VOLUME 22, 1960/61, No. 1 pp. 1-28 Published 4th November 1960

Philips Technical Review

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

# A MAGNETIC WHEEL STORE FOR RECORDING TELEVISION SIGNALS

by J. H. WESSELS.

621.397.6:621.395.625.3

*The magnetic recording of signals on the periphery of a wheel or drum, coated with a magnetic material, has been known for some considerable time: it is used, for example, to delay acoustic signals for producing reverberation effects (ambiophony), and for information storage in certain electronic computers. The article below describes a system by which television signals can be magnetically recorded on the rim of a wheel, and subsequently displayed as and when required. The system is likely to find an important application in radiology.*

## Video tape recording

The first point to be noted about the magnetic recording of television picture signals is that it involves frequencies very much higher than are encountered in sound recording. In the latter the top frequency is about 20 kc/s, whereas for television pictures (with 625 lines) it is necessary to go up to about 5000 kc/s.

A sinusoidal electrical signal is impressed on to the magnetic tape with a particular wavelength. Between this wavelength $\lambda$, the frequency $f$ and the writing speed $v_s$ (the speed of the tape relative to the recording head) the following relation exists:

$$v_s = \lambda f.$$

In order to keep down the writing speed the minimum wavelength $\lambda_{\min}$ must be chosen as short as possible. A lower limit is set to $\lambda_{\min}$ by the reproduction quality, which declines rapidly when $\lambda_{\min}$ drops below 5 μ. At $\lambda_{\min} = 5$ μ and $f = 5$ Mc/s the writing speed $v_s$ is 25 metres per second. If the recording head is stationary, as it always is in sound recording, the writing speed is the linear speed of the tape past the head. There are obvious objections to a tape speed of 25 m/sec. One method of video tape recording [1]) gets around this difficulty by making the magnetic head rotate, thus allowing the tape speed to be reduced considerably below the writing speed.

## Recording on a magnetic wheel

Where the object is not to record a whole series of television pictures (part of a television programme, for example) but only to "store" one or a few television frames in a memory device, the latter can be given a form that readily allows a high writing speed to be used. We refer to the form of a wheel, analogous to the acoustic delay wheel used to produce ambiophonic and other sound effects [2]). The periphery of the wheel is provided with a coating of magnetic material, and at a very short distance from it a recording or "writing" head is mounted which, for video purposes, also serves as the playback or "reading" head. Such a wheel, which we shall presently describe *in extenso*, can easily be given a peripheral speed of some scores of metres per second. This ensures a sufficiently faithful reproduction of the *high* frequencies.

The trouble here arises at the other end of the spectrum, at the *low* frequencies. For example, at $f = 100$ c/s and $v_s = 25$ m/sec, the wavelength is 25 cm. This is much longer than the length $l$ of the head, which is about 1 cm (*fig. 1*). The result is that the magnetic flux of the magnetized layer — in so far as it corresponds to the low frequencies — no

[1]) C. P. Ginsburg, Comprehensive description of the Ampex video tape recorder, J. Soc. Mot. Pict. Telev. Engrs. **66**, 177-182, 1957.

[2]) Philips tech. Rev. **17**, 259, 1955/56, and **20**, 325, 1958/59.

longer passes through the head but around it. Thus, as the signal frequency decreases, the output signal not only shows the familiar gradual drop of 6 dB per octave (because with increasing wavelength the contained magnetic flux varies more slowly with time) but also, when $\lambda$ is several times longer than $l$, it begins to fall very rapidly to zero.

The difficulty, then, is that in video recording the ratio of the highest to the lowest signal frequency is particularly large, being about $10^5$ — against $10^3$ in the case of sound. The difficulty can be circumvented, however, by recording instead of the video signal itself a "carrier wave" modulated by the video signal.

Before describing this system we should point out that direct (i.e. unmodulated) recording on a magnetic wheel does produce useful results for certain purposes. Visitors to the 1958 Photokina Exhibition at Cologne who accepted the invitation to be photographed in the Philips stand, saw themselves appear, at the moment the shot was taken and for some time after, on two television screens. A magnetic wheel store played an essential part in this stunt, as explained in the caption to *fig. 2* [3]).

In 1959 the magnetic wheel was demonstrated at the 9th Radiological Congress in Munich (see end of this article) and at the Radio and Television Show in Brussels. At the latter the visitors saw a moving picture of themselves on one television screen and a stationary picture on another. The first set was connected directly to a television camera, the second via a magnetic wheel.

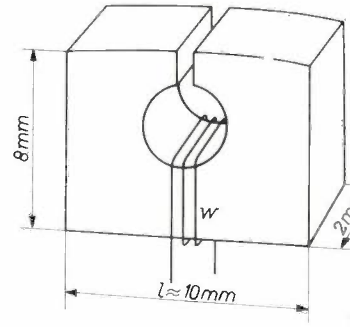


Fig. 1. Form and general dimensions of a write-read head, with schematically-represented winding *w*. The core is of ferroxcube IV. The magnetic layer travels in the direction of the length dimension *l*.

### Frequency modulation system

To avoid the difficulties involved in recording a signal having a frequency ratio of $10^5 : 1$, the signal can be made to modulate a carrier wave. As we shall see, the spectrum of the modulated carrier then has a much smaller frequency ratio.

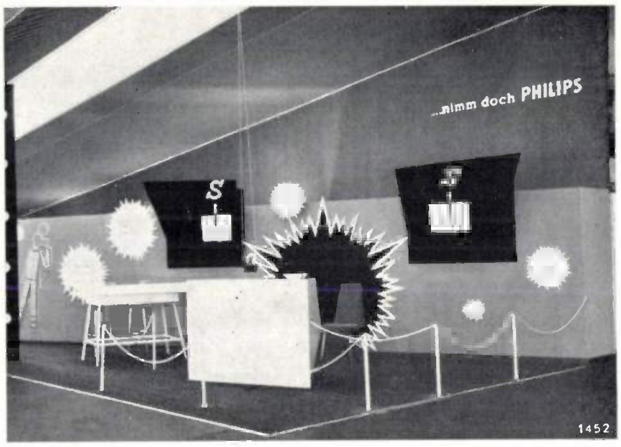


Fig. 2. Philips stand at the 1958 Photokina Exhibition in Cologne. Members of the public were invited into the stand to have a flash-photograph taken of themselves, the photograph to be posted on to them later. Immediately after the flash, a picture of the subject appeared simultaneously on two television screens *S*, and remained there for some time after the subject had left the stand.

How this television picture was produced was not disclosed, and must have mystified many. It was done as follows. When the subject took his seat in front of the photographer's camera, he was at the same time within the field of view of a concealed television camera. The light from the flash bulb made a phototransistor conductive, which in turn actuated a relay, leading successively to the erasure of the picture already stored on a concealed magnetic wheel, and to the recording and display of the new picture [3]).

