\,(\cos\beta - \cos\alpha)\,, \quad . \quad . \quad (1)$$

where G is the weight of the hammer (in our experiments 10 or 30 kg, or 98 and 294 newtons respectively) and l is the distance between the pivot and the centre of gravity of the pendulum.

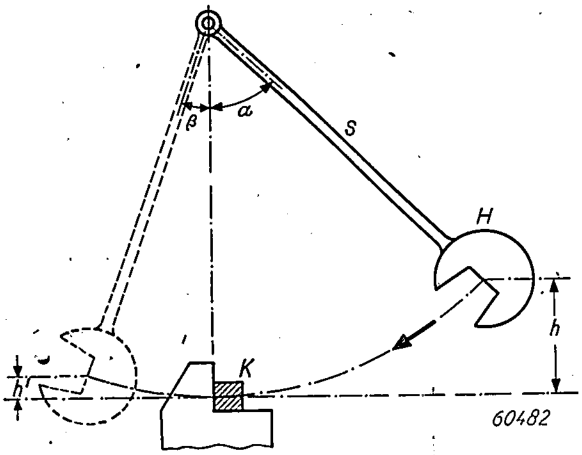


Fig. 2. Diagrammatic represention of the Charpy machine for determining the impact value. The pendulum hammer H is so constructed that its so-called centre of percussion coincides with the point where it strikes the notched bar K. Given the angle α, the work performed by the pendulum hammer can be determined by measuring the angle β.

The impact work A (in kgm), converted per cm^2 of the plane of fracture, in our case for a plane of 10 mm × 5 mm, multiplied by two, is called the notched-bar impact strength (or impact value). This value has no other direct physical significance but gives a good technical measure of the resistance of the material against fracturing. If the break is preceded by an appreciable plastic deformation the impact value is high and the metal usually shows a "fibrous" ("mat") fracture. If there is practically no plastic deformation then the impact value is low and as a rule the metal shows a "granular" ("bright") fracture. Between these two cases are "mixed fractures", having partly a "fibrous" and partly a "granular" appearance.

Significance of the notched-bar impact test

The phenomenon that one and the same material at the same temperature shows ductility under one test (e.g. the tensile test) and brittleness under another (e.g. the impact test) is mainly due to the fact that under these tests the material is subjected to different states of stress. (The difference in the rate at which the test is carried out is of less importance; see also pages 307-308.) The state of stress in any particular point of a body can be described by imagining a large number of planes drawn through that point and indicating the magnitudes and directions of the forces acting upon these differently orientated planes per unit area. Each of these stresses can be represented by a vector. This vector can be resolved into a component lying in the particular plane, the shear stress, and a component perpendicular to that plane, called the normal stress. It can be proved that under any load (however complicated) in each point of a body three mutually perpendicular planes can be indicated in which the shear stresses are nil. The normal stresses on these planes are called the principal stresses. Once the magnitudes and the directions of these three principal stresses are known, then the state of stress in the particular point is fully known, for the stresses on all the plane elements differently orientated through that point can be found by calculation. These three principal stresses will be denoted by the symbols σ_1, σ_2 and σ_3, the first one denoting the greatest and the last one the smallest (in the algebraical sense). Tensile stresses are reckoned as being positive and compressive stresses negative, so that a compressive stress is always smaller than a tensile stress. The principal stress σ_1 is also the greatest of the normal stresses acting in the particular point:

$$\sigma_{\max} = \sigma_1 . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

The maximum shear stress is given by:

$$\tau_{\max} = {}^1/_2\,(\sigma_1 - \sigma_3) \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

It acts in each of the two mutually perpendicular planes bisecting the right angles between the directions of σ_1 and σ_3 and containing the direction of the principal stress σ_2. In the ordinary tensile test ($\sigma_2 = \sigma_3 = 0$) the maximum shear stress acts in all planes making an angle of 45° with the direction of the tensile force.

The normal stresses tend to tear the body apart along the planes on which they are acting, whereas the shear stresses tend to shear the adjacent planes of the body along each other, thus being responsible for plastic deformations, where these occur. To be able to predict whether a body can bear a certain state of stress without plastic deforming and without fracturing, we must know its resistance to shear and its resistance to cleavage.

The significance of these two factors is most readily seen in the case of single crystals: when the shear stress acting in a certain crystallographic plane and in a certain direction exceeds a critical value (the resistance to shear, or shear strength, for that plane and that direction), then permanent deformation takes place owing to the planes shearing one over the other. When the normal stress acting perpendicular to a certain crystallographic plane exceeds another critical value (the resistance to cleavage, or cleavage strength, for that plane) then fracturing takes place.

It is obvious that even in the case of a single crystal there is already a complexity of shear and cleavage strengths to be distinguished. Still more complicated, at least in principle, is the situation in polycrystalline metals, with which one always has to deal in technical engineering. If, however, the crystals are rather small and contain no preferred orientations then certain mean properties can be ascribed to the polycrystalline material, thus one particular shear strength τ_{cr} and one particular cleavage strength σ_{cr} independent of position and direction. These concepts now have the significance of certain critical values of the maximum shear stress τ_{max} and maximum normal stress σ_{max} given by the load. Plastic flow begins to take place as soon as

$$\tau_{max} > \tau_{cr}\,, \quad \ldots\ldots\ldots \quad (4)$$

and rupturing takes place as soon as

$$\sigma_{max} > \sigma_{cr}\,. \quad \ldots\ldots\ldots \quad (5)$$

When, as the load is increased, σ_{cr} is exceeded before τ_{cr} then the material breaks without any previous plastic deformation taking place. If, on the other hand, τ_{cr} is first exceeded then fracturing is preceded by deformation, to a greater or lesser extent.

According to a criterion which in most of the cases investigated proves to be somewhat more accurate than that of the maximum shear stress τ_{max}, flow begins in an isotropic polycrystalline metal as soon as a quantity taking the place of τ_{max} exceeds a critical value depending not only upon σ_1 and σ_3, as is the case with τ_{max} (see formula (3)), but also upon the principal stress σ_2. For our considerations, which are for the greater part qualitative, however, nothing is to be gained by going into this point any farther; neither shall we go farther into the fact that in certain cases the occurrence of fracture is determined by a criterion other than that of the maximum normal stress.

Now in order to understand how it is possible that in certain cases brittle fracture may occur with the impact test and not during the ordinary tensile test, we shall first consider the extreme case (impossible of practical realization) of a body subjected to a uniform triaxial tension ($\sigma_1 = \sigma_2 = \sigma_3$). According to formula (3) no shear stresses then arise in any of the intersecting planes of the body. Under such a loading with continuously increasing tensile stresses all bodies would break without any plastic deformation previously taking place, even bodies of a material that would appear to be highly ductile in a normal tensile test.

During the impact test the three principal stresses in the material underneath the notch are not, it is true, mutually equal, but when the loading is applied then there arises in that part of the material a considerable triaxial tensile stress, in contradistinction to the uniaxial stress arising in the normal tensile test. This can be explained in the following way.

The material adjacent to and lying underneath the base of the notch is locally elongated in the longitudinal direction of the test bar on account of a large tensile stress arising there as the bar is bent. This is accompanied by a contraction (Poisson contraction) in all directions perpendicular to the direction of the tensile stress. The notch, however, forms an interruption in the continuity of the bar, so that the material adjacent to the sides of the notch does not come under load and does not take part in the deformation. This forms a hindrance for the Poisson contraction in the deformation zone and results in tensile stresses perpendicular to the direction of the primary tensile stress. According to formula (3) the occurrence of these secondary tensile stresses implies that the maximum normal stress σ_{max} ($= \sigma_1$) arising in the impact test must be greater than that arising in the ordinary tensile test in order to get the same value of the maximum shear stress τ_{max}.

In the ordinary tensile test σ_{max} is twice as great as τ_{max} under any loading. If the load on the test bar is gradually increased from zero, σ_{max} and τ_{max} increase (without their ratio $\sigma_{max}/\tau_{max} = 2$ being changed), and it will depend entirely upon the values of the cleavage strength σ_{cr} and the shear strength τ_{cr} whether the material will ultimately fracture in a brittle way or whether a larger or smaller plastic deformation will precede the rupture. It goes without saying that the latter will be the case when $\sigma_{cr}/\tau_{cr} > 2$.

According to the foregoing, in the impact test $\sigma_{max}/\tau_{max} > 2$. Assuming, for the sake of argument, that in the case of the notched bar used by us $\sigma_{max}/\tau_{max} = 3$, then if it is to show ductility it will be necessary that $\sigma_{cr}/\tau_{cr} > 3$, which means to say that the cleavage strength of the material of the bar must be more than three times its shear strength. If in a certain material σ_{cr}/τ_{cr} were to have the value 2.5 then that material would show ductility in the ordinary tensile test and brittleness in the impact test; the absorption of energy in the tensile test would then be relatively great but in the impact test relatively small (low impact value), because comparatively little energy is required to bring about the fracture itself.

If for a certain material subjected to the impact test the relation $\sigma_{cr}/\tau_{cr} > \sigma_{max}/\tau_{max}$ is satisfied, then there is still no certainty that a high impact value will be found, or, in other words, that considerable deformation will precede the fracture, since from what follows later it will be seen that σ_{cr}/τ_{cr} decreases as the deformation increases. If, therefore, at the beginning of the test this ratio is only little greater than σ_{max}/τ_{max} then fracturing will take place after comparatively little deformation, so that the impact value will be low, though higher than it would be in the case of fully brittle fracturing. For a metal like iron the quotient σ_{cr}/τ_{cr} also varies with the temperature, so that it depends to a high degree upon the temperature at which the test is carried out whether brittle fracturing will occur or not. We shall defer the discussion of this influence of temperature till later, when dealing with the experiments that we carried out.

Influence of oxygen and carbon on the impact value of iron at 20 °C

Much technical research work has already been done with regard to the impact value of iron and steel, but the influence of each impurity and each component of an alloy separately has not yet been investigated, nor even has the impact value of pure or practically pure iron ever been measured as a function of temperature. In the following pages, therefore, results will be given of some experiments relating thereto. The preparation of the iron and the various alloys needed for these experiments has already been described in this journal [2]).

The impact value of pure iron (carbon and nitrogen contents less than 0.001 wt %, oxygen content about 0.001%) at 20 °C was found to be 17 kgm/cm². Relatively small additions of oxygen were sufficient to reduce considerably the impact value at that temperature, as demonstrated by the lower curve in *fig. 3*, which shows, for instance, that an oxygen content of 0.018% caused the impact value to drop to 3.3 kgm/cm². On the other hand, such an oxygen content has but very little effect upon the behaviour of iron during the normal tensile test.

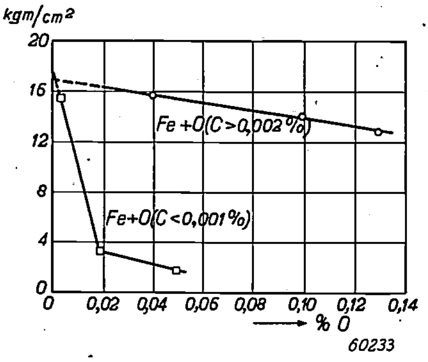


Fig. 3. Influence of the oxygen content of iron upon the impact value. When the iron contains less than 0.001 wt % carbon the impact value drops to low values upon the addition of only a few hundredths wt % of oxygen (lower curve). If, however, the iron contains a small quantity of carbon (e.g. 0.002%) the influence of the oxygen disappears almost entirely (upper curve). For those who prefer to use the Giorgi units, it is to be noted that 1 kgm/cm² ≈ 10^5 Nm/m².

The great influence of oxygen upon the impact value of iron points, according to what has been stated in the foregoing, to a lowering of the quotient σ_{cr}/τ_{cr} by oxygen. In many cases the fracture of impact-test bars containing oxygen was found to follow, for a considerable part, the crystal boundaries, so that presumably the cleavage strength σ_{cr} is reduced at the crystal boundaries. At first sight this seems to be peculiar, since a metallographic examination shows that the greater part of the oxygen in the metal is not along these boundaries but contained inside the crystals in the form of roughly spherical inclusions of iron oxide, and it is not to be supposed that these inclusions have any great effect upon the impact value. (It is known, for instance, that spherical inclusions of manganese sulphide have no appreciable effect upon impact strength, not even when they are present in large numbers.) Apparently it is to be assumed that some iron oxide is also present along the crystal boundaries, causing a considerable reduction of the intergranular cohesion. Remarkably enough, carbon has no great influence upon the impact value at 20 °C, even if this element is present at the crystal boundaries. This might be explained by assuming the oxide to be present as an almost continuous skin along the crystal boundaries and thus interrupting the metallic continuity over great distances, whilst the carbide may perhaps form isolated islands at the crystal boundaries. There are indications that something similar is

[2]) J. D. Fast, Philips Techn. Rev. 11, 241-244, 1950 (No. 8).

the case in molybdenum[3]), which has the same crystal structure as iron.

Still more remarkable is the fact that mere traces of carbon (0.002 wt % or more) already reduce very considerably the harmful effect of oxygen. This is demonstrated by the upper curve in fig. 3, which shows, inter alia, that when the metal contains only a few thousandths per cent of carbon even an amount of 0.13% oxygen only reduces the impact value of iron from 17 to 13 kgm/cm². It is not yet clear how this action of traces of carbon is to be explained.

Impact value as a function of temperature

For determining the impact value of iron as a function of temperature an iron was used that had not such a high degree of purity as that of the "pure" iron referred to above; it was found to take too long to remove completely the last 0.01% of oxygen, with the aid of hydrogen, from the rather large quantities of iron (about 2 kg) needed for such measurements. The oxygen remaining after a comparatively short reduction time was therefore chemically bound by adding to the liquid metal the calculated quantity of pure zirconium, pure titanium or pure aluminium (see the article quoted in footnote [2])).

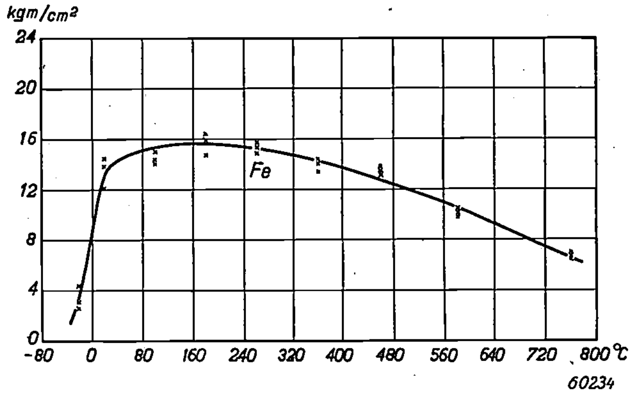


Fig. 4. Impact value of "pure" iron as a function of the temperature. The 0.012% oxygen contained in the iron has been bound with the equivalent quantity of zirconium to form zirconium oxide. There is a sharp rise in the impact-value curve at about 0 °C.

As an example *fig. 4* gives the impact value, as a function of temperature, of iron in which the remaining 0.012% oxygen (all other impurities having been removed) was converted into zirconium oxide with the aid of 0.035% zirconium. The zirconium oxide remained in the metal in a finely divided state.

As is seen from fig. 4, the impact-value curve shows a comparatively sharp transition from high to low values on the side of the low temperatures. For technical mild steel this sharp drop is well known. On the other hand the curve does not show the minimum that is known to occur with mild steel between 400 °C and 500 °C.

The drop round about 0 °C can be explained by the almost generally accepted assumption that as the temperature falls τ_{cr} increases at a relatively higher rate than σ_{cr}. As already mentioned, plastic deformation will take place so long as the relation $\sigma_{cr}/\tau_{cr} > \sigma_{max}/\tau_{max}$ is satisfied. According to the experiments (fig. 4) this is certainly the case for iron at high temperatures, but according to the above-mentioned assumption the ratio σ_{cr}/τ_{cr} decreases with falling temperature and ultimately becomes less than σ_{max}/τ_{max}. Roughly speaking, brittle fracturing will occur below the temperature at which the two relations become equal. The temperatures at which this will be the case will be all the higher as σ_{max}/τ_{max} is greater. Now from our previous discussions it appeared that this relation is much greater in the impact test than in the tensile test. Consequently the transition from deformation fracturing to brittle fracturing is to be expected at higher temperatures in the case of the impact test than in that of the tensile test. Such is indeed in agreement with what is found experimentally.

The quicker execution of the impact test compared with the tensile test is also to a small extent responsible for the fact that the transition range is shifted to higher temperatures, since the internal friction (τ_{cr}) increases with the rate of flow, so that under the conditions of the impact test one has to deal with smaller values of σ_{cr}/τ_{cr} than in the case of the tensile test.

The change from high to low impact values does not take place at any sharply defined transitional temperature but in a more or less wide range of temperatures. This can be explained by the fact that τ_{cr} increases with the degree of deformation, at least if it is permitted to assume (on which point there is no general agreement) that σ_{cr} changes less with deformation. Then, the ratio σ_{cr}/τ_{cr} decreases not only with falling temperature but also with increasing deformation. Consequently, at temperatures that are not much higher than the temperature T_k at which σ_{cr}/τ_{cr} and σ_{max}/τ_{max} in the non-deformed metal are equal, this equality will be reached and fracturing will take place after comparatively small deformations. According to these arguments, with rising temperature the impact value will change not abruptly but gradually from low to high values.

[3]) C. A. Zapffe, F. K. Landgraf and C. O. Worden, Metals Techn. **15**, T. P. 2421, Aug. 1948.

The fact that the transition range appears to be much narrower than is the case with ordinary tensile tests and with notched-bar bending tests carried out slowly can be understood, we believe, when it is borne in mind that the process of deformation in the impact test is almost adiabatic, owing to the test being carried out so quickly (time of deformation about 1/200 sec). As a result in the course of that process there is a not inconsiderable elevation of temperature in the deformation zone and this has an effect upon τ_{cr} opposed to that of the mechanical deformation. Therefore, at temperatures not much higher than the aforementioned temperature T_k larger deformations may take place than would be expected when not taking the adiabatic temperature increase into account. This results in a narrower transition zone (see *fig. 5*).

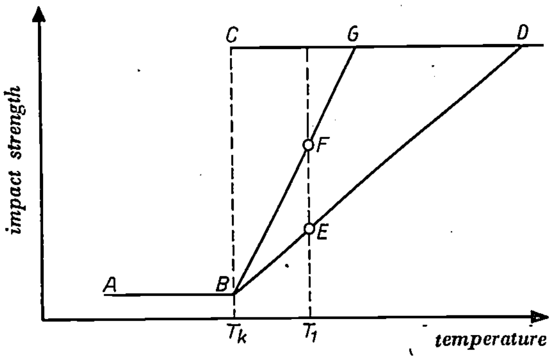


Fig. 5. If no account were taken of the effect of deformation then iron might be excepted to show high impact values at temperatures higher than a sharply defined temperature T_k and low values at temperatures below T_k. The lines *CD* and *AB* corresponding to these high and low values respectively have been drawn as horizontal straight lines only for the sake of clarity. Taking into account the effect of deformation (increase of τ_{cr}) it is understandable that at a temperature T_1 not much higher than T_k fracturing will take place after comparatively little deformation. The point *E* indicates the impact value corresponding to this little deformation. When, however, the adiabatic elevation of temperature is likewise taken into account then it is to be understood that the deformation preceding the fracturing will be greater, as a consequence of which the impact value will rise, for instance, to a level corresponding to the point *F*. This narrows down the transition range (in our simplified diagram from *CD* to *CG*).

In the case of "pure" iron according to fig. 4 the impact value remains almost constant between about 20 °C and 460 °C. This constant value amounts to approximately 14 kgm/cm^2, corresponding (given the dimensions of the test bar) to an impact work of 0.5×14 kgm $= 0.5 \times 14 \times 2.34$ cal $= 16.4$ cal. The specific heat of iron in the range from 20 °C to 460 °C averages about 0.13 cal degree^{-1} gram^{-1}. The deformed mass is of the order of 2 grammes. The average increase of temperature $\varDelta T$ in the deformed zone, disregarding for a first approximation the increase of the potential energy in the metal, therefore amounts to:

$$\varDelta T \approx \frac{16.4}{0.13 \times 2} \approx 60\ ^{\circ}\mathrm{C}.$$

Also the great spread of values in the transition zone as found experimentally can be understood to a certain extent, considering that a slight deformation taking place at the beginning of the impact increases the change of further deformation owing to the accompanying elevation of temperature. On the other hand, when a crack begins to form at the beginning of the impact this does not cause any rise in temperature and it will tend to propagate. Small statistical fluctuations will therefore be intensified by the temperature effect just discussed.

At the temperatures in question above the transition range, where considerable deformation precedes the fracturing, the impact value is determined almost exclusively by the shear strength and the deformed volume [4]). By "shear strength" it is not the initial value that is meant here but a mean value (taking into account the strain-hardening of the material during the deformation), whilst the expression "deformed volume" is to be understood as comprising not only the number of volume units over which the deformation extends but also the degree of the deformation of each part of this volume.

As the temperature rises from 20 °C to 460 °C the mean shear strength certainly decreases very considerably. The fact that in the case of "pure" iron the impact value in this temperature range nevertheless appeared to be almost constant is to be accounted for by the deformed volume increasing at the same time to such an extent that these two changes practically compensate each other. This increase of the deformed volume could be observed qualitatively. At temperatures above about 460 °C the reduction in the mean shear strength begins to predominate, and thus the impact value decreases. This decrease of the mean shear strength is due partly to the fact that at high temperatures the recrystallization (the formation and growth of new, non-deformed, crystals at the cost of the deformed metal) takes place so quickly that the strain-hardening of the metal during the deformation plays a smaller and smaller part until ultimately it is even entirely absent.

Influence of oxygen, carbon and nitrogen upon the trend of the impact-value curve

Oxygen

If the oxygen present in the iron is not bound, as was the case with the "pure" iron previously

[4]) This can be understood when it is borne in mind that the work W required to bend the bar can be represented by the symbolic equation $W =$ force $\times$ path.

dealt with, to an active element like zirconium, titanium or aluminium, then this element has a much more detrimental influence upon the mechanical properties of the metal. This is to be seen when comparing *fig. 6*, for iron containing 0.015% oxygen as the only impurity, with fig. 4 relating to iron containing 0.012% oxygen in the form of ZrO_2. The curve in fig. 6 follows the same trend as that of the curve in fig. 4 except that the transition range is shifted to higher temperatures. This agrees with the finding that oxygen present in the form of iron oxide reduces the σ_{cr}/τ_{cr} ratio of iron. Consequently as the temperature falls this ratio drops below the critical value at a higher value of the temperature. In view of the foregoing we presume the favourable effect of zirconium, aluminium, etc. to be due to these elements forming oxides which, contrary to the case with iron oxide, do not form continuous skins along the crystal boundaries.

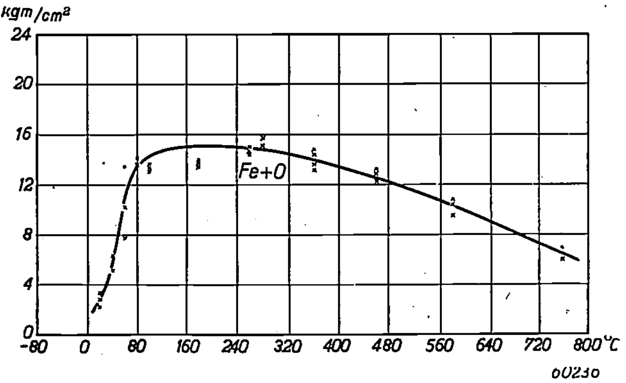


Fig. 6. Impact value of iron with 0.015% oxygen as the only impurity, as a function of temperature. (Here the oxygen is not bound as zirconium oxide.) The steep transition in the curve now lies at about 50 °C.

Carbon

Fig. 7 relates to an iron containing 0.020% carbon as the only impurity. (Here the oxygen is fairly easy to dispel, since when melting in vacuum it disappears in the form of CO.) Practically the same curve was found for iron with a carbon content four times as high. Contrary to what was found for "pure" and oxygenous iron, here the impact value between 20 °C and 580 °C still increases. Owing to more and more carbon passing into solution (solubility practically nil at 20 °C and equal to about 0.01% at 600 °C) apparently the shear strength decreases less quickly with rising temperature than it does in "pure" iron, so that the influence of the increasing deformed volume upon the amount of work taken up becomes predominant. Ultimately, above 580 °C the effect of recrystallization gains the upper hand.

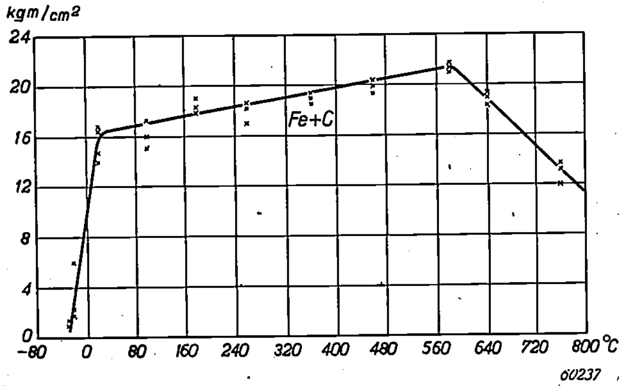


Fig. 7. Impact value of iron with 0.02% carbon as the only impurity, as a function of temperature. The steep transition lies at the same temperature as in fig. 4 but the impact value continues to rise, though slowly, at higher temperatures. Above 580 °C the curve descends again as a result of the rapid recrystallization taking place during the deformation in the impact test.

Nitrogen

Nitrogen is but very slowly absorbed by "pure" iron, whereas it appears to be more quickly absorbed by oxygenous liquid iron and still more quickly by carbonaceous liquid iron (see the article quoted in footnote [2])). The influence of nitrogen upon the temperature-dependency of the impact value of iron has therefore not been separately investigated but deduced from the behaviour of iron containing O + N and that of iron containing C + N.

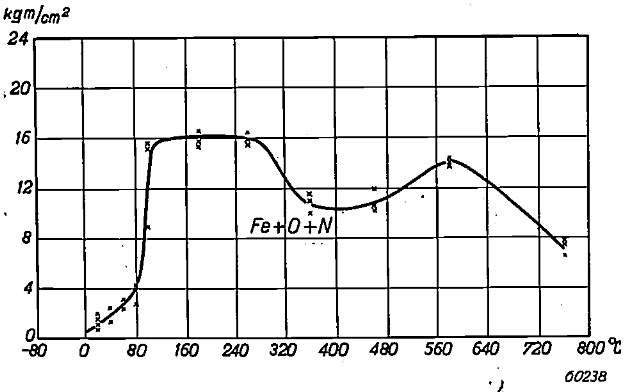


Fig. 8. Impact value, as a function of temperature, of iron containing 0.015% oxygen and 0.012% nitrogen as impurities.

Fig. 8 gives the impact value, as function of temperature, of iron containing 0.015% oxygen and 0.012% nitrogen, whilst *fig. 9* relates to an iron containing 0.06% carbon and 0.018% nitrogen. Both curves show a flat minimum between 360 °C and 460 °C, which is apparently to be ascribed to the presence of nitrogen, since oxygen and carbon, each separately, do not give rise to any such a minimum (cf. figs 6 and 7).

Although carbon and nitrogen, when present in steel as impurities, generally behave as equivalent elements, they appear to show a striking difference here. No entirely satisfactory explanation can yet be given for the occurrence of the minimum in the impact-value curve due to the presence of nitrogen.

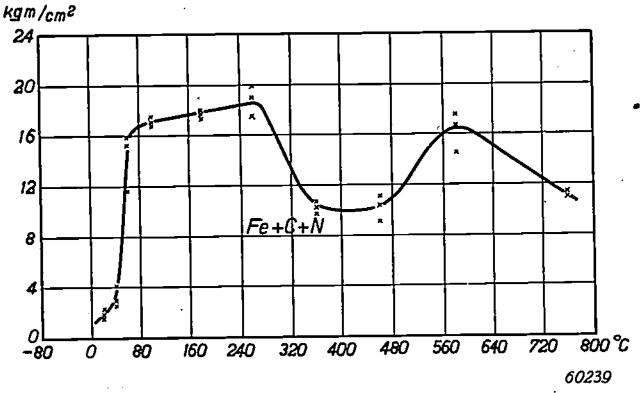


Fig. 9. Impact value, as a function of temperature, of iron containing 0.06% carbon and 0.018% nitrogen as impurities. As in the previous curve, a minimum occurs between 360 °C and 460 °C, which is to be ascribed to the presence of nitrogen.

Furthermore, experiments carried out with notched bars of technical mild steel (i.e. iron containing, among others, manganese, sulphur, phosphor, carbon and oxygen as admixtures) likewise showed a minimum in the impact-value curve, also in the case where the mild steel was entirely freed of nitrogen. Experiments with mild steel from which both the carbon and the nitrogen had been driven out gave a flat impact-value curve. Whereas, therefore, in the case of pure iron the presence of nitrogen seems to be essential for the occurrence of a minimum, in the case of mild steel apparently carbon is able to take over the part played by nitrogen.

Summary. The results are discussed of experiments carried out to investigate the notched-bar impact strength of pure iron and of iron to which impurities were purposely added. This is preceded by a description of the impact test according to the Charpy method and a review of the simplest concepts that can be formed according to known theories about the relation between the state of stress in a loaded metal and the possible occurrence of plastic deformation and/or rupture. According to this simplified picture, plastic deformation of the material takes place when under increasing load the shear strength of the metal is exceeded before the cleavage strength; in the opposite case the metal shows a brittle fracture. In the impact test the state of stress during the loading is such that there is a greater chance of brittle fracture than under the conditions of the ordinary tensile test. — It appears that oxygen reduces the impact value of iron at 20 °C considerably. Traces of carbon may greatly diminish this harmful effect of oxygen. As a result of the marked increase of shear strength with falling temperature (assuming the cleavage strength to remain constant) the impact value of pure iron when the temperature drops to around 0 °C shows the same transition from high to low values as is already known in the case of normal technical mild steel. Due to an adiabatic elevation of temperature taking place in the deformation zone during the bending of the test bar, this transition extends over a shorter temperature range than is the case with tensile or bending tests carried out more slowly. When oxygen is present in the iron the transition range is shifted to higher temperatures: oxygen reduces the cleavage strength of iron. Above the transition range carbon causes the impact-value curve to rise gradually with increasing temperature, whilst nitrogen gives rise to a flat minimum in the curve at temperatures of about 400 °C. Such a minimum is shown by the impact-value curve of mild steel, also when it does not contain any nitrogen. No really satisfactory explanation for this phenomenon can yet be given.

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

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

1870*: G. Diemer and K. S. Knol: Measurements on total-emission conductance at 35 cm and 15 cm wavelengths (Physica **15**, 459-462, 1949, No. 5/6).

From Q-measurements on a circuit containing a disc-seal diode, the electronic conductance was calculated and plotted as a function of $-V_a$. The experimental results do not at all agree with the linear field theory. The influence of space charge is discussed. In the space-charge-limited region the total emission conductance does not play such an important part as might be excepted from an extrapolation of the *g*-values measured beyond cut-off.

1871*: K. S. Knol and A. Versnel: Suppression of shot effect noise in triodes and pentodes (Physica **15**, 462-464, 1949, No. 5/6).

It is shown experimentally that the presence of a capacitance between the control grid and the anode results in a much larger suppression of shot effect noise by detuning, as suggested by Strutt and Van der Ziel, than is possible in the absence of such a capacitance, as may be expected from a theoretical point of view.

1872: J. de Jonge and R. Dijkstra: The photo-decomposition products of some diazonium salts (Rec. Trav. chim. Pays-Bas **68**, 424-429, 1949, No. 6).

By irradiation of solutions of some benzene diazonium salts (including one p-hydroxybenzene diazonium salt) in water the corresponding phenols could be isolated with a high yield. The photo-decomposition of two p-diazo-oxides seems to have a more complicated character.

1873: J. de Jonge, R. Dijkstra and P. B. Braun: The thermal decomposition of o-hydroxy-diazonium compounds (Rec. Trav. chim. Pays-Bas **68**, 431-432, 1949, No. 6).

Experimental evidence has been found that the thermal decomposition of o-hydroxy-diazonium salts may be identical with the photo-decomposition.

1874/75: B. H. Schultz: On the application of Reynolds' analogy and the heat-exchange factor to the design of heat exchangers, I-III (Appl. sci. Res. **A1**, 387-416, 1949, No. 5/6).

Part. I. Reynolds's analogy between heat transfer and friction leads to the well-known equation $h = \frac{1}{2} fC_p\varrho v$. According to improved theories, for turbulent flow of gases in long straight ducts, the factor $\frac{1}{2}$ is to be replaced by $c = 0.55$. It is shown that for linear flow in long ducts c is equal to 0.3 for a circular and 0.4 for a flat rectangular cross-section. Further, in very short ducts an approximate theory is shown to give values that are of the same order of magnitude. Though the equation $h = cfC_p\varrho v$ has an approximative character, it has the advantage of being simple and generally valid for gases flowing in ducts.

Part II. For heat exchangers in which the temperature differences between gas and wall decrease considerably it proves to be useful to introduce a "heat-exchange factor" K. A simple relation between K and the pressure drop can be derived from the modified form of Reynolds's analogy given in part I. This relation is applied to some problems of optimum dimensions for heat exchangers.

Part III. The value of c in the relation between K and the pressure drop for turbulent flow is taken from the experimental data given in the literature. The author's measurements in the transition region show that, though both heat exchange and friction may vary considerably with changing flow conditions at (or before) the entrance of the ducts, the factor c does not vary appreciably.

In those cases where the curve for f shows a pronounced "dip" this is also found in the curve for the heat exchange factor.

1876: J. A. Haringx: Elastic stability of helical springs at a compression greater than the original length (Appl. sci. Res. **A1**, 418-434, 1949, No. 5/6).

A former statement that, as to the behaviour of helical compression springs in respect of elastic stability, it is quite immaterial whether the ends of the springs are hinged, constrained parallel or clamped, appears to be true only for compressions less than the original spring length. For greater compressions, on the other hand, the differences occurring for the various end conditions are found to be quite characteristic. Since a (cylindrical) spring could be realized with a compression greater than the original length, simply by turning it inside out, the new theoretical results found for this region of compressions could be verified by experiment. These experiments demonstrated the remarkable toppling over of the coils which accompanies the transition from the stable to the unstable state. Moreover they confirmed the existence of the characteristic differences in the behaviour of the spring for the various end conditions, though qualitatively only. An explanation for the numerical deviations occurring is mentioned.

1877: A. Bremmer: The propagation of electromagnetic waves through a stratified medium and its W.K.B. approximation for oblique incidence (Physica **15**, 593-608, 1949, No. 7).

The plane-wave solution of Maxwell's equations for a stratified medium is split into a series of terms with a simple geometrical .meaning, the first of which constitutes the W.K.B. approximation. The original vector problem is reduced to a scalar problem by introducing a convenient hertzian vector. The application of the saddle-point method to the individual terms of the series mentioned leads to simple geometric-optical approximations.

1878: F. A. Kröger: Sodium and lithium as activators of fluorescence in zinc sulphide (J. Opt. Soc. Amer. **39**, 670-672, 1949, No. 8).

Lithium and sodium dissolved in zinc sulphide act either as activators of fluorescence or as quenchers, according to whether chlorine is present or not. The fluorescence occurs only at low temperatures and is excited by short-wave ultraviolet. The lithium band has a maximum at 4380 Å in the sphalerite modification of ZnS; the sodium band has a maximum at 3940 Å in the sphalerite and at 3800 Å in the wurtzite modification.

The difference in the position of the bands for wurtzite and sphalerite and a shift found upon incorporation of CdS proves that we are dealing with electron-transfer bands. The centers of fluorescence are supposed to consist of lithium or sodium and chlorine ions occupying normal lattice sites. The levels involved in the fluorescence transitions are levels due to lattice ions, changed by the presence of the monovalent ions (indirect activation).

The quencher centers are supposed to consist of Li_2S and Na_2S incorporated in ZnS in such a way that the alkaline ions occupy two cation sites, while of two anion sites one is occupied by a monovalent sulfur ion and the other by an electron.

1879: G. W. Rathenau and H. de Wit: Indicator for small amounts of oxygen in reducing furnace atmospheres (Metallurgia **40**, 114, 1949, June).

When annealing metals or alloys in pure hydrogen a rough determination of the oxygen pressure of the protecting gas is often desirable. This is done by means of an electrically heated Cr-Ni-steel strip, which shows surface oxidation below the temperature at which the oxygen pressure of the atmosphere equals that of the chromium oxide.

VOL. 11 No. 11, pp. 313-340 MAY 1950

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

GYROMAGNETIC PHENOMENA OCCURRING WITH FERRITES

by H. G. BELJERS and J. L. SNOEK.

538.245:538.69:621.3.017.39

Non-metallic magnetic materials (ferrites marketed by Philips under the trade name "Ferroxcube") have recently attracted attention on account of their favourable properties, such as low eddy-current losses, hysteresis losses and after-effect losses. The upper limit of the useful frequency range, from the point of view of losses, varies between 10^5 and 10^8 c/s according to the material. When this limit is exceeded losses arise which increase rapidly with the frequency. It is probable that these additional losses are to be ascribed to gyromagnetic phenomena inherent in all magnetic materials. The theoretical and practical significance of these phenomena is explained.

When a ferromagnetic material is placed inside a coil carrying a constant current, then, as is known, the magnetic flux through the coil is increased. Upon the coil being connected to a source of alternating current the inductance is found to have increased. As a rule, however, even when the resistance of the coil may be ignored, the alternating voltage across the coil is no longer exactly in quadrature with the alternating current, owing to the losses occurring in the material.

These losses may in the first place be due to eddy currents in the ferromagnetic material, in which case one speaks of eddy current losses. These increase with the frequency and depend not only upon the properties of the material (conductivity and permeability) but also upon its dimensions [1]).

By making a suitable choice of the dimensions (lamination) and of the properties of the material the eddy current losses can often be made negligible, but still other losses remain which, apart from the frequency, depend only upon the properties of the material and not upon the dimensions. First of all there are the hysteresis losses, which are connected with the fact that the induction B, plotted as a function of the field strength H, shows what is called a hysteresis loop. The hysteresis losses per cycle and per unit of volume of the material are equal to the area of the hysteresis loop.

In the second place there are the magnetic after-effect losses. These losses have been dealt with at length in an article by Snoek and Du Pré [2]). They are usually small and are presumably to be regarded as connected with the gyromagnetic losses dealt with in the present article.