Frequency modulation is the most appropriate system for the purpose [4]). In contrast to amplitude modulation, it is possible in this system to suppress modulation noise to a great extent, and moreover the noise present is so distributed over the spectrum as to cause much less interference in a television picture. It will be useful to deal at greater length with these two reasons for preferring frequency modulation.

The output signal of a magnetic recording is always modulated in amplitude by noise. This modulation noise is partly due to the fact that the distribution of the grains of the magnetic coating is not perfectly uniform [5]). Other causes are variations in the thickness of the coating, dust particles between the coating and the head, and — in the present case— imperfect roundness of the wheel. All these imperfections, then, give rise to undesired amplitude modulation. If the video signal were also present as amplitude modulation on the

[3]) The credit for this idea, and for part of its technical realization, goes to J. F. van Oort of the Philips Exhibition Department.

[4]) C. E. Anderson, The modulation system of the Ampex video tape recorder, J. Soc. Mot. Pict. Televׅ. Engrs. **66**, 182-184, 1957.

[5]) See e.g. D. A. Snel, Magnetic sound recording, Philips Technical Library 1959.

carrier, it would not be possible to separate these two modulations from one another. When frequency modulation is used, however, nearly all the amplitude modulation can be removed by means of a limiter, leaving an almost purely frequency-modulated signal and eliminating the above-mentioned noise contributions.

The second advantage of frequency modulation again relates to noise. In *fig. 3a* the centre frequency of the frequency-modulated carrier is denoted by $f_c$, whilst $\Delta f$ denotes a small band of the noise spectrum about an arbitrary frequency $f_n$. After detection, this band comes within the video spectrum, as shown in fig. 3*b*, i.e. around the video frequency $f_c—f_n$. Now, the noise power $\Delta P_n$ in this band is proportional to $(f_c - f_n)^2$ (in contrast to amplitude modulation, where $\Delta P_n$ is independent of $f_c—f_n$). This proportionality is a favourable circumstance, experiments having shown that noise in a television picture is more troublesome the lower are the frequencies of the noise components for the same $\Delta P_n$ [6]).

The frequency-modulation system should be arranged such that the ratio of the highest to the lowest frequency in the signal to be recorded is very much smaller than in the video signal, without the highest frequency appreciably exceeding the highest video frequency. These requirements are fulfilled if the instantaneous frequency of the recorded signal is, say, 7 Mc/s for the brightest white in the picture, 5.5 Mc/s for black, and 5 Mc/s for the peak of the synchronizing signals ("blacker than black"); see *fig. 4*.

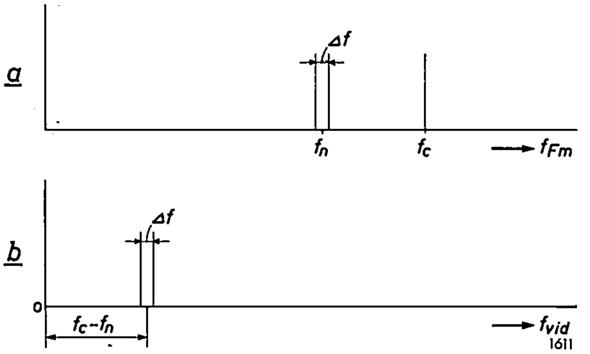


Fig. 3. *a*) Of the noise spectrum of a frequency-modulated signal, with centre frequency $f_c$, the diagram shows a narrow band $\Delta f$ about an arbitrary frequency $f_n$. After detection, this band appears in the video spectrum as illustrated in (*b*). The noise contribution of $\Delta f$ is proportional to $(f_c—f_n)^2$, and is hence smaller for low video frequencies than for high. As a consequence, the noise in the picture is less troublesome.

What the frequency spectrum of a sinusoidally frequency-modulated signal will look like depends to a large extent, of course, on the modulation index *m*, defined as the ratio of the frequency deviation $\Delta f$ to the modulation frequency [7]). As a rough approximation, let us assume that the video signal is sinusoidal; given 5 and 7 Mc/s as the extreme values of the instantaneous frequency the centre frequency will then be 6 Mc/s and the frequency deviation 1 Mc/s. If *m* is greater than about 25 (i.e. in our case $f_{vid} < 40$ kc/s) the spectrum will then consist mainly of numerous weak lines within the extremes 5 and 7 Mc/s (*fig. 5a*). For $m = 3$ ($f_{vid} = \frac{1}{3}$ Mc/s) the lines are much farther apart, but those beyond 5 and 7 Mc/s are of little consequence (fig. 5*b*). At $m = 0.2$, however, corresponding to the highest video frequency $f_{vid} = 5$ Mc/s, the spectrum is made up virtually of only three lines, at 6, 6 + 5 and 6 — 5 Mc/s (fig. 5*c*), two of which are thus far outside the 5 and 7 Mc/s limits.

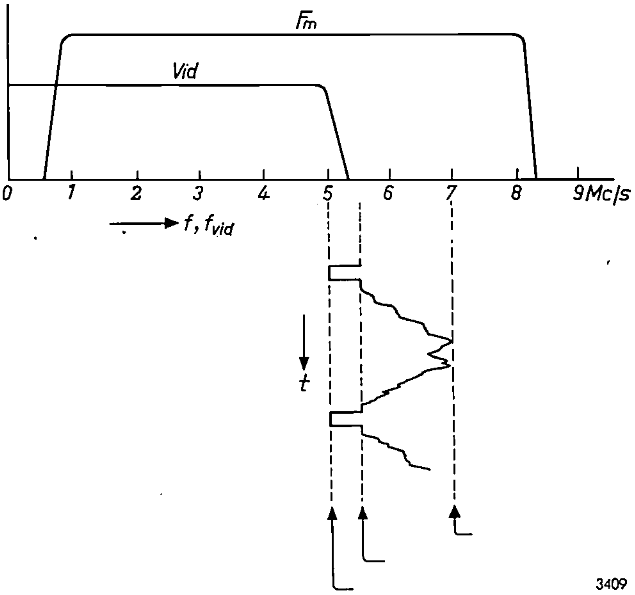


Fig. 4. *Vid* frequency range of the video signal for a television picture with 625 lines. *Fm* frequency range of the signal to be recorded, whose instantaneous frequency is 7 Mc/s in the brightest white, 5.5 Mc/s in the black, and 5 Mc/s in the "blacker-than-black" (peak of sync signals).

The latter does not mean that, for recording, we must reckon with 6 + 5 = 11 Mc/s as the highest frequency. It is sufficient if we take the lower sideband plus part of the upper sideband, up to about 8 Mc/s (curve *Fm* in fig. 4). The lower limit can

[6]) Amongst the extensive literature on the noise nuisance in television pictures, mention may be made of: L. Goussot, Le brouillage des images de télévision par les signaux parasites, Onde électr. **39**, 352-361 and 690-700, 1959 (No. 386 and No. 388/389). This also quotes references to other articles on the subject.

[7]) See e.g. Th. J. Weijers, Frequency modulation, Philips tech. Rev. **8**, 42-50, 1946, in particular fig. 4.

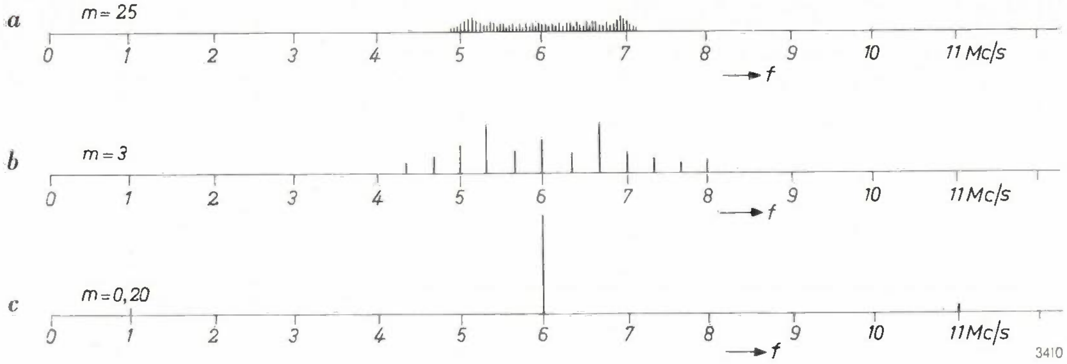


Fig. 5. Frequency spectra of a sinusoidally frequency-modulated signal, with centre frequency 6 Mc/s, frequency deviation 1 Mc/s and modulation index $m = 25$, 3 and 0.20, respectively.

usefully be chosen between 0.5 and 1 Mc/s. The frequency ratio of the modulated signal is then of the order of 10 : 1, instead of the ratio of $10^5$ : 1 for the video signal. Expressed in octaves, the frequency range to be recorded is thus reduced from 17 to 3 or 4 octaves. True, the highest frequency for recording is now 8 Mc/s instead of 5 Mc/s, but this presents no major difficulties for recording on a wheel store.

The quality at present achieved with the aid of frequency modulation appears from *fig. 6*, which shows a photograph of a resolution chart obtained on a monitor via a magnetic wheel. Further improvements in quality may be expected in the near future.

### Mechanical features of the magnetic wheel

To keep the wheel store as small as possible, it was decided in the design stage to record one television frame on the periphery; one frame lasts $^1/_{50}$ second, hence the wheel must turn at 3000 revolutions per minute.

As we have seen, the highest frequency in the signal to be recorded is 8 Mc/s. The shortest useful

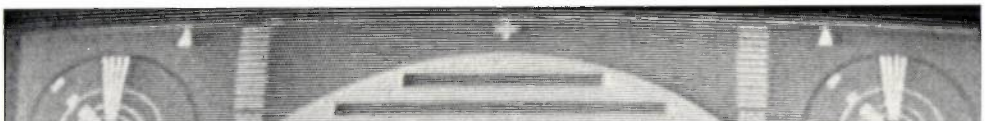


Fig. 6. Picture of a resolution chart picked up by a television camera, stored on a magnetic wheel and displayed on a monitor. For this recording, use was made of frequency modulation as in fig. 4. The raster consisted of over 300 lines.

wavelength was taken to be 8 μ. It therefore follows from $v_s = \lambda f$ that the peripheral speed of the wheel must be 64 m/sec. To achieve this at 3000 r.p.m. the wheel must have a diameter of 40 cm.

One of the wheels used in numerous experiments is shown in *fig. 7*. Two heads are disposed around the rim: an erasing head and a head serving alternately for writing and reading. Neither of the heads must touch the wheel, otherwise rapid wear of the heads and of the magnetic coating would result. Provided the erasing current is strong enough, the distance *d* between erasing-head and rim need not be extremely small. The distance between the write-read head and the rim is critical, however, and must not exceed about 1 μ. In the reading process the following relation exists between the attenuation *a* of the signal, the distance *d* and the wavelength $\lambda$ [8]):

$$a = 55 \frac{d}{\lambda} \text{ dB}.$$

For $\lambda = 8$ μ and $d = 1$ μ, *a* is as much as 7 dB. The writing process, which is difficult to express in a formula, is even more critical in this respect.

The mechanical construction therefore calls for the highest precision. As can be seen in fig. 7, the rim of the wheel has the form of a flange, 30 mm in width; without this the wheel would suffer too much deformation at high speeds. The magnetic layer is applied over the whole width of the flange, thus providing space for numerous tracks side by side. The wheel is mounted on a thick shaft (to withstand bending moments) and is machined in its bearings (journal bearings of exceptionally high quality) to within a tolerance of better than 1 μ [9]).

Since temperature variations may easily cause a change of a few microns in the radius, the write-read head cannot be rigidly mounted in a fixed position, otherwise, with a change in temperature, it would either touch the wheel or be too far away from it. In the construction chosen the head is arranged to ride on an air cushion when the wheel is turning at the right speed. The head *2* (*fig. 8*) is attached to an arm *6* which pivots about the point *7*. The air dragged round with the wheel pushes the head outwards against the pressure of a spring *8*, thus ensuring that a certain distance is maintained between the head and the turning wheel.

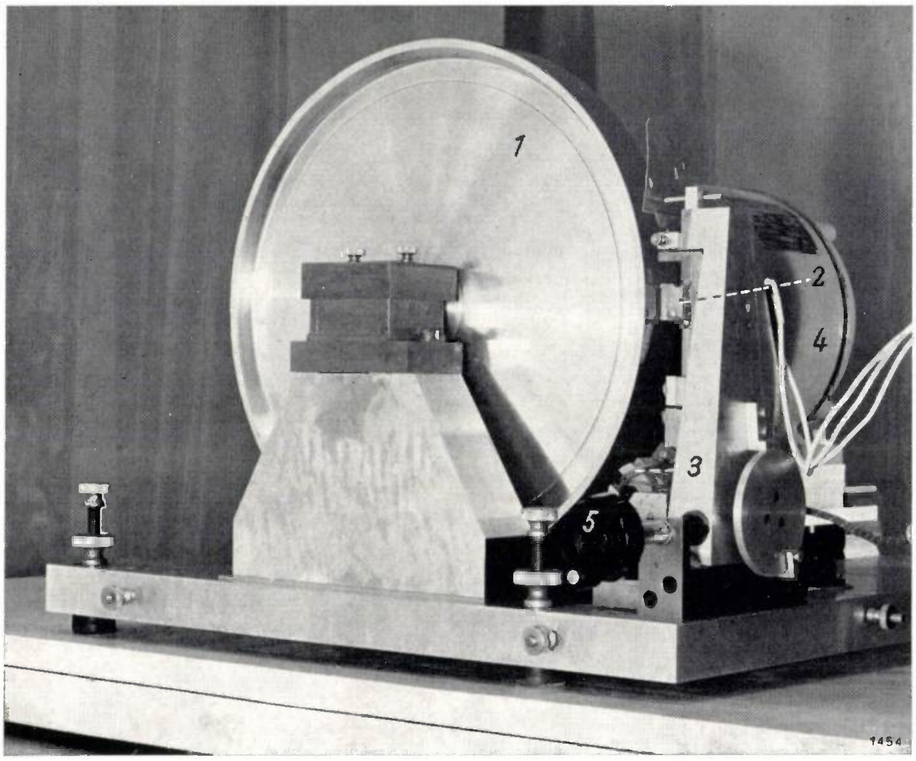


Fig. 7. *1* magnetic wheel. *2* write-read head. *3* erasing head. *4* motor. *5* knob for axially displacing the heads (the 30 mm wide rim flange can accommodate a large number of tracks side by side).