Let the magnetic alternating field in the material be $H = H_m \cos \omega t$. Owing to after-effect and hysteresis [3]) the induction B will be shifted in phase with respect to H, so that we may write $B = B_m \cos (\omega t - \delta)$. Applying the usual complex method of expression, these formulae become:

$$H = H_m \,\mathrm{e}^{\mathrm{j}\omega t} \quad \text{and} \quad B = B_m \,\mathrm{e}^{\mathrm{j}(\omega t - \delta)} \,. \quad . \quad (2)$$

Thus the permeability μ, which equals the quotient B/H, is complex, and naturally the relative permeability μ_r is likewise complex ($\mu_r = \mu/\mu_0$, where $\mu_0 = 4\pi \times 10^{-7}$ H/m). Thus for μ_r we have the formula

$$\mu_r = \frac{B}{\mu_0 H} = \frac{B_m \,\mathrm{e}^{-\mathrm{j}\delta}}{\mu_0 H_m} = \frac{B_m}{\mu_0 H_m} (\cos \delta - \mathrm{j} \sin \delta).$$

Denoting $(B_m/\mu_0 H_m) \cos \delta$ by μ_r' and $(B_m/\mu_0 H_m)$

[1]) See, e.g., J. L. Snoek, Philips Techn. Rev. 2, 77-83, 1937.

[2]) J. L. Snoek and F. K. Du Pré, Several after-effect phenomena and related losses in alternating fields, Philips Techn. Rev. 8, 57-64, 1946.

[3]) In the case of hysteresis the induction is not purely sinusoidal as a function of time; higher harmonics also occur. As regards the losses, however, we have only to do with the fundamental oscillation, and this shows a phase shift with respect to H.

sin δ by μ_r'' we can write for the complex (relative) permeability:

$$\mu_r = \mu_r' - j\mu_r'' \quad . \; . \; . \; . \; . \; . \; . \quad (3)$$

The absolute value $|\mu_r|$ is $\sqrt{\mu_r'^2 + \mu_r''^2}$; when δ is a small angle then $|\mu_r| \approx \mu_r'$. Further, for the loss angle δ we have the formula $\tan\delta = \mu_r''/\mu_r'$.

Attention has recently been drawn to a certain kind of non-metallic magnetic materials, namely the ferrites, marketed by Philips under the trade name "Ferroxcube". These materials have a low conductivity ($\sigma = 10^{-3}$ to 10^{-5} Mho/m ($=$ ohm^{-1} m^{-1})), and consequently the eddy-current losses are negligibly small, whilst also the hysteresis losses are generally relatively small. In a large frequency range (upper limit 10^5—10^8 c/sec, according to the material) the values of the quantity $\tan\delta$ are therefore quite small (e.g. 0.01). The losses occurring in this frequency range are to be ascribed to after-effect. However, in contrast to other cases where this has been thoroughly investigated (see footnote [2])), little is as yet known with certainty regarding the nature of this after-effect in ferrites. When the limit just mentioned is exceeded, in the case of ferrites there is a considerable increase in the value of μ_r'', accompanied by a decrease of μ_r', as is to be seen from *fig. 1* where the quantities μ_r' and μ_r'' have been plotted for a number of kinds

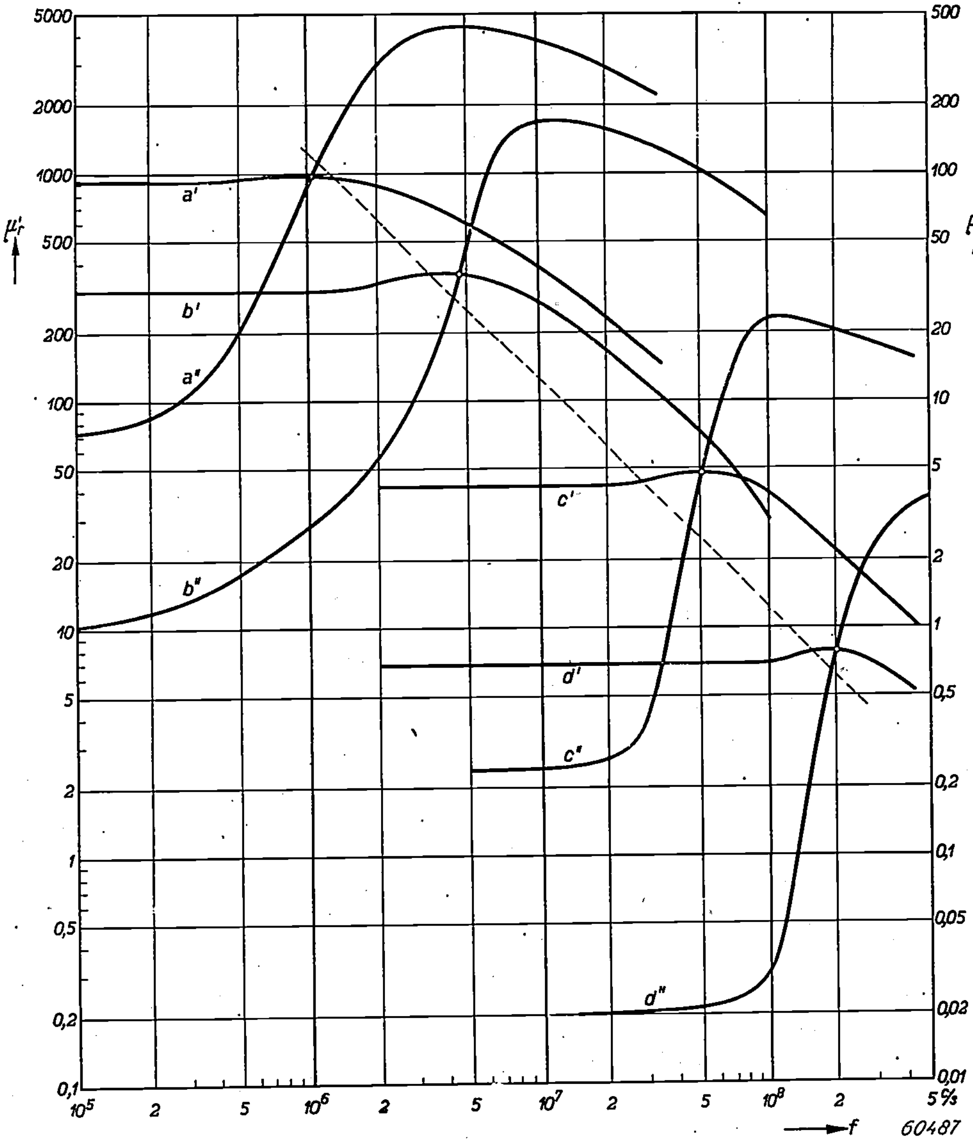


Fig. 1. The quantities μ_r' (left-hand scale) and μ_r'' (right-hand scale) for a manganese-zinc ferrite (a) and for a number of nickel-zinc ferrites of different composition (b-d). The small circles denote the points where $\tan\delta = 0.1$. The dotted line indicates the mean relation between the value of μ_r' (in the low-frequency range) of a certain ferrite and the critical frequency setting a limit to the range within which that ferrite can be used in filter coils ($\tan\delta = 0.06$).

of "Ferroxcube" as functions of frequency on a logarithmic scale [4]).

In the case of filter coils the critical frequency is generally assumed to be that at which tan δ reaches the value 0.06. From fig. 1 it appears that this critical frequency increases according as μ_r' in the low-frequency range is smaller. For application in transformers the material can be used at much higher frequencies.

The sharp increase of tan δ at high frequencies depends upon various factors. For instance, if the material shows internal stresses the increase of tan δ begins at lower frequencies than is the case with the same material free of internal stresses. After annealing, the high values of tan δ have shifted to higher frequencies. When a new material is prepared by sintering together a mixture of different ferrites it is often found that the large losses appear at lower frequencies, this being particularly the case if the ferrites have not been intimately mixed prior to the sintering; more thorough mixing shifts the region of high losses to higher frequencies.

It appears that the cause of the additional losses found in ferrites at very high frequencies is to be sought in a phenomenon theoretically predicted by Landau and Lifshitz in 1935, when no practical examples of it were known to those authors. They called this phenomenon gyromagnetic resonance [5]). The nature of this gyromagnetic resonance, as subsequently found in various magnetic materials, can be explained with the aid of a simple mechanical model.

A model illustrating gyromagnetic resonance

According to the modern theory of matter, magnetism is ascribed mainly to the fact that an electron may be regarded as a charged sphere rotating about an axis and thus showing a magnetic moment, or "spin". In a magnetic field this spin tends to follow the direction of the field and endeavours to take up a position where the potential energy is a minimum, just as is the case, for instance, with a pendulum in a gravitational field.

In addition to being charged, however, the electron also has mass, and as a result of its behaving like something that rotates it is not only a carrier of a magnetic moment but also a carrier of a mechanical angular momentum.

The picture of the pendulum, therefore, has to be completed by imagining a spinning-top to be contained in the body of the pendulum with its axis of rotation directed towards the point of suspension. Such a pendulum is schematically represented in *fig. 2a*. This "spinning-top pendulum" behaves quite differently from an ordinary pendulum, in that when the pendulum arm is moved an angle ϑ away from the vertical and then released, the pendulum does not begin to oscillate in the normal way in a plane determined by the starting position and the vertical through the point of suspension. Instead of that, the released pendulum arm begins to describe the surface of a cone with ϑ as half the apex angle (fig. 2*b*), each point along the arm thereby describing a horizontal circle. The spinning-top pendulum is then said to make a precessional movement.

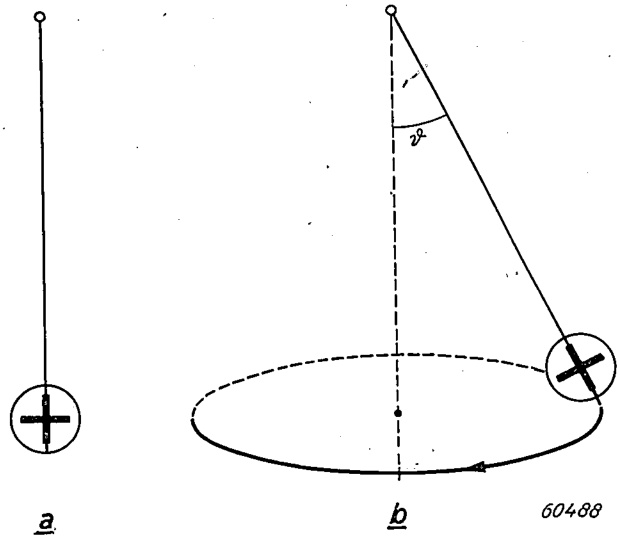


Fig. 2. Spinning-top pendulum (*a*) in the position of minimum potential energy and (*b*) making a precessional movement at an angle ϑ with the vertical.

The angular frequency (= angular velocity) of this precession is

$$\omega_p = M/J , \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

where $M = Gl$ (G = weight of the pendulum, l = distance from the centre of gravity to the point of suspension; the couple driving the pendulum towards the position of equilibrium is therefore $Gl \sin \vartheta = M \sin \vartheta$). Further, $J = I_r \omega_r$ and is the angular momentum of the top spinning around its axis (I_r = moment of inertia about the axis of rotation, ω_r = the angular velocity of that rotation). It is seen that the angular velocity ω_p is independent of the angle of deflection ϑ.

[4]) The data of fig. 1 are taken from unpublished measurements by C. M. van den Burgt and M. Gevers. See also J. L. Snoek, Philips Techn. Rev. 8, 353-360, 1946.

[5]) L. Landau and E. Lifshitz, Phys. Z. Soviet Union 8, 153-169, 1935. J. H. E. Griffiths, Nature **158**, 670, 1946. C. Kittel, Phys. Rev. 71, 270-271, 1947. See also D. Polder, Physica The Hague 15, 253-255. 1949 (Nos 1/2) and Phil. Mag. London 40. 99-115, 1949 (No. 1).

Obviously what has been stated above no longer holds when ω_r, and thus J, is made to approach zero, since in that case the top comes to a standstill and the spinning top pendulum is no longer to be distinguished from an ordinary pendulum. The manner in which the movement of the released pendulum changes into the precessional movement as the value of J increases is to be seen from *fig. 3*, where *a* represents the movement (seen from above) of a point along the pendulum arm when $J = 0$. If the value of J is small then a figure will be described as indicated by *b*. As J increases *b* changes into *c*, until eventually a sort of cycloid like that at *d* is described which ultimately becomes the precessional movement along the circle.

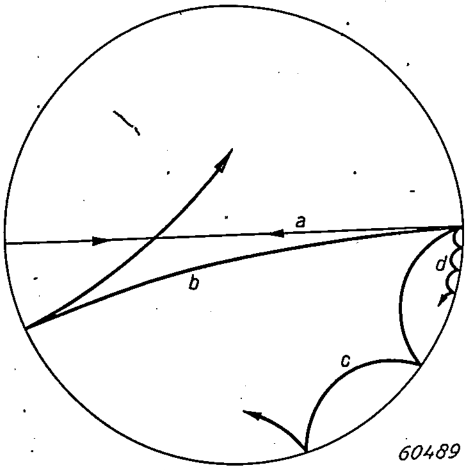


Fig. 3. Transition from the normal pendulum movement to the precessional movement as the speed of rotation ω_r of the spinning top increases.

From what has been said it will be evident that the precessional movement of the spinning-top pendulum must not be confused with the movement made by a so-called "conical pendulum"; in fact this has an entirely different frequency, viz. the same as that of an ordinary pendulum (case *a* in fig. 3).

In the transition from *a* to *d*, not only the quantity J but also the mass of the pendulum itself plays a part. But since the mass of the electron as such is of no influence it is assumed that, with our model, the above-mentioned limit has been reached.

Let us suppose that the top is rotating with its axis vertical, thus in the lowest position (as in fig. 2*a*), and that we then suddenly turn the direction of the force of gravity over a small angle ε in the plane of the drawing (*fig. 4*). It is clear that at that moment a precessional movement will begin with ε as half the apex angle and 0 and $+ 2\varepsilon$ as the extreme positions. When, after half a cycle has been completed, we suddenly cause the force of gravity to make an angle $-\varepsilon$ with the vertical, the half apex angle becomes $3\,\varepsilon$ and the extreme positions $+2\varepsilon$ and $-4\,\varepsilon$. Upon the force of gravity being caused to return to the first position ($+\varepsilon$) after another half cycle, the half apex angle becomes 5ε. Continuing in this way, at the n^{th} half cycle the apex angle becomes $(2n-1)\varepsilon$.

Thus we see that, when the force of gravity is caused to oscillate in the manner described at a rate determined by the precession, a uniform increase of the precession takes place which is to be compared to a "resonance", especially when the deflection in the plane of the interference is considered. In the case we have just been dealing with, where no account has been taken of a possible damping, the deflection may ultimately become very great, however small ε may be. As will be explained below for the case of a sinusoidal interference, for oscillations of the force of gravity at a frequency differing from ω_p one always finds a finite deflection.

Gyromagnetic resonance of electrons

The mental experiment that we have just made is not easy to carry out mechanically because we cannot govern the force of gravity, but in the magnetic case, with the means that are now at our disposal, it is fairly easy to perform.

Imagine that we have a free-spinning electron, the angular momentum of which is

$$J = \tfrac{1}{2}\,(h/2\pi)\,,$$

where h represents Planck's constant ($h = 6.6 \times 10^{-34}$ kg·m²/s). This is accompanied by a magnetic moment (Bohr magneton)

$$\mu_B = \frac{e}{m_0}\,J = \frac{eh}{4\pi m_0} \quad . \; . \; . \; . \; . \quad (5)$$

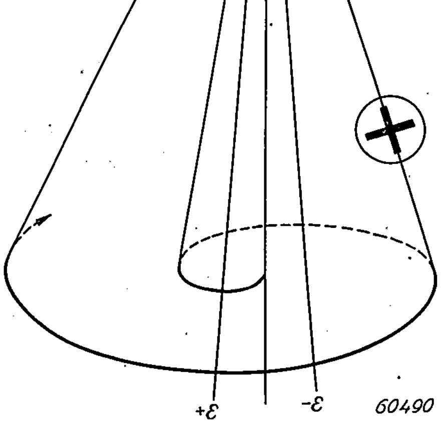


Fig. 4. Behaviour of the spinning-top pendulum when the direction of the force of gravity is changed step by step to make the angles $+\,\varepsilon$ and $-\,\varepsilon$ with the vertical, at the rate of the precession.

(e = charge of the electron = 1.6×10^{-19} Asec, m_0 = mass of the electron at rest = 9×10^{-31} kg, $eh/4\pi m_0 = 9.3 \times 10^{-24}$ A·m^2).

Let us suppose that this minute electron magnet is placed in a constant magnetic field H. It will then first try to point with its north pole in the direction of H, thus tending to take up a position where the momentum is the minimum. The situation would then correspond to that of the spinning-top pendulum in fig. 2*a*.

We know, however, that when the axis of the spin makes an angle ϑ with the field the electron comes under the influence of a couple $M \sin \vartheta = \mu_B \cdot \mu_0 H \sin \vartheta$, as a result of which, like the case with the spinning-top pendulum in fig. 2*b*, the axis describes a precessional movement about the direction of H with an angular velocity

$$\omega_p = \frac{M}{J} = \frac{e}{m_0} \mu_0 H \quad . \; . \; . \; . \; . \; . \quad (6)$$

Now it is easy to cause the direction of the magnetic field to oscillate over a small angle ε_1. All that need be done is to apply an alternating field

$$h\ (t) = \varepsilon_1 H \cos \omega t$$

perpendicular to H.

Calculation for a sinusoidal interference

Let us go back for a moment to the model of the spinning-top pendulum. We take the direction of the vertical as the z-axis of a rectangular system of coordinates and let the force of gravity oscillate in such a way that the z-component is a constant (value g) and the y-component remains zero, whilst the x-component, $u(t)$, becomes:

$$u\ (t) = \varepsilon_1 g \cos \omega t \qquad (\varepsilon_1 \ll 1)\ .$$

If, now, ξ and η represent the angles enclosed by the z-axis and the projections of the axis of the spinning-top on the XOZ and YOZ planes respectively (*fig.* 5) then as long as $\xi \ll 1$ and $\eta \ll 1$, and in the absence of frictional forces, we have

$$\frac{d\xi}{dt} = -\omega_p\, \eta\,, \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad (7)$$

$$\frac{d\eta}{dt} = \omega_p\, (\xi - \varepsilon_1 \cos \omega t) \quad . \; . \; . \; . \; . \quad (8)$$

These equations signify that the point (ξ, η), the point where the pendulum arm intersects a horizontal plane at a distance 1 below the point of suspension, is describing a circle around the point $(\mu, 0)$, i.e. the point where the direction of the force of gravity through the point of suspension intersects the said plane, and at a constant angular velocity ω_p. The point (ξ, η) must thereby always "follow" the variable point $(\mu, 0)$.

The stationary solution of the equations (7) and (8) reads:

$$\xi = \frac{\omega_p{}^2}{\omega_p{}^2 - \omega^2}\, \varepsilon_1 \cos \omega t, \quad . \; . \; . \; . \quad (9)$$

$$\eta = \frac{\omega_p \omega}{\omega_p{}^2 - \omega^2}\, \varepsilon_1 \sin \omega t\,. \quad . \; . \; . \quad (10)$$

Thus the axis of the spinning top describes an elliptical cone, the apex angles of which are greater according as the difference between ω and ω_p is smaller. When the value of ω lies very close to ω_p these formulae no longer hold, because then the values found for ξ and η are no longer $\ll 1$. It is to be noted that ξ is in phase with the interference $\varepsilon_1 \cos \omega t$, as a result of which no work will be performed by the alternating field on the spinning top. Thus there is no dissipation of energy, in accordance with the assumption made that frictional forces are absent.

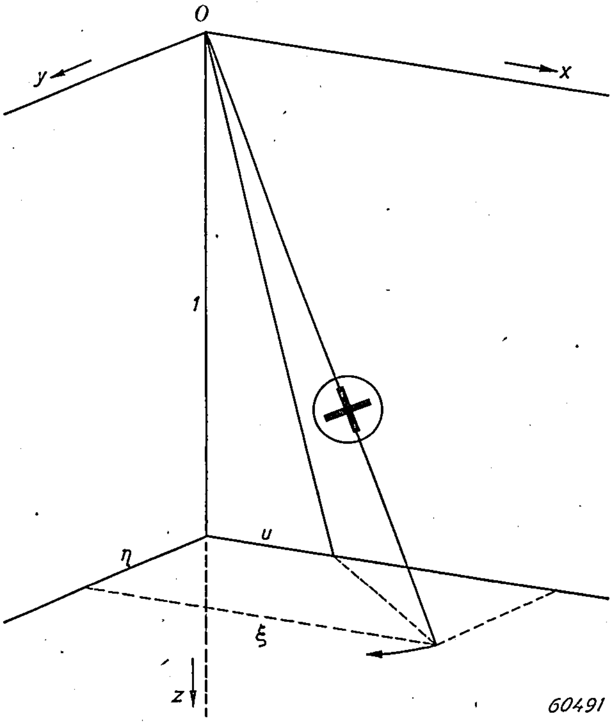


Fig. 5. The behaviour of the spinning-top pendulum when the direction of the force of gravity oscillates sinusoidally between $+\varepsilon_1$ and $-\varepsilon_1$.

When, in the casc of the electron, we take H parallel to the z-axis of a rectangular system of coordinates and h parallel to the x-axis, a calculation shows that the components μ_x and μ_y of the magnetic moment of the electron in the x- and y-directions respectively are given by:

$$\frac{\mu_x}{\mu_B} = \frac{\omega_p^2}{\omega_p^2 - \omega^2}\,\varepsilon_1 \cos \omega t \quad . \; . \; . \quad (11)$$

$$\frac{\mu_y}{\mu_B} = \frac{\omega_p \omega}{\omega_p^2 - \omega^2}\,\varepsilon_1 \sin \omega t \, . \quad . \; . \; . \quad (12)$$

In these expressions we recognize the equations (9) and (10). Naturally, limitations similar to those applying in the case of the mechanical model also apply here ($\mu_x/\mu_B \ll 1$, $\mu_y/\mu_B \ll 1$, absence of damping).

From the formulae (11) and (12) it appears that under the given conditions an alternating magnetic field directed along the x-axis causes not only an alternating magnetic moment directed along the x-axis and in phase with the field, but also an alternating magnetic moment which is directed along the y-axis, thus perpendicular to the magnetic alternating field, and which lags 90 ° in phase behind that field.

Experimental evidence for the precessional movement

The following experiment [6]) directly proves the existence of the magnetisation component μ_y.

An arrangement of wave guides forming a so-called "magic tee", as indicated in *fig. 6*, is placed in an adjustable constant magnetic field H_z. A sphere of ferrite material is placed inside the wave guide *a-b*, in the origin of the system of coordinates *x*, *y*, *z*. A microwave ($\lambda = 3.2$ cm, $\omega = 2\pi \times 9.4 \times 10^9$ rad/s) enters at *c*.

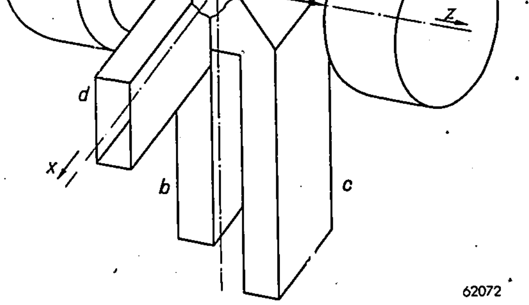


Fig. 6. The wave guides *a*, *b*, *c* and *d*, forming a so-called "magic tee", are placed in an adjustable constant magnetic field H_z. At the origin of coordinates a small sphere of ferrite is placed. A microwave enters at *c*. The fact that a wave is observed in *d* proves the existence of the magnetisation component in the "third direction" μ_y in the ferrite material.

Let us first consider the case $H_z = 0$. Provided a rigorous symmetry of the whole arrangement is established (e.g. by adjusting the position of the sphere), the wave in *c* splits into equal waves travelling in *a* and *b*, but no wave is observed in the limb *d*. This is due to the fact that the possible mode of a wave in *d* can be launched only by an oscillating magnetic field component in the *y*-direction, while the arriving wave from *c* does not possess such a component. A wave in *d* is found, however, when the constant field H_z is given a finite value. This is explained by the gyromagnetic effect in the ferrite sphere, giving rise to a magnetization component in the "third direction", μ_y.

When H_z is varied the output measured in *d* shows a rather sharp maximum at $H_z = m_0/e\,\mu_0 \cdot \omega$, which is in agreement with eq. (6).

The formulac (11) and (12) derived above apply for a single, isolated electron and also, as the theory further shows, for a homogeneous, spherical, ferromagnetic body magnetized to the point of saturation by an external magnetic field. This is subject to the condition that the dimensions of the sphere are small in relation to the wavelength $\lambda = 2\pi v/\omega$, where $v = c/\sqrt{\varepsilon\mu}$ represents the velocity of propagation of electromagnetic waves in the ferromagnetic medium.

In a magnetic material of less symmetrical shape, for instance an ellipsoid, a somewhat different result is found for ω_p, namely:

$$\omega_p = (e/m_0)\,\mu_0\,[H + (N_x - N_z)M]^{1/2}\,[H + (N_y - N_z)M]^{1/2},$$

where M represents the magnetic moment of the body per volume unit (in A/m) and N_x, N_y and N_z are numerical factors determining the demagnetizing forces in the three principal directions. For a sphere $N_x = N_y = N_z = 1/3$, so that we again obtain formula (6) unchanged.

Influence of damping

Just as in the discussion of the case of exact resonance, in the formulae (9) and (10) no account has been taken of the damping which always exists in nature. Let us consider the case of a spinning-top pendulum making a precessional movement about the direction of minimum energy at an angle ϑ to the vertical, as in fig. 2*b*, and let there be a force which does not influence the speed of rotation of the spinning top but which tends to retard the movement of the pendulum — just as it would do in the case of an ordinary pendulum — and which is, let us say, proportional to the (precessional) velocity. Obviously such a frictional force reduces the energy of the pendulum and, since we are assuming that the spinning top itself continues to rotate at an undiminished velocity, as a result the centre of gravity must in time fall,

[6]) H. G. Beljers, Physica The Hague, **16**, 75-76, 1950 (No. 1)

whilst further, in the absence of any external interference, the precession gradually comes to an end and the pendulum takes up its lowest position (as in fig. 2a).

In the case of an external interference $\varepsilon_1 \cos \omega t$ as dealt with above, such a frictional force will likewise make itself felt. It will result in the ultimate amplitude of the precession being smaller than the value given by (9) and (10) and the x-component lagging behind the interference. This gives rise to a continuous dissipation of energy, which, if the friction in question were caused, for instance, by a viscous medium, would result in that medium being heated.

In particular, in the case of exact resonance ($\omega = \omega_p$) the expression for the amplitude of the precession will no longer become infinite but will ultimately assume a finite value, so that the half-value width [7]) of the precession, which in the absence of damping is nil, has a value differing from zero.

Just as with the spinning-top pendulum, also in the case of the electron a damping force leads to a reduction of the amplitude and at the same time a relative widening of the magnetic resonance interval, accompanied by dissipation of energy. Such a damping cannot arise when an electron is free but it may well occur in the case of electrons forming part of a crystal. It will result, on the one hand, from interaction with the free electrons in the material and, on the other hand, from an interaction between the electron in question and the ions of the crystal lattice. The energy dissipation accompanying this leads to the crystal being heated (magnetic losses). Although, therefore, we have regarded the damping here more or less as a secondary phenomenon, it is of very great importance in practice.

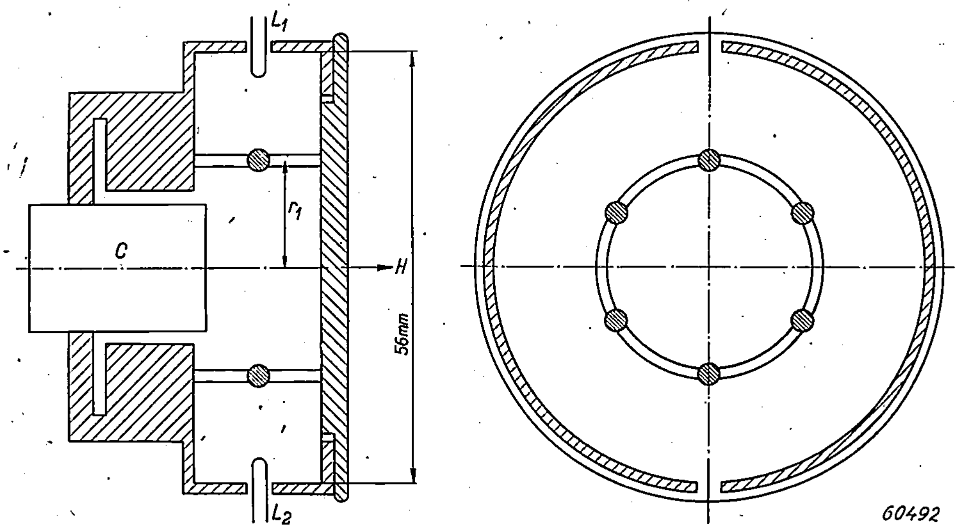


Fig. 7. Box-shaped cavity resonator for studying gyromagnetic resonance. The ferrite is applied in the form of small spheres on a cylinder of insulating material. The movable metal cylinder C serves for tuning. One of the two loops, L_1, serves for generating the oscillation energy and the other, L_2, for measuring it.

Further experimental study of magnetic resonance

For a closer investigation of the consequences of gyromagnetic resonance of ferrites, experiments were carried out with a cylindrical cavity resonator (*fig.* 7) containing small samples of ferrites at given places[8]). The magnetic material is introduced in the form of small spheres placed on a circle situated in a plane perpendicular to the axis of the box and at half its height. With the aid of a coupling loop passed through an opening in the side of the box it is possible to generate in the cavity resonator a standing electromagnetic wave of a frequency corresponding to a certain form of oscillation of the cavity resonator. The mode of oscillation and the radius of the circle mentioned above in relation to the dimensions of the cavity resonator can be so chosen that the magnetic vector h of the alternating field at the place where the spheres are situated lies in the plane of the circle and is tangential thereto, whilst

[7]) The half-value width is the difference between the two frequencies at either side of the resonance frequency for which the energy equals half the energy in the point of resonance.

[8]) H. G. Beljers, Measurements on ferromagnetic resonance using cavity resonators, Physica The Hague 14, 629-641, 1948 (No. 10).

at the same time the electric vector at those points is just zero or approximately zero.

A constant magnetic field H is applied parallel to the axis of the box. Since the frequency of the cavity resonator, partly in connection with the above-mentioned requirements, is fixed (in the case investigated it was $f = 10^{10}$ c/s, $\omega = 2\pi f = 6 \times 10^{10}$ rad/sec), H has to be so chosen that according to formula (6) ω_p comes to lie round about ω (in our case $\omega_p = \omega$ when $H = 3 \times 10^5$ A/m or $\mu_0 H = 0.4$ Wb/m^2).

In the bottom of the box is a movable, threaded metal cylinder C, by means of which the cavity resonator, the natural frequency of which depends partly upon the presence of the magnetic material, can be brought to resonance at the fixed frequency ω. The presence of such a resonance can be determined with the aid of a second coupling loop inserted through an opening in the side of the box in the same way as is done with the driving loop. The influence of the presence of the ferrite in the box is then studied by first tuning the empty cavity resonator to the fixed frequency ω and then determining how far the cylinder C has to be moved in order to bring the cavity into resonance again at the frequency ω after the ferrite has been put in place and the field H imposed. At the same time measurements are taken, in the known manner, of the resonance width, and thus of the reduction in quality of the cavity resonator.

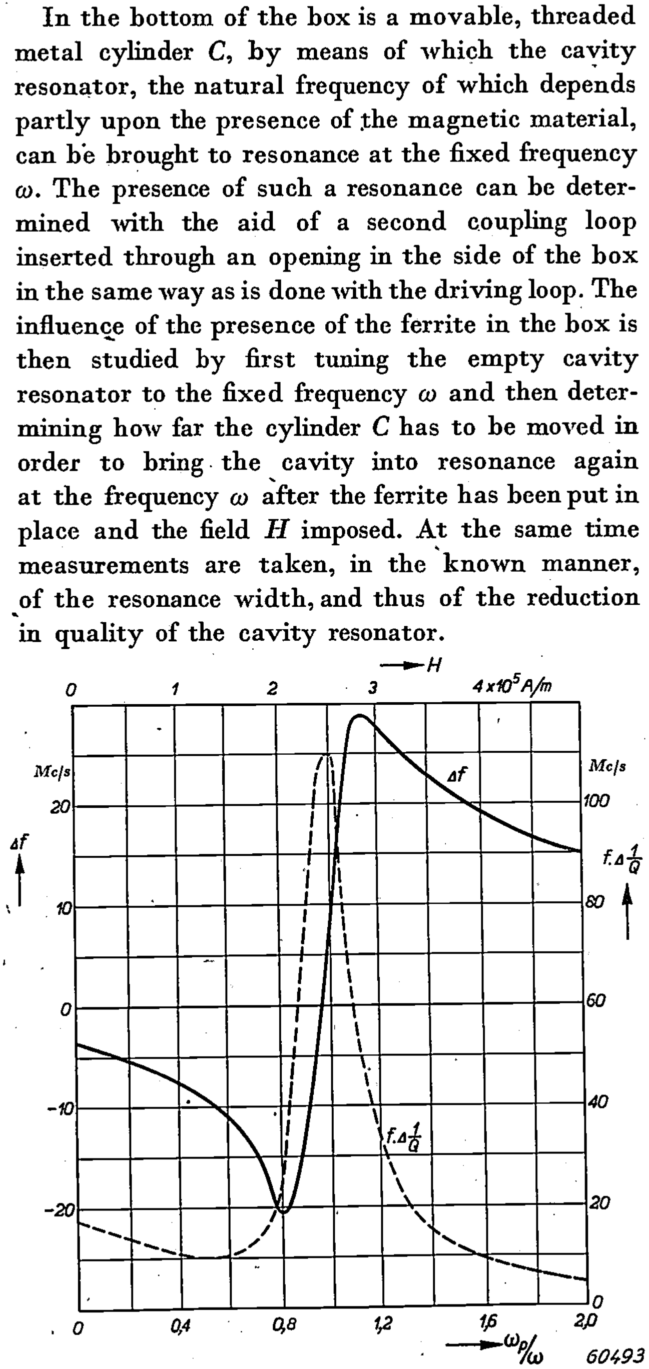


Fig. 8. The detuning Δf of the cavity resonator and the quantity $f \cdot \Delta(1/Q)$ as functions of the field strength H and of the quantity ω_p/ω respectively (ω is the practically fixed angular frequency of the cavity resonator).

Let x be the reading obtained when C is in a certain position. Then, in the absence of the ferrite, with a certain reading x_0 one will find a sharp resonance with the fixed frequency ω. When ferrite is present and at the same time a homogeneous magnetic field H is applied parallel to the axis of the box, resonance is found when C gives a reading x_1. The quantity $x_1 - x_0$ is a measure for the detuning of the cavity resonator due to the gyromagnetic effect. One can also determine the readings x_2 and x_3 of C where the energy, with the drive kept constant, is half as great as that for the reading x_1. The quantity $x_2 - x_3$ is then a measure for the quality of the cavity resonator when this is affected by the presence of the ferrite.

In *fig. 8* the detuning Δf, i.e. the frequency change which the displacement of the body C would have caused in the case where the box is empty, has been plotted on the left as a function of H, whilst the change $f \cdot \Delta(1/Q)$ — where Q represents the quality of the cavity resonator — has been plotted on the right. The quantities Δf and $f \cdot \Delta(1/Q)$ correspond to the quantities μ_r' and μ_r'' introduced at the beginning of this article.

Discussion of the results

The trend of both these functions answers the theoretical expectation if allowance is made for a certain amount of magnetic damping. As in the first experiment described (fig. 6), the value of ω_p for which $f\Delta(1/Q)$ is the maximum ($\omega_p = \omega = 5.8 \times 10^{10}$ rad/sec) agrees well with the value which, according to formula (6), follows from the corresponding value of H (2.5×10^5 A/m), namely 5.6×10^{10} rad/sec.

It appears that the general trend of the curve as found from experiments agrees with formula (11) for μ_x/μ_B if we add to that formula the damping term $\mathrm{j}d\omega$, thus:

$$\frac{\mu_x}{\mu_B} = \frac{\omega_p^2}{\omega_p^2 - \omega^2 + \mathrm{j}d\,\omega}\,\varepsilon_1\,\mathrm{e}^{\mathrm{j}\omega t}.$$

The real part of this expression is proportional to $x_1 - x_0$, the imaginary part to $x_2 - x_3$. It appears that for $\omega_p > \omega$ the real part increases as ω_p decreases and rather suddenly changes sign when the quotient ω_p/ω becomes less than 1. The peak value for $(x_2 - x_3)$ is a measure for the gyromagnetic damping factor d, in that the smaller the value of d the greater is this peak value (thus the maximum decrease in quality of the cavity resonator).

For a study of the gyromagnetic losses it is desirable to know the trend of μ_r' and μ_r'' as a function of ω or of $f = \omega/2\pi$ with a constant field H, and preferably for small values of H. It is very tempting to read fig. 8 — where a scale is also given for the ratio ω_p/ω (where ω_p is proportional to H, and ω represents the fixed frequency of the cavity resonator) — from right to left and to regard ω_p as a constant and ω as a variable frequency. Since in the formulae (11) and (12) the ratio ω_p/ω occurs as the only variable factor, figure 8 would then have to apply for an arbitrary value of ω_p. However, this is not permitted, because the damping, which has not been taken into account in the formulae (11) and (12), is a function of ω that presumably increases strongly with diminishing ω, in a manner so far unknown.

Therefore, until further experiments have been made (with smaller values of ω and with correspondingly smaller values of ω_p and thus of H) nothing can be said about this with any certainty. But it is possible to predict qualitatively what is to be expected. With constant H and increasing ω first an increase of μ_r' is to be expected, followed by a sharp decline, whereby μ_r' may even become negative. In the interval where μ_r' is already falling there will be a sharp rise in μ_r'', after which it will likewise pass through a maximum. This is exactly in accordance with the behaviour found for various ferrites according to fig. 1 when there was no constant magnetic field present; only the maximum of μ_r' is not very clear there and no negative values of this quantity are found. This will be considered somewhat more closely in what follows.

Resonance in the absence of a constant external field

We revert to the losses occurring in ferrites at frequencies of about 10^6 c/s and higher when these ferrites are not placed in a constant external magnetic field.

In the case of a material that is not placed in an external field and in which the losses are therefore measured in the usual way, it must be borne in mind that gyromagnetic resonance may still play a part, since in a magnetic material there are always some internal forces present which tend to give certain directions to the spinning vectors [9]). The greater these anisotropic forces the more difficult it is to change the direction of the spinning vectors, and thus the smaller will be the permeability μ_r. Although these coercive forces need not necessarily be of a magnetic nature, it may be assumed that they are capable of causing a precession of the spinning vectors, just like an external field H. Expressing these forces as an equivalent field H that would cause the same coercive effect, we arrive at values of 10^{-3} to 10^{-2} Wb/m² for $\mu_0 H$, corresponding to a frequency $\omega_p/2\pi$ of 2×10^6 to 2×10^7 c/s. This does indeed already bring us very close to the frequency range in which the phenomena described in the beginning of this article occur. The fact, expressed in fig. 1, that the abnormal behaviour of μ_r' and μ_r'' occurs at higher frequencies according as μ_r' is smaller in the low-frequency range, agrees well with what has just been stated, namely that a small value of μ_r signifies strong anisotropic forces, thus a high value of the equivalent H and the corresponding ω_p.

The fact that the ageement of the curves of fig. 1 with fig. 8 read from right to left is only qualitative, and that the abnormal behaviour of μ_r'' already begins at frequencies which are still a factor 10 or 20 smaller than what follows from the estimation made by means of the anisotropic forces, is most probably to be ascribed to the internal stresses in the material. The accompanying internal field strengths may weaken or compensate the anisotropic fields locally, resulting in a general widening of the resonance figure and a flattening of the maxima.

An analysis shows that for forces originating from elastic stresses the magnetic energy, which in the case of an external field H_p assumes the form

$$E = \tfrac{1}{2} H_p \Theta^2 \,,$$

in which Θ is the angle of magnetization by H_p (assumed to be small), is often to be generalized to the expression

$$E = \tfrac{1}{2}\,(H_x \Theta_x^2 + H_y \Theta_y^2)\,,$$

in which H_x and H_y are the "effective fields" for deviations in the x and y directions respectively. It is easily calculated that in that case the resonance frequency ω_0 assumes the form:

$$\omega_0 = \gamma (H_x H_y)^{1/2}, \text{ in which } \gamma = \frac{e}{m_0}\,\mu_0 \,.$$

For stresses of a special type H_y may, for instance, sometimes be zero, and then this leads in principle to a resonance frequency zero. It will not be so bad as this in practice, but still it is to be expected that certain parts of the material will have a lower resonance frequency than others [10]).

To sum up, it may be said that a deeper insight has been obtained into the nature of the losses occurring in ferrites at high frequencies, but that it is not yet possible to account for all the details of the phenomena observed. The future will certainly show some progress, in which the perfecting of the material itself will play a part.

The gyrator

Finally, attention is drawn to a possibility of making use of the gyromagnetic effect for constructing a so-called gyrator.

In the foreging (see equations (11) and (12)) it has been shown that the gyromagnetic resonance differs from the normal resonance in that an alternating field in the x-direction

[9]) See, e.g., J. J. Went, Philips Techn. Rev. 10, 246-254, 1948 (No. 8).

[10]) See J. L. Snoek, Physica The Hague 14, 207-217, 1948 (No. 4).

causes an alternating polarization not only in the x-direction but also in the y-direction, thus at right angles to it. This alternating polarization, which has also been observed experimentally, can be used in electro-technical engineering.

When an alternating current passes through a coil the axis of which is parallel to the x-axis of a system of coordinates then, without the presence of ferrite, no voltage is induced in another coil the axis of which is parallel to the y-direction. When, however, there is in the field a material which shows gyromagnetic resonance and is magnetized in the z-direction then a voltage will indeed be induced in the latter coil. Tellegen[11]) has shown that such a combination of two crossed coils forms a four-terminal network, called a gyrator, differing from the known four-terminal networks, such as the transformer, in that it does not comply with the law of reciprocity (the law of reciprocity for an electrical network states that the current in one branch of this network, generated by an alternating voltage in another branch, corresponds in amplitude and phase to the current generated in the latter branch as a result of an equally large alternating voltage in the branch first mentioned). Two of such four-terminal networks connected in cascade form a four-terminal network obeying the law of reciprocity. We cannot here enter further into the interesting practical possibilities offered by this new element for electrical networks.

Summary. The material named "Ferroxcube" is characterized by very low eddy-current and hysteresis losses. In the frequency range of 0-10^5 c/s some after-effect is observed. Above a certain critical frequency, differing between one material and another, additional losses occur which may be ascribed to gyromagnetic resonance. The theory of this phenomenon is explained with the aid of a mechanical model. Further, measurements of the gyromagnetic resonance are described which have been carried out with a material placed in a constant, strong magnetic field polarizing the material in a direction perpendicular to the direction of observation. These measurements confirm in the main the theoretical expectation. It is indicated briefly how the losses observed can be explained qualitatively from the gyromagnetic phenomena. Finally it is pointed out that with the aid of gyromagnetic effect a four-terminal network with new properties (a gyrator) can be constructed.

[11]) B. D. H. Tellegen, The gyrator, a new electric network element, Philips Res. Reports 3, 81-101, 1948 (No. 2).

THE MANUFACTURE OF QUARTZ OSCILLATOR-PLATES

I. HOW THE REQUIRED CUTS ARE OBTAINED

by W. PARRISH [*]).

549.514.51:621.396.611.21

The quartz plate as used for the frequency control of radio transmitters is a device requiring extreme precision in its preparation. The manufacture of such plates is an interesting problem, especially so when mass production is envisaged. This problem was tackled and solved in an impressive way by the United States during the war employing the combined efforts of crystallographers, physicists, experts in electronics, mechanical engineering and allied fields.

Among the more than 100 factories which took part in the mass production, the North American Philips Company was able to make several not unimportant contributions to the development of the methods of manufacture. These contributions, as well as the development referred to in general, have been described since the war in various American periodicals. Nevertheless the editors of this Review feel that it may still be of interest to many readers, especially non-Americans, to learn something about these methods of manufacture.

Introduction

In 1921 W. G. Cady showed that the vibration of a plate cut from a piezo-electric crystal such as quartz may serve to control the frequency of a vacuum tube oscillator. This method gradually found its way into the practice of short wave radio transmitters. Thus, for example, in the United States in 1939 about 50,000 quartz plates were manufactured for this purpose.

World War II gave a strong impulse to short wave communication technique. Upon the entrance of the United States into the war a program was set up in that country which provided for the equipment of huge numbers of aircraft, tanks, small infantry units, etc., with a transmitting-receiving apparatus. The use of quartz oscillator-plates made possible precise frequency control and push-button tuning on a scale never before attempted. The production of quartz oscillator-plates rose from about 100,000 in 1941 to about 6 million in 1942 and then to about $28^1/_2$ million in 1944. The following is a rough breakdown of the last figure [1]): 8 million oscillator-plates were made for aircraft; $4^1/_2$ million for tanks and artillery units, 11 million for walkie-talkies and handie-talkies, the rest for various vehicles and naval vessels. At the end of the war the production rate increased to about 60 million per year and the price had dropped to only a small fraction of pre-war prices.

This rapid rise in the production of quartz plates was based on the development of methods of manufacture which are interesting because of the remarkable combination of extreme precision required, with the speed and simplicity made necessary by wartime demands and limitations. A concise, but practically complete survey of these manufacturing methods was published in a symposium of the Mineralogical Society of America [2]) in the early part of 1945. In this periodical we must restrict ourselves to several phases of the manufacturing process, viz., those in whose development the Philips concern, in particular the crystal factory at Dobbs Ferry, the X-ray factory at Mt Vernon and the laboratory at Irvington, have had an active part. This article will give a broad survey of the problem of cutting the quartz plates out of the natural raw crystals; in a subsequent article a discussion will be given of the accurate determination of the

[*]) Philips Laboratories, Inc., Irvington-on-Hudson, New New York, U.S.A.

[1]) See: C. Frondel, History of the quartz oscillator-plate industry 1941-1944, Amer. Mineralogist 30, 205-213, 1945.

[2]) The symposium number of the American Mineralogist (May/June 1945) contains in addition to the article quoted in footnote [1]) the following:
K. S. van Dyke, The piezo-electric quartz resonator (p.214-244).
R. E. Stoiber, C. Tolman and R. D. Butler, Geology of quartz crystal deposits (p. 245-268).
S. G. Gordon, The inspection and grading of quartz (p. 269-290).
J. S. Lukesh, The effect of imperfections on the usability of quartz for oscillator-plates (p. 291-295).
W. Parrish and S. G. Gordon, Orientation techniques for the manufacture of quartz oscillator-plates (p. 296-325).
W. Parrish and S. G. Gordon, Precise angular control of quartz-cutting by X-rays (p. 326-346).
W. Parrish and S. G. Gordon, Cutting schemes for quartz crystals (p. 347-370).
W. Parrish, Methods and equipment for sawing quartz crystals (p. 371-388).
W. Parrish, Machine lapping of quartz oscillator-plates (p. 389-415).
C. Frondel, Final frequency adjustment of quartz oscillator-plates (p. 416-431).
C. Frondel, Effect of radiation on the elasticity of quartz (p. 432-446).
C. Frondel, Secondary Dauphiné twinning in quartz (p. 447-460).

desired cutting directions with the help of a special X-ray diffraction apparatus designed by Philips; a third article will treat the methods employed in lapping and finishing the oscillator plates.

The reader who would like to refresh his memory of several general concepts in the field of the application of the piezo-electric effect should réfer to a short survey article [3]) which appeared in this periodical several months ago and to Heising [4]) and Cady [5]).

Low temperature coefficient cuts

The resonance frequency of a quartz plate of given dimensions and vibrating in a given mode of oscillation depends upon the temperature. This dependence may be very slight in certain temperature ranges, when the crystallographic axes of the plate make certain, accurately determined angles with the boundaries of the plate. This is usually formulated the other way round: it is said that the oscillator plate must be cut from the crystal at definite angles to the crystallographic axes of the crystal. Examples of these "low temperature coefficient cuts" are the AT, BT, GT cut, etc.

It is obvious that quartz plates cut to be independent of temperature should be preferred for radio transmitters. The object of using such plates is the stabilization of the transmitter frequency, and by the use of the cuts in question, without the necessity of a thermostat for the quartz plate, the frequency is made also highly insensitive to variations in the temperature of the surroundings, which e.g. in aircraft may be very great. The low value of the temperature coefficient with these cuts is obtained only in a limited temperature range: the resonance frequency as a function of temperature passes through a maximum (or through a flat turning point). Optimum frequency stability is realized by operating the crystal around the temperature of this maximum. The broadness of the maximum, i.e. the width of the useful temperature range, depends on the type of cut. In the case of the often used BT cut the frequency does not vary more than 0.02 % in a temperature range of about 140 °C. With other cuts an even greater constancy of frequency can be obtained.

The various cuts differ in the oscillation mode for which the temperature coefficient becomes small. As the resonance frequency of a plate vibrating in a given mode chiefly depends on its dimensions, and as it is not practicable to make the plates either very thin and small or very large, each cut has a specific frequency range in which it can be used. Furthermore there is the difference in width and position of the useful temperature region. A further factor in the choice of cut to be employed is the degree to which undesired modes of oscillation (with nearby resonance frequencies) are excited by the desired vibration. This would lead to an extra damping, decrease of "activity" at some frequencies, heating of the plate and other disturbing effects. Roughly it may be said that in the frequency regions of 2-5 Mc/sec only AT cuts and of 5-9 Mc/sec only BT cuts are employed, both vibrating in the fundamental thickness shear mode shown in *fig. 1.* For frequencies up to 100 Mc/s AT- or BT-plates vibrating in higher harmonics of this mode are used. With the BT-cut the useful temperature region can be varied within limits by changing the cutting angle of the plate with respect to the optic axis. For example, a change of 30 minutes of arc in cutting angle may shift the useful range as much as 10-20 °C.

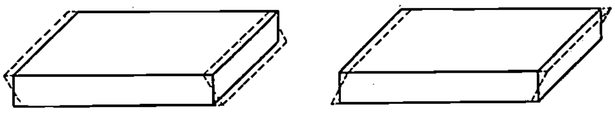


Fig. 1. AT an BT cut oscillator plates are used in a thickness shear vibration mode. The deformation involved in this modc is indicated by dotted lines.

The slightness of this angle variation gives at the same time an idea of the high precision required in the orientation of the plates. As to the thickness of the plates, the tolerances there are also extremely small; the plates for two communication channels adjacent to each other in frequency may for example differ in thickness by as little as one ten-millionth of an inch.

What is now the position in the quartz crystal of the different low temperature coefficient cuts?

Quartz crystallizes in the symmetry class D_3 (Schoenflies notation) which possesses one threefold axis of symmetry and perpendicular to that three twofold axes of symmetry. Some of the faces which may be found on a freely growing quartz crystal are shown in *fig. 2.* The vertical axis *Z* is the threefold, so-called optic axis. The six side faces parallel to *Z* are called the prism faces (m). The three horizontal twofold axes *X*, each of which is parallel to two prism faces, are the so-called electric axes. For the sake of completeness wc

[3]) J. C. B. Missel, Piezo-electric materials, Philips Technical Review 11, 145-150, 1949 (No. 5).
[4]) R. A. Heising, Quartz crystals for electrical circuits, Van Nostrand, New York 1946.
[5]) W. G. Cady, Piezoelectricity, McGraw-Hill, New York 1946.

also mention the three intermediate horizontal axes, the Y-axes, each perpendicular to an X-axis (and therefore perpendicular to the prism faces). The crystal boundaries at the pyramidal ends of the crystal drawn are formed chiefly by the three so-called major rhombohedron-faces, indicated by r, the three so-called minor rhombohedron faces, indicated by z, and the small faces s and x.

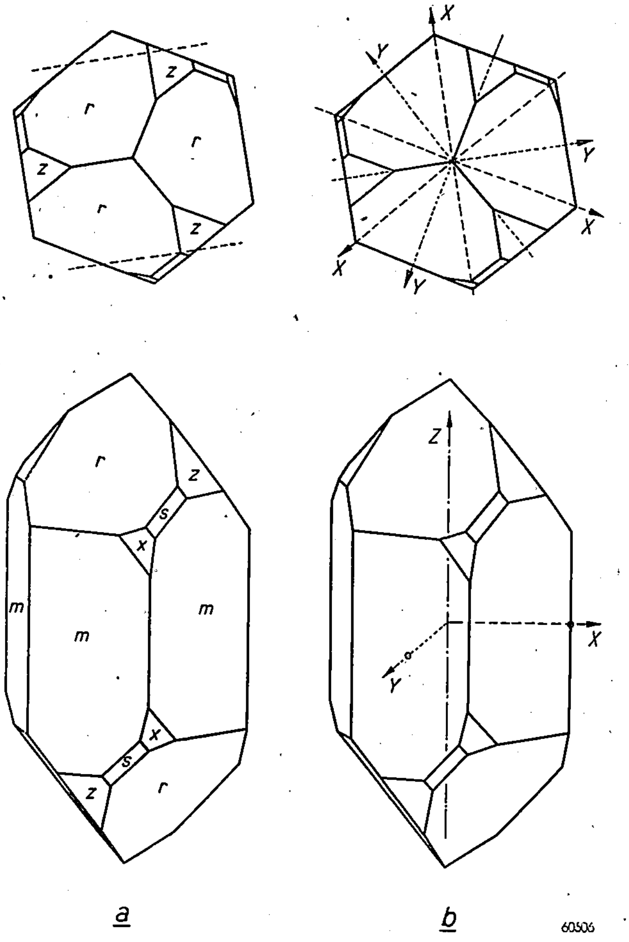


Fig. 2. Ideal quartz crystal. In (*a*) the customary indications for the various faces are given; *m* are the prism faces, *r* major and *z* minor rhombohedron faces. In (*b*) the directions are given of the *Z*-axis (optic axis), the three *X*-axes (electric axes) and the three *Y*-axes.

The dotted lines in (*a*) above are the cutting lines along two planes parallel to *X* and *Z*, as described in the first step of the *X*-block method of cutting the quartz crystal.

Most of the low temperature coefficient cuts developed until now are parallel to one of the X-axes and therefore perpendicular to a Y-Z-plane. They differ only in the angle which they make with the Z-axis. In *fig. 3*, which shows a cross section of the crystal in the Y-Z-plane, these cuts are indicated. They are all perpendicular to the plane of the drawing, like the rhombohedron faces r and z crossed by the section shown. It may be seen that the AT and the BT cuts, in which we are mainly interested, run roughly parallel to a z and an r face, respectively.

The orientation of the crysta!s under the saw

The quartz crystals found in nature usually have little resemblance to the ideal crystal drawn in fig. 2. *Fig. 4* shows several natural crystals. On these crystals some prism and rhombohedron faces may be seen and the type of face can in general readily be established, since the angles between faces are always the same. But the various faces are by no means developed to the same extent, so that the crystal does not in general show any macrosymmetry. Crystals are frequently found which show no clear boundary faces at all (defaced crystals).

In such a more or less irregular crystal the determination of the planes along which it must be sawed, e.g. for an AT cut, is made possible with the required very small tolerances by two fundamental means: the phenomena of double refraction and X-ray diffraction. The way in which these means, along with other ones (as the natural faces on the crystal and the light figures on etched surfaces), are applied can most easily be explained by following in detail one of the several cutting methods. For that purpose we have chosen the

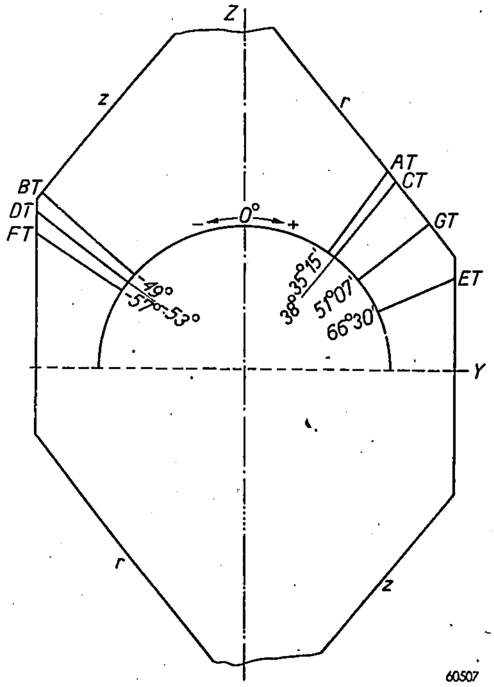


Fig. 3. Cross section of a quartz crystal parallel to a Y-Z-plane (i.e. perpendicular to an X-axis). Most of the low temperature coefficient cuts are perpendicular to this cross section, at angles with the Z-axis as indicated here. AT, CT, GT and ET cuts are roughly parallel to the z face, and BT, DT an FT cuts roughly parallel to the r face.
Some oscillator plates are made circular or rectangular but most are made square. One side is then perpendicular to the plane of the drawing (thus parallel to X), except in the case of the GT cut, where the sides must make angles of 45° with this plane. The length of the sides varies between 0.1 and 2 inches, a normal value being 0.5 inch.

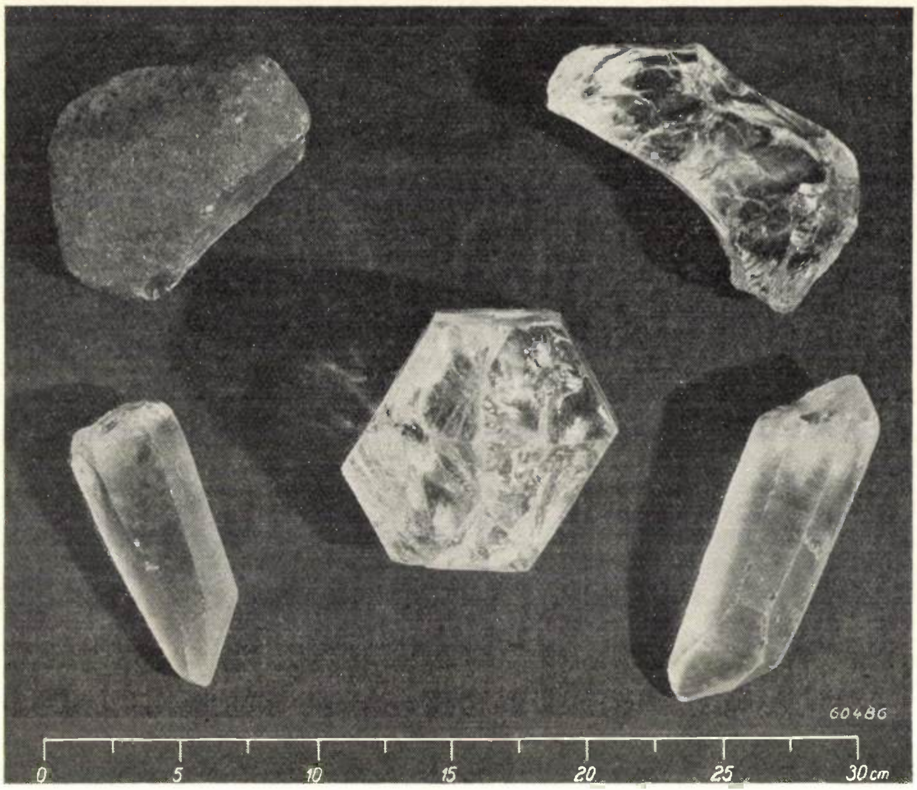


Fig. 4. Several typical natural quartz crystals. Top, two defaced crystals; the one on the left was rolled around in a stream bed in nature and natural faces abraded off, whereas the faces of the crystal on the right were cobbed off at the mines to remove defective parts. Lower, three faced crystals; the one in the middle is set with the optic axis vertically to show the hexagonal outline while the crystals on the sides are "candle" shape with optic axis parallel to plane of photograph. (This shape with prism faces not parallel to the Z-axis is caused by these faces not being perfectly smooth but having a very fine step-like structure, composed of alternating prism and rhombohedral faces.)

so-called X-block method, which is the most generally successful method for crystals weighing less than about 1000 gms.

Suppose that the crystal to be sawn, although irregular, possesses at least one well developed prism face (m). First, a plane perpendicular to this prism plane and parallel to the Z-axis is cut at either side of the crystal. The planes cut are parallel to a Y-Z-plane (cf. *fig. 5*). The "X-block" thus obtained is placed on the table of the quartz saw (*fig. 6*) so that the saw blade is perpendicular to the Y-Z-planes, cut previously, and thus parallel to an X-axis of the crystal. According to fig. 3 it is now only necessary to rotate the crystal about this X-axis until the angle between the Z-axis and the saw blade has the value required for the type of cut and indicated in fig. 3. One may then saw a wafer out of the crystal (*fig. 7*), of the thickness desired for the oscillator-plates (plus the necessary extra thickness for lapping), and this wafer may then be cut into a number of oscillator-plates.

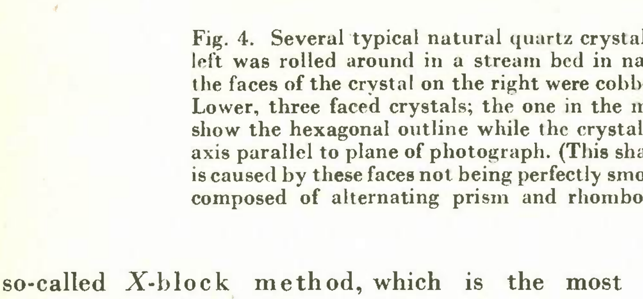


Fig. 5 When cutting crystal blocks according to the X-block method, preliminarily planes perpendicular to a prism face and parallel to the Z-axis are cut from both sides of the block. These are Y-Z planes (parallel to the cross-section in fig. 3).

It is seen that in order to obtain the desired low temperature coefficient cut by this method two different orientations of the quartz crystal with respect to the saw are involved:

In the first orientation, for cutting a Y-Z-plane, the plane of the saw must be set perpendicular to the prism face and parallel to the Z-axis of the crystal. This is accomplished as follows. The crystal is placed with the well-developed prism face on a thick glass plate with a ground reference edge. On the saw table the glass plate can be fastened with the help of a reference edge on the table so that the

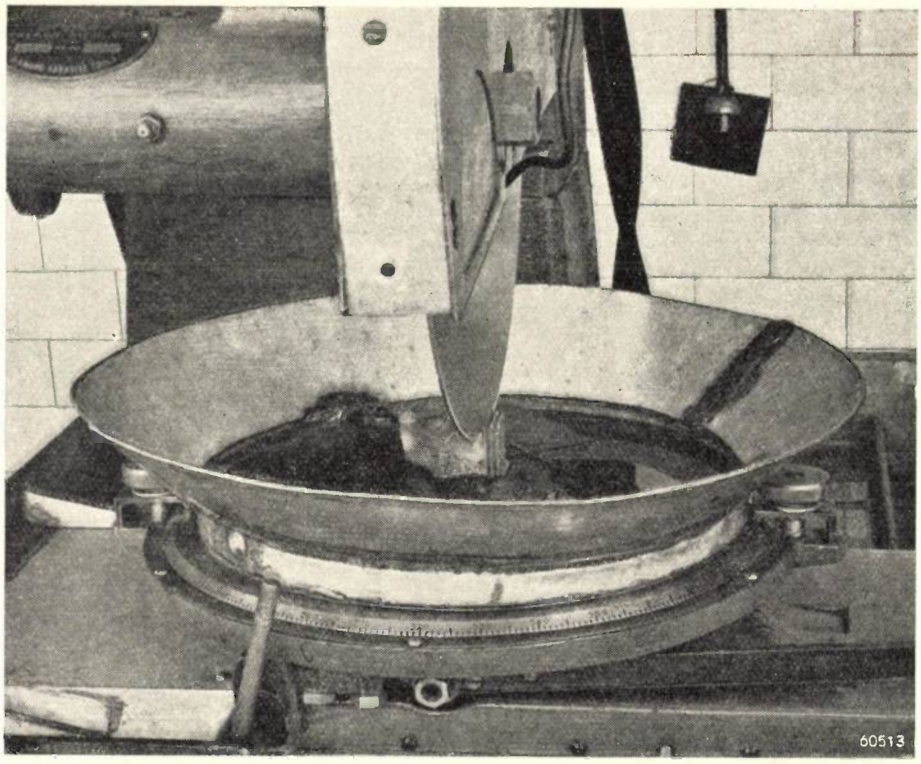


Fig. 6. Arrangement of a circular saw for sawing quartz crystals. A commonly used type of saw blade consists of sintered bronze containing diamond powder. The disc rotates with a peripheral velocity of about 30 m/sec, and can cut a surface area of for instance 30 cm^2 in two to three minutes. The quartz block is placed on the saw table, which can be rotated around a vertical axis and tilted around a horizontal axis; for this purpose the circumference is provided with a scale division in degrees. Moreover the mounted quartz block can be displaced parallel to itself over a distance in order to saw a series of parallel wafers of a given thickness out of the block.

reference edge on the glass is parallel to the saw blade. This gives the desired orientation, provided the crystal was placed on the glass plate in such a way that the *Z*-axis is exactly parallel to the reference edge [6]). A rough degree of parallelism can usually be accomplished by eye. The setting is then improved with the help of a simple instrument, the stauroscope (see *fig. 8*): in this instrument a beam of light travels perpendicularly through the glass plate and the crystal mounted on it while the crystal lies between crossed Polaroids. In the absence of the crystal there would be complete extinction; due to the double refraction of the crystal perpendicular to the optic axis a certain transmission of light is obtained, unless the "main directions" of the crystal (in this case the *Z*-axis and *X*-axis) are parallel to the directions of polarization of the Polaroids. Since the reference edge of the glass plate is set parallel to one of these directions of polarization by means of a reference edge on the instrument stage, by shifting the crystal slightly until extinction is re-established, the *Z*-axis can be made parallel to the reference edge with an accuracy of about 1°. In this position the crystal is cemented to the glass plate and a plane is cut from one side of the block. This, however, does not yet conclude the first orientation, since the accuracy mentioned is not sufficient. The X-ray method must now be applied to the cut surface enabling the crystallographic angles of the plane to be determined with an accuracy of a few minutes. The necessary corrections in the position

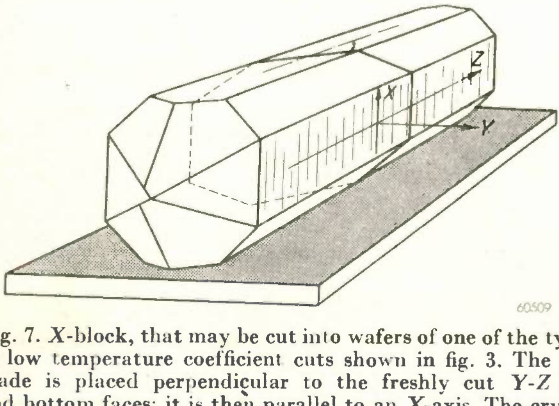


Fig. 7. *X*-block, that may be cut into wafers of one of the types of low temperature coefficient cuts shown in fig. 3. The saw blade is placed perpendicular to the freshly cut *Y-Z* top and bottom faces; it is then parallel to an *X*-axis. The crystal block is turned around this *X*-axis in order to obtain the desired angle between the saw blade and the *Z*-axis.

[6]) The method is not impaired by a "candle" shape of the crystal. In that case the prism face used is not perpendicular to a *Y*-axis, but it is only the parallelism of the face to an *X*-axis that matters. It is therefore even possible to use a rhombohedral face instead of a prism face in the mount for making an *X*-block.

of the crystal with respect to the saw blade can be deduced very simply from this and then realised by rotating (and if necessary slightly tilting) the saw table, whose circumference is provided with a degree scale with vernier reading. Again a plane is cut and X-ray tested, and if necessary a second correction must be applied. Usually, however, after the first correction the cut is already inside the tolerances of 15′ to 30′ parallel to the Y-Z-plane.

The Y-Z-plane at the other side of the mounted crystal block is cut in exactly the same way.

The second orientation described above consists in bringing the plane of the saw into a position perpendicular to the Y-Z-plane, and at the desired angle with the Z-axis. The first is accomplished by mounting the X-block, after it has been removed from its glass plate, with one of the fresh cut Y-Z-planes on another glass plate, again with the Z-axis parallel to the reference edge. After the plate has been fastened to the saw table the desired angle with the Z-axis is obtained by rotating the saw table into the position required, which can be read off on the degree scale of the table. In doing this, however, it is necessary to know whether to turn to the right or to the left. In fact, fig. 3, the cross section perpendicular to an X-axis in which the low temperature coefficient cuts have been indicated, could also be viewed from the other side: in that case left and right are reversed. Therefore, when examining an X-block it is necessary to ascertain the position of the r and z faces; then the rule can be applied that the AT cut is approximately parallel to z and the BT cut approximately parallel to r. The identification of r and z is quite easy by the method of etch figures. The crystal is etched in a concentrated water solution of ammonium bifluoride. When one cut surface of the etched X-block is placed over a point source of light one sees an unsymmetrical light figure on the top cut surface (due to refraction and reflection in the myriads of etch pits), from whose orientation the directions of r and z can immediately be deduced. This is explained by *fig. 9* [7]).

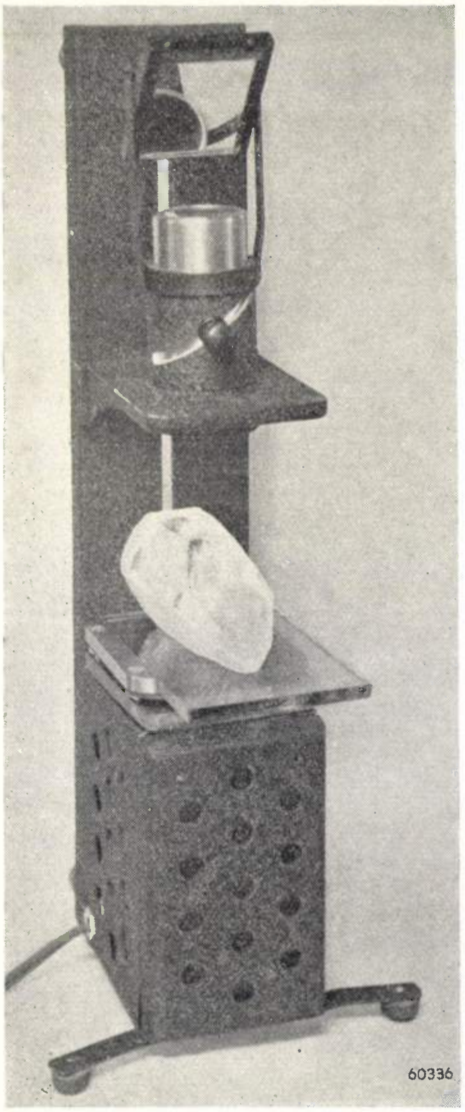


Fig. 8. Stauroscope. With the help of this instrument the Z-axis of a quartz block is made parallel to the reference edge of the glass holder. (The instrument shown is produced by The Polarizing Instrument Co., Inc.)

The adjustment to the required angle between Z-axis and saw blade by means of the scale division on the saw table is not accurate enough. Here also a test cut is made, the X-ray method applied to it for exact determination of the crystallographic angles actually obtained and the necessary corrections deduced. The tolerances here are 10′ to 15′.

When a test cut finally has been made which satisfies the requirements the crystal is sliced into wafers parallel to the test cut and about 1 to 1.2 mm in thickness. With poorly constructed saws, after each three or four cuts the orientation must be checked and if necessary corrected again.

The wafers obtained (see *fig. 10*) have two edges remaining from the Y-Z-planes which are accurately perpendicular to the X-axis. The square AT and BT oscillator-plates must also possess two sides perpendicular and two sides parallel to the X-axis

[7]) The ambiguity in viewing fig. 3 can also be eliminated by the statement that in the quartz crystal considered the positive direction of the X-axis, to which the cross section drawn is perpendicular, must point towards the observer. To this corresponds a method which was formerly used to solve the question about the direction of rotation, viz. by direct determination of the polarity of the X-axis by an electrical method. The crystal was compressed in the X-direction and the side was ascertained on which the charge caused by the piezo-electric effect was positive. This method is not only more elaborate than the one described in the text but it also has the disadvantage that it can in some cases be subject to serious errors entailing the loss of valuable material. This will be explained later (footnote [8])).

and it is therefore quite simple to mark off the desired blanks on the wafers and cut them out (see *fig. 11*).

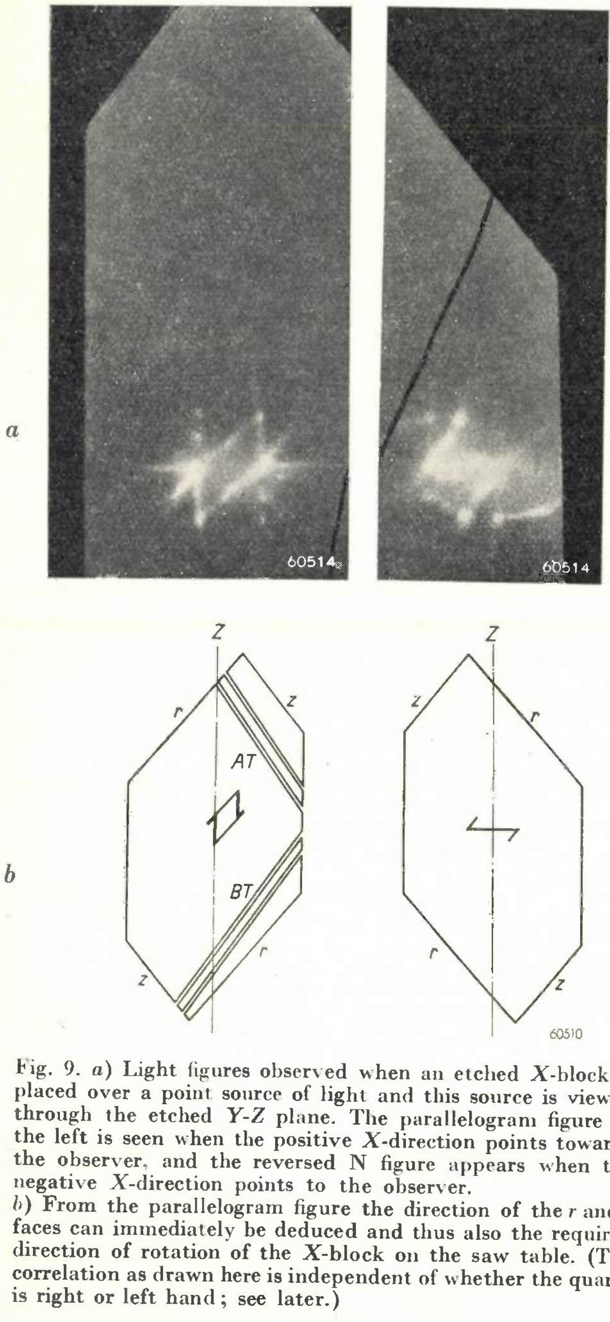


Fig. 9. *a*) Light figures observed when an etched X-block is placed over a point source of light and this source is viewed through the etched Y-Z plane. The parallelogram figure on the left is seen when the positive X-direction points towards the observer, and the reversed N figure appears when the negative X-direction points to the observer.
b) From the parallelogram figure the direction of the *r* and *z* faces can immediately be deduced and thus also the required direction of rotation of the X-block on the saw table. (The correlation as drawn here is independent of whether the quartz is right or left hand; see later.)

Cutting strategy

The above description still does not give the reader a good picture of the process of cutting quartz crystals because a very important complication has not been considered — a complication which overshadows the whole process, that is, twinning in natural crystals.

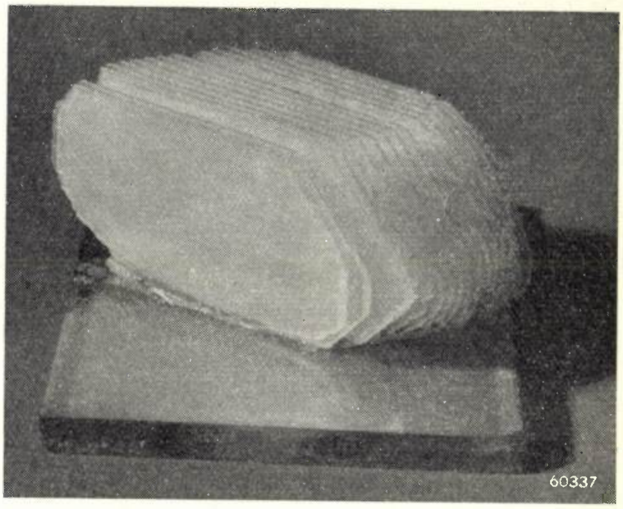


Fig. 10. A quartz block cut into wafers on its glass-holder. The photograph gives an idea of the large percentage of weight of quartz which is lost due to the thickness of the saw blade. It is furthermore of importance to note that the upper face of the quartz block, in the X-block method here used, is horizontal. With cutting methods where this is not the case there is a disturbing tendency of the saw blade to drift when entering the slanting surface, thus making the angles inaccurate.

In a single quartz block, even one having practically the ideal shape of fig. 2, two or more crystals differing from each other in certain respects may be grown together. In the case of quartz there are two types of such twinning: electrical (Dauphiné twins) and optical (Brazil twins).