[8]) See e.g. H. G. M. Spratt, Magnetic tape recording, Heywood, London 1958, p. 84.

[9]) This precision work was done very skilfully by L. M. Leblans of this laboratory.

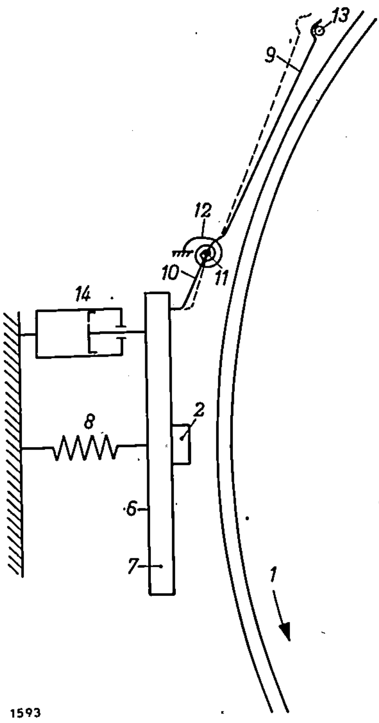


Fig. 8. Illustrating the method of maintaining a constant distance of about 1 μ between the magnetic wheel *1* and the writing-reading head *2*. *6* arm with pivot *7*. *8* compression spring. *9* vane, with arm *10* and pivot *11*. When the wheel is stationary the coil spring *12* forces the vane against the stop *13*; the arm *10* then prevents the spring *8* from pressing the head against the wheel. When the wheel is turning at full speed the air stream around the wheel pushes the vane *9* outwards (to position of dashed line) against the action of the coil spring *12*; arm *10* then releases arm *6*, and spring *8* pushes the head towards the wheel. The air film dragged round by the wheel forms an air cushion on which the head rides at a constant distance of about 1 μ from the wheel. *14* oil damping.

An optical test, made by passing light through the gap, demonstrated with a high degree of probability that the distance *d* remains roughly 1 μ. Oil damping *14* protects the head from vibrations, without preventing it from following slow changes due to temperature variations.

Special measures are needed to prevent the spring *8* from pressing the head against the wheel when the wheel is stationary, not turning fast enough or slowing down to a standstill[10]). For this purpose, use is made of a vane *9* connected to an arm *10*, pivoting about point *11* (see fig. 8). If the wheel stops or turns too slowly, a coil spring *12* holds the vane *9* against a stop *13*, and the arm *10* prevents arm *6* from moving the head into contact with the wheel. When the wheel comes up to full speed, the dragged air current pushes the vane *9* back and arm *10* releases arm *6* (dashed line in figure), thus restoring the above-mentioned equilibrium between the force exerted on the head *2* by spring *8* and that exerted by the air current.

## Circuit for writing and reading a single frame

To write a single frame on the wheel, the writing head must be supplied with the frequency-modulated television signal for the duration of one frame ($^1/_{50}$ sec). This calls for a circuit that will keep the recording amplifier preceding the write-read head normally blocked but will unblock it during the first frame to begin after the depression of a push-button.

The block diagram of this arrangement is shown in *fig. 9*. $A_2$ is the recording-head amplifier. $F_1$ and $F_2$ are flip-flops, each with two stable states, *I* and *II*. Both receive at their inputs, *1*, a continuous train of (positive) frame-synchronizing pulses — sync pulses — which are separated in the usual way from the video signal in the circuit *S*. These sync pulses, represented in *fig. 10a*, keep the two flip-flops in state *I*, whilst $F_2$ delivers a biasing voltage which blocks the amplifier $A_2$. When the button *B* is depressed the discharge of capacitor *C* causes a pulse (fig. 10*b*) to appear at the input *2* of $F_1$; this pulse brings $F_2$ into state *II*. The first sync pulse now to arrive returns $F_1$ to state *I* (fig. 10*c*) which, via the coupling between $F_1$ and input *2* of $F_2$, has

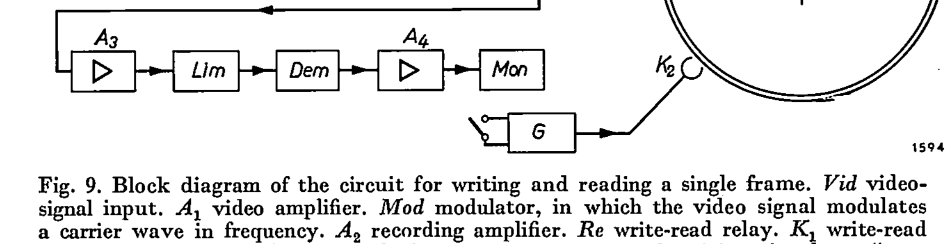


Fig. 9. Block diagram of the circuit for writing and reading a single frame. *Vid* video-signal input. $A_1$ video amplifier. *Mod* modulator, in which the video signal modulates a carrier wave in frequency. $A_2$ recording amplifier. *Re* write-read relay. $K_1$ write-read head. *W* wheel store. *S* circuit which separates picture-synchronizing signals — "sync pulses" — from video signal. $F_1$ and $F_2$ bistable flip-flops, with inputs *1* and *2*. *B* push-button, which, by discharging capacitor *C*, produces the "display" pulse. $A_3$ read amplifier. *Lim* limiter. *Dem* demodulator (frequency detector). $A_4$ video amplifier. *Mon* monitor. $K_2$ erasing head. *G* erasing-current generator.

[10]) The solution found for this problem was devised by J. F. van Oort of the Philips Exhibition Department.

the effect of causing $F_2$ to change from *I* to *II* (fig. 10*d*); the next sync pulse to arrive returns $F_2$ to state *I*. During exactly one frame, then, $F_2$ is in state *II*, and the recording amplifier is opened. This single frame, the first to follow the depression of the button, is therefore recorded on the wheel.