If we rotate the crystal cross-sectioned in fig. 3 180° around the Z-axis, the positive directions of all Y- and X-axes will point in the opposite directions, but the faces which exchange places in the figure make the same angles with the Z-axis. Two crystal individuals rotated in this way with respect to each other can therefore grow together without the intergrowth necessarily being visible to the eye: the faces *m* and *r* of the one individual are then "coplanar" with the faces *m* and *z*,

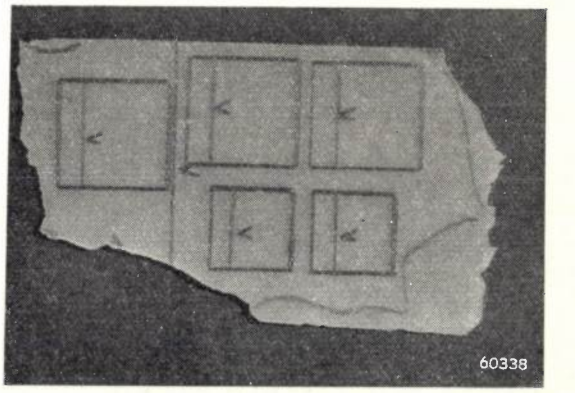


Fig. 11. Quartz wafer on which the blanks to be cut out are indicated with a rubber stamp. The straight edges on top and bottom of the wafer remaining from the X-block stage are perpendicular to an X-axis and are used as an accurate guide in cutting the blanks from the wafer. On the whole wafer and the separate blanks, the X-axis and the positive direction of the projection of the Z-axis (cf. the arrow heads) are indicated for later control measurements of the crystallographic angles of each oscillator plate. The determination of the Z-direction is carried out with an instrument similar to that of fig. 8 (angular view stauroscope).

respectively, of the other. This is called an electrical twin. It can sometimes be recognized by the occurrence of extra s and x faces and by other minor morphological features. The boundaries between the two or more individuals in an electrical twin are generally irregular as shown in *fig. 12.*

The symmetry class D_3 to which quartz belongs possesses no plane of symmetry. If the crystal drawn in fig. 2 is mirrored at any plane through the Z-axis, a crystal is obtained which is not identical with the original one, i.e. cannot be made to coincide with it by rotations and/or translations, but which nevertheless possesses all the properties of the first and is thus a possible crystal form of quartz. The two crystal modifications are in the same relation to each other as a right- and left-rotating threedimensional system of coordinates (or like a right and left glove). A crystal of the left-hand modification can grow together with one of the right-hand modification in such relative positions that the Z-axes are parallel and every r face of the one is parallel to a z face of the other. This is optical twinning, called thus, because the twins may be distinguished optically: they rotate the plane of polarization of a light beam in opposite directions. The boundaries between the two or more optical twins are always plane (parallel to an X-axis) and remarkably straight-edged as shown in *fig. 13.* Very often both types of twinning can be found in the same crystal block. (In rare cases twinning of a combination of both types is encountered, thus an intergrowth of a right-hand crystal with a left-hand crystal rotated 180°.)

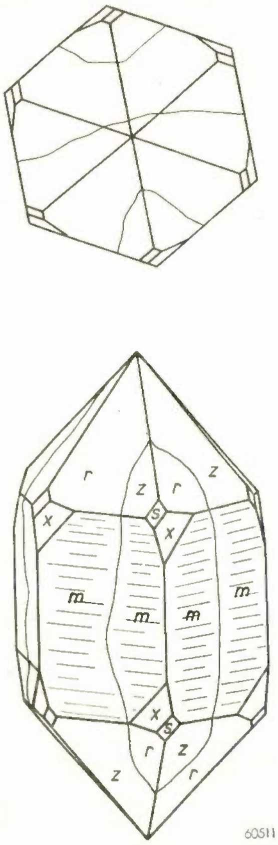


Fig. 12. Electrical twinning, drawn for a crystal with ideal faces. In well-developed crystals, such a twinning may be revealed by the presence of extra s and x faces (compare with fig. 2a). The irregular boundaries, drawn with thin lines, between the differently oriented crystal individuals in general are hardly if at all visible. They may be traced by the different brightness of the adjacent areas on the r and z faces and/or as sutures interrupting the horizontal striations that in general appear on the m faces. The only certain way to trace the twin boundaries, however, particularly if only a few or no faces are present, is to sandblast the crystal to roughen the smooth surfaces and then etch the crystal for about 10 hours in cold concentrated ammonium bifluoride solution.

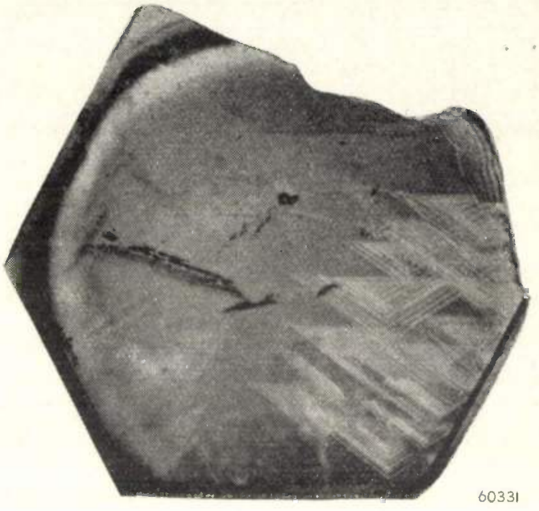
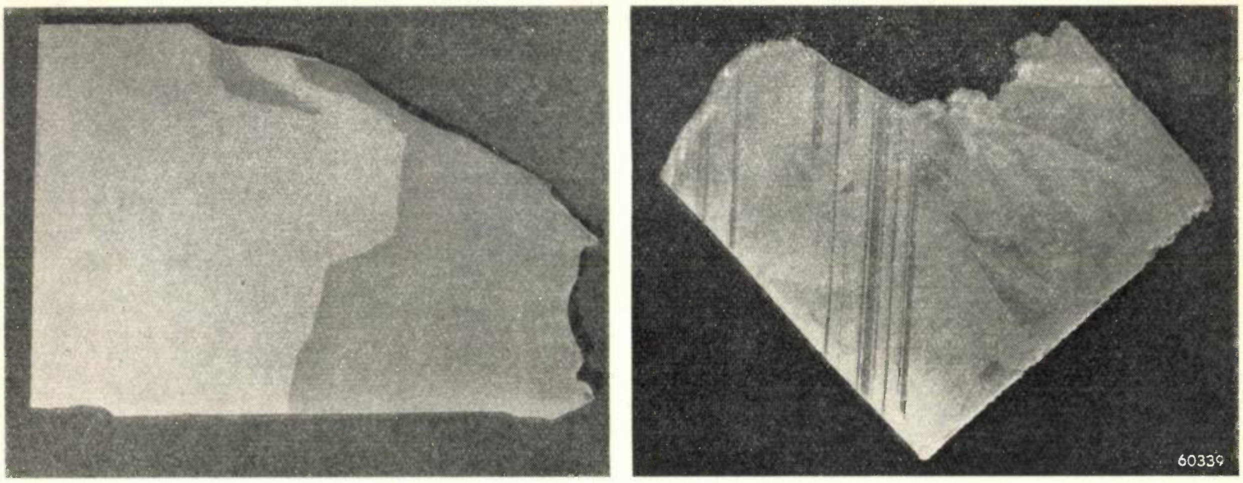


Fig. 13. Optical twinning. With this type of twinning a right-hand quartz contains more or less regularly bounded inclusions of left-hand quartz (or vice versa), which become visible on observation in polarized light. The inclusions often occur as thin regularly spaced laminae. In the photograph shown the regularity of the twin boundaries is truly remarkable.

Fig. 2, with the positive axis directions assigned by convention, represents a right-hand crystal. For a crystal of the left-hand modification the steps in cutting are exactly the same as those described above. The angles of the various low temperature coefficient cuts with the Z-axis are opposite for the two kinds of quartz, but the above-mentioned r-z rule still holds. When the same saw must be used for right-hand and left-hand quartz, in order to prevent confusion it is simplest to keep to the same direction of rotation in adjusting the saw table and to fasten right-hand X-blocks to the glass plate with the positive X-direction upwards and left-hand X-blocks with the positive X-direction downwards [8]).

[8]) At this place, the importance of the r and z rule and its application by means of etch light figures should again be stressed. With the formerly used method of orienting the crystal, mentioned in footnote [7]), it was evidently necessary to determine also the hand of the crystal, in addition to the positive X-direction. Moreover, in cases of twinning errors were often made. For example the crystal shown in fig. 12 might be oriented on the basis of the polarity determined on the edge which is a region of electrical twinning, and hence the entire crystal would be cut incorrectly. — In view of the advantages of the r and z rule, it can be stated with confidence that the simple technique of the etch light figures has been one of the conditions for a successful mass production of oscillator-plates.



a *b*

Fig. 14. *a*) Etched wafer cut from a crystal block with electrical twinning. By means of the difference in the reflection of the two parts it is not only easy to ascertain the position of the boundary line, but also to discover which part is usable. In the case of a BT cut that part can be used which, upon being viewed at an angle of about 22° with respect to a perpendicular incident beam of light, reflects brilliantly once in a 360° revolution of the wafer in its own plane. *b*) Etched wafer cut from a crystal block with optical twinning. Due to the unfavorable position of the twin laminae no blanks can be cut from this wafer. If in cutting and mounting the quartz block we had started from a different *X*-axis, so that the twin laminae (which are always parallel to one of the *X*-axes) were perpendicular to the saw table, it would have been possible to cut some usable wafers from between the laminae.

An oscillator plate cut out of a quartz block in such a way that it consists of parts of different crystal individuals is not in general usable (unless the twin structure is limited to a fraction and/or to certain parts of the plate). This is easy to understand: the parts of a plate belonging to an electrical twin show an opposite charge upon identical deformation; those made of an optical twin show the same charge but resonate at different frequencies, since their crystallographic orientation differs (interchange of *r* and *z* in fig. 3).

Now, it must be realized that practically no crystals from the quartz mines are entirely free of twinning. In very many cases one encounters an intergrowth of many individuals. The sawing of oscillator plates thus takes on an entirely new aspect: it becomes a problem of prime importance, on the one hand, to cut out as many oscillator plates from a raw crystal block as possible without touching twin boundaries, and on the other hand to use as little sawing time as possible in cutting these plates [9]).

We can here only offer a few remarks about the means by which one attempts to attain these objects, the cutting "strategy".

Every wafer thas has been cut is etched to make the electrical or optical twinning boundaries on it visible (*fig. 14a, b*). Only one of the two orientations occurring in the wafer corresponds to the desired low temperature coefficient cut. The other orientation, represented in fig. 3 by the mirror image of the desired cut with respect to the *Z*-axis, does not correspond to any low temperature coefficient cut (at least not nearly enough to make it possible to obtain such by lapping off at an angle); the parts of the wafer having this orientation are thus worthless, and need not be sawed into blanks.

In order to avoid as much as possible loss of material by such useless parts of wafers, measures are taken before wafering. In fact it should be pointed out that in the first step of the *X*-block method chosen here as example, placing the crystal block with a prism face on the glass plate, generally a choice must be made among three prism faces, corresponding to cutting planes perpendicular to the three *X*-axes. This choice is decided by etching the whole crystal block and observing it in ordinary light.

The regions and boundaries of electrical twinning are clearly traced in this way. If possible the crystal block is laid on a prism face on which no twinning boundaries occur, since in that case the twinning boundaries will in general become visible on the *Y*-*Z* planes to be cut (after etching these too), and it will then be possible to decide whether the *X*-block will yield sufficient twin-free wafers in the direction of the AT cut, or will perhaps give a better yield

[9]) In addition to twinning, various kinds of inclusions and crystal defects also often occur in natural crystals, for which similar considerations hold.

in the direction of the BT cut (or other cuts), and whether some parts of the X-block cannot be used at all. A beautiful example is shown in *fig. 15.*

As in the case of optical twinning the crystal individuals of one kind are generally present in the form of thin inclusions, this type of twinning is not revealed by etching the crystal faces, but it becomes visible when the whole crystal block is viewed in polarized light (while immersed in a suitable liquid of the same refractive index as quartz to reduce reflection and refraction effects at the crystal bounraries). The twinning structure thus revealed must in a similar way serve as a guide in mounting and wafering the crystal.

It will be evident that due to the inspection at each stage of the cutting process much loss of material and much useless sawing can be avoided. This applies to the *X*-block method as well as to other cutting schemes developed for cutting raw crystals. Most of these cutting schemes have one or more disadvantages as compared to the *X*-block method, so there is no need to mention them in this short article. However, we must make an exception for the scheme called the *Z*-section, *Y*-bar method. This method, which is sketched in *fig. 16*, is ideally suited for crystals that are large (say more than about 1000 gms) and of exceptional quality. A description of the method is given in the legend of the figure.

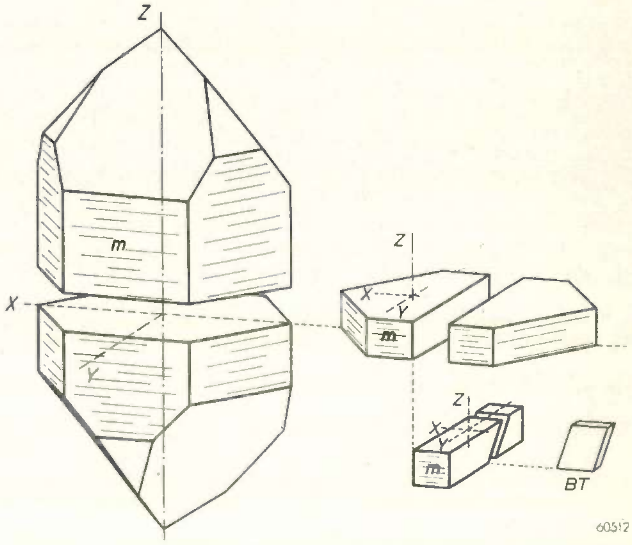


Fig. 16. In the case of large quartz crystals of exceptional quality, instead of the *X*-block method, the so-called *Z*-section *Y*-bar method can better be applied. First, sections are sawed out of the block perpendicular to the *Z*-axis, each section with a thickness about equal to the width of the desired square oscillator-plates. From the section, bars are cut whose direction of length is parallel to the Y-axis, and whose thickness is also equal to the width of the desired oscillator plates. A bar is now turned 90° from the position shown in the figure and mounted on the saw table. As an *X*-axis is then vertical, the bar can be turned to the correct angle and cut in the same way as an *X*-block, but instead of wafers, slices are obtained having roughly the dimensions of the final oscillator plates, and which need only be lapped and finished.

A saving in sawing time compared with other methods is realized because the sawing of blanks out of wafers is eliminated.

If there is an electrical twin boundary approximately parallel to the *Z*-axis, both parts of the Y-bar can be used; for the same type of cut they must be sliced at opposite angles (cf. fig. 15).

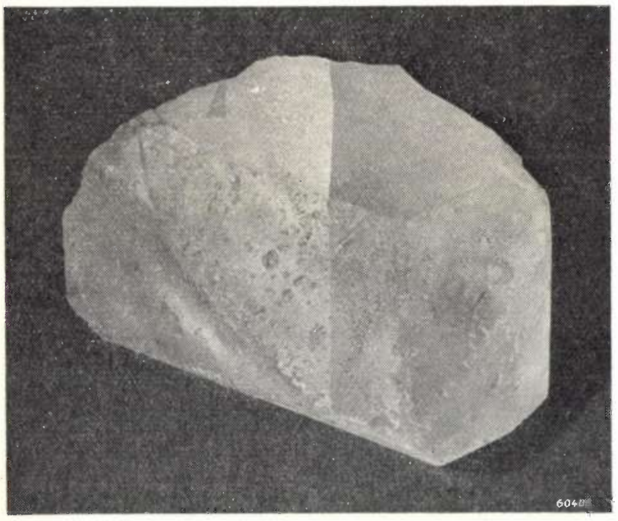


Fig. 15. Etched *X*-block showing electrical twin boundary in middle, with boundary surface approximately perpendicular to crystal surface. If this block is cut along the twin boundary, each piece can then be wafered separately and there will be no loss of material. If wafered prior to cutting apart, then one half of each wafer will be useless (cf. fig. 14*a*).

Summary. In cutting an oscillator plate out of a quartz crystal according to one of the so-called low temperature coefficient cuts, such as the AT, BT cut etc., the angle tolerances for the orientation of the crystal are very small, for example only 10 minutes. The method of obtaining the desired orientation of the crystal block under the saw is explained here, using the so-called *X*-block method for illustration. For a rough adjustment, methods based on the double refraction, amongst others, are used, while all ambiguity as to the direction of rotation of the saw table is eliminated by the observation of light patterns on etched crystal surfaces. Then the crystallographic position of a test cut is determined very accurately by means of an *X*-ray diffraction measurement and from this the required correction in the position of the saw table is deduced. The process of cutting is greatly affected by the twinning almost universally present in natural crystals. The types of twinning are briefly discussed and their influence on the "strategy" of the cutting is explained. Etching and examination at every stage of the cutting process gives greater yields and eliminates useless sawing.

PERCEPTION OF CONTRASTS WHEN THE CONTOURS OF DETAILS ARE BLURRED

by A. M. KRUITHOF.

612.843.355

The degree of contrast sensitivity is important when judging the performance of the human eye. Various investigators have made a study of contrast sensitivity in order to ascertain how far it is governed by such factors as the size and shape of the details observed and the brightness of the background. Little is known, however, about the influence that a blurring of the contours of the detail to be observed has in this respect. It is of importance to know more about this in connection with radiological, photometrical and pyrometrical investigations and for observations in the free field. To this end some experiments have been carried out whereby attention has been paid not only to the organs of the eye used in daytime (the cones) but also to those used at night (the rods).

Introduction

Visual perception is a complicated process. The eye is called upon to fulfil a great variety of tasks, whilst moreover it is expected to perform these duties with sufficient accuracy under greatly differing conditions of brightness.

One of the most important of these tasks is the perception of contrasts. In many cases it is a matter of distinguishing spots having a brightness differing but very little from that of the background against which they are seen.

Many investigators have taken measurements in respect to this perception of contrasts, whereby, inter alia, the size and shape of the spots, the brightness of the background and even the brightness of the surroundings of the actual field of observation have been varied [1]). In these investigations the measurements have mostly been confined, as far as brightness is concerned, to observations where almost exclusively the cones are used.

In some investigations the size and shape of the spot to be observed have been mainly left unchanged, only the contour being altered, for instance by giving it a saw-tooth or sinusoidal form. Little attention has been paid so far to the question as to what the influence may be of a gradual change of the brightness in the area bordering upon the spot to be observed. Yet this is a problem that is encountered in several methods of observation.

We have in mind here, in the first place, radiological examinations. In an X-ray image there is blurring due to the manner in which the image is formed [2]). Obviously this blurring affects perception when the spots to be studied are of the order of size of the blurred area. The same applies in regard to the blurring that often occurs as a result of the structure of the object (e.g. in the case of a gradually spreading pulmonary process). These two kinds of blurring may also affect the perceptibility of larger spots, and it is likely that the extent of their effect is related to the brightness of the image [3]).

This problem is also encountered in photometry. In visual photometry it is often a matter of observing the contrast between two degrees of brightness differing but little one from the other. One must then know whether, in order to get accurate results, it is necessary that the fields of different brightness are sharply defined.

In pyrometry, too, such a problem arises. In the case of the pyrometer working according to the principle of the disappearing wire, the brightness of an incandescent filament is made visually equal to that of the image of the object. For accurate measurements of low temperatures high brightness of the image is desired, and this is favourably influenced by a large aperture of the pyrometer. An objection against a large aperture is that blurring may occur along the filament [4]). The question now arises how far this affects the accuracy of the measurements.

[1]) A summary of these investigations has been given, e.g. by A. A. Kruithof and H. Zijl, Illumination intensity in offices and homes, Philips Techn. Rev. 8, 242-248, 1946.

[2]) G. C. E. Burger, B. Combée and J. H. van der Tuuk, X-ray fluoroscopy with enlarged image, Philips Techn. Rev. 8, 321-329, 1946; H. A. Klasens, The blurring of X-ray images, Philips Techn. Rev. 9, 364-369, 1947.

[3]) The problem of blurring of contours is encountered also when taking observations with a radar screen.

[4]) See C. O. Fairchild and W. A. Hoover, J. Opt. Soc. Amer. 7, 543-579, 1923.

Middleton [5]) has investigated the effect of blurred contours upon contrast sensitivity, being led thereto when considering a problem differing greatly from that just mentioned above, namely the visibility of objects in a mist. In a fog the shape of objects cannot be sharply discerned owing to phenomena of scattering along the boundaries. Middleton's measurements were taken with a field 7.5 cm × 12 cm divided into two areas with a gradual transition of brightness, and with a variable width of the blurred "edge". The brighter part of the field had a brightness of 32 cd/m^2 [6]). The surroundings were dark or had a brightness amounting to one-fifth of that of the actual field of vision. It was found that the contrast threshold suddenly increased when the transitional area was made so wide as to be seen at an angle of 7′ to 8′. At the brightness of 32 cd/m^2 (the only level at which the investigation was carried out) the organs of the eye in action are the cones.

In view of the possible fields of application mentioned above it is desired to investigate this effect more fully, it not being sufficient to confine the investigations to brightnesses where the day organs are used. It is also necessary to ascertain what happens when the rods are in use, and in particular when having regard to the application of the results to radiological observations. An X-ray image on a fluorescent screen has a brightness of 0.0003 to 0.006 cd/m^2, which is so low that only the elements for night sight are brought into action when one studies that image. When, however, a radiograph is made and the photographic plate or film is studied in the light of a viewing box, the brightness of the same image is so much greater that the cones of the eye are used. It is therefore of importance to investigate whether the effect of a blurred contour upon visibility is the same in both cases.

It is not likely that this blurring will have the same effect for high and for low levels of brightness, since it is common knowledge that whereas contours can quite well be sharply (accurately) observed with the organs used in daytime such is practically precluded with the organs for night sight; as everyone can easily determine for himself, at night it is very difficult to discern sharply the outlines of an object. This is accounted for by the construction of the retina of the eye. The part of the retina normally used brings far more cones into play for daytime sight than the number of rods that come into action for night sight. This is because the cones lie much closer together than the rods, whilst moreover the rods are connected in groups (with a diameter corresponding to an angle of about 20′) to one nerve fibre whereas the majority of the cones are connected to a nerve fibre of their own. As a consequence of this structure of the retina one may expect for cone sight a resolving power such that details of 0.5-1′ can still be perceived, whereas for rod sight the limit will certainly not be lower than 20′.

The effect of blurred boundaries can be studied for all levels of brightness of the background and for different dimensions and shapes of the spot. It should be possible to carry out these measurements not only with white light but also with coloured light, but it was not our intention to extend our investigations so far as that. We have confined our experiments to the case of circular spots of one certain diameter and to a high and a low level of brightness of the background, for which 50 and 0.0025 cd/m^2 respectively were chosen. The colour of the light was determined by the colour temperature of the incandescent lamp, viz. approx. 2800 °K.

Principle of the experiments

When we have a background with a brightness B and a spot is projected thereon in such a way that its brightness is ΔB greater (or smaller), then we call $\Delta B/B$ the contrast. If ΔB is the smallest still perceptible difference in brightness we call $B/\Delta B$ the contrast sensitivity.

If we now denote the contrast sensitivity by $B/\Delta B$ for the case where the spot is sharply defined and by $B/\Delta'B$ for the case where the spot has a blurred contour, then the ratio of the contrast sensitivity in the case of blurring to that in the case of a sharp delineation is given by $\Delta B/\Delta'B$. If the sharply defined spot can be seen better than the blurred spot then $\Delta B/\Delta'B$ is less than 1.

The experiments to be described here furnish a relation between $\Delta B/\Delta'B$ and the width of the blurred boundary zone. It is to be expected a priori that for very narrow zones $\Delta B/\Delta'B$ will be equal to 1 and for wider zones less than 1. For the lower level of brightness $\Delta B/\Delta'B$ will presumably assume values less than 1 at a greater width of the boundary zone than will be the case for the higher level of brightness, owing to the fact that at a low level of brightness the rods are used, with which we cannot see so sharply.

We shall now first describe the set-up used for these experiments for measuring $\Delta B/\Delta'B$ as a function of the boundary width.

[5]) J. Opt. Soc. Amer. **27**, 112-116, 1937.

[6]) cd = candela, the now internationally accepted denomination for the standard candle; see Philips Techn. Rev. **10**, 150-153, 1948 (No. 5).

Description of the set-up

The experimental apparatus employed is represented diagrammatically in *fig. 1.* A white screen *S* (30 cm × 30 cm) with a reflection coefficient of 0.80 is illuminated with an incandescent lamp (not drawn) to a brightness of 0.0025 or 50 cd/m^2. The surroundings of the screen are not entirely dark but have a brightness of one-fourth to one-third of that of the screen. A circular, screened, piece of frosted glass *M* receives light from the lamp *P* and is projected with the aid of a three-fold lens *L* onto the screen *S*, the dimensions being so chosen that the image has a diameter of 3.5 cm. The lamp *P* can be moved along rails, so that *M* can be given the desired brightness by varying the distance between the lamp and the glass.

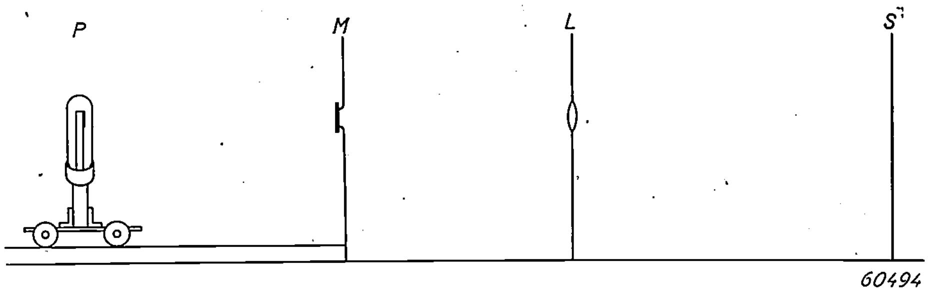


Fig. 1. Diagrammatic representation of the experimental apparatus. *P* the movable lamp, *M* the frosted glass, *L* the lens, which can be moved either in the direction of the frosted glass or towards the screen *S*.

The image of the piece of frosted glass forms on the screen a spot having a greater brightness than the background. The observer takes up a position 60 cm away from the screen and sees a spot with a diameter of 3.5 cm viewed from an angle of 3.5 °. Of course it depends upon the difference in brightness between the spot and the screen whether he does indeed see the image or not.

Let us suppose that the distances are so chosen that the lens casts a sharp image of the frosted glass on the screen. Upon the lens being moved, either in the direction of the screen or in that of the frosted glass, the image on the screen will be blurred, and the greater the displacement of the lens the wider will be the area of confusion.

When, after focusing, the lens is moved towards the screen the passage of the rays will be as indicated in *fig. 2*. From the similarity of triangles it follows that $p - q = (d/b)\cdot 2r$, where $p - q$ denotes the width of the area of confusion, *d* the distance of the screen behind the image plane, *b* the image distance and $2r$ the diameter of the lens. For the case where the lens is moved towards the frosted glass we find in the same way $q - p = (d/b)\cdot 2r$, where *d* is then the distance of the screen in front of the image plane.

The distance between the frosted glass and the screen is kept unchanged. From this distance, the focal length of the lens, the diameter of the lens, the object distance *ML* and one of the above formulae one can then calculate the width of the area of confusion in mm. From this width one can calculate the visual angle from which the observer sees the area of confusion, since the distance between the observer and the screen is known. In *table I* the lens displacements are given which were needed in our experiments to get an area of confusion of a certain width (expressed in arc units).

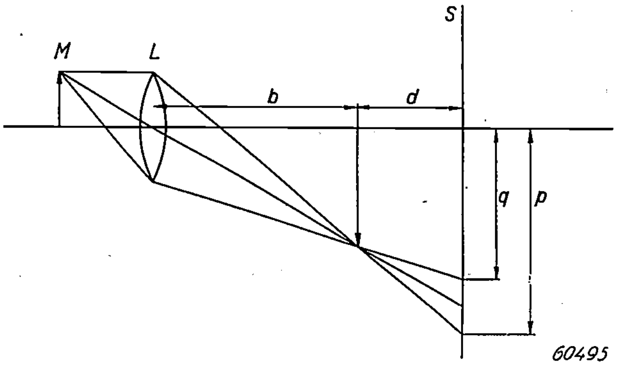


Fig. 2. Passage of the rays in the formation of an image with an area of confusion. *M* the frosted glass, *L* the lens, *S* the screen, *p*—*q* the width of the area of confusion, *b* the image distance and *d* the distance of the screen behind the image plane.

Table I. The lens displacements which were necessary to get an area of confusion of a certain width (expressed in arc units) and the extra brightness ΔB of the central part of the spot on the screen with a brightness $B = 50$ cd/m^2 at these distances from the lens.

Width of area of confusion (in ′)	Lens displacement in the direction of *S* (in mm)	Lens displacement in the direction of *M* (in mm)	Extra brightness ΔB (in cd/m^2) of the central part of the spot when shifting the lens in the direction of	
			S	*M*
0.0	—	—	2.92	2.92
2.4	2.7	2.3	2.93	2.88
6.0	6.8	5.0	3.01	2.84
12	14.1	10.0	3.07	2.77
24	29.2	19.0	3.22	2.67
36	60.0	27.0	3.54	2.63

The variation of the brightness in the area of confusion is of a simple nature if it is permitted to assume that the frosted glass and the screen have a distribution of radiation according to Lambert's law at least for directions making a small angle with the normal, and if, moreover, the distances from M to L and from L to S are great with respect to the diameters of M, L, the spot and its area of confusion. To get an idea of this variation of the brightness, let us take a look at *fig. 3*, where PQ represents the width of the said area.

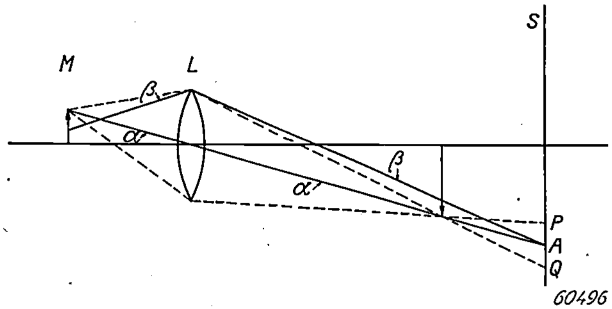


Fig. 3. Diagram illustrating the change of brightness in the area of confusion PQ. α and β are the boundary rays of the beam determining the brightness at A.

Above P one observes the full brightness of the spot, and below Q the brightness is zero. Halfway between P and Q is the point A where the rays merge that are contained in the beam bounded by the rays α and β. In the same way one can draw in the cross section under consideration for any point of the spot and its boundary zone the marginal rays of the beam producing the brightness at that point.

By carrying out the following experiment in our imagination we can form a picture of the cross section of the conical beam reaching the screen at any arbitrary point A' between P and Q. We put our eye at a point on the screen above P and direct it upon the image of the frosted glass formed by the lens, which it is assumed could be perceived in the image plane. If this "image of the frosted glass" were made visible we should see it as a very large circle. Our eye, placed at a point of the screen above P, sees the lens as a small circle lying entirely within the "image of the frosted glass". Now we move our eye in the direction towards P and then via A' to Q. The small circle (the lens) will then first approach the boundary of the large circle (the image of the frosted glass), which is practically a straight line, and then cross it. This is schematically represented in *fig. 4*. The lens is only visible in its entirety so long as the cone with the eye as apex passing through the circumference of the lens falls within the cone having the same apex and passing through the circumference of the image of the frosted glass. When we reach with our eye a point on the screen where the first cone falls only partly within the second one, that point will receive less light. When the eye has reached Q the lens circle crosses the boundary of the image of the frosted glass; one cone then lies outside the other, and at this point the frosted glass no longer gives any brightness.

If this can be understood then it is clear how the brightness of the points between P and Q can be calculated. This brightness is proportional to the area of the segment of the lens circle observed in the points of this zone above the edge of the image of the frosted glass. A segment with the height a (fig. 4) corresponds to a point A' at such a distance between P and Q that $QA' : A'P = a : (2\varrho - a)$, where ϱ is the radius of the image of the lens.

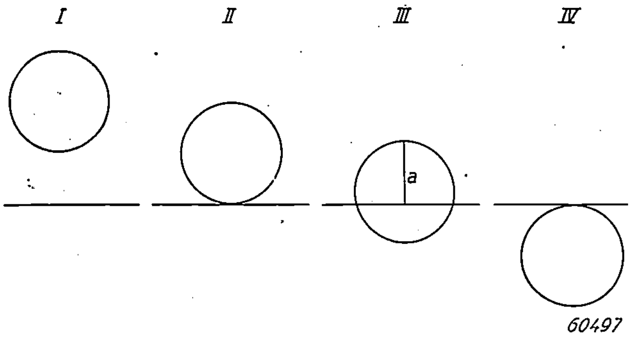


Fig. 4. Position of the image of the lens with respect to the "image of the frosted glass" in an imaginary experiment where the eye is moved to different points of the screen. The horizontal line represents the edge of the (very large) image of the frosted glass and the circles represent the image of the lens. The four illustrations relate to the cases where the eye is situated: I above P (see fig. 3), II at P, III at a point A' between P and Q, and IV at Q.

The calculation shows that the brightness in the area of confusion varies approximately in a straight line. It is found that the distribution of brightness of the spot with its area of confusion is as represented diagrammatically in *fig. 5*. To a first approximation this figure is a trapezium. From the description of the set-up of the apparatus depicted in fig. 1 it follows immediately that the brightness of the central part of the spot undergoes a change when the lens is displaced either in one direction or in the other. Thus the height of the trapezium in fig. 5 depends upon the position of the lens. Consequently, when measurements are taken after the lens has been displaced, a correction has to be made.

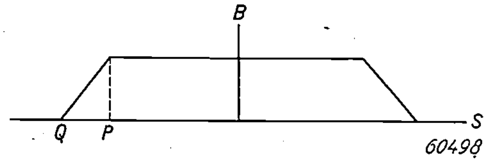


Fig. 5. Diagram showing the change of brightness B in the spot and in the area of confusion. P and Q relate to the points denoted by these letters in fig. 3.

With a view to making this correction the brightness of the central part of the spot was determined for the various positions of the lens before lighting the lamp illuminating the whole of the screen. With the lamp P at a certain distance from the frosted glass M the illumination in the spot on the screen S was measured with the aid of a selenium cell, and from that the brightness was calculated. This was repeated after the lens (with P at the same distance) had been shifted to one of the positions given in table I. The increase of the brightness of the spot in cd/m^2 with the respect to the brightness of 50 cd/m^2 of the screen is also shown in table I. The position of the lens is characterized by the angle at which the observer sees the boundary zone of the spot with the lens in the respective position [7]).

For the lower brightness level of the screen (0.0025 cd/m^2) the brightness of the spot was reduced to the desired level with the aid of neutral filters of a known transmission.

How the measurements were taken

In carrying out these measurements two people worked together, one of whom will be called the observer and the other the investigator.

The observer is the one who took up a position 60 cm away from the screen and during the experiments adjusted the contrast so that he could just perceive the spot; he does this by placing the lamp P in such a position that he sees the spot appear in the background or disappear.

For our experiments we had two male observers, one 26 and the other 35 years of age, whose eyes showed no particular aberrations and whose contrast sensitivity may be taken as normal.

The investigator noted the position of the lamp after each adjustment, without communicating this to the observer.

A series of experiments begins with the observing of a spot that is sharply delineated. Then by gradually increasing the distance between the lamp and the frosted glass the observer makes 10 adjustments in succession until he can just no longer see the spot, after which he makes 10 more adjustments by drawing the lamp towards the frosted glass until the spot just becomes visible. The arithmetical mean of the 20 brightnesses of the spot corresponding to these adjustments gives the contrast sensitivity of the observer with the given brightness of the background and a sharply delineated spot seen at an angle of 3.5 degrees.

Next, the lens is moved, for instance first towards the screen, over such a distance that the spot has an area of confusion of 2.4′, at which distance the 20 measurements are taken again. These are repeated once more after the lens has been moved in the opposite direction so far that the spot again has an area of confusion of the same width. After a correction has been made for the changed brightness of the spot as indicated in table I, these 40 measurements give the contrast sensitivity of the same observer for the same brightness of the background and a spot of 3.5° diameter with an area of confusion of 2.4′.

The contrast sensitivity for the other widths of the area of confusion given in table I is determined in the same way. It is then possible to calculate the ratio $\Delta B/\Delta' B$ as function of the width of the area of confusion.

Care has to be taken that the part of the spot with the maximum brightness always has a diameter of approximately 3.5 degrees, this being done by adjusting the diaphragm in front of the frosted glass.

It was found impracticable to carry out all these adjustments after the other because it was too tiring for the observer. When a series had to be interrupted it was resumed next day by starting again with observations of the sharply delineated spot and then carrying on with the remaining observations of the spot with blurred edge.

It has to be added that when the adjustments were made the observer had to rest his head in a certain way against a support, so that both the distance from the screen and the position of the head were fixed.

In these experiments one has to rely upon the observer's promise not to fix his eye outside the spot. Fixing of the eyes upon the right place is particularly necessary when the experiments are carried out at the very low brightness of 0.0025 cd/m^2. At such a low level there is a natural tendency not to fix the eye on the spot, because the sensitivity of the eye outside the central part of the retina is then greater than the sensitivity in that part itself, whilst with higher levels of brightness the reversc is the case. It goes without saying that the room in which the experiments were conducted was otherwise quite dark, and that the observer was given ample time (about 15 minutes) to adapt his eyes to the brightness of the screen and his surroundings.

Results of the experiments

Extensive experiments with the aid of the apparatus described and two observers each completing the full series of adjustments with a background brightness of 50 cd/m^2 yielded the results tabulated in *table II* and graphically represented in *fig. 6.*

[7]) The changes in brightness can also be found by calculation. The corrections determined in this way were in agreement with the results of the measurements.

Table II. Change of the contrast sensitivity of two observers as function of the width of the area of confusion of a spot against a background brightness of 50 cd/cm².

Width of area of confusion (in ′)	$\frac{\Delta B}{\Delta' B}$ first series	$\frac{\Delta B}{\Delta' B}$ second series	$\frac{\Delta B}{\Delta' B}$ average
0	1.00	1.00	1.00
2.4	1.055	1.02	1.04
6.0	0.99	1.03	1.01
8.4	—	0.86	0.86
12	0.69	0.77	0.73
24	0.585	0.775	0.68
36	0.625	0.715	0.67

It appears that up to a width of about 7′ for the area of confusion there is no decline in the contrast sensitivity, but that it then diminishes rapidly to about two-thirds of the original value until a width of 12′ is reached, after which it remains practically constant.