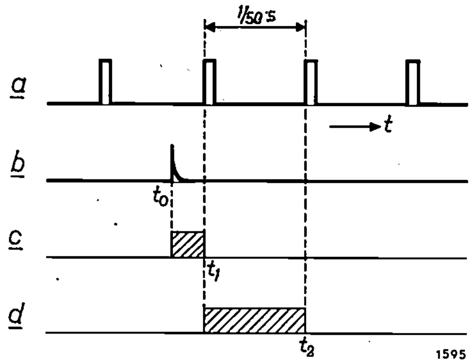


Fig. 10. *a*) Pulse train at the inputs *1* of flip-flops $F_1$ and $F_2$ in fig. 9. *b*) "Display" pulse at input *2* of $F_1$, produced when button *B* is depressed. *c*) During the interval $t_0$-$t_1$, $F_1$ is in state *II*. *d*) During the interval $t_1$-$t_2$ (= duration of one frame) $F_2$ is in state *II*. The recording amplifier $A_2$ (fig. 9) is then unblocked and the picture is recorded on the wheel.

## Synchronizing the wheel with the video signal

An important application of the magnetic wheel store is that in which the recorded picture originates from a television camera at the same location. This is the case in the radiological application presently to be discussed. The frame frequency of the camera and the monitor, and the speed of revolution of the wheel, can then be governed by the frequency of the local electricity mains. As far as the wheel is concerned, this amounts to the use of a synchronous motor.

Cases also arise, however, where the frame frequency of the television picture to be recorded is not exactly equal to that of the local mains. Steps must then be taken to synchronize the wheel with the given frame frequency. The system which we have devised for this purpose consists of a flexible eddy-current coupling between the (non-synchronous) motor and the wheel, in combination with an electrical control system for governing the speed of revolution of the wheel.

### *Eddy-current coupling*

*Fig. 11* shows two cross-sections of the eddy-current coupling, which consists of the following components:

*a*) a flanged aluminium pulley *1*, at the rim of which is fitted a hollow steel cylinder *2* with copper lining *3*;

*b*) a sectored rotor *4*, keyed to the shaft *5* that drives the magnetic wheel;

*c*) a steel housing *6* containing a field coil *7*.

In relation to the pulley *1* and the housing *6*, the shaft *5* can rotate freely in ball-bearings. The pulley is motor-driven by means of a belt; the housing and the field coil are stationary. A variable direct current is passed through the field coil. This excitation current produces a magnetic flux through the components *6*, *4* and *2*, which concentrates in the sectors of *4*. When the cylinder *2* rotates and the rotor *4* is still stationary, eddy-currents are induced in the copper lining *3*. As a result, a torque is exerted on the rotor, causing the rotor — and hence the magnetic wheel — to rotate in the same sense as the pulley, but with a smaller angular velocity. When the angular-velocity difference $\omega_1 - \omega_2$ is small, the torque is proportional to $\omega_1 - \omega_2$ (*fig. 12*); for large differences in angular velocity, a phase shift exists between the pulsating magnetic field produced by the sectors and the induced eddy currents, causing the curve to bend over and a maximum to appear. When the wheel is started up, $\omega_2$ is initially zero, hence $\omega_1 - \omega_2 = \omega_1$; if the maximum lies approximately at this value, the starting torque is high and the wheel quickly reaches full speed.

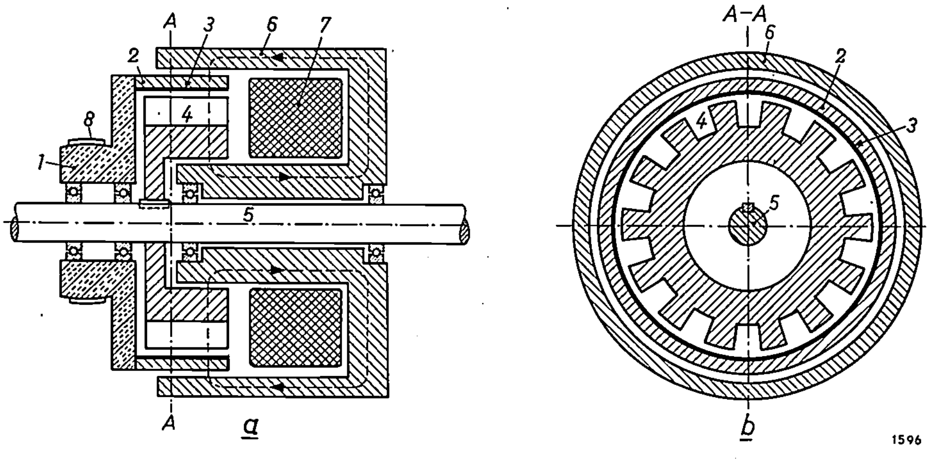


Fig. 11. Eddy-current coupling. *a*) axial cross-section, *b*) transverse section through *A*-*A* in (*a*). *1* flanged aluminium pulley with steel cylinder *2* and copper lining *3*. *4* sectored rotor, keyed to shaft *5* which, via a flexible coupling, drives the shaft (not shown) of the magnetic wheel. *6* steel housing. *7* field coil. *8* driving belt.

### *The control system*

Synchronism between the speed of revolution of the wheel and the frame frequency of the video signal is obtained by causing any deviation from synchronism to react on the excitation current of the coupling. This is done by comparing the phase of two pulse trains: the sync pulses and wheel pulses.

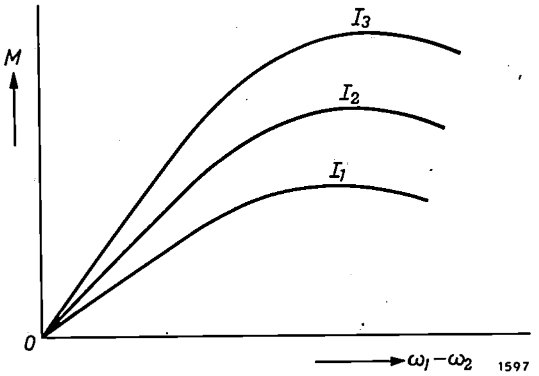


Fig. 12. Torque $M$, transmitted by eddy-current coupling, as a function of the angular-velocity difference $\omega_1$—$\omega_2$ between the pulley *1* and the shaft *5* (in fig. 11), for three values of exciting current ($I_3 > I_2 > I_1$).

As mentioned above, the sync pulses are derived from the video signal; the wheel pulses are induced in an appropriate pick-up head by a small magnet of ferroxdure fixed to the wheel. Any phase difference arising between the two pulse trains changes the excitation current, via a special circuit, in such a way as to make the phase difference smaller.

The circuit diagram of the control system is shown in *fig. 13*. The triodes $T_1$ and $T_2$ form part of a bistable flip-flop with a common cathode resistor $R_2$. The grid of the cathode follower $T_3$ is coupled to the anode of $T_2$. When the flip-flop is in state *1* ($T_1$ conducting, $T_2$ cut-off) the anode potential of $T_2$ is high and so too, therefore, is the grid potential of $T_3$ with respect to earth; $T_3$ therefore passes current, and thus the potential $v_{\mathrm{k}}$ across the cathode resistor $R_3$ is high ($v_{\mathrm{k}} = V_1$). In state *2*, on the other hand, $v_{\mathrm{k}}$ is low ($= V_2 < V_1$). When the potential $v_{C1}$ across $C_1$ drops below a certain critical value $V_{1,2}$, the flip-flop changes from state *1* to state *2*. For the change from *2* to *1*, $v_{C1}$ must exceed another critical value, $V_{2,1}$. The levels of $V_{1,2}$ and $V_{2,1}$ in relation to $V_1$ and $V_2$ are indicated in *fig. 14*.