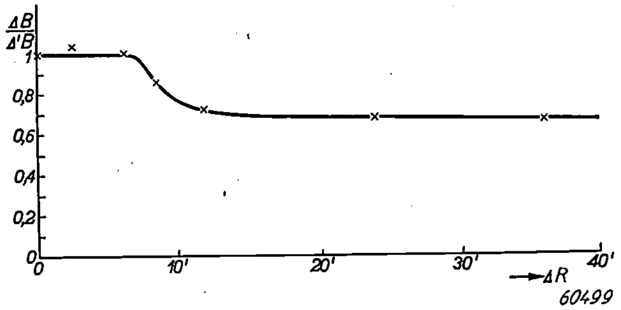


Fig. 6. The ratio $\Delta B/\Delta' B$ of the contrast sensitivities when observing a sharply delineated spot and a spot with an area of confusion, as function of the width of that zone against a screen brightness of 50 cd/m². The width of the area of confusion ΔR has been plotted along the horizontal axis. Each point of the curve is the result of 160 adjustments.

The width of the blurred edge where contrast sensitivity begins to diminish according to our experiments is the same as that which Middleton found in his investigations mentioned above. His measurements too give an indication that with greater widths of the area of confusion there is a less pronounced decline in contrast sensitivity, but there was not such a marked return to a practically constant value, and the width of the area of confusion where this began was greater. This may be due to the fact that Middleton used a test object of larger dimensions and different shape. We are of opinion, however, that the decline found from both investigations at 7′ to 8′ is connected with the nature of the retina, having regard to the distance between the cones and the nerve connections at and possibly between the cones.

The experiments described above were also carried out with the two observers when the screen had a lower level of brightness (0.0025 cd/m²).

Table III. Change of contrast sensitivity (average of two observers) as function of the width of the area of confusion of a spot against a background brightness of 0.0025 cd/m².

Width of the area of confusion (in ′)	$\frac{\Delta B}{\Delta' B}$
0	1.00
2.4	1.08
6.0	0.95
12	0.97
17	0.97
20.5	1.08
24	0.82
36	0.83

The average of the results of their observations is given in *table III* and represented in *fig. 7*.

It appears that with this background brightness the ratio of the contrast sensitivity for sharp delineation and for a blurred contour is constant until an area of confusion of about 20′ is reached, after which it drops rather suddenly to about 0.82. Since the angle of 20′ also plays a part in the structure of the retina there is reason to assume that the cause of the reduced contrast sensitivity at this width of the area of confusion is connected with the composition of the rod system.

It is seen that both when the cones are used and when the rods are brought into action there is a decline in contrast sensitivity as soon as the area of confusion exceeds a certain width. Under otherwise the same conditions the reduction when the rods are in action is less than that when the cones are used, but then it must of course be borne in mind that when the organs of day sight are in use contrast sensitivity is very much greater (25 to 50 times) than when the organs of night sight are used.

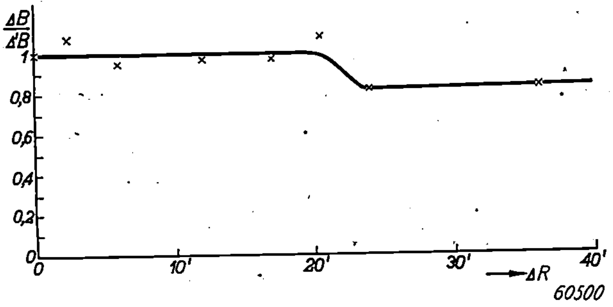


Fig. 7. The same function as plotted in fig. 6, but for a screen brightness of 0.0025 cd/m². Each point in the curve is the result of 80 adjustments.

Some conclusions

From the experiments described here it follows that in radiological examinations, both with high and with low brightness levels, blurred contours of the image will only slightly affect visibility, at least where it is a matter of observing contrasts in relatively large objects.

It appears that in photometry a sharply defined boundary is not strictly necessary. If a photometer is so constructed that the transition between two areas of almost the same brightness is gradual this need not prevent accurate results being obtained.

In the case of optical pyrometers the effect observed means that larger apertures can be used, and thus lower temperatures measured, without any noticeable loss of accuracy.

Summary. The observing of contrast is an important duty of the eye. In most cases it is a matter of perceiving spots differing only slightly in brightness from the background against which they are seen. In connection with various applications (radiological examinations, photometry, pyrometry) it is of importance to investigate experimentally in how far blurring of the contours of a spot that is to be observed affects the contrast sensitivity of the eye. A description is given of the apparatus used for such an investigation. Experiments have been carried out with a round spot (diameter 3.5 cm) projected in such a way that its brightness was slightly greater than the background; two cases were taken, viz. with screen brightnesses of 50 and 0.0025 cd/m^2. It appeared that when the daytime organs of sight are used (the cones) contrast sensitivity is not reduced by a blurred contour until the width of the area of confusion has grown to an angle of about 7′, when it rapidly declines to about two-thirds of its original value at an area of confusion having a width of 12′, after which it remains practically constant. When the organs of night sight are used (the rods) there is a slight decline in contrast sensitivity when the width of the area of confusion is extended to about 12′. From these results some conclusions are drawn for the above-mentioned technical applications.

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

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

1880: J. ter Berg and G. J. van Wijnen: The porosity of welds (Welding J. **28**, 269 S-271 S, 1949, No. 6).

A study of external porosity in welds, as caused by sulphur. By relatively slight alterations in the composition of the coating it is possible to obtain welds free from this porosity, even with a considerable sulphur content of core wire and coating. Two methods were applied:

1) The coating was made more basic in order to absorb more S in the slag.

2) The oxidizing power of the coating was increased in order to get rid of sulphur in the form of SO_2.

1881: N. W. Smit and F. A. Kröger: The luminescence of zinc sulfide activated by lead (J. Opt. Soc. Amer. **39**, 661-663, 1949, No. 8).

Zinc sulfide activated by lead shows fluorescence in various bands, two of which can be attributed to lead. The first band has a maximum at 4850 Å, the second one at 6100 Å. The orange band is favoured by sulfurizing conditions and appears irrespective of whether halide ions are present or not. This band is attributed to a characteristic electronic transition in divalent lead ions, occupying normal lattice sites. The green band is favoured by reducing conditions and only appears when chlorine ions are present. This band is attributed to electron transfer transitions between positive and negative ions of configurations formed by monovalent lead and chlorine ions. situated at normal lattice sites, together with the neighbouring lattice ions. A green band appearing in reduced zinc sulfide is attributed to a stoichiometric excess of zinc.

1882: R. van der Veen: Induction phenomena in photosynthesis, I (Physiologia Plantarum **2**, 217-234, 1949).

Adaptation phenomena, which occur in photosynthesis when leaves are suddenly illuminated, in an atmosphere containing CO_2 can be divided into an initial sudden CO_2-uptake (I.U.), followed by a light adaptation period with constant negative slope of CO_2-content (L.A.S.). The I.U. is independent of temperature and of the length of the preceding dark period, while the L.A.S. is strongly influenced by these factors.

When leaves are exposed to a high temperature (± 48 °C) for a few minutes and afterwards examined at normal temperature, the capacity of CO_2 assimilation is irreparably damaged, but the I.U. still exists. When the illumination is stopped the amount of absorbed CO_2 is released by such leaves.

At low temperature CO_2 assimilation is also inhibited, but the I.U. is still quite strong. At the end of the illumination the absorbed CO_2 is not released, but after a short period of darkness (less than 15 minutes) the I.U. is not maximal. The duration of the I.U. is about 20 seconds.

So in normal leaves at low temperature the I.U. is not reversible, while in heat-treated leaves it is made reversible.

In the discussion a theoretical scheme is developed to explain these facts.

1883: J. A. Haringx: Elastic stability of flat spiral springs (Appl. sci. Res. The Hague **A2**, 9-30, 1949, No. 1).

This paper deals with the elastic stability of the flat spiral spring, that is a structure consisting of a wire coiled into a spiral lying in a plane. A rough calculation, valid for spiral springs having a large number of coils, shows that the critical number of turns which the spring has to be wound or unwound to reach the state of instability is determined only by the ratio of the sides of the rectangular wire section. It depends neither upon the number of coils nor upon Young's modulus of the wire material.

In order to verify in how far this holds for a spring with a small number of coils a second and more accurate calculation is given, though only for a (fictitious) spiral spring having circular and identical coils. Further, it is shown that the realization of a spiral spring able to undergo more than, say, three turns without becoming unstable is impossible.

R 116: G. Diemer and K. S. Knol: Measurements on total-emission conductance at 35-cm and 15-cm wavelength (Philips Res. Rep. **4**, 321-333, 1949, No. 5).

For the contents of this article see these abstracts, No. **1870***.

R 117: L. J. Dijkstra and J. L. Snoek: On the propagation of large Barkhausen discontinuities in Ni-Fe alloys (Philips Res. Rep. **4**, 334-356, 1949, No. 5).

The propagation of the Bloch boundary between two macrodomains under the influence of an external magnetic field H has been investigated for Ni-Fe wires, of the composition 60-40 and 50-50, subjected to a tensile stress. The interesting quantities involved are the effective length λ of the discontinuity and the rate of propagation v. The length λ is proportional to the diameter of the wire and independent of the temperature and the working conditions. In thick, well annealed wires, at low temperature, the velocity v is determined chiefly by the counteracting field of the eddy currents. With thin wires and at high temperatures, especially in cold-worked material, another limiting factor comes to the fore, which, however, is as yet not sufficiently understood.

R 118: J. L. H. Jonker: The computation of electrode systems in which the grids are lined up (Philips Res. Rep. **4**, 357-365, 1949, No. 5).

Formulae are developed describing the path of the electrons and the position of the focus in a system of electrodes in which the grids are lined up. These are then applied to the calculation of a plane arrangement such as to possess prescribed characteristics and to have zero screen-grid current when the control grid is at zero potential.

R 119: B. D. H. Tellegen: Complementary note on the synthesis of passive, resistanceless four-poles (Philips Res. Rep. **4**, 366-369, 1949, No. 5).

By the removal of a pole at a finite value of the frequency a passive, resistanceless four-pole of order n may be split up into a second-order four-pole of order n—2 (see also **R 89**).

R 120: J. D. Fast: An apparatus for preparing small samples of pure iron to which fixed quantities of impurities can be added (Philips Res. Rep. **4**, 370-374, 1949, No. 5).

A high-frequency melting apparatus is described for preparing alloys of well-defined compositions in quantities up to 2000 grams (see Philips Techn. Rev. **11**, 244-247, 1949, No. 8).

R 121: J. A. Haringx: On highly compressible helical springs and rubber rods, and their application for vibration-free mountings, V (Philips Res. Rep. **4**, 375-400, 1949, No. 5).

This paper deals with the behaviour in space of the different types of vibration-free mounting, starting from the simple construction of a resiliently supported body up to the damped dynamic vibration absorber provided with an auxiliary mass. In order to simplify the problems of forced and free vibrations it is aimed at splitting up the movements in space into a number of one- and two-dimensional movements to be treated independently. In this connection it appears to be necessary to confine the present considerations to constructions where the spring and damping systems show some special properties with regard to their so-called principal axes, of elasticity and of damping. These systems of principal axes, which exist in each two-dimensional case, only exceptionally occur in space. The respective requirements can best be met by introducing in the construction a certain degree of symmetry, as is elucidated for the two- and the three-dimensional case separately. (See these abstracts, Nos. **R 94, R 101, R 109,** and **R 113.**)

VOL. 11 No. 12, pp. 341-372 JUNE 1950

Philips Technical Review

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

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

A NEW THERMIONIC CATHODE FOR HEAVY LOADS

by H. J. LEMMENS, M. J. JANSEN and R. LOOSJES.

621.3.032.213.2:
621.385.13.029.6

The fact that the science of electronics — that branch of electrotechnical engineering dealing with free electrons *— has become of such great importance is for the greater part due to the ease with which free electrons can be* produced*: a conductor or semi-conductor is heated and electrons "evaporate" from it. The thermionic cathode based upon this principle has become an integral part of innumerable types of electronic valves, a part to which little attention is now usually paid in the descriptions of the valves because in their construction the cathode does not give rise to any particular difficulties. Various modern types of valves, however, especially those for ultra short waves, have to satisfy such extreme demands that the cathode has again become an "interesting" part. A new cathode developed in the Philips Laboratories at Eindhoven (Holland), designated as the L cathode, will be able to answer these demands in many cases, due, in part, to its high maximum emission; it can yield some hundreds of amperes per sq.cm of its surface and still retain a reasonable length of life.*

Different types of thermionic cathodes

Nowadays three types of thermionic cathodes are being widely used in electronic valves: the tungsten cathode, the thoriated tungsten cathode and the oxide-coated cathode. *Table I* gives a survey of a number of characteristic features of these cathodes.

The oxide-coated cathode has by far the best thermal efficiency, requiring the least heating power for a given electronic emission. Consequently this cathode is indicated for radio receiving valves, where the filament current consumption is an important point of consideration. Adverse properties of the oxide-coated cathode, such as its susceptibility to „poisoning". (reducing the emission) through traces of oxygen or other gases in the valve, and its evaporation of barium and strontium, which may cause the grids and the anode to emit electrons too, are not very objectionable in this case. For transmitting valves and X-ray tubes, however, the last-mentioned drawback is a very great objection. Moreover, for these valves and tubes the cathode must be able to stand up against the electrostatic forces of attraction of the anode, which is at a high potential, and this is not a property

Table I. Properties of some common types of cathodes.

Type		Maximum useful thermionic emission in A/cm²	Maximum useful thermal efficiency in A/W	Susceptibility to poisoning	Proof against high tension	Proof against high-velocity gas ions
Tungsten		1	0.006	slight	good	good
Thoriated tungsten		2	0.070	great	good	poor
Oxide-coated cathode *)	D.C.	0.5	0.25	great	poor	during a short time good
	pulse emission	50	20	great	fair	

*) $BaCO_3$ and $SrCO_3$ applied in a ratio of about 1:1 by wt.

possessed by the normal oxide-coated cathode. Since in these cases the thermal efficiency of the cathode is of less importance, tungsten or thoriated tungsten cathodes can quite well be used.

Whereas for the valves and tubes mentioned, and also for cathode-ray tubes, high-tension valves, iconoscopes, etc., one of these three types of cathodes has therefore always provided a more or less satisfactory solution, this has not always been found possible for modern varieties of electronic valves such as are now used for generating or amplifying ultra short waves, because in some cases entirely new combinations of properties are demanded. In the case of magnetrons for radar installations for instance the cathode is required to yield momentarily an emission with current densities [1]) of some tens of amperes per sq.cm, whilst at the same time it must be proof against high tensions and also against a bombardment by accelerated electrons returning to the cathode. To a certain extent this has been provided against by reinforcing the oxide coating with metal, mostly nickel, applied in the form of a gauze or of a coarse powder sintered onto the support [2]). Another, fundamental, difficulty that arises with valves for ultra short waves concerns their degassing: such valves usually contain relatively large parts made of copper, which may not be heated to such a high temperature as is desired for driving out the last traces of gas. Consequently the activation of an oxide-coated cathode in such a valve is hampered on account of the gases thereby released from the copper, poisoning this type of cathode and more or less permanently damaging it. On the other hand cathodes of tungsten or thoriated tungsten are as a rule unsuitable for use in ultra-short-wave valves because in order to obtain a reasonably high power the cathodes must have a very high specific emission (sometimes in the form of pulses) on account of the very small dimensions prescribed by the wavelength.

Having regard to the requirements for a particular kind of valve, Philips Laboratories at Eindhoven have recently developed a new type of cathode combining great mechanical strength with favourable thermionic properties and with high resistivity against impurities. From these and some other good properties it is to be expected that this cathode, which is to be designated as the L cathode, will be excellently suitable for some modern types of valves and that the limitations hitherto set by the cathode in the construction of these valves will now be removed. Experience with several types of valves has in fact already confirmed this expectation.

The construction of this L cathode will now be described, and this will be followed by an account of its main properties and an explanation of the mechanism of its emission.

Construction of the L cathode

Figs 1a and *1b* represent two forms of the L cathode in cross section. Cathodes of the design shown in fig. 1*a* have a cylindrical emitting surface and can be used for instance in magnetrons, whilst those of the design in fig. 1*b* have a flat, circular, emitting surface and are suitable for velocity-modulation and reflex-oscillator valves ("klystrons"), disc-seal diodes and triodes, and also, for example, for cathode ray tubes and iconoscopes. For the sake of simplicity the further description will be confined to the form of cathode depicted in fig. 1*b*.

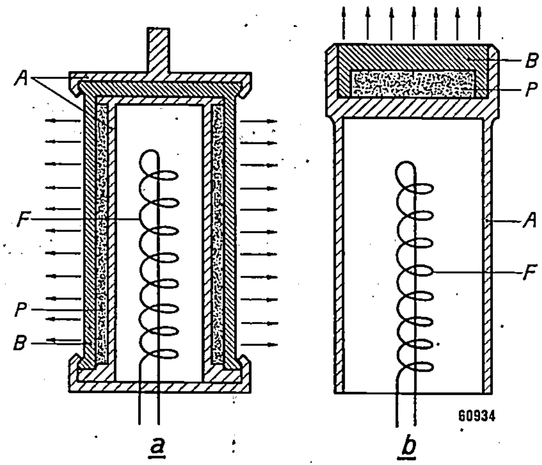


Fig. 1. Cross section of two basic forms of the L cathode. *a*) With cylindrical emitting surface. *b*) With flat, circular, emitting surface. *A* = wall of molybdenum, *B* = wall of porous tungsten, *P* = tablet of barium-strontium carbonate, *F* = filament.

The cathode consists of two chambers shaped out of one piece of molybdenum. The lower chamber is open at one end and contains an insulated filament for indirect heating of the cathode. The upper chamber is closed with a cap of porous tungsten, underneath which is a tablet of barium-strontium-carbonate (in the proportions of about 1 : 1 by weight); the only connection with the outside from this upper chamber is via the pores of the tungsten. Supporting rods can be welded to the bottom end of the cathode for fixing it in the right place.

[1]) The emission need not be purely thermionic, but can be obtained partly as secondary emission; this will be dealt with farther on.

[2]) J. B. Fisk, H. D. Hagstrum and P. L. Hartman, Bell Syst. Techn. J. 25, 167-348, 1946.

The cathode of the design according to fig. 1*a* is likewise divided into two chambers, that containing the barium-strontium-carbonate being connected with the outside only via a jacket of porous tungsten.

All sorts of variations of these two basic forms of the construction of the L cathode are possible, some of which are illustrated in *fig. 2.*

This temperature-dependency is graphically represented in *fig. 3* for the L cathode and for the three types of cathodes given in table I, by plotting the logarithm of the saturation emission as a function of the temperature. The saturation emission is about the maximum emission to be obtained from a cathode at a given temperature.

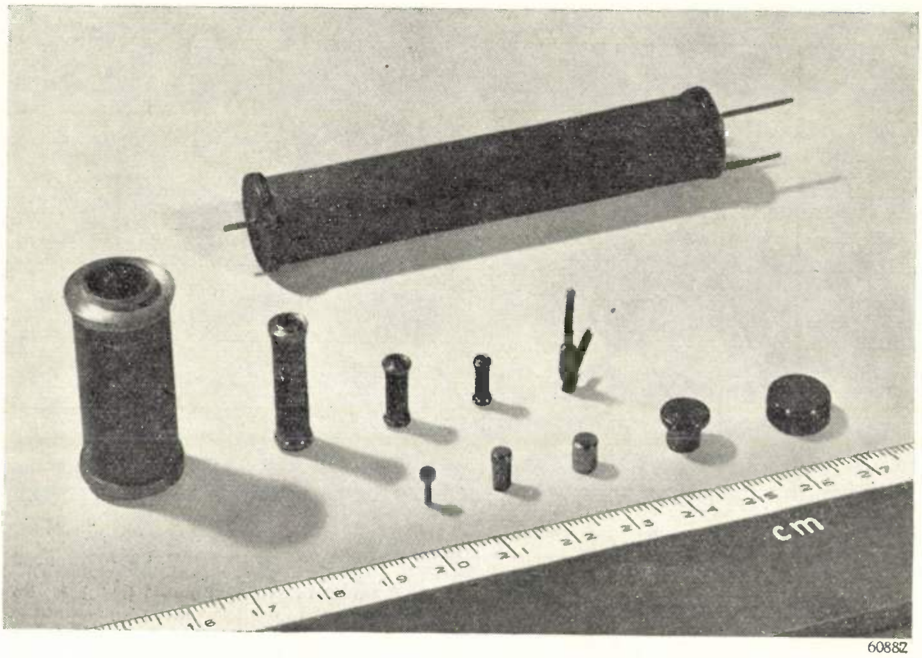


Fig. 2. Various forms of execution of the L cathode, for different tubes.

The cap (fig. 1*b*) or the cylinder (fig. 1*a*) of porous tungsten is made by compressing tungsten powder under high pressure and then sintering it at high temperatures.

After it has been mounted in a valve the cathode is activated by heating it in vacuo first to a temperature of about 1100 °C, whereby the barium-strontium carbonate gives off carbon dioxide and is converted into barium-strontium oxide, and then, after the carbon dioxide escaping through the porous tungsten wall has been pumped off, to a higher temperature. During this further heating the cathode begins to emit electrons thermally and sooner or later, depending on the temperature, the emission reaches an almost constant value. The cathode is then ready for use.

Thermionic emission of the L cathode

What is primarily of interest when comparing different cathodes is the temperature-dependency of the emission (by emission is to be understood here the density of the electron current emitted, in A/cm^2).

As regards the temperature required to obtain a given saturation emission, fig. 3 shows that the L cathode occupies a place between the oxide-coated cathode and the thoriated tungsten cathode. In this connection it is to be noted that when an oxide-coated cathode is used continuously no more than about one hundredth part of the saturation emission may be taken from it, as otherwise the emitting layer would suffer damage. The other types of cathodes can operate continuously with an emission close to the saturation value. Consequently, for a certain desired emission the working temperature of the oxide-coated cathode cannot be chosen quite so low as might be concluded from fig. 3.

Turning now to the question as to what is the maximum emission that can be obtained, i.e. the extreme point to which the curves in fig. 3 have been drawn, it is seen that with the L cathode we have got much farther than with any other type of cathode. The supremacy of the L cathode is even more impressing when it is realized that the termination of the curves is not due to the same causes in all cases. The curves for the three types of cathodes previously mentioned do not extend any farther because at higher temperatures the cathodes are very soon destroyed; the extreme end of the curve for the oxide-coated cathode, extending to about 120 A/cm², can even only be reached with pulse emission. In the case of the L cathode, however, the extreme point of the curve, lying at about 300 A/cm², which in principle is to be reached equally well with continuous as with pulse emission, is in point of fact not determined by the cathode itself but by difficulties in measuring. At a pulse emission of about 300 A/cm² in the measuring tubes used, a gas discharge took place, presumably due to the formation of gas from the anode material as a result of the sudden heating: during a pulse the anode dissipates several megawatts per sq.cm. When measuring with continuous emission this anode dissipation gives rise to still greater difficulties. Naturally the measurements have to be taken with a system of electrodes in which an accurately known and homogeneously loaded cathode area takes part in the emission. We used for this purpose a flat cathode, with an emitting surface of say 6 mm², and a flat anode in parallel. The distance between cathode and anode was made very small, so that for an emission of 40 A/cm² (the heaviest D.C. load of the L cathode so far reached) the anode voltage required in view of the counteracting space charge amounted to only 400 volts. With that emission this still means a dissipation of 1 kW, on a small anode surface; even with water-cooling it was not easy to carry off such a power.

It is quite possible that in practice better solutions can be found for these difficulties of dissipation than were possible with the tubes built for carrying out the measurements, and that still higher emissions can then be obtained from the L cathode. Anyhow the foregoing shows that, thanks to the L cathode, the emission attainable need no longer constitute an obstacle for the development of new valves.

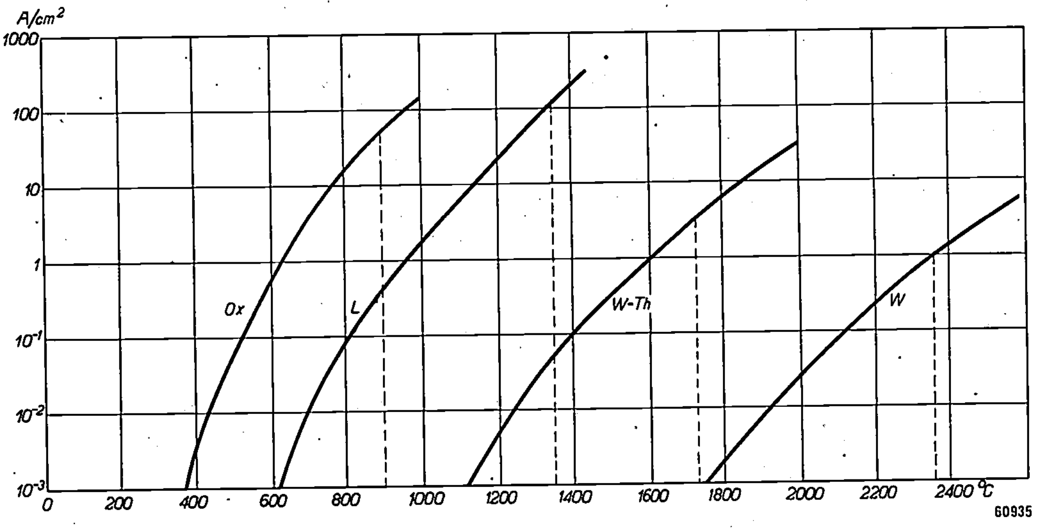


Fig. 3. Saturation emission J_s (in amperes/cm²) as function of the temperature (in °C) for the L cathode and three other types of cathodes. (The curves for these three types and the data for the corresponding curves in figs 4 and 5 and for table II have been taken from: G. Hermann and S. Wagener, Die Oxydkathode, publ. J. A. Barth, Leipzig 1944, Vol. II, pp. 79-80). Along the abscissa, as also in figs 4 and 5, the true temperature has been plotted (not the "black" temperature, as is often taken). For each cathode a vertical dotted line has been drawn to indicate approximately the maximum temperature at which the cathode can be used if it is desired to have a life of at least some hundreds of hours. In the case of the curve for the oxide-coated cathode it is to be borne in mind that these emission values can only be used under pulse conditions; for D. C. emission only about one-hundredth part can be used. The curves end at the temperature where the respective cathode would very soon break down; the end point for the L cathode, however, is determined rather by the difficulties encountered in measuring, since in this case exceptionally high emissions are reached. It is this high maximum emission (together with other properties which are dealt with in the text but cannot be expressed in the emission curve) that typifies the L cathode.

In addition to the maximum emission and the temperature-dependency of the emission there is also the thermal efficiency of the cathode to be considered, by which is to be understood the maximum thermionic current emitted in amperes per watt heating power. The heating power required is in the first instance determined only by the thermal losses due to radiation and conduction at the temperature at which the cathode yields the desired electron emission. Therefore the lower the temperatures of the emission curve in fig. 3, and the more the cathode differs from the black body — that is to say, the less the radiation from the cathode at a given temperature — the smaller will be the heating current required. As far as the thermal losses through conduction (via the supporting rods, etc.) are concerned, these are not usually taken into account when comparing different types of cathodes, because these losses depend to a high degree upon the construction of the valve. Also the radiation from parts of the cathode not contributing towards the thermionic emission is disregarded.

The so defined theoretical thermal efficiencies, where only the temperature-dependency of the emission and the radiation properties of the cathode surface play a part, have been plotted in *fig. 4* as a function of the saturation emission, again both for the L cathode and for the three other types of cathodes. From this diagram it can be seen that the theoretical efficiency of the L cathode is greater than that of the tungsten and thoriated tungsten cathodes but less than that of a normal oxide-coated cathode. It is not surprising that the latter should be the case, considering that the temperature of the L cathode is higher than that of the oxide-coated cathode, with equal saturation emission, whilst moreover the porous metal surface of the L cathode is a good radiator of heat.

From the foregoing it can already be concluded that for valves for which relatively small current densities are needed, say less than 0.25 A/cm^2, and for which the smallest possible heating power is required (as is the case for normal radio receiving valves), the oxide-coated cathode is to be given preference over the L cathode [3]). In cases where higher emission currents are required, however, it is necessary to consider the thermal efficiency somewhat more closely before drawing any conclusions.

Anticipating the discussion of the emission mechanism it can be stated that there is a relation between the thermionic properties of the various types af cathodes and their work function φ; before an electron can leave the cathode and enter the vacuum it has to overcome at the surface a potential difference φ, the value of which is characteristic for the type of cathode. At room temperature only few electrons possess the necessary energy $e\varphi$ (e = the charge of the electron) to overcome this potential difference. This number rapidly increases as the temperature rises and a measurable quantity of electrons are emitted. Of course the same number of electrons are restored to the cathode from the supply lead. Since the electrons possessing most energy are selected for the emission, whilst this is not the case with the electrons fed to the cathode, the result is that in the emission the cathode loses energy, or in other words it cools down. Where a large number of electrons are emitted this effect is quite noticeable, as may be illustrated by an example: when working with a D.C. emission of 40 A/cm^2 and taking for φ a value of 1.8 V (a normal value for the L cathode) the electrons emitted carry 40 × 1.8 = 72 watts/cm^2 out of the cathode (actually a little more than this, because the electrons leave the cathode at a fairly considerable velocity). The heating power must therefore be increased by a like amount in order to keep the cathode surface at the temperature required for the emission. Without emission, thus for covering the radiation losses only (theoretical efficiency), a current of 20 W/cm^2 would be sufficient for this temperature.

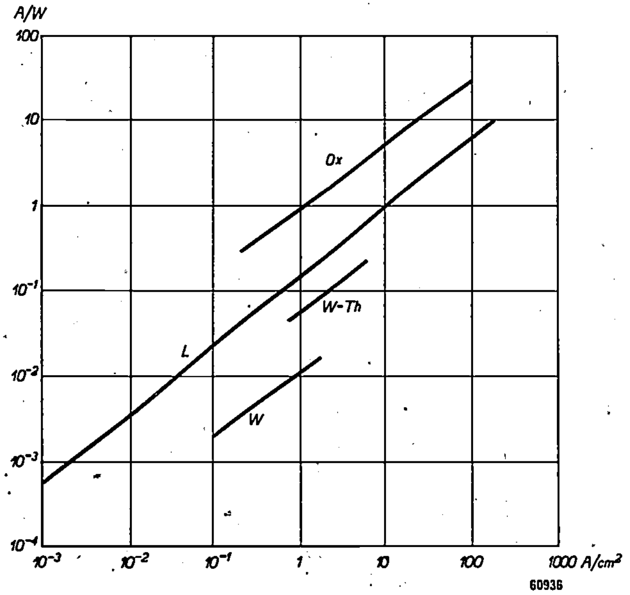


Fig. 4. Theoretical thermal efficiency (in amperes per watt heating power) as function of the saturation emission J_s for the L cathode and three other types of cathodes.

[3]) This still holds when it is taken into account that only a fraction of the saturation emission of the oxide-coated cathode may be used for D.C. loads.

In the case of the oxide-coated cathode the cooling effect is about $1^1/_2$ times smaller, since the work function φ = 1.0 to 1.5 V. At the same time, however, with this cathode another effect arises, which acts in the opposite direction. The emitting oxide coating is a semi-conductor, thus having a fairly high resistance R for the electron current (anode current) that has to traverse it before it can be emitted. Consequently in the oxide coating, owing to the Joule effect, the emitted current i develops a certain amount of heat $i^2 \times R$, which with high currents, such as are drawn from the oxide-coated cathode in the case of pulse emission, assumes very large values [4]). The cooling effect, which increases only in proportion to i, may be entirely overshadowed by the heat due to the Joule effect and in some cases there may be a rise in temperature of 100 °C. (With an emission of about 1 A/cm^2, thus near the limit for D.C. emission of the oxide-coated cathode, these two effects are about equal.)

It would be wrong to regard this Joule effect as another favourable factor for the thermal efficiency of the oxide-coated cathode. Although, as a consequence of the Joule effect, for the temperature desired a still smaller heating current can (or rather, must) be applied to the oxide-coated cathode, this only means that part of the power for the heating is drawn from the source of supply for the anode current instead of from the source of the cathode current, whilst the total power remains unchanged. Thus no advantage whatever is derived from the Joule effect and in fact it has even a disadvantage: the part of the energy contributed by the anode-current source depends upon the resistance of the oxide coating and may therefore vary slightly for different valves, whilst it also varies with the load of the valve, so that it is less easy to control the cathode temperature than when the cathode is heated exclusively by the filament current.

Looked at in this light, the absence of this Joule effect is to be regarded as an important advantage of the L cathode. It is truc that, as a result of the cooling effect, which is not compensated in this case, the thermal efficiency reached in practice is even lower than the theoretical efficiency, but in those cases where a large emission is required, and thus the cooling effect bccomes noticcable, the efficiency of the cathode will not usually be a factor of any great importance. Difficulties with dissipation do not arise from the additional heating current required, because this only leads to a very small percentual increase of the anode dissipation.

Other properties of the L cathode

In outward appearance the most striking feature of the L cathode is its smooth, metallic, emitting surface. This gives the cathode great mechanical strength, it is not liable to damage while being mounted, and there is no risk of particles peeling off under the influence of electrostatic forces of attraction. Thus a great drawback of the oxide-coated cathode, with its mechanically weak layer of carbonate or oxide, is avoided. Furthermore, the emitting surface of the L cathode, which is turned on a lathe or moulded, can easily be made perfectly flat and exactly to given dimensions, within tolerances of only a few microns. For valves where it is of importance that the cathode should be at a very short distance (of the order of some tens of microns) from the other electrodes in order to limit the transit times of the electrons, as is the case with disc-seal valves, this offers very good possibilities for their construction, better even than with the finest-grained oxide-coated cathodes, because these have to be handled with the utmost care and are therefore difficult to mount with the necessary precision.

It is also to be pointed out that for the kind of ultra-short-wave valves just mentioned the high emission that can be obtained continuously with the L cathode yields a two-fold advantage. Not only can the cathode surface be made very small, thus giving extremely small valve capacitances — which is of importance for a good quality of the oscillatory circuits at very high frequencies — but a great density of current requires a high control voltage, which further shortens the transit times of the electrons and thus reduces the "transit-time damping". Also the fact that the L cathode has no cross resistance worth mentioning — in contrast to the emitting layer of the oxide-coated cathode, see above [5]) — helps to reduce the damping of the circuits, at least at wavelengths of the order of 1 m (at shorter wavelengths the effect of the resistance of the oxide-coated cathode again decreases owing to the capacitive component in the impedance of the oxide coating shunting the resistive component).

[4]) This generation of heat, as also, inter alia, phenomena of electrolysis of the coating due to the passage of the current, account for the difference between the permissible D.C. and pulse emission of the oxide-coated cathode; cf table I.

[5]) Cf also R. Loosjes and H. J. Vink, Conduction processes in the oxide-coated cathode, Philips Techn. Rev. 11, 271-278, 1949/1950 (No. 9).

As was to be expected, the L cathode has proved to be fully capable of withstanding the heavy electron bombardment taking place in magnetrons and also in reflex valves [6]), in which respect it compares very favourably with the oxide-coated cathode. In two similar reflex valves working on a wavelength of 10 cm and with a continuous cathode load of up to about 2 A/cm^2 an oxide-coated cathode lasted no longer than about 100 hours, whereas an L cathode continued to work for more than 1000 hours.

When the electrons returning to the cathode in magnetrons and reflex valves have a sufficiently high velocity, then the secondary emission from the L cathode is such that the ratio of the number of emerging secondary electrons to the number of incident primary electrons is appreciably greater than 1, as is the case with most metals. In so far as it may be of importance to effect a saving in heating power or to increase the thermionic emission beyond what is thermionically possible with the L cathode, advantage can be taken of this property (which in the case of the oxide-coated cathode for magnetrons is practically a conditio sine qua non).

As regards the behaviour of the L cathode in a magnetron, provisional tests show that good results are to be expected also with this type of valve. Other investigators in the Philips laboratories will deal with this application of the L cathode on a later occasion.

The drawback of the oxide-coated cathode that small quantities of barium and strontium evaporate from it applies likewise to the L cathode, as will be shown at the end of this article. The L cathode also tends to be poisoned to a certain extent by oxygen or oxygen compounds (e.g. carbon monoxide) present in the valve, as is also the case with the oxide-coated cathode, but when the normal vacuum has been restored then the thermionic emission of the L cathode recovers more quickly and more easily than that of the oxide-coated cathode. This gives the L cathode the advantage in valves that are difficult to degas. Such applies all the more in cases where the cathode is exposed to a bombardment by high-velocity gas ions and to sparking. After a short bombardment the oxide-coated cathode recovers, but when the bombardment lasts for any length of time the whole of the emitting coating may be lost through atomization, whilst sparking makes holes in the coating. The L cathode, on the other hand, can withstand both sparking and a lengthy bombardment by gas ions. Although the thermionic emission is greatly reduced during the bombardment, as soon as the cause of the bombardment has been removed the emission recovers and it is found that no permanent damage has been suffered. Clear proof of this has been obtained for instance with certain types of velocity-modulated valves, in which, notwithstanding the fact that the load was not very high (0.8 A/cm^2), oxide-coated cathodes broke down after about 25 hours, whereas an L cathode lasts some thousands of hours.

The high degree of reliability resulting from these properties is of particular importance for the manufacture of very expensive valves and tubes (e.g. iconoscopes and super-iconoscopes). Even if these can quite well be made with an oxide-coated cathode, an L cathode is likely to be preferred in order to minimize the risk of failure.

As the last and highly important feature, the life of the cathode under normal working conditions may be discussed. Whereas the life of the oxide-coated cathode depends not only upon the temperature at which it is used but also upon the emission current taken from it, the latter factor does not play a part in the case of the L cathode. As might be expected, the higher the working temperature the shorter is the life of this cathode, but the temperature range within which the L cathode can be worked with a useful emission and a reasonable life is very extensive, much more so than in the case of the oxide-coated cathode, viz. 900 to 1350 °C as compared with 700 to 900 °C. To give some further data: at 1000 to 1100 °C the life of the L cathode is some thousands of hours, at 1250 °C some hundreds of hours, and even at 1350 °C it is still some tens of hours. For a proper appreciation of these figures it should be borne in mind that at the temperatures mentioned the saturation emission amounts respectively to 3, 100 and 250 A/cm^2. Such performances are not equalled by any other type of cathode.

The mechanism of the emission

From the fact that both in the case of the L cathode and in that of the oxide-coated cathode a mixture of barium and strontium carbonates is employed, from which the carbon dioxide is driven out, it might be thought that the L cathode could be regarded as being merely a variation of the oxide-coated cathode. A closer investigation into the mechanism of the emission, which will

[6]) See, e.g., F. Coeterier, The multireflection tube, a new oscillator for very short waves, Philips Techn. Rev. 8, 257-266, 1946.

now be dealt with, shows however that this is not the case but that the L cathode represents a new type. First we shall briefly review the laws which any material emitting electrons obeys.