Roughly, the circuit functions as follows. When synchronism is approximately achieved the flip-flop is brought into state *1* by every (positive) wheel pulse, and into state *2* by every (negative) sync pulse. The potential $v_{\mathrm{k}}$ of the cathode $K$ is thus alternately high and low, remaining longer high the longer state *1* lasts, i.e. the more time it takes before a wheel pulse follows a sync pulse.

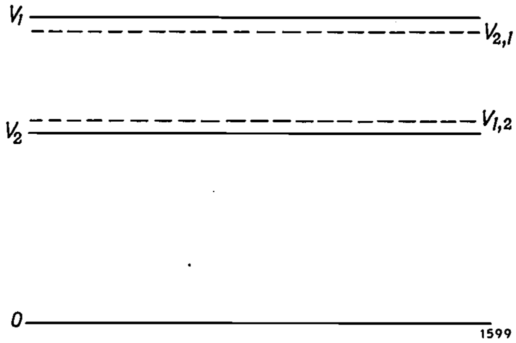


Fig. 14. $V_1$ potential of point $K$ (fig. 13) in stable state *1*; $V_2$ *idem* in state *2*. $V_{1,2}$ value which $v_{C1}$ must reach to cause state *1* to change to state *2*; $V_{2,1}$ *idem* for transition from *2* to *1*.

The high voltage $V_1$ of $K$ appears each time across the capacitor $C_2$ via the diode $D_1$. This capacitor discharges gradually through the resistor $R_4$. The longer the interval $\tau$ between a sync pulse and a wheel pulse, the lower is the final value to which the voltage $v_{C2}$ across $C_2$ decreases. This final value is taken over via an "overflow diode" $D_2$ by capacitor $C_3$ and determines the current that flows through the triode $T_4$. Between each two decreases in $v_{C2}$ the voltage $v_{C3}$ across $C_3$ is

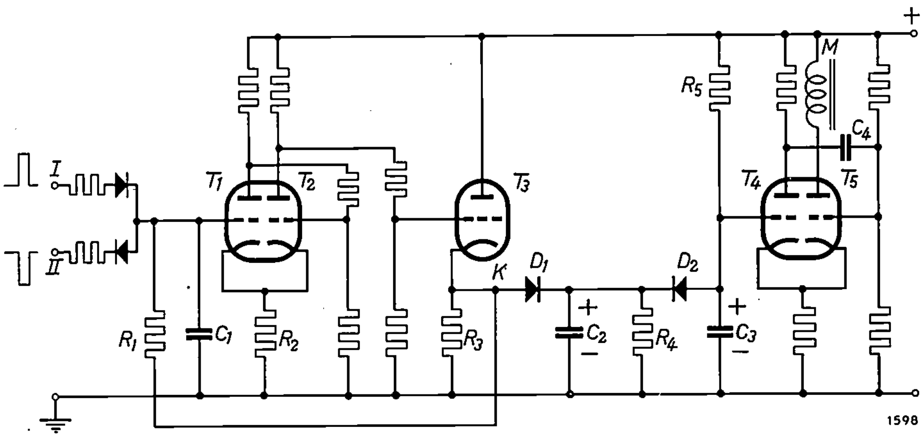


Fig. 13. Circuit diagram of control system. *I* input for (positive) wheel pulses. *II* input for (negative) sync pulses. $M$ field coil of eddy-current coupling (*7* in fig. 11). If the wheel turns too slowly, the circuit energizes the electromagnet more strongly, and if the wheel turns too fast, less strongly. For explanation, see text.

able to rise again slightly as $C_3$ is charged up via $R_5$.

Let us assume that the wheel is turning a little too slowly. The interval $\tau$ will then show a tendency to increase. The potential $v_{C3}$ therefore drops, the current through $T_4$ decreases, and, since $T_4$ and $T_5$ have a common cathode resistor, the decrease in the the current through $T_4$ causes an increase in the current through $T_5$. The latter current energizes the electromagnet in the eddy-current coupling. In the case under consideration, then, the excitation current rises and so, too, does the torque exerted on the rotor, causing the wheel to turn faster. Conversely, if the wheel starts to turn too fast, the excitation current is reduced, resulting in a drop in speed.

The capacitor $C_4$ between the anode of $T_4$ and the grid of $T_5$ influences the frequency response of the system in such a way as to prevent instability occurring.

A difficulty in phase control systems is often the starting-up process, when the phase relation between the two pulse trains is completely irregular. The usual practice is to switch-off the control system before starting-up, and to work with maximum torque until an electrical tachometer shows that the right speed has been reached; only then is the control system put into operation. This precaution is unnecessary with the arrangement in fig. 13; the control system can be switched on right from the beginning and will always ensure that synchronism is reached.

Each pulse from the wheel causes the voltage $v_{C1}$ to jump by $V_1 - V_2$, and each sync pulse causes an equal downward drop. When the wheel is started up, the wheel pulses occur at first much more slowly than the sync pulses. The waveform of $v_{C1}$ is then as shown in *fig. 15a*. The flip-flop is in state *2* ($T_1$ cut-off, $T_2$ conducting) and remains for a while in that state, because $v_{C1}$ cannot reach the critical value $V_{2,1}$ as long as the wheel is turning much too slowly. In this state, $v_k$ $v_{C2}$ and $v_{C3}$ have the low value $V_2$, which corresponds to strong excitation of the eddy-current coupling.

When the wheel approaches the correct speed of revolution the flip-flop will change its state now and then, but only when a wheel pulse is followed very quickly by a sync pulse, as in the intervals $t_1$-$t_2$ and $t_3$-$t_4$ in fig. 15*b*. During these intervals state *1* predominates, and $v_k$ and $v_{C2}$ have the high value $V_1$. When $v_k$ jumps back to the low value $V_2$, capacitor $C_2$ begins to discharge through $R_4$ and therefore $v_{C2}$ starts gradually to drop. In the case shown in fig. 15*b*, $v_{C2}$ again reaches the value $V_2$, so that $v_{C3}$ — which follows the lowest values of $v_{C2}$ — still retains the value $V_2$ and thus the eddy-current coupling remains strongly energized.

The wheel now begins to turn rather too fast. Consequently the wheel pulses become somewhat more frequent than the sync pulses (fig. 15c) and a state is soon reached where $v_{C1}$ and $v_k$ jump continuously to and fro between the values $V_1$ and $V_2$. The voltage $v_{C2}$ continually jumps up again before it can drop to $V_2$, and $v_{C3}$ follows the successively somewhat

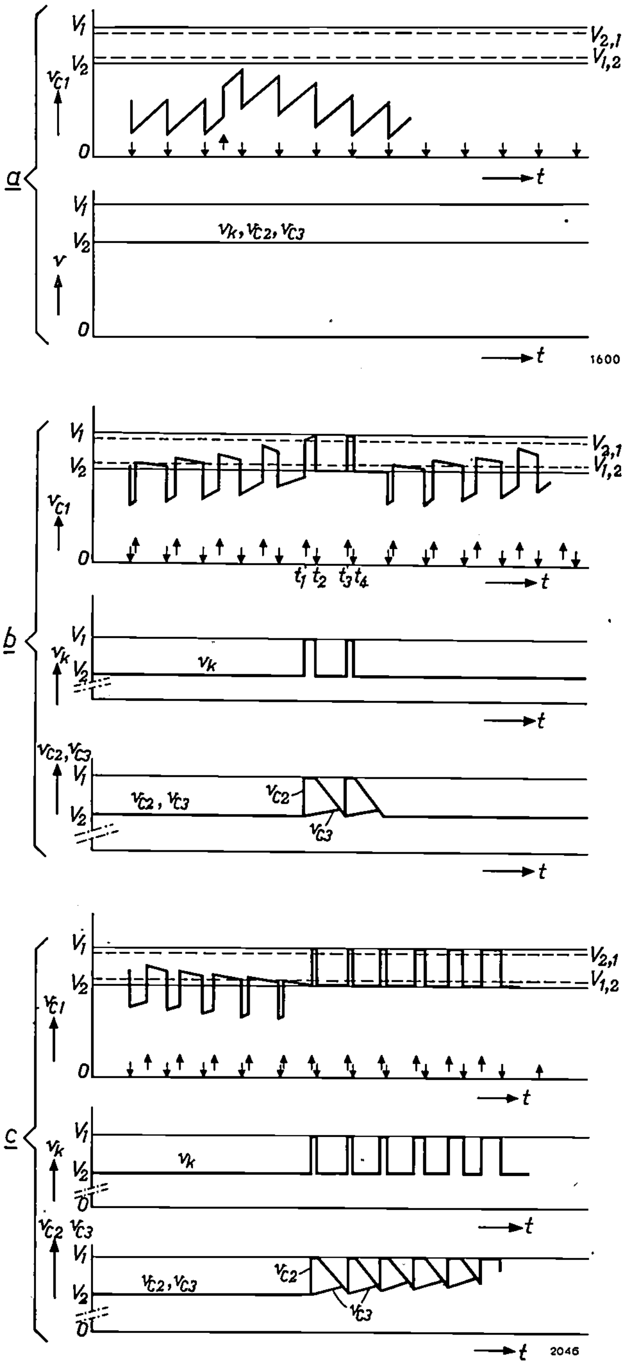


Fig. 15. Waveform of voltages $v_{C1}$, $v_k$, $v_{C2}$ and $v_{C3}$ in the control circuit (fig. 13). The arrows ↓ denote sync pulses, arrows ↑ wheel pulses.

*a*) The wheel has just been started and is turning much too slowly. The wheel pulses are consequently much less frequent than the sync pulses. $v_{C1}$ does not attain the critical value $V_{2,1}$. The flip-flop $T_1$-$T_2$ therefore remains in state *2*; $v_k$, $v_{C2}$ and $v_{C3}$ are constant ($= V_2$), and the eddy-current coupling is strongly energized.

*b*) The wheel still turns rather too slowly. At $t = t_1$, $v_{C1}$ reaches the critical value $V_{2,1}$, so that $T_1$-$T_2$ changes to state *1*, but the sync pulse at $t_2$ restores state *2*. The same happens at $t_3$-$t_4$. From $t_1$ to $t_2$ and from $t_3$ to $t_4$, $v_k = V_1$ and $v_{C1} = V_1$, but $v_{C3}$ remains unchanged, and the coupling stays strongly energized.

c) The wheel begins to turn rather too fast, and locks back into synchronism. $v_{C1}$ and $v_k$ jump to and fro between $V_2$ and $V_1$; $v_{C2}$ no longer drops to the level $V_2$; $v_{C3}$ gradually rises and the coupling becomes less strongly energized, until the steady state is reached at synchronism.

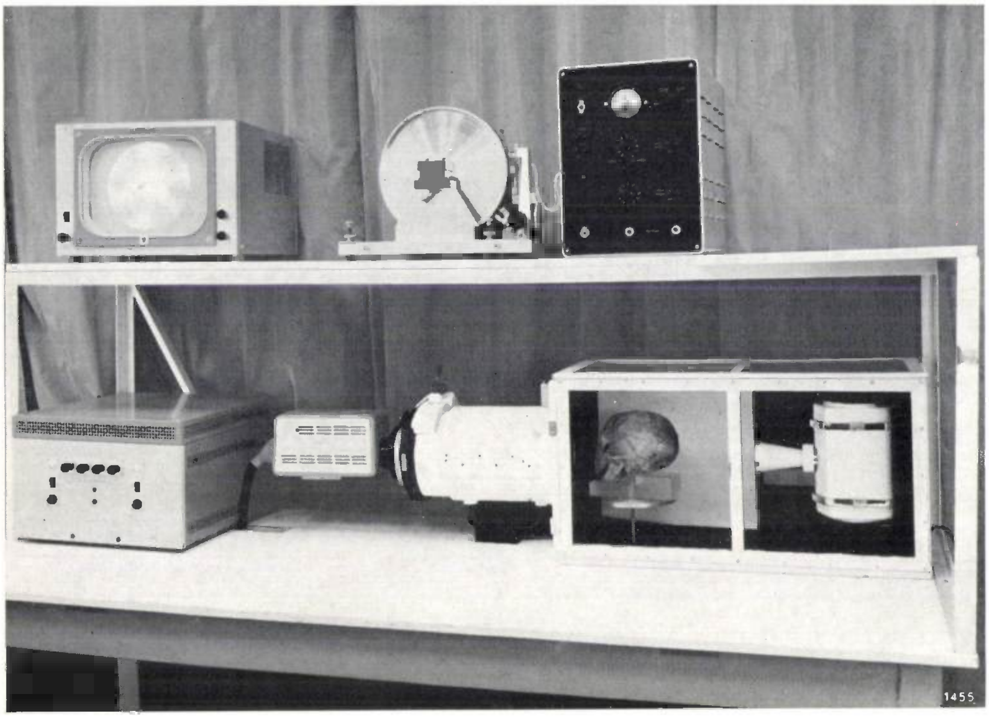


Fig. 16. Equipment for demonstrating the radiological application of the magnetic wheel store (as yet without frequency modulation). From right to left, below: X-ray tube, subject (a skull), X-ray image intensifier, television camera, auxiliary apparatus for television system; above: amplifying apparatus for recording and reproduction, magnetic wheel, monitor.

higher minima of $v_{C_2}$. (The charge needed to effect this increase in $v_{C_3}$ is supplied to $C_3$ through $R_5$.) As $v_{C_3}$ rises it increases the current through $T_4$ and decreases the current through $T_5$. The eddy-current coupling is therefore less strongly energized, the wheel turns more slowly and enters into synchronism.