The temperature-dependency of thermionic emission (i.e. the function represented in fig. 3) can be generally described by Richardson's formula:

$$J_s = AT^2 \exp\left(- e\varphi_0/kT\right),$$

where J_s represents the saturation emission in A/cm^2, T the absolute temperature, φ_0 the work function in volts at absolute zero, e the charge of the electron = 1.60×10^{-19} coulombs, k Boltzmann's constant = 1.38×10^{-23} joule/degree, and A a constant of the emitting surface expressed in A/cm^2/degree^2.

Since in the temperature range occurring in practice in all cases $e\varphi_0 \gg kT$, it is clear that the variation of J_s as a function of T is determined almost entirely by the exponential function. This implies (1) that J_s varies greatly with the temperature, the more so the larger the value of φ_0, and (2) that for different materials, but with equal temperature, the value of J_s will be determined by the magnitude of φ_0, such that the smaller the value of φ_0 the greater will be the value of J_s (the values of A for various materials do not differ so much as to have any appreciable influence affecting this conclusion).

From this it appears that the value of φ_0 for a type of cathode can be used for indicating roughly the behaviour of that type by a single figure. It is then obvious that also differences in the mechanism of the emission will find expression in differences in the value of φ_0.

In the case of the tungsten cathode the emission takes place from the surface of the pure metal, the φ_0 value of which reaches the very high figure of 4.5 V. The thoriated tungsten cathode has a monoatomic layer of thorium on the surface of the tungsten, and such monoatomic layers greatly reduce the high work function of tungsten: the value of φ_0 for the thoriated tungsten cathode is about 2.7 V. In the case of the oxide-coated cathode we have to do with an emitting semi-conductor, and not with an emitting metal. The emission mechanism in this case is rather complicated and we cannot enter into it here, but it results in an exceptionally low value of the work function, viz. 1.0 to 1.5 V.

Once the emission J_s of a cathode has been measured at different temperatures it is easy to determine φ_0 by plotting log (J_s/T^2) as a function of $1/T$. According to Richardson's formula the measuring points should lie on a straight line — which as a rule appears to be fairly well the case —, and the slope of the line indicates the value of φ_0, apart from a factor $k/(e \log \mathrm{e}) = 1.98 \times 10^{-4}$V/degree. *Fig. 5* shows the Richardson lines for four cathodes, the older types already mentioned and an L cathode, whilst *table II* gives the values of φ_0 (and of A) derived from such lines for a

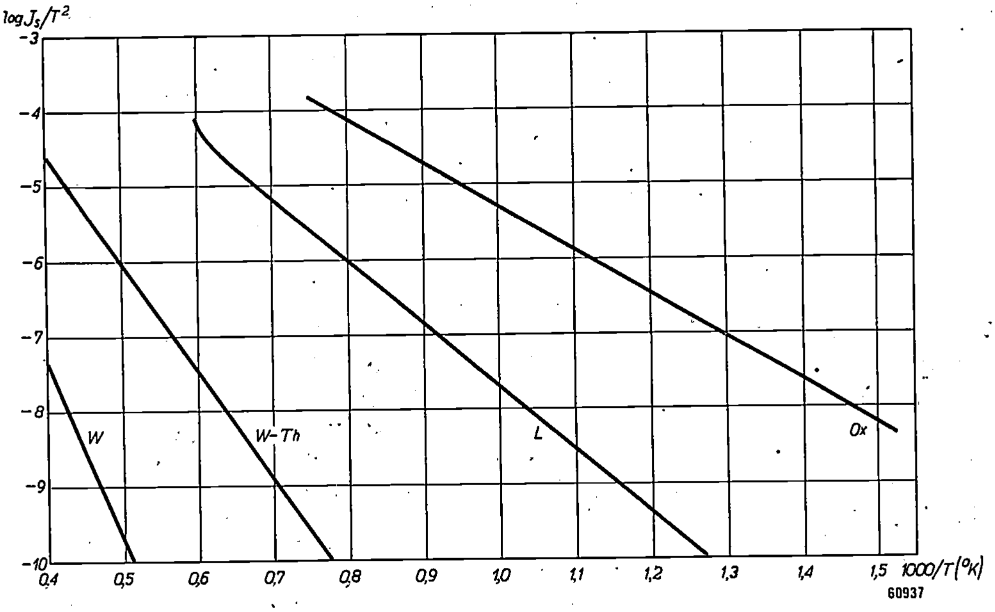


Fig. 5. Richardson lines for the L cathode and three other types of cathodes.

Table II.

Type of cathode	φ_0 in volts	A in A/cm²/degree²
Tungsten	4.44-4.63	22 -210
Thoriated tungsten	2.6 -2.9	3 - 15
Oxide-coated cathode	1.0 -1.5	0.01- 5
L cathode	1.6 -2.0	1 - 15

series of various specimens of the four types of cathodes.

Although the extreme values for φ_0 of the L cathode and the oxide-coated cathode closely approximate each other, the difference is still too great for these types to be regarded as being identical. To put it in other words, it is unlikely that the emission of the L cathode takes place from a layer of barium-strontium oxide. On the other hand, considering the relatively low value of φ_0, there cannot be any question of emission from the porous tungsten surface itself.

What, then, is the mechanism of the emission of the L cathode?

During the very first heating in vacuo the barium-strontium carbonate in the closed chamber is dissociated according to:

$$\mathrm{Ba(Sr)CO_3} \rightarrow \mathrm{Ba(Sr)O} + \mathrm{CO_2}\nearrow .$$

The carbon dioxide is pumped off. During the next heating the barium oxide is partly reduced to barium metal:

$$\mathrm{BaO} + \mathrm{Me} \leftrightarrows \mathrm{Ba}\nearrow + \mathrm{MeO}.$$

By Me is meant one of the surrounding metals (for the sake of simplicity taken to be bivalent). At the temperatures used (900 to 1350 °C) the vapour pressure of the barium metal is very high. Consequently this metal will escape from the reaction equilibrium and the reaction will move gradually to the right, notwithstanding the fact that the heat of formation of BaO, which is great compared with other metals, tends to force the reaction to the left. Thus, upon the cathode being heated, barium vapour will be formed in the closed chamber under a certain, very small, pressure that is determined by the speed of the reaction just mentioned. The same applies for the strontium [7]). Further, there will be also a noticeable amount of BaO vapour, since at these temperatures also BaO has a rather considerable vapour pressure; the vapour pressure of SrO is negligible. See *fig. 6.*

The mixture of Ba(Sr) and BaO vapours passes out through the pores of the tungsten, and in those pores it will form a monoatomic layer on the tungsten (a multiatomic layer would evaporate again owing to the high vapour pressure, whilst the first atomic layer is retained by the forces of adsorption). Now the barium in this form of an adsorbed layer moves about over the surface [8]), so that after a time the whole of the surface of the tungsten, inside and outside, will be covered with a monoatomic layer of barium, mixed with some oxygen. This layer, just like the layer of thorium in the case of the thoriated tungsten cathode, results in a considerable reduction of the work function. Thus the high thermionic emission of the L cathode is made understandable [9]).

The representation of the mechanism as developed above is supported by a number of experimental facts. For instance it has been possible to prove by direct means that barium, barium oxide and

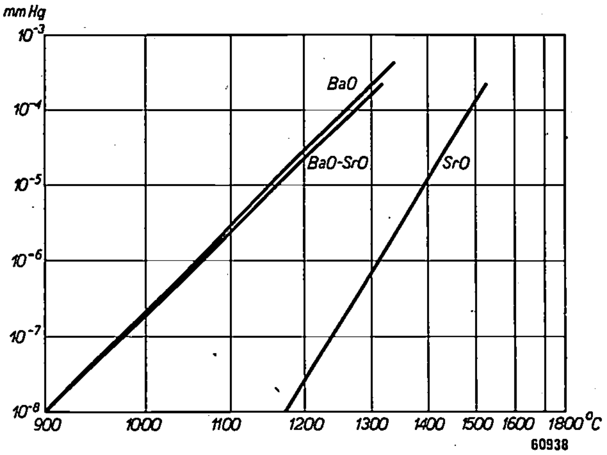


Fig. 6. Vapour pressure of BaO, BaO-SrO and SrO as function of the temperature (taken from A. Claassen and C. F. Veenemans, Z. Physik 80, 342-351, 1933). The mixture of BaO and SrO has a molecular ratio of 42.8 : 57.2. The vapour of this mixture does not contain any measurable quantity of SrO.

[7]) The heating also reduces the strontium oxide to strontium. SrO has an even greater heat of formation than BaO, but the vapour pressure of Sr is also greater than that of Ba, so that the reaction with SrO is still noticeably directed to the right. — Exactly what part the strontium plays in the functioning of the cathode cannot yet be sufficiently explained. The strontium cannot be dispensed with since it has been proved empirically that the life of the cathode is then considerably shortened.

[8]) This, as well as the formation of a layer as mentioned above, has been demonstrated by, among others, J. A. Becker in Trans. Far. Soc. 28, 148-158, 1932, and J. A. Becker and G. E. Moore in Phil. Mag. 29, 129-139, 1940.

[9]) The adsorbed oxygen atoms, which in themselves obstruct the escape of electrons from the metal, have also a favourable effect in that, as negative centres, they promote a stronger binding of the adsorbed barium atoms on the surface. This favourable action predominates so long as there is not too much oxygen present. Cf: J. H. de Boer, Elektronenemission und Adsorptionserscheinungen, J. A. Barth, Leipzig 1937, page 113.

strontium evaporate from the L cathode. In the pores of the activated cathode a quantity of barium is found which corresponds fairly well to the quantity needed for the formation of a continuous monoatomic layer. An experiment carried out with a cathode without the tablet of barium-strontium carbonate showed that the cathode could be activated by evaporating upon it a monoatomic layer of barium from the outside (most probably also traces of oxygen then reach the cathode), and it was then found to have a work function $\varphi_0 = 1.7$ volts, the same as that of a normal L cathode! When, however, a layer of barium oxide about 1 micron thick is applied to the metal surface then, at the same temperature, the emission is greater and $\varphi_0 = 1.4$ volts, which is a noticeable difference compared with the normal L cathode.

It is therefore quite evident that the L cathode is not to be regarded as a kind of oxide-coated cathode with a thin coating of barium oxide on a support of tungsten, but rather as showing a strong resemblance to the thoriated tungsten cathode, with the part played by the thorium taken over in the L cathode by the barium. The special properties described in the foregoing as characterizing the L cathode are due to the fact that the barium in a monoatomic layer reduces the work function of tungsten to a much greater extent than thorium does, and to the special construction which allows of the monoatomic layer being continually renewed automatically.

Summary. A new thermionic cathode developed in the Philips Laboratories at Eindhoven (Holland), designated as the L cathode, has a mixture of barium and strontium oxides contained behind a wall of porous tungsten. At the working temperature of the cathode (900-1350 °C) the barium and strontium oxides are gradually reduced; owing to the fairly high vapour pressures of Ba, Sr and BaO these substances escape through the pores of the wall and form there on the tungsten surface a monoatomic layer of barium and strontium with some oxygen in between. This layer reduces the work function, which for pure tungsten is 4.5 V, to 1.6-2.0 V. This value is not quite so low as that of the oxide-coated cathode (1.0 to 1.5 V), so that the required temperature is higher and the thermal efficiency lower than in the case of the oxide-coated cathode, but owing to its construction the L cathode is much better able to withstand heavy loads (even at the higher temperature). The maximum useful emission, with a reasonable life, amounts to some hundreds of amperes per cm^2 (measured under pulse conditions; the maximum useful emission in the case of a D.C. load is in principle the same, but difficulties then arise in the measurement). Further the L cathode is proof against the electrostatic forces of attraction at high tensions, it easily recovers after a possible poisoning by oxygen or other gases and after a bombardment by high-velocity gas ions, and it can quite well withstand a bombardment by electrons, whereby a considerable emission of secondary electrons takes place. All these properties make the L cathode highly suitable for all types of modern valves and tubes for generating ultra short waves, and in general for those where the cathode is required to have great mechanical strength and a high degree of reliability in addition to a high emission.

THE MANUFACTURE OF QUARTZ OSCILLATOR-PLATES

II. CONTROL OF THE CUTTING ANGLES BY X-RAY DIFFRACTION

by W. PARRISH *).

549.514.51:621.369.611.21:
537.531:535.4

During the war X-ray diffraction, which had been used only in research in some scientific and technical laboratories, also proved its usefulness as an aid in mass production manufacturing. It was introduced into the quartz oscillator-plate industry for the measurement of the crystallographic angles of cuts from quartz crystals. The diffraction apparatus developed for this purpose makes it possible for unskilled people to measure deviations with respect to specified crystallographic angles with an accuracy of a few minutes of arc in a time of 10 to 15 seconds.

The tolerances of oscillator plates

As an introduction to this second article on the manufacture of oscillator plates it is desirable to consider more closely the question of tolerances in the orientation and dimensions of these plates.

It has been stated [1]) that the deviations in the crystallographic angles with respect to the prescribed values for a low temperature coefficient cut, a BT cut for example, may not amount to more than 10 minutes, while the thickness of a plate is restricted to a tolerance of, for instance, 10^{-5} mm. Furthermore the length of the sides is often also prescribed, with a tolerance of 0.03 mm or less.

The crystallographic angles involved are illustrated in *fig. 1,* for the case of the square AT and BT plates. This figure shows how these two cuts are orientated in a quartz crystal. An edge X' of the plate must be parallel to an X-axis (electrical axis) of the quartz, thus the angle $XX' = 0°$. The projection Z' of the Z-axis (optic axis) on the face of the plate must make an angle ZZ' with the Z-axis amounting to 35° 15′ in the case of the AT cut and of —49° 20′ in that of the BT cut. Generally the tolerance for XX' is somewhat larger than for ZZ'. During the war it was in most cases 20′ for XX' and 10′ for ZZ'. In the case of the higher precision AT plates made to-day, the ZZ' angle is often held to smaller tolerances, perhaps 3 or 4 minutes.

These tolerances in the crystallographic angles are dependent upon the requirements regarding the variation of the resonance frequency as a function of the temperature. The turning point of this curve may be displaced to an undesirable temperature region or it may disappear due to a slightly incorrect orientation of the cut. This is illustrated by *fig. 2.* The tolerance in the thickness depends upon the accuracy with which the desired resonance frequency must be realized. In the case of AT and BT plates where the desired mode of vibration involves a thickness shear, the frequency is inversely proportional to the thickness. The tolerance in the length of the sides is finally prescribed by the requirement that a coupling of the desired vibration with undesired modes of vibration of the plate must be avoided as much as possible. In the case of AT and BT plates, the high

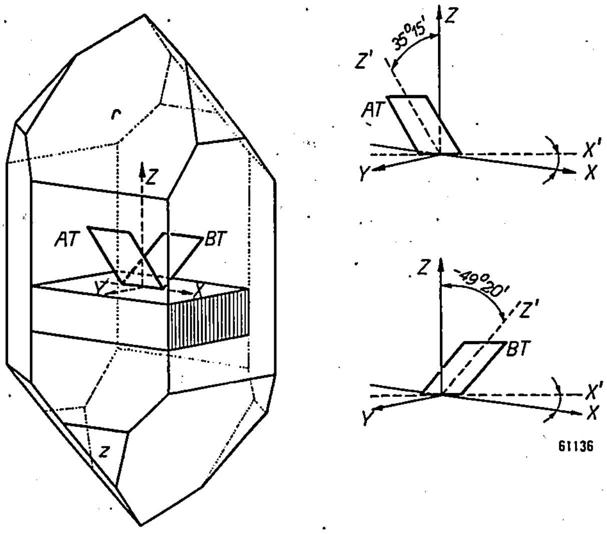
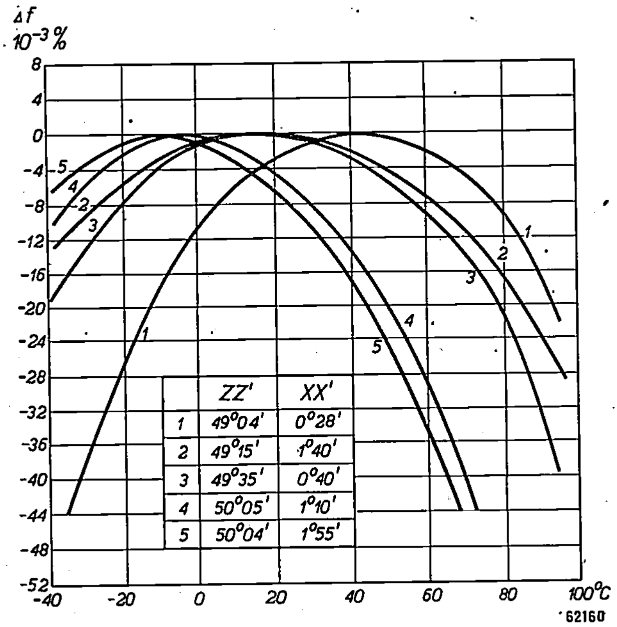


Fig. 1. Orientation of AT and BT cuts in an ideal quartz crystal (natural crystals do not often have all the regularly developed faces shown here). Both cuts, as well as most other low-temperature coefficient cuts, are perpendicular to a Y-Z plane, at an angle of 35° 15′ and —49° 20′, respectively, to the Z-axis. For the BT cut the desired angle varies between —49° 0′ and —49° 30′, dependent on the temperature range in which the oscillator plate is to be used; in the following we shall base our discussion on the first-mentioned intermediate value.

*) Philips Laboratories, Inc., Irvington-on-Hudson, N.Y., U.S.A.

[1]) W. Parrish, The manufacture of quartz oscillator-plates, I. How the required cuts are obtained, Philips Techn. Rev. 11, 1949 (No. 11). This article is referred to in the following as I.

harmonics of a fundamental low frequency flexural mode about a Y-axis, which runs diagonally through the plate, often cause the most trouble.



	ZZ'	XX'
1	49°04'	0°28'
2	49°15'	1°40'
3	49°35'	0°40'
4	50°05'	1°10'
5	50°04'	1°55'

Fig. 2. Variation of resonance frequency with temperature, for a number of BT cuts with slightly differring crystallographic angles. The ordinate scale gives the deviation of the resonant frequency from the maximum value, in 10^{-3} %. The turning points (maxima) of the curves lie near 42, 18, 16, —2 and —10 °C respectively. Besides the position of the turning point, the steepness of the curve must be taken into account, in order to get a broad range of working temperatures.

The check on whether the plate has been brought within the tolerances of thickness and length is not carried out by measuring directly these dimensions but indirectly (or rather even more directly) by investigation of the behaviour of the plate in the oscillator circuit of a valve oscillator. In this way it is easy to determine the mechanical resonance frequency of the plate. The "activity" of the plate, i.e., the amplitude at which the plate vibrates, is also measured and serves as a criterion of an adequate limitation of undesired couplings: excessive couplings cause activity "dips" at some frequencies. The lapping (and etching) of the plate is continued until it satisfies the specifications as to frequency and activity. In the last article of this series we shall have an opportunity of going into this more deeply.

The proportionality constant in the relation between resonance frequency and thickness (i.e. the product of frequency times thickness) depends closely upon the orientation of the cut in the crystal. The dimensions required for the avoidance of couplings are also very much influenced by the orientation. In the case of the "indirect" empirical process of finishing the oscillator plates little attention was given to these influences. Nevertheless, the last mentioned influence is of practical importance. Several manufacturers endeavor to minimize the coupling with undesired modes of vibration by giving the plates well defined and previously determined optimal dimensions (predimensioning). These dimensions vary with the cutting angles and hence it is necessary to maintain small tolerances for these angles to make the scheme practical.

During the war predimensioning was not widely practiced in production. Activity dips over the temperature range were then one of the principal causes of rejection of plates.

The checking and correction of the cutting angles was formerly accomplished by a similar "indirect" procedure: a plate was cut from the crystal and lapped to the desired resonance frequency. The frequency was measured as a function of the temperature and from the curve obtained it was deduced as best one could by what amount the cutting angles were incorrect. It was not possible from this method to separate the corrections for ZZ' and XX', but, moreover, it was a quite lengthy and elaborate test. To be sure, checking the orientation is not required for every blank separately; it is to be performed on a test piece, parallel to which a whole series of wafers or blanks can be cut out of a crystal block. The result of the former test method would however be that after cutting the test piece the quartz saw would be idle for quite a long time during the lapping and performance of the test.

An economical mass production of oscillator plates with the present extremely strict specifications only became possible upon the introduction of the X-ray diffraction method, with which the crystallographic angles such as XX' of ZZ' can be measured directly, rapidly and very accurately. With the apparatus designed for this purpose by the North American Philips Co., which was used on a large scale in the American quartz industry during the war, the required corrections for a test cut could be determined by relatively unskilled help with an accuracy of a few minutes of arc and the whole measurement required only 10 to 15 seconds time [2]).

Principle of the angle measurement by X-ray diffraction

The phenomenon of X-ray diffraction in a crystal can be described as a reflection of an X-ray beam at the lattice planes, an appreciable reflection at a set

[2]) W. Parrish and S. G. Gordon, Precise angular control of quartz-cutting by X-rays, Amer. Mineralogist **30**, 326-346, 1945. See also W. L. Bond and E. J. Armstrong, Use of X-rays for determining the orientation of quartz crystals, Bell System Tech. J. **22**, 293-337, 1943; V. Petržilka and J. Beneš, A method for the determination of crystal cuts by applying the reflection of X-rays from a known lattice plane, Phil. Mag. (7) **37**, 399-410, 1946.

of parallel lattice planes only being possible when Bragg's condition,

$$\sin \Theta = \frac{n\lambda}{2d}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

is satisfied (see *fig. 3*). In this relation d is the spacing of the parallel lattice planes, Θ the angle of incidence and emergence between X-ray beam and lattice plane, λ the wavelength of the X-rays, n a whole number (order of the reflection).

Each set of lattice planes in a quartz crystal makes definite angles with the crystallographic axes (X, Y, Z-axes), with the natural faces of the crystal (prism faces m, major and minor rhombohedron faces r and z, respectively, etc.) and therefore also with the various low-temperature coefficient cuts and any auxiliary planes used in the manufacture. Hence, the orientation of a test cut can be checked by measuring the angle between the plane of that cut and a suitably chosen lattice plane. The principle of the arrangement used for this purpose is shown in *fig. 4*.

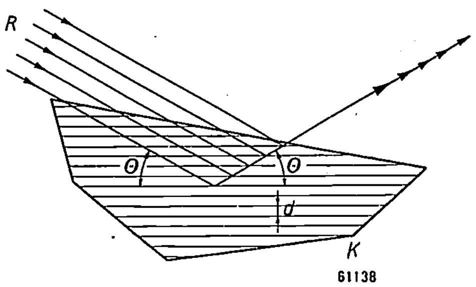


Fig. 3. Diffraction of an X-ray beam R in a crystal K. The beam is reflected at a set of parallel lattice planes provided the spacing d of the planes, the angle Θ between the beam and the lattice plane and the wavelength λ of the X-rays satisfy Bragg's condition, $n\lambda = 2d \sin \Theta$.

The freshly cut flat surface of the piece or wafer cut from the quartz crystal is pressed against the reference surface of the specimen holder which has been ground perfectly plane. The holder has an opening which exposes part of the cut surface. A narrow beam of nearly monochromatic X-rays from the X-ray tube B passes through collimator slits S_1 and S_2 and impinges on the exposed surface of the wafer. The specimen holder can be rotated about the vertical axis P. The angle α between the reference surface and a chosen fiducial or reference position, which for the sake of convenience in this discussion may be chosen in the extension of the primary X-ray beam, can be read off on the goniometer scale. After a lattice plane whose spacing d and orientation are known has been chosen as a plane of reference, the crystal holder is turned until at the point 2Θ in the goniometer scale (Θ being calculated according to (1) from the known values of d and λ) a reflected X-ray beam is observed. The difference between the position α of the crystal holder and the angle Θ is, as may be seen immediately in fig. 4, the angle between the cutting plane and the lattice plane of reference. (This is rigorously correct only if the lattice plane of reference also is perpendicular to the plane of the drawing; we shall return later to this essential condition.)

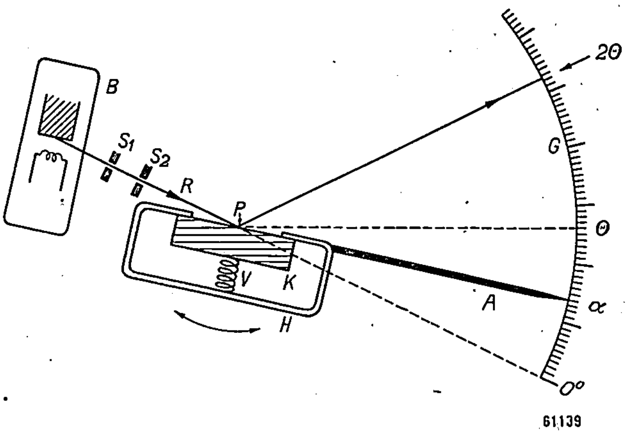


Fig. 4. Principle of the measurement of crystallographic angles by X-ray diffraction. H specimen holder with reference surface, against which the surface of the crystal piece sawed off as a test cut, or of the test wafer K is pressed by the spring V. The holder is fastened to a goniometer arm A and can be rotated around an axis which at point P is perpendicular to the plane of the drawing. B X-ray tube, S_1, S_2 collimator slits, R X-ray beam which falls at P on the crystal surface to be examined, G goniometer scale.

The X-ray diffraction apparatus

In the practical execution of the measurement by this principle the way in which the reflected X-ray beam is observed requires the most attention. In the Philips apparatus a Geiger counter tube is used, in a form which was developed by Friedman[3] and is shown in *fig. 5*. The slit in front of the window of the tube is brought to the desired position of the goniometer scale (2Θ) and the tube remains fixed during the measurement. The X-ray quanta entering the slit are detected by current pulses in the tube and the X-ray intensity is indicated directly by use of circuits measuring the average current developed in the tube. We need not go more deeply into the construction or mechanism of the counter tube since these

[3] H. Friedman, F. K. Kaiser and A. L. Christenson, Applications of Geiger-Muller counters to inspection with X-rays and gamma rays, J. Amer. Soc. Naval Eng. **54**, 177-209, 1942, H. Friedman, Geiger counter tubes, Proc. Inst. Rad. Eng. **37**, 781-808, 1949.

have already been discussed by Friedman[3]) and in this periodical[4]).

This method of indication has been essential for the success of the apparatus. Photographic detection of the diffraction (which would have required a different X-ray method) would have meant a serious obstruction in the smooth flow of the mass production process because of the time lost in developing the film. The counter tube on the other hand records directly the X-radiation falling on the entrance slit. Compared with the fluorescent screen (which was used in only one plant) as means of indication, the counter has the advantages that it is much more sensitive and convenient and that it can be used in a normally lighted room. The ionization chamber which is also used in some apparatus as a direct reading instrument is also less sensitive and less easy to use than the Geiger counter.

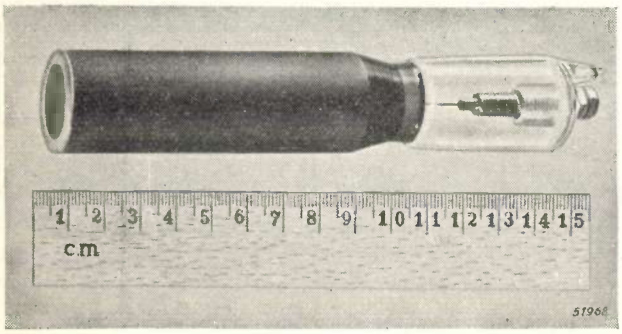


Fig. 5. Geiger counter tube of a recent design. The form originally used in the diffraction apparatus here described was somewhat different but the principle is the same: the X-ray beam to be detected, coming from the left, enters through a thin window of material with low absorption (Lindemann glass, at present mica), and traverses the tube in longitudinal direction running parallel to an anode wire stretched along the axis. The fraction of the radiation absorbed in the argon with which the tube is filled is dependent upon the pressure. Each X-ray quantum absorbed results in a current impulse through the tube between the anode wire and the surrounding cathode cylinder forming the wall of the tube.

In this case the counter tube was used as a so-called proportional counter. The amplification factor of the counter tube in a certain region of voltages (the "plateau") is quite independent of the voltage applied between the wire and cylindrical shell. In a lower voltage region the size of the current impulse caused by one absorbed X-ray quantum is much smaller and increases with the voltage applied and therefore it is possible to regulate the magnitude of the average current obtained through the counter tube simply by changing the supply voltage. This is a convenience in this application because it makes it possible to keep the response to the wide range of reflection intensities from different lattice planes within the range of the meter. In practice it is not necessary to adjust the voltage when making successive measurements on the same type of cut, using always the same lattice plane of reference.

This method of operation is possible here because a single crystal produces very intense reflections, making it unnecessary to resort to the maximum gain of the counter tube, which is attained when working on the "plateau". (Strictly speaking, the name "Geiger counter" is historically correct only under the latter working condition.) Moreover, in this problem it is only a question of detection of a single diffraction peak. For the apparatus described in [4]), with which intensities of a number of (much weaker) diffraction lines of powder specimens must be compared, it is important not only to have maximum gain but also to keep the gain of the counter as constant as possible, and hence the "plateau" is used.

Incidentally, it should be noted that the proportional counter has a much shorter "dead time" than has the Geiger counter proper. Therefore, individual quanta may be detected at much higher counting rates, which means that the response of the proportional counter is linear to higher X-ray intensities than that of the Geiger counter.

The slit mounted in front of the window of the counter tube, through which the X-rays pass, is quite wide, so that the setting of the counter tube on the goniometer scale is not critical. This is important because the diffraction spots given by the quartz crystals are very sharp (compared with the diffraction lines of powders) and thus with a slightly incorrect placing of a narrow slit the operator would run the risk of turning the crystal holder past the reflecting position without observing any reflection at all. The slit may not of course be so wide that in certain positions it would allow the passage of two neighboring diffraction spots simultaneously and thus lead to mistakes as to the reflecting lattice plane. But a slit width of $^1/_4{}^\circ$, measured along the goniometer scale, is quite permissible for the lattice planes normally used in this work, as may be seen from the diffraction spectrum of a quartz sample given in *fig. 6*.

The apparatus is equipped with an X-ray tube of low power, operating with 3 to 4 mA at a peak voltage of 35 kV, and is air cooled. This is feasible because the intensities of the diffraction spots of single crystal plates are considerably higher than those of the diffraction lines of a polycrystalline specimen. Moreover, the counter tube is much more sensitive than the photographic film usually used in dif-

[4]) J. Bleeksma, G. Kloos and H. J. Di Giovanni, An X-ray spectrometer with Geiger counter for measuring powder diffraction patterns, Philips Tech. Rev. **10**, 1-12, 1948 (No. 1). The spectrometer described in this article was developed from the special apparatus designed for quartz oscillator-plate manufacture.

fraction research, so that a relatively low intensity of the primary X-ray beam is sufficient. A copper anode is used giving the characteristic copper radiation, of which the Kα lines are used: Kα_1 = 1.54050 Å, Kα_2 = 1.54434 Å; Kα (weighted) = 1.5418 Å. The CuKβ-radiation, which gives diffraction spots at slightly different angles, is sufficiently weakened by a nickel foil in front of the window of the X-ray tube, so that it is not detected.

The opening in the specimen holder must be large enough that the metal edges are not struck by the primary X-ray beam, otherwise X-radiation scattered at the edges would cause a troublesome background intensity in all positions of the counter tube. Furthermore it is important that the reference surface, against which in a smoothly running manufacturing process a block of extremely hard quartz is laid several hundred times daily, should not gradually be worn off and lose its plane surface or its precise setting with respect to the goniometer arm. The holder is therefore made of the hardest available hardened tool steel or of boron carbide, and the reference surface is checked every few days.

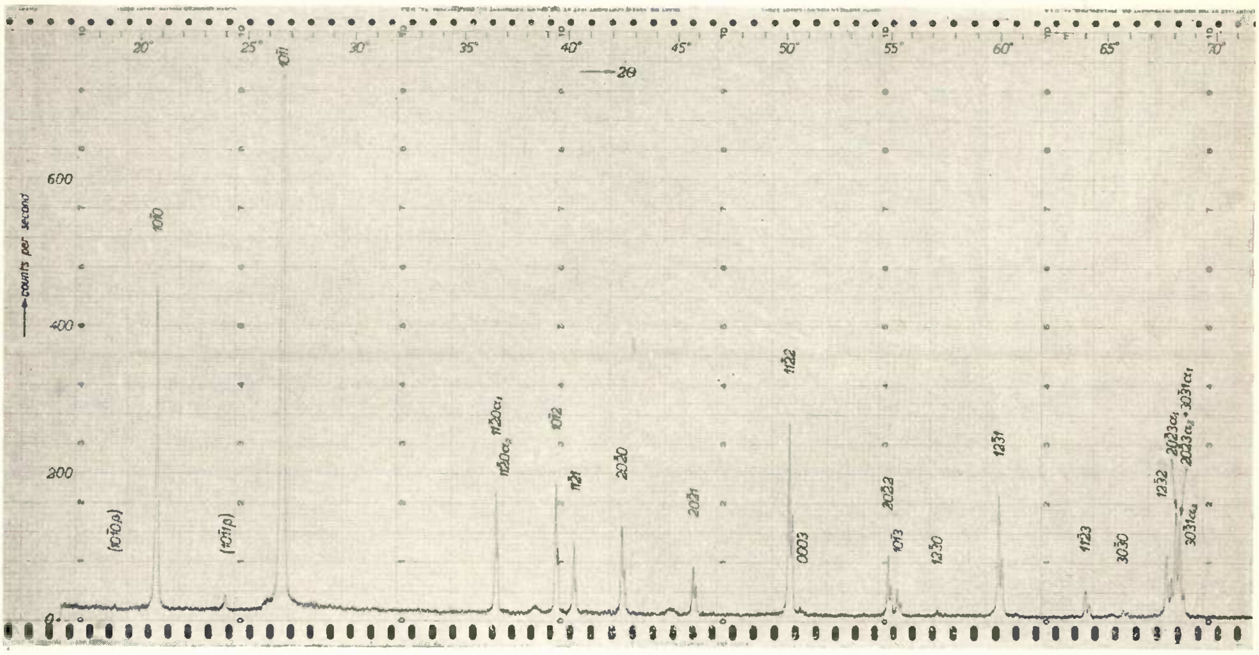


Fig. 6. X-ray diffraction spectrum of the front reflection region of a quartz powder sample made with the Geiger counter X-ray spectrometer [5]) (improved form of the instrument previously described in this periodical [4])). The diffraction spots from single crystal quartz plates are even sharper than the lines of the polycrystalline specimen. This pattern was recorded automatically at a rate of $^1/_4$° of 2Θ per minute, with the copper target X-ray tube operated with full-wave rectification at 40 kVp, 20 mA, 0.015 mm nickel filter. The goniometer radius was 170 mm, the angular aperture of the incident beam 1° and width of the receiving slit 0.08 mm. (The 1011-peak is fas off the chart.)

The goniometer scale is graduated in whole degrees, and a fine-adjustment knob which drives the arm of the specimen holder carries a scale graduated in minutes. In order to avoid repeated subtraction of the angle Θ for a reference lattice plane the goniometer has a second sliding scale whose zero point can be set at the angle Θ or any other desired fiducial point; this is best done empirically, by means of a standard crystal which has a face lapped parallel to the desired lattice plane. The minute scale on the adjustment knob can be set at the desired position at 0′ by temporarily uncoupling it from the driving gear. The specimen holder can also be unlocked and rotated to any position with respect to the arm which rocks it on the goniometer scale so that one can make measurements on any convenient part of the scale.

A photograph of the arrangement, including the X-ray tube, slit system, shutter, specimen holder, counter tube, direct beam shield, goniometer, meter for reading the X-ray intensity, and supply voltage control for counter tube is shown in *fig.* 7. The X-ray tube has two windows. Two complete measuring setups can therefore be used, one on either side of the tube. The complete apparatus is shown in *fig. 8*.

[5]) W. Parrish, X-ray powder diffraction analysis: film and Geiger counter techniques, Science **110**, 368-371, 1949.

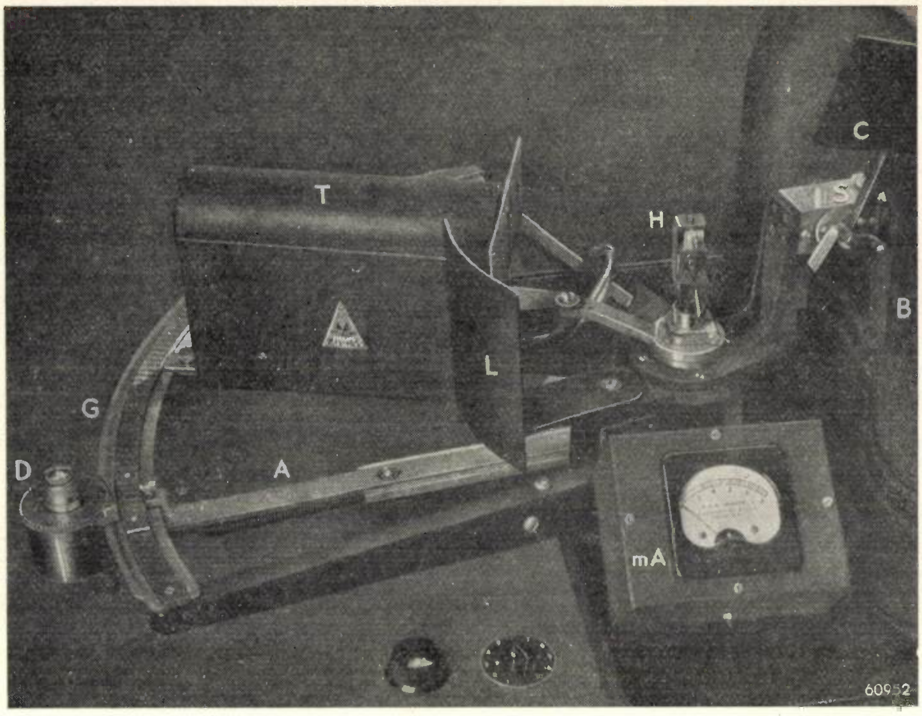


Fig. 7. Measuring table of the X-ray diffraction apparatus. On the right is the X-ray tube housing *B* with slits *S* and a hinged shutter in front of the X-ray tube window. When this shutter is opened a cover *C* descends in front of the crystal holder, so that the operator cannot place his hands in the path of the X-ray beam. On the left is the goniometer scale *G* along which the Geiger counter tube *T* and the goniometer arm *A* bearing the specimen holder *H* can be rocked. The counter tube remains radially directed since it is fastened to an arm which rotates around the same axis (*P* in fig. 4) as the goniometer arm. The crystal holder visible in the middle is intended for measuring blanks; for *Y-Z* test cuts or for test wafers from an *X*-block, etc. different holders are used. The undiffracted portion of the primary beam is absorbed by a lead plate *L* behind the crystal holder. On the goniometer scale there is a short sliding auxiliary scale whose position is set with a standard crystal and which permits direct reading of the desired angles. At the end of the goniometer arm is the knob *D* for fine adjustment with scale in minutes. In front is the milliammeter *mA* which indicates the intensity by reading the average current flowing through the counter tube. (The arrangement in this case is the mirror image of that in fig. 4.)

Recently, a number of refinements have been incorporated in the instrument that markedly increase the precision. The width of the diffraction spot and hence the accuracy of the angle measurement, in case of a nearly perfect crystal such as quartz, depends on the width of the source slit S_1 (fig. 4) and the divergence of the X-ray beam falling on the crystal surface. In the instrument described above the slit-width was 0.39 mm and the divergence 0.9°. In the improved arrangement an X-ray tube with a smaller focal spot is used so that it can assume the function of the source slit. This focal spot is 3 mm wide and the beam is obtained at an angle of $^1/_4$° with the anode surface, the projected width thus being 0.04 mm. The divergence is limited to less than 1′.

These changes make it possible to measure angles accurately to within 0,002° to 0,003°, i.e. about 10 seconds. In order to enable the angles to be read with such a precision, the minute scale has been provided with a vernier.

This great accuracy is not necessary for the production of quartz plates. However, it affords the possibility of using the instrument for studying the degree of perfection of crystalline surfaces, say the effect of lapping a crystal plate, and for similar problems.

Performance of the measurements

In article I it was shown that the desired low-temperature coefficient cut in a quartz block was not obtained in a single operation but in several

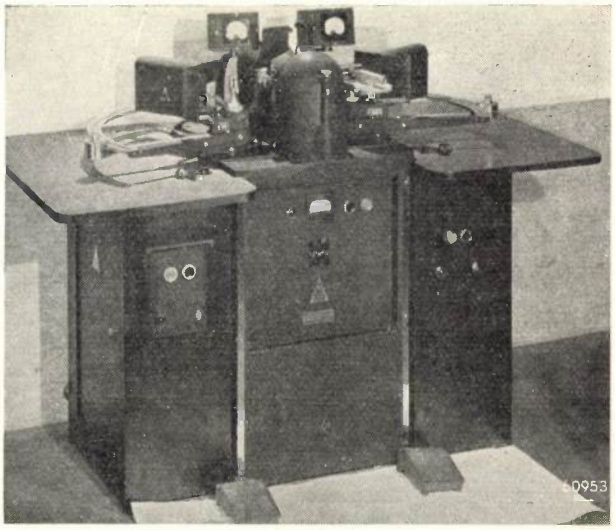


Fig. 8. The complete X-ray diffraction apparatus of the North American Philips Co. for quartz-plate manufacture. The instrument is provided with two independent measuring tables which can be seen here to the left and right of the X-ray tube. Under each measuring table is the supply apparatus for the counter tube, in the middle cabinet the supply and controls for the X-ray tube.

steps. We shall briefly review the way in which the angle measurement with X-ray diffraction enters into the procedure, using the X-block method as an example.

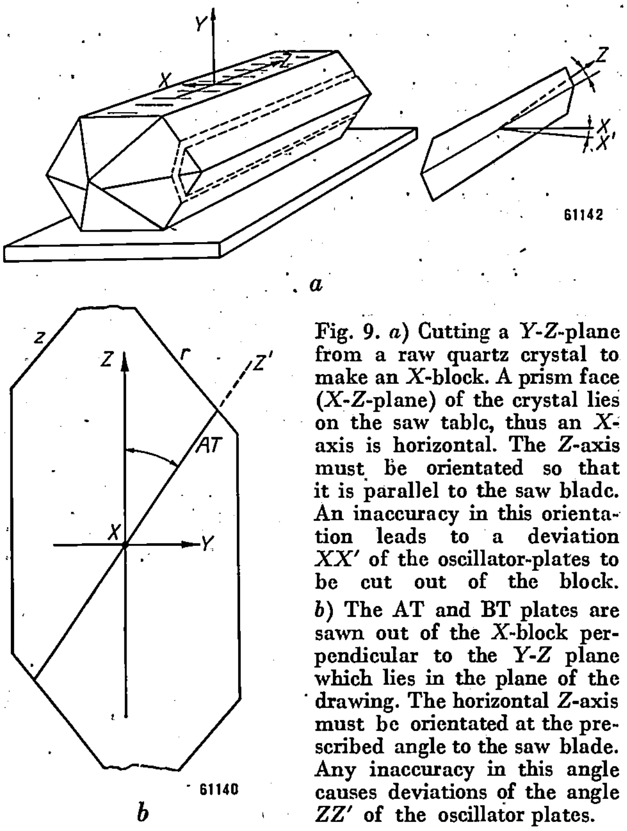


Fig. 9. *a*) Cutting a Y-Z-plane from a raw quartz crystal to make an X-block. A prism face (X-Z-plane) of the crystal lies on the saw table, thus an X-axis is horizontal. The Z-axis must be orientated so that it is parallel to the saw blade. An inaccuracy in this orientation leads to a deviation XX' of the oscillator-plates to be cut out of the block. *b*) The AT and BT plates are sawn out of the X-block perpendicular to the Y-Z plane which lies in the plane of the drawing. The horizontal Z-axis must be orientated at the prescribed angle to the saw blade. Any inaccuracy in this angle causes deviations of the angle ZZ' of the oscillator plates.

In the X-block method a plane which should be perpendicular to an X-axis (YZ-plane) is first cut off the quartz crystal. One of the prism faces of the crystal (XZ-plane) is laid on the horizontal saw table (so that an X-axis is horizontal) and a segment is sawn off as nearly as possible parallel to the Z-axis, see *fig. 9a*. A deviation of parallelism to Z entails a deviation of the normal of the cut surface from the X-axis in the horizontal plane and is found later as angle XX' in all the blanks obtained from this X-block. It is thus necessary even at this stage to measure the cutting direction accurately and to correct according to the results of the measurement.

The cut YZ-plane of the quartz block is then laid on the saw table (X-axis vertical), which is rotated through an angle read from the degree scale on the table until the saw blade makes the desired angle with the horizontal Z-axis of the quartz block; see fig. 9*b*. The inaccuracy of this angle setting, when the block is cut into parallel wafers, would occur in all the blanks as an error in the angle ZZ'. Therefore in this step a test wafer must also be cut, the angle in question measured and the saw should be corrected if necessary.

Checking a YZ-plane

The YZ-plane which is first cut in the X-block method is itself a lattice plane of the crystal, namely the $11\bar{2}0$-plane; see *fig. 10*. The measurement is in this case relatively simple. The counter tube is set at the Bragg angle on the goniometer scale $2\Theta = 36° 34'$ for this reference plane. When reflection occurs (more precisely: when maximum reflection occurs), the reflecting $11\bar{2}0$-plane will be oriented so as to intersect the goniometer scale at the point $\Theta = 18° 17'$, whatever may be the orientation of the actual cut we have made. If the surface was cut exactly parallel to a $11\bar{2}0$-plane the goniometer arm which bears the specimen holder will be found at the angle $\alpha = \Theta$ and it will be independent of any rotation of the cut crystal segment in the holder (with the cut surface always against the vertical reference surface). If the cut plane deviates by a small angle ε from the $11\bar{2}0$-plane, an angle α is found which is not in general equal to Θ (deviation $\Delta\alpha$) and which more-

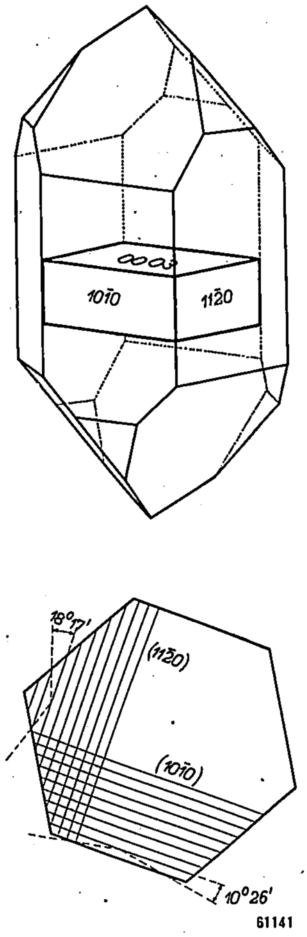


Fig. 10. Position of the $10\bar{1}0$, $11\bar{2}0$, and 0003-planes in a quartz crystal; these lattice planes are identical, respectively, with an X-Z plane (prism face), Y-Z plane and X-Y-plane. In the cross section dotted lines show the direction of incidence and emergence of X-rays upon reflection for each set of lattice planes.

over varies upon turning the crystal in the holder. In turning the crystal the normal of the reflecting $11\bar{2}0$-plane describes a cone of half apex angle ε about the horizontally directed normal to the cutting plane (reference surface). For the highest and lowest positions of the normal, $\Delta\alpha$ equals 0 (unless ε is so large that in these positions no reflected intensity at all is observed in the goniometer plane and hence measurement is impossible); for the extreme right- and left-hand horizontal positions of the normal (thus where the $11\bar{2}0$-plane of reference is vertical), $\Delta\alpha = +\varepsilon$ and $-\varepsilon$ respectively (*fig. 11*).

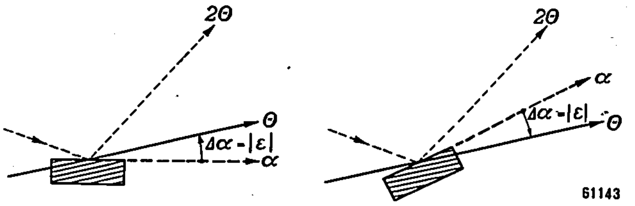


Fig. 11. X-ray reflection is observed when the lattice plane of reference is in the position Θ. This corresponds to a position α of the cut surface of the crystal segment (reference surface of the crystal holder), if the cut surface does not coincide with the lattice plane. The deviation $\Delta\alpha$, by which α differs from Θ, varies when the crystal is rotated in the holder; the maximum value of $\Delta\alpha$, which is found in two opposite positions of the crystal, equals the correction ε of the cutting angle required under the saw to make the cut coincide with the lattice plane.

From the above it will be clear that the angle $\Delta\alpha$, which is read off the scale, is equal to the required horizontal correction angle under the saw only when the test cut is placed in the specimen holder in the same orientation it had on the saw table. This is most easily accomplished by drawing a vertical arrow pointed up on the outer surface of the test cut before removing it from the saw table. The test cut is placed in the specimen holder with the fresh cut surface toward the X-ray beam and the arrow again vertical and pointed up. It is then easy also to correlate the direction of the correction: if the goniometer arm must be rotated clockwise from the position Θ in order to bring the $11\bar{2}0$ reference plane of the test cut into a reflecting position, the crystal on the saw table must also be rotated clockwise in order to make the $11\bar{2}0$-plane parallel to the saw block.

It is easy to verify that the correlation thus expressed is valid for the right-hand as well as for the left-hand goniometer in fig. 8, the two scales of which are mirror images of each other.

If in cutting a crystal for making an X-block the saw blade was not exactly perpendicular to the saw table, the normal of the cut surface will deviate from the X-axis in the vertical plane. This vertical deviation can be measured in exactly the same way as above, with the only difference that the arrow painted vertically on the test cut must be placed in a horizontal position. Correlation of the direction of the measured deviation and the required correction is simple also in this case, provided care is taken to note if the test cut was made from the right or left side of the crystal.

Such simple correlation rules, which apply to all types of test cuts, are of great importance for mass production, where the saws and X-ray machines are operated by relatively unskilled persons.

Checking AT and BT cuts

In the previous section the desired cutting plane coincided with a lattice plane, but in checking the low temperature coefficient cuts proper this is not the case. In checking the ZZ' angle of an AT or BT test cut from an X-block therefore a choice must first be made of the lattice plane to be used as plane of reference.

For AT test cuts the $01\bar{1}1$-plane (minor rhombohedron face) is chosen as plane of reference. This plane, like the AT cut and most other low-temperature coefficient cuts, is perpendicular to the YZ-plane (see *fig. 12*), and makes an angle of +38° 13′ with the Z-axis. The ideal AT cut, with $ZZ' = +35° 15'$, thus deviates by the angle $\gamma = -2° 58'$ from the plane of reference. The fact that this deviation

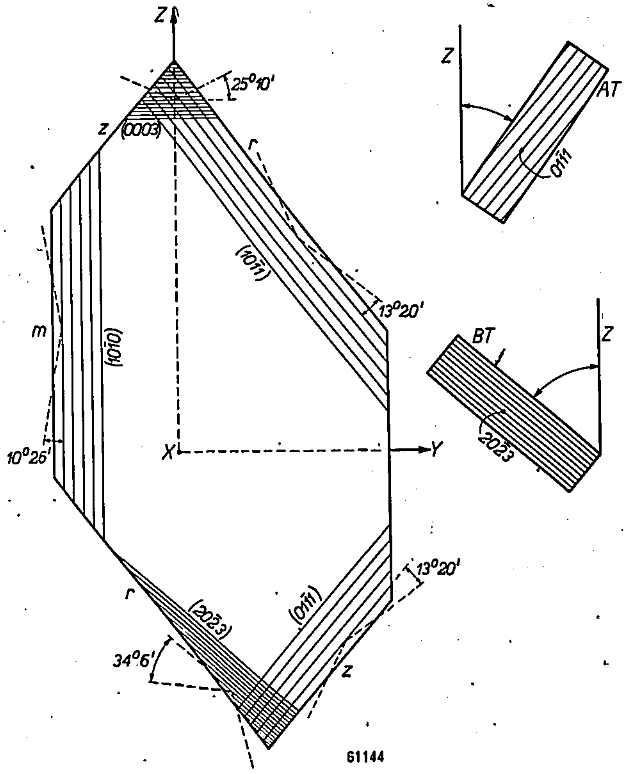


Fig. 12. Position in the quartz crystal of a number of lattice planes all of which are perpendicular to a YZ-plane. The $10\bar{1}1$-plane is the major rhombohedron face, the $01\bar{1}1$-plane the minor rhombohedron face. The latter is suitable as lattice plane of reference for AT cuts. For BT cuts the $20\bar{2}3$-plane is usually used for reference, since it is almost parallel to the BT cut.

Table I. Crystallographic and X-ray data for checking various cuts.

Name of cut	ZZ' of cut	Reference lattice plane	ZZ' of reference lattice plane	Angle γ	Counter tube setting (2Θ) for CuKα *)
XZ-planc	0°	$10\bar{1}0$	0°	0°	20° 52½′
YZ-planc	0°	$11\bar{2}0$	0°	0°	36° 34¾′
AT	+35°15′	$01\bar{1}1$	+38°13′	−2°58′	26° 39¾′
BT	−49°20′	$10\bar{1}1$	−38°13′	−11°07′	26° 39¾′
BT	−49°20′	$20\bar{2}3$	−49°44′	+0°24′	68° 12¾′
CT	+38°0′	$01\bar{1}1$	+38°13′	−0°13′	26° 39¾′
DT	−51°58′	$20\bar{2}3$	−49°44′	−2°14′	68° 12¾′
GT	+51°08′	$02\bar{2}3$	+49°44′	+1°24′	68° 12¾′

*) Based upon CuKα = 1.5418 Å. Calculated from the dimensions $a = 4.9131$ Å and $c = 5.4046$ Å of the unit cell of the quartz crystal, at 18.0 °C (measurements by A. J. C. Wilson and H. Lipson, Proc. Phys. Soc. **53**, 245-250, 1941).

is so small is very important for the practical measurements, as will be seen below. The chief advantages are that even fairly large errors in the cutting plane become accessible to measurement and the accuracy with which the correction can be determined is high.

Imagine the test cut turned in all directions in the crystal holder, as in the previous section, with the cut surface always against the reference surface and with the counter tube fixed in the 2Θ position corresponding to the reflection of the chosen lattice plane of reference. The normal to this reflecting plane again describes a cone but now with a half apex angle $|\gamma - \varepsilon|$, where ε is the error in ZZ' of the test cut. The position a of the goniometer arm necessary for maximum reflection then varies around the theoretical angle Θ to an amount between the maximum values $+|\gamma - \varepsilon|$ and $-|\gamma - \varepsilon|$. In order to measure the correct value of $|\gamma - \varepsilon|$, an arrow must be used again to preserve the orientation of the test wafer in transferring it from the saw table to the specimen holder of the X-ray machine. On placing the arrow point in the same direction in the holder as under the saw, the simple correlation rule for the direction of the required correction may also be applied [6]). Since, however, the half apex angle $|\gamma - \varepsilon|$ of the cone may be much larger (nearly equal to γ in the case of small ε) than that with which we were concerned in the previous section, small inaccuracies in the position of the arrow when measuring have greater consequences and may cause appreciable errors in the derived correction, errors increasing as γ becomes larger. Moreover, if γ is large, the normal to the lattice plane of reference may deviate so far from the horizontal plane when the arrow is slightly oblique that no reflection is observed in the horizontal goniometer plane and thus no measurement can be made.

From this it follows that it is important to have a small γ, and thus to choose a lattice plane of reference that is as nearly as possible parallel to the cut being checked. For BT cuts, the $20\bar{2}3$-plane is excellently suited, as this has an angle $ZZ' =$ −49° 44′; thus γ is only 0° 24′. Some other lattice planes of reference used for different types of cuts are indicated in *table I.*

Primarily, the X-ray measurement described above gives only the angle $|\gamma - \varepsilon|$, indicating the amount and, by the correlation rule, the direction of the correction Δa that would be necessary under the saw to make the test cut coincide with the lattice plane of reference. Now, in order to be able to make the test cut coincide with the ideal AT cut, we must calculate the angle ε. This is not possible unambiguously from the data obtained: although the value of the angle γ between lattice plane and AT cut is known ($\gamma = -2° 58'$ in our case), we do not know whether γ should be added to or subtracted from Δa, as the measurement cannot reveal whether $\varepsilon > \gamma$ or $\varepsilon < \gamma$. This difficulty, which is illustrated by *fig. 13 a, b*, does not present itself in the case that γ is rather large (lattice plane of reference far from parallelism to AT and to test cut), as in that case there will be no doubt that the orientation error ε is smaller than γ. If, however, for the sake of precision as explained above, γ is chosen rather small, the ambiguity can be solved only by repeating the measurement of the test cut (in a rough way) using another

[6]) This is the reason for determining on the test wafer the line of the X-axis, which is vertical in the process of cutting, and the direction of the Z-axis, indicating the latter by an arrowhcad, as was mentioned in article I. The line of the X-axis is made to stand vertically in the crystal holder, while the Z-arrowhead pointing to the right or to the left gives the correct direction. (The separate blanks are marked in the same way and may be similarly measured to make certain that the blanks to be lapped have an orientation within the desired tolerances.)

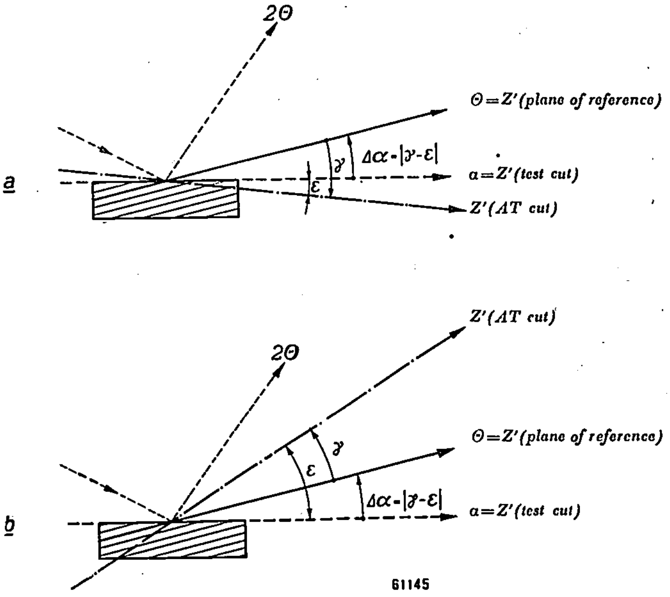


Fig. 13. The X-ray measurement gives the value and direction of the correction angle $\Delta\alpha = |\gamma - \varepsilon|$ that is required to make a test cut coincide with the lattice plane of reference. In order to find the correction ε necessary for obtaining the desired AT cut, it must moreover be known whether $\gamma > \varepsilon$ or $\gamma < \varepsilon$, which is not revealed by the X-ray measurement.

lattice plane of reference, for instance the $02\bar{2}3$-plane ($\gamma' = 35°\,15' - 49°\,44' = -14°\,29'$), the two conditions $\varepsilon = \Delta\alpha \pm \gamma$ and $\varepsilon = \Delta\alpha' \pm \gamma'$ together leaving only one possible value for ε.

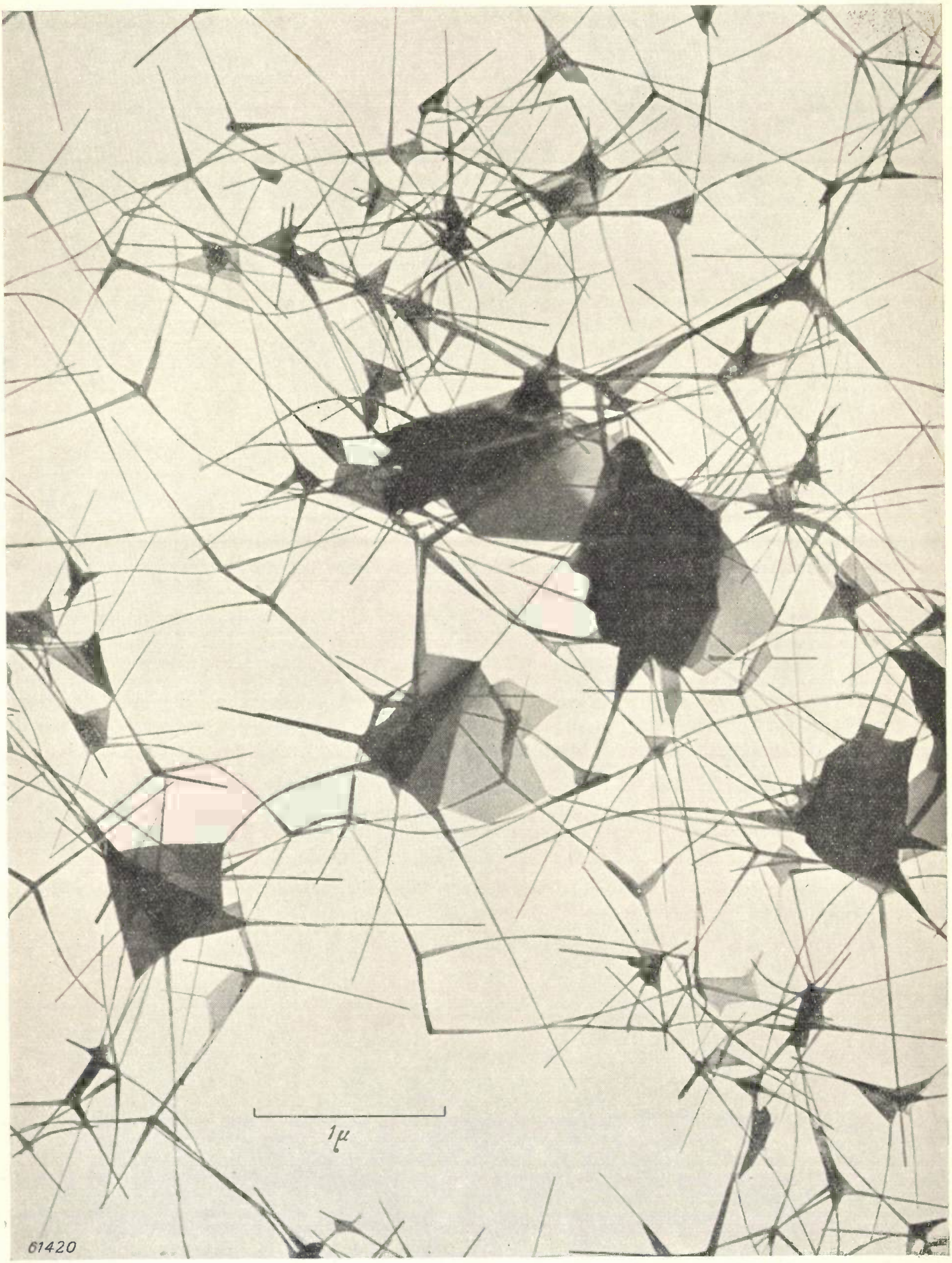
There are various other procedures possible with this equipment which are adaptable to the particular method of cutting (the previously discussed "strategy" of the cutting). For example in cases where it is desired to have a surface either parallel to or making some small angle with a chosen lattice plane, and the degree of precision required is beyond the accuracy easily attainable with the saw, the following method has proven useful. The crystal mounted in a special jig which is adjustable in two mutually perpendicular planes is placed in the X-ray beam and by manipulation of the setting screws of the jig it is tilted until the desired reflection occurs. The crystal is then correctly oriented with respect to the reference edges of the jig. The latter is then transferred to a lapping machine and the crystal surface ground to the orientation set by the X-ray machine. This procedure was used for wafering before accurate sawing methods were introduced. It is apparent that, if large numbers of wafers were cut in this manner, the method would waste a large amount of quartz, because thicker cuts are required than in direct sawing.

The method of measuring crystallographic angles by X-ray diffraction has also assumed some importance in other fields. With some modification it may be applied in the orientation of diamonds for wire drawing, and of sapphire needles used in gramophone pick-ups, in the cutting of barium titanate crystals, etc. [7]).

[7]) Nora Wooster, Orientation of diamonds by X-rays, Ind. Diam. Rev. 3, 1-3, 1943. R. H. Gillette and M. H. Jellinek, An instrument for rapid determination of crystal orientation, Rev. sci. Instr. 20, 480-483, 1949 (No. 7).

Summary. After an introductory discussion of the tolerances in the manufacture of quartz oscillator-plates, the X-ray diffraction apparatus is described which was constructed by the North American Philips Co., Inc. during the war to make possible a rapid and accurate check of the crystallographic angles of test cuts. A Geiger counter tube is used in this apparatus to detect the reflected X-ray beam. After placing the counter tube in the position 2Θ (Θ = Bragg angle) on a goniometer scale, the holder in which the test piece, test wafer or blank to be examined is clamped is turned until reflection occurs. The reflecting lattice plane is then oriented according to the angle Θ on the goniometer scale and from this the angles between this lattice plane and the reference surface of the crystal holder (cutting plane of the test cut) can be derived. The execution of the measurements, in particular the choice of a suitable lattice plane of reference is explained by means of several examples: checking the YZ-cut in making an X-block and checking AT and BT cuts.

ZINC-OXIDE CRYSTALS



When a small piece of zinc is burnt in a flame a white cloud of zinc-oxide is seen to rise, from which flakes up to a few mm in size are precipitated. A very minute flake collected along the edge of a specimen plate shows under the electron microscope a picture of small crystals with a large number of needle-like spurs loosely intertwined. Photograph taken with the Philips 100 kV electron microscope with a magnification of about 32 000 times.

THE PRACTICAL APPLICATION OF SAMPLING INSPECTION PLANS AND TABLES

by H. C. HAMAKER, J. J. M. TAUDIN CHABOT and F. G. WILLEMZE.

620.113.2:658.562

For the practical execution of sampling inspection by unskilled factory personnel the Philips Works in Holland have introduced a sampling table from which a suitable and efficient sampling plan can easily be derived for each particular case.

Recapitulation

The operating characteristic of any sampling plan — single, double, or sequential — can be specified with sufficient accuracy by two parameters. As such the point of control (p_0) and the relative slope (h_0) possess certain advantages, as explained in a previous article[1]). If a set of definite values of these parameters have been prescribed, the corresponding single sampling plan, specified by its sample size n_0 (= number of items to be inspected) and acceptance number c_0 (maximum number of rejects permitted for acceptance of the lot), can at once be deduced by means of the relations (see II):

$$c_0 = \frac{\pi}{2} h_0^2 - 0.73 \,, \quad . \; . \; . \; . \; . \quad (1)$$

$$n_0 = \frac{c_0 + 0.67}{p_0} \quad . \; . \; . \; . \; . \; . \; . \quad (2)$$

Should n_0 turn out to be rather high, this can be remedied by using a double sampling plan with the same characteristic (p_0, h_0). How to find such a double plan in a simple manner has also been explained in II (see table II in that article).

It might seem as if the problem of finding the most suitable sampling plan fitting a practical situation has thereby been solved: prescribe values of p_0 and h_0, and n_0 and c_0 follow automatically from (1) and (2).

To simplify matters still further, we may set up a table from which the values of n_0 and c_0 corresponding to all possible sets of values of p_0 and h_0 can be read at once.

However, in practice the matter is not so simple as suggested here. This will be better understood if we tear ourselves away from the atmosphere of abstract scientific argument in which this research has so far been conducted and transplant ourselves into the factory where sampling inspection has actually to be carried out.

Point of control and sharpness of inspection

At one end of the factory material is received, sometimes as raw materials but more often in the shape of parts or components manufactured elsewhere. This material will then be subjected to further treatment or assembled into higher aggregates, until the finished product is delivered at the other end. Frequently also batches of products have to be passed from one department to another during this production process.

For technical and administrative reasons this flow of material is subdivided into definite lots, and our aim will be from time to time to assess the quality of these lots by sampling inspection. This may be done at the entrance, at the exit, between departments, but also during processing, for example on the machine or the conveyor belt.

This last type of inspection, though of great importance because it serves as a check on the production process in the most direct way, is not considered in this paper, where we shall mainly be concerned with incoming and outgoing inspection procedures. Why these two types of inspection should be separately treated is explained later.

The inspection of the incoming and outgoing material will preferably be entrusted to a separate department with specialized personnel. This has the advantage that the useful information supplied by the inspection records with regard to the quality of the products or the manufacturing process can be systematically collected and evaluated.

In consultation with producer and consumer the inspector has to decide what sampling plan is to be applied for each lot of material submitted

[1]) H. C. Hamaker, The theory of sampling inspection plans, Philips Techn. Rev. 11, 260-270, 1949/1950 (No. 9). See also: H. C. Hamaker, Lot inspection by sampling, Philips Techn. Rev. 11, 176-182, 1949/1950 (No. 6). These two articles will be cited as II and as I respectively.

for inspection. That is to say, he has to choose the point of control and the degree of sharpness (which may suitably be measured by h_0) of the inspection procedure.

The point of control may conveniently be interpreted as the point dividing "good" and "bad" lots. Experience has taught that producer and consumer readily agree as to a suitable choice of this parameter.

But the choice of the sharpness of inspection is less simple and straightforward. A sharp inspection requires, roughly, a large sample and consequently much work, and this factor must be balanced against the evil consequences of the erroneous acceptance or rejection of some of the lots as a result of a less sharp control (see I). Even if we restrict ourselves to one type of product, the successive lots coming up for inspection are not of the same size; it commonly happens that one day a lot of 500 pieces has to be checked, and a few days later one of 5000. And it is obvious that in such cases we should wish to exercise a somewhat closer control over the larger lot. If, for example, the point of control has been fixed at 2% (that is to say, a lot with 2% defectives is given a 50% chance of acceptance), then the probability of accepting a lot with 5% defectives will certainly be small; and we shall require this latter probability to be smaller for a lot of 5000 items than for a lot of 500; this can only be achieved by combining the point of control of 2% with a higher value of h_0 in the case of the larger lot.

When the lots submitted for inspection are of the same size, but the points of control vary, a difference in sharpness of inspection is likewise desirable. If the relative slope h_0 were set at the same value with $p_0 = 2\%$ as with a $p_0 = 4\%$, for example, this would require a sample size half as large in the second case. The following arguments show that this cannot as a rule be the best choice. Frequently rejected lots are subjected to a 100% inspection, and sampling inspection necessarily implies that this occasionally happens to some "good" lots. In appraising the advantages of a sampling procedure the additional labour involved should also be taken into account; that is the total work of inspecting both the samples and the rejected "good" lots should be reduced to a minimum.

Now for two different products inspected with $p_0 = 2\%$ and $p_0 = 4\%$ respectively but with the same h_0, the probabilities of unnecessarily having to inspect a good lot will not on the average differ widely and the labour of inspection will be about the same in both cases. The sample size, however, is half as large in the second case. Consequently if h_0 has been so adjusted that the total amount of inspectional labour (samples + rejected good lots) is a minimum at $p_0 = 2\%$, the balance will be upset at $p_0 = 4\%$ and it will be of advantage to reduce the probability of rejecting a good lot by taking a somewhat higher value of h_0, that is by using a somewhat larger sample size.

The statement regarding the superfluous sorting out of "good" lots requires some further comment. In passing over from an operating characteristic with $p_0 = 2\%$ to one with $p_0 = 4\%$, h_0 being unaltered, the probability of rejecting a good lot, with say 1% defectives, is considerably reduced, as will be seen from *fig. 1a*. But this effect is in practice counterbalanced by the fact that a higher value of p_0 is as a rule assigned to products containing on the average a greater percentage of defectives. It is therefore to be expected that when $p_0 = 4\%$ then lots with 2% defectives will occur with the same relative frequency as lots with 1% defectives occur when $p_0 = 2\%$ (cf fig. 1*b*).

Extended over all percentages less than p_0, the integral of the product of this relative frequency and the probability of acceptance measures the chance that a "good" lot has to be subjected to 100% inspection. This integrated probability will not differ greatly in the two cases depicted in fig. 1.

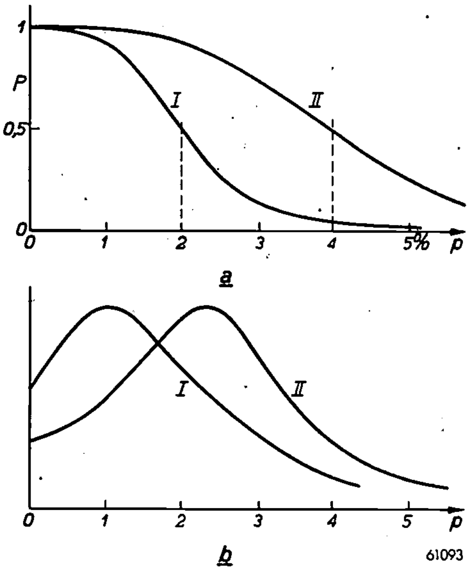


Fig. 1. a) Two operating characteristics with the same relative slope; *I* has a point of control (= percentage defectives corresponding to a probability of acceptance $P = {}^1/_2$) $p_0 = 2\%$, and *II* $p_0 = 4\%$, h_0 being equal; the difference amounts to a change of the scale along the abscissa.
b) A product for which p_0 is fixed at 4% is in general produced with a higher percentage defective than a product for which p_0 is set at 2%. The relative frequency of occurrence of lots containing p% defectives will roughly be as curves I and II in these two cases.

So far we have exclusively been arguing from the inspector's point of view, whose main aim is to reduce the labour involved in inspection as much as possible. From the standpoint of the consumer, however, it is also advisable to increase h_0 as p_0 increases, even for lots of the same size. Leaving h_0 constant, an increase in p_0 signifies a stretching of the operating characteristic by a proportional change in the scale along the p-axis (fig. 1a), thereby increasing the 10% consumer's risk point in the same ratio as p_0. Consequently the consumer who was running a 10% risk of receiving a lot with a certain high percentage defectives x as long as p_0 was 2% will at $p_0 = 4\%$ run the same 10% risk of obtaining a lot with $2x$ percent defectives. But the loss that will be incurred owing to the defective products is in this case twice as large. Though the consumer may concede to an increase in p_0 he will not be prepared to accept a proportional increase in his 10% risk point, so that again the use of a steeper operating characteristic, or of a higher value of h_0, is required.

Summing up, we may conclude that in practice the sharpness of inspection as measured by h_0 must be made a function of both the lot size and the point of control. Such a function, once adopted, might be presented in the form of a table with lot size and point of control as the two main entrances. It is simpler still, however, to combine this procedure with equations (1) and (2); this will lead to another table from which for a given size and point of control the sample size n_0 and acceptance number c_0 can be read off at once.

Only when presented in this form will a sampling table be acceptable in the factory where simplicity of manipulation is a paramount requirement and where even the simplest calculations should preferably be avoided.

The Philips SSS table

How the relative slope, h_0, should be made to depend on lot size and point of control is a matter of experience. Precise rules cannot be laid down owing to the impossibility of determining the influencing economic factors, as discussed in I.

Nevertheless there are many cases in modern mass production where these economic factors do not differ greatly in importance where the occasional acceptance of a bad lot does not involve an excessive loss and where the cost of inspection — for example the checking of dimensions with gauges, or the visual inspection of the finish of machine-shop products — is relatively low. As proved by experience, it is possible to provide a single table that can successfully be applied to all such cases alike. The practicability of sampling inspection is thereby much improved, while, at the same time the uniformity attained in taking decisions greatly contributes towards a good understanding and smooth arrangements between consumer and producer.

The Philips Standard Sampling System is a table of this kind which is employed in the Dutch Philips factories and is reproduced in *table I.*

The increase in the relative slope h_0 downwards and towards the right in this table has been chosen in analogy with existing tables which were known to give satisfactory results in practice, particularly in analogy with the Philips R.M.I. Tables (Raw Material Inspection) in use in our factories in Great Britain. Since approximately the same operating characteristics have been adopted as corresponding to that table, both these tables can be interchanged without fear of running into serious trouble.

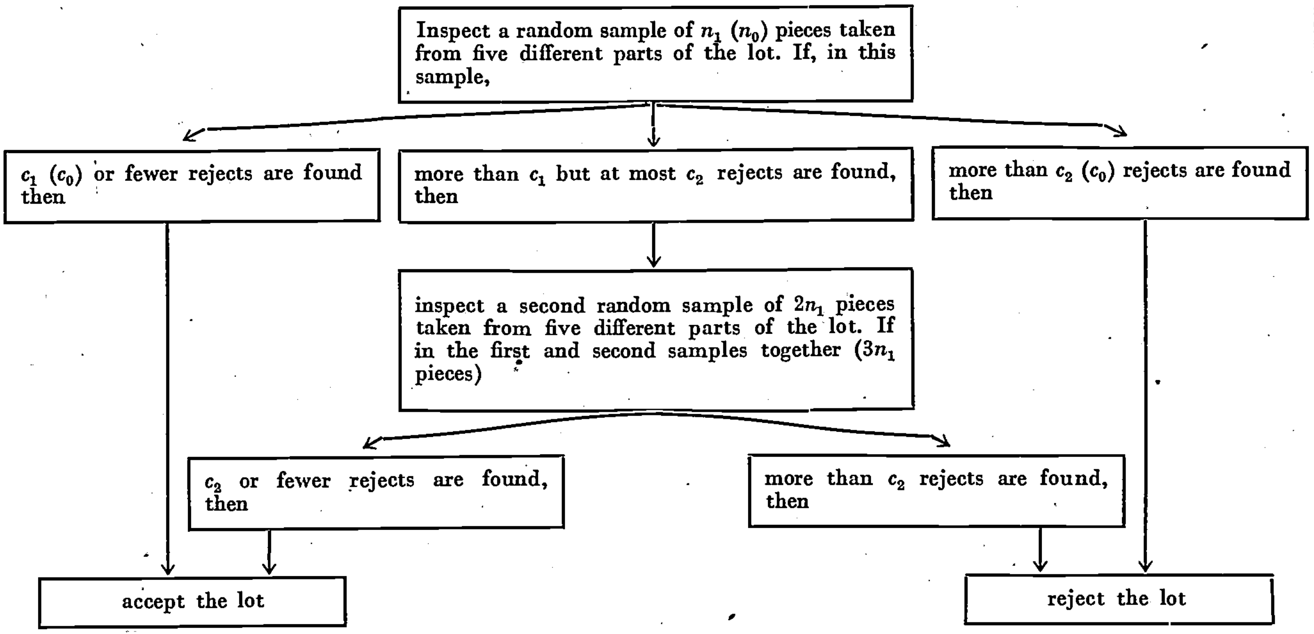
Practically all existing sampling tables are based on the same principle of giving n_0 and c_0 — or in case of double sampling n_1, n_2, c_1, c_2 and c_3 — as a function of two parameters, one of which is always the lot size. Usually the second parameter is not the point of control but some other quantity serving the same purpose of taking into account the requirements of producer and consumer. Our reasons for preferring the point of control have already been explained in II; a further advantage of this parameter is that its meaning can be roughly explained as the boundary between "good" and "bad" lots, a definition that is readily understood in the factory.

The "control" used in the Philips R.M.I. tables is much the same as the "point of control" adopted in the SSS tables, though mathematically not so sharply defined. Against the R.M.I. table, however, some other objections may be raised, as will be explained in the appendix, where various other tables are briefly discussed.

The Philips SSS table contains both single and double sampling plans. Small lots necessarily require small samples, and the advantages of double sampling are then of little importance. Only for lot sizes exceeding 1000 has double sampling been applied.

In accordance with the theoretical arguments advanced in II, the second sample is invariably twice as large as the first, while, to simplify matters still further, we have also adhered to the Dodge and Romig principle ($c_2 = c_3$, see Appendix). In most cases c_3 has been made 5 times c_1, a choice that guarantees a satisfactory efficiency of the double sampling plans.

Tabel I. Philips Standard Sampling System (SSS).



Inspect a random sample of n_1 (n_0) pieces taken from five different parts of the lot. If, in this sample,

c_1 (c_0) or fewer rejects are found then

more than c_1 but at most c_2 rejects are found, then

more than c_2 (c_0) rejects are found then

inspect a second random sample of $2n_1$ pieces taken from five different parts of the lot. If in the first and second samples together ($3n_1$ pieces)

c_2 or fewer rejects are found, then

more than c_2 rejects are found, then

accept the lot

reject the lot

N.B. Where the letter A is given under n_0 this means that no sample is to be taken but that the whole lot has to be inspected.

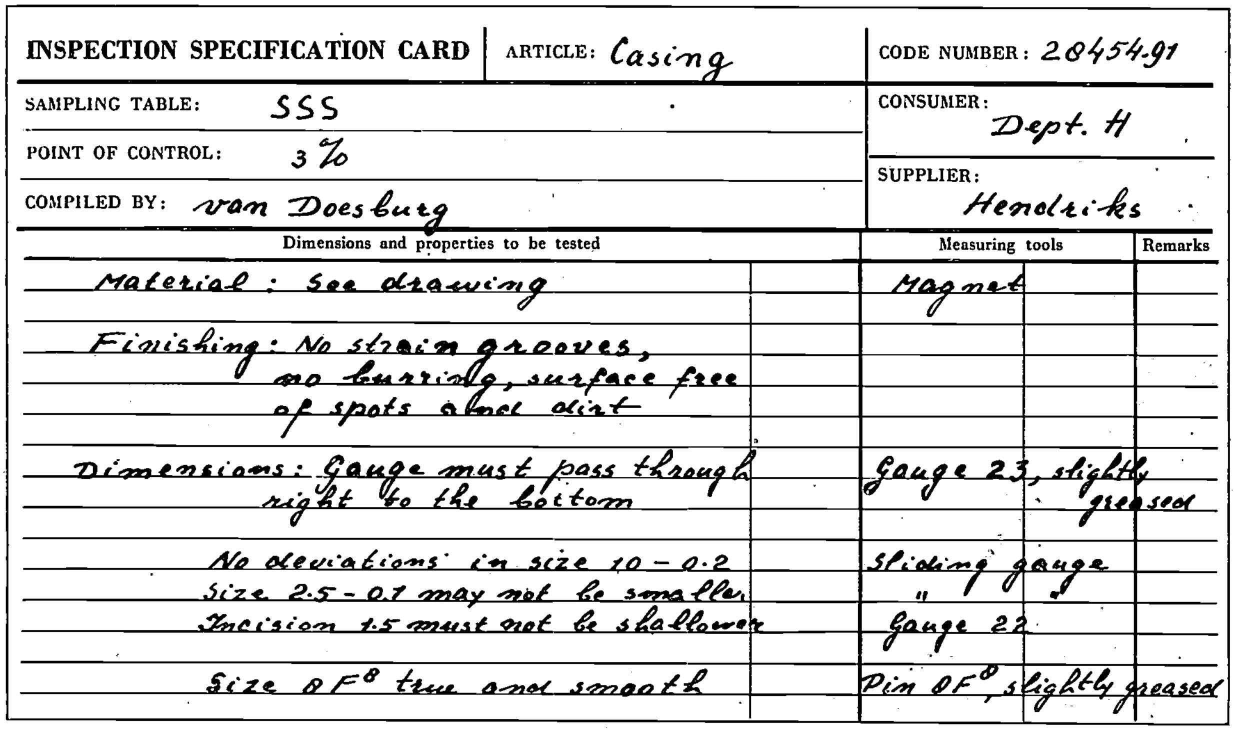
	POINT OF CONTROL LOT SIZE	¼%			½%			1%			2%			3%			5%			7%			10%		
Single-sampling plans		n_0	c_0		n_0	c_0		n_0	c_0		n_0	c_0		n_0	c_0		n_0	c_0		n_0	c_0		n_0	c_0	
	20-50	A	—		A	—		A	—		30	0		20	0		13	0		10	0		7	0	
	51-100	A	—		A	—		60	0		30	0		20	0		13	0		10	0		7	0	
	101-200	A	—		100	0		60	0		35	0		55	1		35	1		25	1		17	1	
	201-500	175	0		100	0		135	1		75	1		55	1		35	1		40	2		25	2	
	501-1000	225	0		225	1		150	1		85	1		85	2		55	2		55	3		35	3	
Double-sampling plans		n_1	c_1	c_2	n_1	c_1	c_2	n_1	c_1	c_2	n_1	c_1	c_2	n_1	c_1	c_2	n_1	c_1	c_2	n_1	c_1	c_2	n_1	c_1	c_2
	1001-2000	330	0	1	150	0	1	110	0	2	55	0	2	45	0	3	25	0	3	30	1	5	22	1	5
	2001-5000	425	0	2	200	0	2	135	0	3	70	0	3	70	1	5	45	1	5	55	2	10	40	2	10
	5001-10000	525	0	3	260	0	3	220	1	5	110	1	5	125	2	10	75	2	10	75	3	15	55	3	15
	10000-20000	875	1	5	440	1	5	380	2	10	190	2	10	180	3	15	110	3	15	100	4	20	70	4	20
	20000-50000	1500	2	10	750	2	10	540	3	15	270	3	15	240	4	20	140	4	20	120	5	25	85	5	25
	50000-100000	2200	3	15	1100	3	15	700	4	20	390	4	20	290	5	25	175	5	25	145	6	30	105	6	30

Slight discontinuities are unavoidable because the acceptance numbers (c_0 or c_1, c_2, c_3) have always to be rounded off to integer values, but this is of no practical consequence since in the choice of a sampling plan a high precision is not required.

For the same reason it is permissible to subdivide the lot sizes and points of control into a few classes only, thus limiting the size of the table so that it can easily be issued in a pocket-size edition. In classifying lot sizes the well-known sequence 1, 2, 5, 10, 20, 50 etc. has been used, a feature that may help to make the table more readily acceptable to the inspector [2]).

In order to prevent inhomogenieties in the lot causing a bias in the sample it has been prescribed

[2]) See next page.



INSPECTION SPECIFICATION CARD	ARTICLE: Casing	CODE NUMBER: 28454.91
SAMPLING TABLE: SSS		CONSUMER: Dept. H
POINT OF CONTROL: 3 %		
COMPILED BY: van Doesburg		SUPPLIER: Hendriks

Dimensions and properties to be tested	Measuring tools	Remarks
Material: See drawing	Magnet	
Finishing: No strain grooves, no burring, surface free of spots and dirt		
Dimensions: Gauge must pass through right to the bottom	Gauge 23, slightly greased	
No deviations in size 10 - 0.2	Sliding gauge	
Size 2.5 - 0.1 may not be smaller	" "	
Incision 1.5 must not be shallower	Gauge 22	
Size $\varnothing F^8$ true and smooth	Pin $\varnothing F^8$, slightly greased	

Fig. 2. "Inspection specification card" on which the requirements to be satisfied by a particular product have been registered.

that each sample should be composed of at least 5 sub-samples drawn from different places in the lot.

Finally it may be mentioned that the table is only intended as a general directive and not as a rigid rule. If experience shows that the lots of a certain product are practically always accepted and never rejected, the inspector is permitted to reduce his inspection procedure. In such cases he is advised to leave the point of control unaltered and to change over to a sampling plan in the same column of the table one or two lines higher up than the plan so far applied.

The use of the table in practice

In addition to one or two copies of the table an inspection department should have at its disposal the complete specification of the requirements to be satisfied by the various articles it may be called upon to inspect.

It is convenient to have these arranged in a card index according to code number, and *fig. 2.* shows a card as in use in some of the inspection departments of our Eindhoven factories. On this card there are also recorded the tools to be used and the point of control required.

It is moreover of importance carefully to register the results of each inspection carried out, and in *fig. 3* a second card designed for this purpose is reproduced. Next to the date and the number of the consignment follow the size of the lot and the result of the inspection by sample, while farther to the right some columns have been provided for recording the decision taken.

Most types of products may show a series of different defects and the actual kinds of defects observed can be specified on the back (fig. 3*b*). In cases of controversy it is then always possible to produce the data that resulted from the actual inspection.

When materials received from outside have to

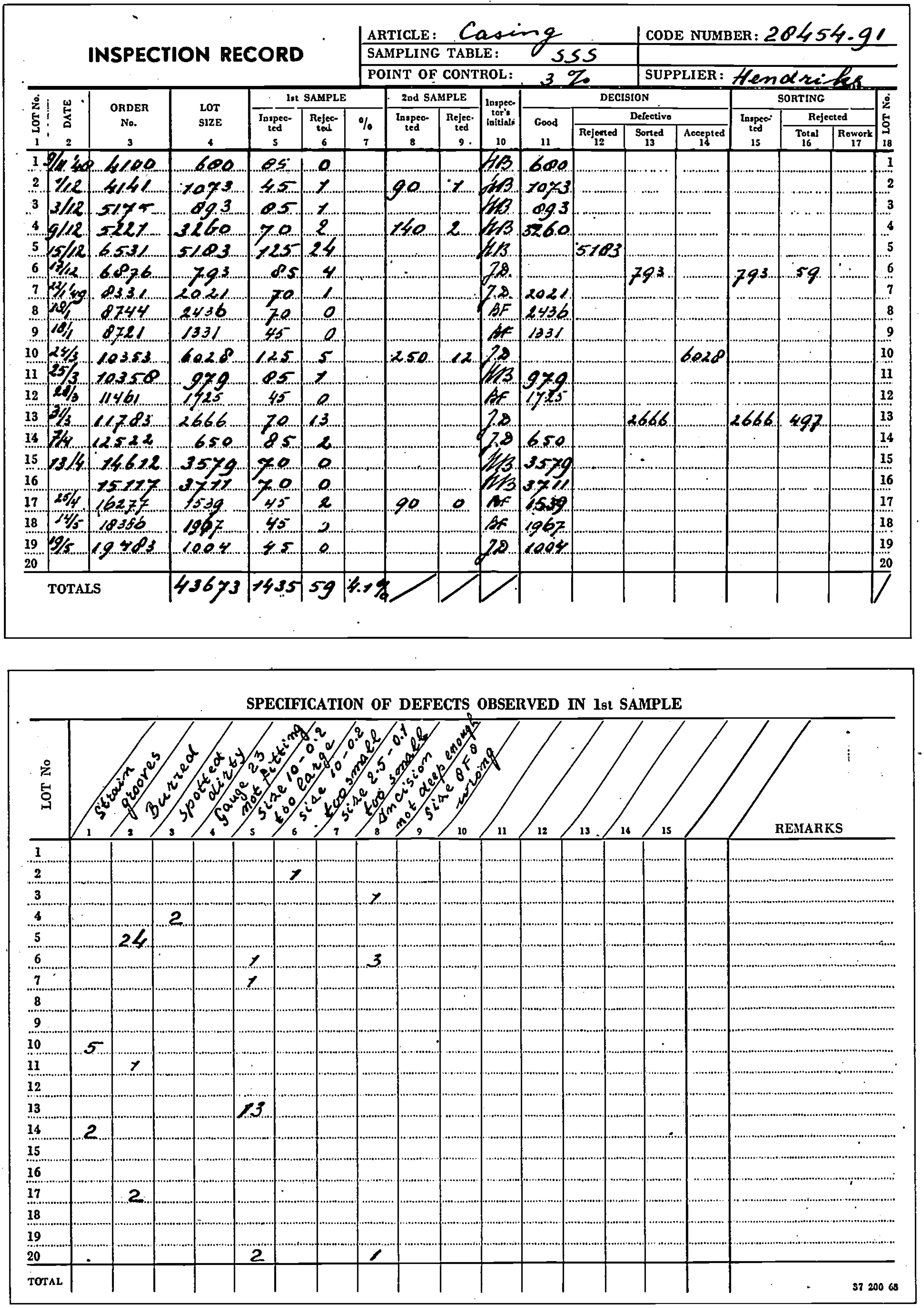
[2] When the lot is of small size the sample will often embrace a large proportion of the whole lot and the conditions under which formulae (1) and (2) are valid (Poisson probabilities) do not apply. By a more detailed analysis it has been found, however, that equations (1) and (2) can be extended so as to provide a reasonable approximation also under these circumstances. We then have

$$n_0p_0 = c_0 + 0.67 - 0.33\, n_0/N,$$

$$(\pi/2)h_0^2 = (n_0p_0 + 0.06)/(1 - n_0/N),$$

where N denotes the lot size. These equations still presuppose that p_0 is small. We see that h_0 increases and the operating characteristic thus becomes steeper as the ratio n_0/N becomes larger; h_0 becomes infinite when $n_0 = N$, that is when the whole lot is inspected.

In setting up the SSS table these extended formulae have been used as far as possible. To take the effect of a finite sample size accurately into account it would, however, be necessary to subdivide the lot size into a larger number of classes, and this did not seem advisable.



INSPECTION RECORD

ARTICLE: Casing
SAMPLING TABLE: SSS
POINT OF CONTROL: 3 %
CODE NUMBER: 28454.91
SUPPLIER: Hendriks

LOT No.	DATE	ORDER No.	LOT SIZE	1st SAMPLE			2nd SAMPLE		Inspector's initials	DECISION				SORTING			LOT No.
				Inspected	Rejected	%	Inspected	Rejected		Good	Defective			Inspected	Rejected		
											Rejected	Sorted	Accepted		Total	Rework	
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	9/11 '48	4100	600	85	0				HB	600							1
2	1/12	4141	1073	45	1		90	1	HB	1073							2
3	3/12	5175	893	85	1				HB	893							3
4	9/12	5221	3260	70	2		140	2	HB	3260							4
5	15/12	6531	5183	125	24				HB		5183						5
6	13/12	6876	793	85	4				J.D.			793		793	59		6
7	24/1 '49	8331	2021	70	1				J.D.	2021							7
8	18/1	8744	2436	70	0				BF	2436							8
9	18/1	8721	1331	45	0				BF	1331							9
10	24/3	10353	6028	125	5		250	12	J.D.				6028				10
11	25/3	10358	979	85	1				HB	979							11
12	28/3	11461	1725	45	0				BF	1725							12
13	31/3	11783	2666	70	13				J.D.			2666		2666	497		13
14	7/4	12522	650	85	2				J.D.	650							14
15	13/4	14612	3579	70	0				HB	3579							15
16		15117	3711	70	0				HB	3711							16
17	25/4	16277	1539	45	2		90	0	BF	1539							17
18	14/5	18356	1967	45	3				BF	1967							18
19	19/5	19483	1004	45	0				J.D.	1004							19
20																	20
TOTALS			43673	1435	59	4.1%											

SPECIFICATION OF DEFECTS OBSERVED IN 1st SAMPLE

LOT No	Strain grooves 1	Burred 2	Spotted dirty 3	Gauge 23 not fitting 4	Size 10-0.2 too large 5	Size 10-0.2 too small 6	Size 2.5-0.1 too small 7	Incision not deep enough 8	Size 8F8 wrong 9	10	11	12	13	14	15		REMARKS
1																	
2						1											
3								1									
4			2														
5		24															
6					1			3									
7					1												
8																	
9																	
10	5																
11		1															
12																	
13					13												
14	2																
15																	
16																	
17		2															
18																	
19																	
20					2			1									
TOTAL																	

37 200 68

Fig. 3. "Inspection record" on which, for a particular product, the data resulting from the samples are registered. *Top:* the front with the results of the successive inspections. *Below:* the back where the defects observed are analyzed in detail.

be inspected it is of some importance to set up a separate record for each product and for each of the sources of supply if there are more than one. When the record (fig. 3*a*) has been completed the sample data (in the case of double sampling only those of the first sample) can be added to give a final result. In fig. 3*a*, for example, 59 rejects have been detected in a total of 1435 items inspected, that is 4.1 %. This provides a measure of the average percentage defectives produced or received during the period under consideration. The final data can be entered on a separate list, and after a certain lapse of time the original records can be destroyed in order to avoid a harassing accumulation of paper.

If the amount of material needed diminishes or increases it may be of advantage to take into account the quality as well as the price of the products from different suppliers, and the collective records will furnish the data required. They will likewise provide useful information for the producer as to what types of defects are of most frequent occurrence and should be particularly attended to.

In conclusion it may be worth while mentioning the limits and restrictions to the use of the SSS table.

One important category to which the table does not apply are life tests, and destructive tests in general. It will be understood that owing to the high cost of inspection the interplay of various economic factors lead in this case to different requirements regarding point of control and relative slope than those on which the SSS table has been founded.

The control of the production process is another case outside the scope of our table. Here the flow of material passing a machine or a conveyor belt is subdivided into small batches consisting, for instance, of half an hour's output. These are tested by small samples, so that serious derangements in the production process, such as may result from sudden misalignment of machines, deterioration of tools or lack of attention of personnel, can be quickly detected and corrected. The discovery of these serious errors does not require a very sharp inspection, so that the sample sizes may be much smaller than in the SSS table. If subsequently the output of say a day or a week is pooled to form one lot, the percentage defectives in this total lot can be judged by adding together the results of all the samples inspected during the corresponding period; in this way a final and sharp control of the lot as a whole can be achieved without additional inspection.

Appendix: Other sampling tables

Dodge and Romig were the first to publish, in 1941, a set of four different sampling inspection tables, two for single and two for double sampling. In these tables, or rather series of tables, the choice of a sampling plan is determined by three parameters. It is assumed that by previous observation the average percentage defectives in the lots submitted for inspection has been determined. This so-called "process average" serves to discount the quality that can be produced; it is one of the two parameters functioning as main entrances in each table, the lot size being the other. As third parameter the average outgoing quality limit (AOQL) is employed in two of the series of tables (one for single and one for double sampling), and the 10% consumer's risk point (see II, p. 267/268), called "the lot tolerance percent defective", in the other two. In the double-sampling plans the ratio of the sample sizes varies, but the criteria always satisfy the condition $c_2 = c_3$, which we have consequently designated as the Dodge and Romig principle.

Lot size, process average, and 10% consumer's risk point (or AOQL) have been classified into 20, 6 and 8 different classes respectively: Consequently each series of tables comprises 20 × 6 × 8 = 960 different cases for each of which a sampling plan is prescribed; hence these tables cover many pages.

Since these tables were developed it has been realized that the differentiation had been carried one or two stages further than needed in practice. This at least explains that the "Army Service Forces Tables" — set up during the second world war by a team of statisticians including Dodge and Romig, and later taken over by the Statistical Research Group, Columbia University — are simpler and much more concise than the original tables of Dodge and Romig. Here only two parameters are employed, namely the 5% producer's risk point and the size of the lot. To simplify matters still further the sample size has been once and for all coupled with the lot size; when arranged as in table I the Army Service Forces Tables contain one sample size for each line of the table which has to be used in all columns. Separate tables have been provided for single and for double sampling, while in the latter the second sample is invariably twice as large as the first. A subdivision according to 7 different lot sizes and 14 values of the producer's risk point leads to a table of in all 98 combinations which may be printed on a single page [3]).

Despite the simplifications, certain objections can still be brought forward against these tables, which, as far as we can see, are mainly a consequence of the rigid connection between lot size and sample size.

As may be inferred from equations (1) and (2), this coupling signifies that for a given lot size (and consequently a fixed value of n_0) the acceptance number c_0 and, with it, h_0 are completely prescribed in their dependence of p_0. For two different lot sizes this functional relation between h_0 and p_0 is represented by the curves CRG in *fig. 4*.

That h_0 increases with p_0 is in keeping with the principles explained earlier in this paper. In our opinion, however, the

[3]) 500 is the smallest value of lot sizes in the "Army Service Forces Tables". They have been extended by the Statistical Research Group, Columbia University, towards the side of smaller lots, which has raised the number of lot-size classes to 14.

increase of h_0 with p_0 is in the "Army Service Forces Tables" more pronounced than is justifiable on these grounds.

In the R. M. I. tables of our English factories, and consequently also in our SSS tables, a more gradual increase of h_0 with p_0 has been adopted, which is expressed in table I as a decrease of the sample sizes with increasing p_0 for a fixed lot size. The relation between h_0 and p_0 as adopted in the R. M. I. and SSS tables has also been represented in fig. 4, each by two curves holding for lot sizes approaching those of the CRG curves as closely as possible.

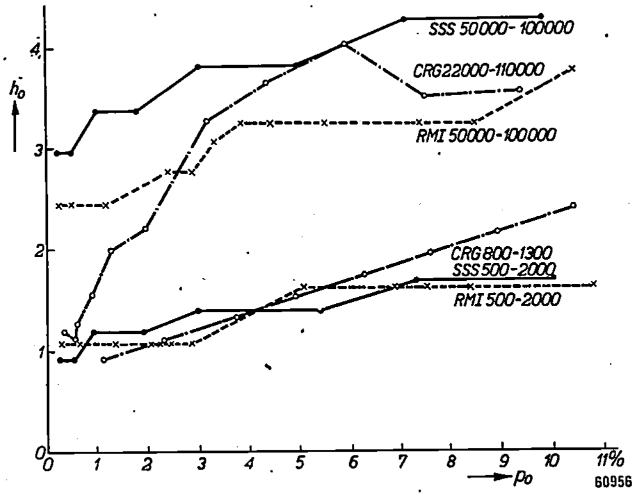


Fig. 4. Variation of h_0 in various sampling tables, plotted as a function of p_0 for a class with a medium and for a class with a large lot size.
CRG: tables of the Columbia Research Group.
RMI: Raw Material Inspection table of the English Philips' factories.
SSS: Philips Standard Sampling System.
The irregularities in the curves are due to the fact that only whole numbers could be chosen for the criteria. The sharp bend in the CRG curve for large lots arises from the fact that in the CRG table at high values of p_0 the coupling of lot size and sample size referred to in the text is no longer adhered to and smaller samples are used.

More serious objections may be raised against the double sampling plans adopted in the "Army Service Forces Tables". As inferred in II from the random walk diagram, a reasonable efficiency requires that the criteria of a double-sampling plan shall satisfy the relation.

$$c_1 < \frac{n_1}{n_1 + n_2} c_3 . \; . \; . \; . \; . \; . \; . \; . \; . \quad (3)$$

If this condition is not fulfilled we obtain double-sampling plans which operate with an efficiency even worse than that of an equivalent single-sampling plan, as was demonstrated in fig. 6 of II. Now it has been found that out of the 40 different combinations of criteria found in the Army Service Forces Tables no less than 27 do not satisfy equation (3), which in this particular case requires $c_1 < {}^1/_3\, c_3$. In constructing this double-sampling table the purpose has been to obtain double-sampling plans with operating characteristics closely approximating those of the corresponding single-sampling plans in the single-sampling table. Since, however, the sample sizes for double sampling were prescribed and coupled with the lot size, the requisite variation in the operating characteristic could only be achieved by having recourse to inefficient combinations of the criteria.

Apart from this, in our opinion the sample sizes have not been correctly fixed. For example, corresponding to a single-sampling plan with

$$n_0 = 150, \quad c_0 = 6$$

the Columbia Research Group arrives at a double sampling plan with

$$n_1 = 100, \; n_2 = 200, \; c_1 = 3, \; c_2 = c_3 = 9,$$

whereas by means of table II in (II) we find

$$n_1 = 61, \; n_2 = 122, \; c_1 = 1, \; c_2 = c_3 = 7$$

as the most suitable approximation; h_0 is 2.07 for the single sampling and 2.08 for this second double-sampling plan, so that the operating characteristics lie very close together. The ratio c_1/c_3 is 1/3 in the first and 1/7 in the last of the double-sampling plans, the higher efficiency of which is demonstrated by a reduction of about 40 % in the sample size.

Consequently, corresponding to a single sample of 150, the double-sample sizes $n_1 = 100$, $n_2 = 200$ are too large, and if we try to adjust the two plans to one operating characteristic we are necessarily forced to adopt an inefficient combination of criteria in the latter case.

It will be appropriate to conclude this paper with a brief discussion of the Philips R.M.I. tables introduced in our English factories by A. S. Wharton. Part of this table is reproduced in *Table II.*

The R.M.I. system prescribes that not one but five different samples shall be taken all of the same size, N, and inspected separately. If the number of defectives recorded in each of these five samples is less than the criterion K the lot is accepted straightaway; if in one sample the number of defectives is greater than K, or if in two samples it is equal to K, the lot is rejected. But if in one out of the five samples K defectives are found, the other samples containing less than K, then a

Table II. Part of the R.M.I. table as used in the Philips factories in England.

Point of control Lot size	¼%		½%		1%		2%		2½%		3%	
	N	*K*	*N*	*K*	*N*	*K*	*N*	*K*	*N*	*K*	*N*	*K*
500 - 2000	65	1	30	1	15	1	10	1	9	1	8	1
2000 - 5000	65	1	30	1	15	1	30	2	28	2	25	2
5000 - 15000	290	2	140	2	50	2	30	2	28	2	45	3
15000 - 25000	540	3	275	3	120	3	60	3	50	3	45	3
25000 - 50000	540	3	275	3	120	3	90	4	70	4	60	4
50000 - 100000	800	4	420	4	180	4	120	5	100	5	110	6
100000 - 250000	1050	5	560	5	280	5	180	6	140	6	145	7

second set of five samples of the same size N must be drawn, and the lot is finally accepted if all these contain less than K defectives.

Thus in principle we have again to do with a double-sampling table, judgement being based on the maximum number of defectives observed in a set of five equal samples. This is the weak point; the numbers of defectives found in the other four samples furnish a certain amount of information with respect to the percentage defectives in the lot, and this information is discarded. It is therefore to be expected that the sampling plans of the R.M.I. tables will be relatively inefficient[4]); this is corroborated by the inverse efficiency characteristics in *fig. 5*, which have been obtained by the methods developed in II.

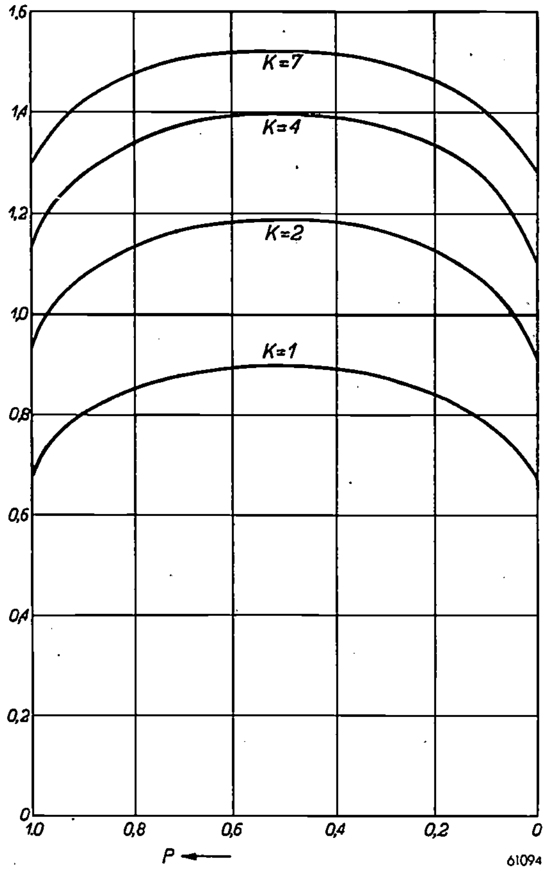


Fig. 5. Inverse efficiency characteristics of the sampling plans of the R.M.I. table. For $K>1$ efficiency is less than that of the equivalent single-sampling plans (which, by way of definition, has been given the efficiency 1).

For $K>1$ the efficiency is reasonably good owing to the fact that under these conditions the R.M.I. system operates as a double-sampling plan: whether we require that the number of defectives in every one of five samples shall be less than 1, or that this shall hold for the total number observed in the five samples together, is just the same thing.

For $K>1$ the efficiency of the R.M.I. table is not as satisfactory as it might be, which shows that by resorting to single or double-sampling plans it should be possible to diminish the average sample size without impairing the degree of control exercised. This possibility of reducing the average number of observations required, while retaining the practical features of the R.M.I. table, has led us to develop the SSS table described above; as already stated and illustrated in fig. 4, the relation between p_0 and h_0 is approximately the same in both tables and the operating characteristics of corresponding sampling plans consequently lie close together.

[4]) To avoid misunderstanding it should be borne in mind that the "efficiency" is a mathematical concept referring only to the number of observations needed and not to the other factors, such as conciseness and simplicity, which also contribute towards the practical success of a table in the factory and which have been duly considered above.

Summary. If we specify the operating characteristic of a sampling plan by its point of control, p_0, and relative slope, h_0, a suitable single or double sampling plan corresponding to any prescribed course of the characteristic can be found in a simple manner. In general it will be advantageous to use a steeper characteristic (a higher value of the relative slope) the larger the size of the inspection lot and the higher the point of control. If a certain functional relation between h_0, p_0 and lot size is adopted then a table can be compiled from which a sampling plan can at once be derived as a function of lot size and point of control.

The Philips SSS table, which was mainly developed on these principles, has in other respects been adjusted to practical requirements with a view to its applicability in the factory by unskilled personnel. In an appendix various other sampling tables are compared with the SSS table.

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

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

1884: J. L. H. Jonker and A. J. W. M. van Overbeek: Control of a beam of electrons by an intersecting electron beam (Nature London **164**, 276, 1949, Aug. 13).

A ribbon-shaped beam of electrons (100 V), emerging from an oblong, concave cylindrical oxide-coated cathode, is crossed by a second ribbon-shaped beam (16 V). This second beam is influenced by the space charge of the first one. In consequence of this the current I_2 in the second beam can be made to decrease from 10^{-5} A to 10^{-6} A by varying the potential on the control grid of the first beam. If the anode of the second beam is replaced by a secondary emission multiplier, slopes of several amps per volt can be attained at anode currents of the order of 10^{-2} A.

1885: C. J. Bakker: Valve noise and transit time (Wireless Eng. **26**, 277, 1949, Aug.).

The writer gives an answer to criticism by Campbell, Francis and James and by Holding (Wireless Eng. **25** (1948), pages 148 and 372 respectively), regarding his 1941 article on valve noise (Physica **8**, p. 23). Further the reader is referred to a letter to the Editor of the Wireless Eng. by F. L. H. M. Stumpers (see these abstracts, No. **1886**).

1886: F. L. H. M. Stumpers: Measurements of induced grid noise (Wireless Eng. **26**, 277-278, 1949, Aug.).

Short description of an arrangement for measuring induced grid noise after Bakker (Physica **8**, 23, 1941) and report of some measurements in the frequency range 20-100 Mc/s. The measurements are in satisfactory agreement with Bakker's theory.

1887: J. J. Went: Relation between the thermal expansion, the Curie temperature and the lattice spacing of homogeneous ternary nickel-iron alloys (Physica The Hague **15**, 703-710, 1949, No. 8/9).

The importance of more experimental data concerning the exchange energy in ferromagnetic alloys is pointed out. The Curie temperature, the expansion anomaly below the Curie temperature and the lattice spacing of 15 only slightly different homogeneous ternary Ni-Fe alloys with about 50% Fe and 50% Ni are measured. A close relationship exists between the change in Curie temperature and the change in the expansion anomaly between different alloys. The value of this change in the Curie temperature depends upon the position of the third alloying element in the periodic system of elements with respect to the position of Ni in this system. There is no direct relation at all between the change in Curie temperature and the lattice spacing.

1888: F. M. Penning and H. J. A. Moubis: On the normal cathode fall in neon (Physica The Hague **15**, 721-732, 1949, No. 8/9).

Applying the same method which had furnished reproducible results for molybdenum, the normal cathode fall V_n in neon has been determined for a number of other materials (plate cathode, 40 mm pressure). The results are given in table III. The values of V_n are compared with those of the work function φ and large deviations have been found from the relation $V_n = C\varphi$ (C a constant).

1889: Th. P. J. Botden and F. A. Kröger: Energy transfer in tungstates and molybdates activated with samarium (Physica The Hague **15**, 746-768, 1949, No. 8/9).

Earth alkali tungstates and molybdates activated with samarium show fluorescence in different bands, which can be correlated with the tungstate group or molybdate group and with trivalent samarium. Upon excitation with λ 3560 Å in one of the absorption bands that are characteristic of samarium, only the orange-red samarium fluorescence is emitted. Upon excitation with short-wave ultra-violet in the tungstate or molybdate absorption bands, both types of fluorescence bands are emitted, in a proportion that is dependent on temperature. At low temperatures the efficiencies of both types of fluorescence are constant. At the temperature at which the tungstate or molybdate fluorescence is quenched or at a slightly higher temperature, the intensity of the samarium fluorescence increases with increasing temperature up to a constant value; at a still higher temperature the samarium fluorescence is also quenched.

The increase of the samarium fluorescence is due to the transfer of energy from tungstate or molybdate groups to samarium ions. The results are discussed and calculations have been made on the basis of the theories of Mott-Seitz and of Möglich-Rompe for the quenching of fluorescence. The results constitute an argument in favour of the model of Mott and Seitz.

1890: F. A. Kröger: A proof of the associated-pair theory for sensitized luminophors (Physica The Hague **15**, 801-806, 1949, No. 8/9).

For $Ca_3 (PO_4)_2$ and $Sr_3 (PO_4)_2$ containing manganese as an activator and tin or cerium as a sensitizer the spectral distribution of the manganese fluorescence, the temperature-dependence of its intensity and the temperature-dependence of its decay have been measured. Analysis of the experimental data leads to figures for the transition probabilities of the fluorescence process and the dissipation process in the manganese centres. Comparison shows that these quantities are markedly different for the centres in products containing tin and cerium as sensitizers. This proves that the sensitizer atoms must be close to the activator atoms, as was assumed in the associated-pair theory of sensitization.

1891: F. A. Kröger, J. Th. G. Overbeek, J. Goorissen and J. van den Boomgaard: Bismuth as activator in fluorescent solids (J. Electrochem. Soc. **96**, 132-141, 1949, No. 3).

Trivalent bismuth forms fluorescence centeıs in its own compounds as well as in other systems in which it is present as an activator. The fluorescent emission consists of various bands lying between the ultraviolet and the red end of the spectrum. The relative intensity of these bands depends upon the nature of the lost lattice and the temperature of observation.

Sulfates and phosphates of the alkine earth group show predominantly red fluorescence. The red fluorescence of $Ca_2P_2O_7$-Bi shows the remarkable feature that its temperature-dependence is different for excitation by cathode rays and by X-rays.

R 122: W. Nijenhuis: A note on a generalized Van der Pol equation (Philips Res. Rep. 4, 401-406, 1949, No. 6).

The problem of solving the asymmetrical Van der Pol equation $\ddot{v} - \varepsilon (1 - 2 \beta v + v^2) \dot{v} + v = 0$ is equivalent to that of solving an ordinary Van der Pol equation ($\beta = 0$) with a constant right-hand side. For large values of ε and β a simple approximation of the limit cycle is found from which it follows that the ratio of the maximum amplitudes in the positive and negative directions cannot exceed 3 : 1.

R 123: J. A. Haringx: On highly compressible helical springs and rubber rods, and their application for vibration-free mountings, VI (Philips Res. Rep. 4, 407-448, 1949, No. 6).

This paper (see Nos. **R 94, R 101, R 109, R 113** and **R 121**) deals with the actual construction of some vibration-free mountings, in particular a damped dynamic vibration absorber supported by helical springs and an elementary mounting without auxiliary mass supported by rubber rods. Their design is completely based upon the results found in the preceding papers, special attention being drawn to constructions approximately meeting the demands for the separation of the various degrees of freedom. In the introduction it is shown what interfering vibrations may be expected and what additional advantage is yielded by a large logarithmic decrement of the mounting as regards the free vibrations of an undamped instrument placed upon it. The characteristic differences between helical springs and rubber rods when used as resilient elements are mentioned, the latter requiring a much larger mass for the same low resonant frequency of the system. Further, a comprehensive treatment of the various resilient and damping elements is given in order to arrive at a number of formulae and graphs simplifying their design.

R 124: R. Loosjes and H. J. Vink: The conduction mechanism in oxide-coated cathodes (Philips Res. Rep. 4, 449-475, 1949, No. 6).

For the contents of this article see these abstracts, **No. 1808*** and Philips Techn. Rev. **11**, 275-281, 1949/50, No. 9.