Another virtue of the system is its relative insensitivity to changes in the frequency of the sync pulses; this is a consequence of the fact that a speed measurement (with an electrical tachometer) is not necessary.

## Application in radiology

In the system described, as in all other magnetic recording systems, the recording can be preserved indefinitely or it can be erased; after erasure the magnetic layer can be immediately used again for a fresh recording. Since there is no contact between head and wheel, the recording can be reproduced as often as required. As we have seen, the picture quality is highly satisfactory when use is made of frequency modulation.

These features have led to a promising application fo the magnetic wheel store in radiology [11]). A picture of the image on an X-ray screen is taken with a television camera and one frame of the picture is recorded on the magnetic wheel. The X-ray image can then be displayed immediately on a monitor tube and viewed by the doctor as long and as often as necessary. In this way no time is lost in developing an X-ray film, and precious minutes can be saved during a surgical operation. The X-ray dose required is extremely low — very much lower than in a fluoroscopic examination of a few seconds, and, if an X-ray image intensifier is used, even lower than that needed for a radiograph. A further important advantage is that an unsatisfactory exposure can immediately be retaken.

*Fig. 16* shows a photograph of the equipment used to demonstrate this application (as yet without frequency modulation) at the International Radiological Congress at Munich.

Summary. A magnetic wheel store is described on which a single television frame can be recorded. The picture quality obtained by direct recording of the video signal is quite serviceable for some purposes, in spite of the wide frequency range (50 c/s to 5 Mc/s). Considerable improvement results if the video signal is made to frequency-modulate a carrier and this latter recorded. The signal then recorded can have e.g. a frequency of 7 Mc/s in the white, 5.5 Mc/s in the black, and 5 Mc/s at the peaks of the sync signals, giving a frequency range from about 0.5 Mc/s to 8 Mc/s. In one experimental version the minimum wavelength impressed on the wheel is 8 μ at 8 Mc/s, the peripheral speed is 64 m/sec, the speed of revolution is 3000 r.p.m. and the wheel has a diameter of 40 cm. The rim of the wheel is 30 mm wide and can accommodate numerous tracks side by side. A device is described in which the air film dragged round with the wheel keeps the write-read head at a distance of about 1 μ from the magnetic coating of the wheel, without any contact. Also discussed are a circuit for writing and reading a single frame and a system of synchronizing the wheel with the video signal. Finally, mention is made of a promising application in radiology; advantages over photography are that it dispenses with the need for developing X-ray films, minimizes the X-ray dose and allows the immediate retake of unsatisfactory exposures.

[11]) Th. G. Schut and W. J. Oosterkamp, The application of electronic memories in radiology, Medicamundi 5, 85-88, 1959 (No. 3/4); Th. G. Schut and W. J. Oosterkamp, Die Anwendung elektronischer Gedächtnisse in der Radiologie, Elektron. Rdsch. 14, 19-20, 1960 (No. 1).

# RESONANCE ISOLATORS FOR MILLIMETRE WAVES

by H. G. BELJERS.

621.372.852.223:621.318.134

In microwave equipment frequent use is made nowadays of non-reciprocal devices. Principal among these is the directional isolator, a device that passes waves in the one direction without significantly attenuating them, and attenuates them very strongly in the other [1]).

The so-called resonance isolator makes use of the presence in a rectangular waveguide of a rotating magnetic field at certain places. If a suitable magnetic material is fixed at these places, gyromagnetic resonance occurs in the one direction of propagation and not in the other. Since gyromagnetic resonance — often referred to briefly as magnetic resonance — is attended by losses, the waves propagated in the first-mentioned direction are strongly attenuated. By using material free or almost free of other kinds of losses, it is possible in this way to make a directional isolator. The millimetre-band types discussed in this article are all resonance isolators.

The operation of other kinds of directional isolator depends on the Faraday effect or on so-called field-displacement. These can also be employed in the millimetre bands. Those based on the Faraday effect involve a round section of waveguide and it is necessary to fit transition sections if they are to be incorporated in a system using rectangular waveguides.

Gyromagnetic resonance occurs when a static magnetic field is applied perpendicular both to the direction of propagation and to the magnetic field of the microwaves. The required field $H$ is roughly proportional to the frequency $f$ of the microwaves. Expressing $H$ in A/m and $f$ in Mc/s, we can write:

$$H \approx f/0.035 \, . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

The exact value of $H$ is also governed by the demagnetization, in other words by the shape of the piece of material. This appears from Kittel's formula:

$$f = 0.035 \sqrt{H + M(N_x - N_y)} \times \sqrt{H + M(N_z - N_y)} , \quad . \quad . \quad . \quad (2)$$

where $M$ is the saturation magnetization and $N_x$, $N_y$ and $N_z$ are the demagnetizing factors. (The direction of propagation is the $z$ direction, and the magnetic field is parallel to the $y$ direction.) As can be seen, formula (2) is equivalent to (1) if $N_x = N_y = N_z$, that is if the magnetic material is spherical in shape.

A further consequence of demagnetization is that, to achieve minimum damping in the forward direction, the microwave magnetic field should as a rule be elliptically and not, as might be thought at first sight, circularly polarized. We shall return to this point presently.

From formula (1) we may deduce that to make resonance isolators for wavelengths in the 8.6 and 4.3 mm bands (frequencies of 35 Gc/s and 70 Gc/s), now coming increasingly into use, we should need magnetic fields of about $10^6$ A/m (12 500 oersteds) and $2 \times 10^6$ A/m (25 000 oersteds) respectively. The properties of the materials at present available for permanent magnets do not, however, allow of generating fields as high as $2 \times 10^6$ A/m, and although a field of $10^6$ A/m is possible, it entails an unmanageably large and heavy magnet.

Nevertheless, a much weaker external field may be used, or it may be dispensed with altogether, if the resonance is produced in a hard magnetic material like a crystal-oriented anisotropic ferrite, that is a ferrite in which the preferred directions of magnetization of the crystallites are aligned parallel to each other. The anisotropic ferrites used in this case all possess hexagonal crystal structure, with the c-axis as the preferred direction of magnetization. The electron spins take up their preferred alignment parallel to this axis. It costs a great deal more energy to magnetize such material in a direction other than that of the hexagonal axis. The stiffness with which the spin orientation is bound to this preferred direction is expressed in terms of the magnetic field — the anisotropy field — that would have to be applied to an isotropic material in order to bind the spins with the same stiffness to the direction of that field. It follows from this definition that the field $H$, which, according to eq. (1), gives rise to magnetic resonance at a certain frequency, is simply equal to the sum of the anisotropy field $H_a$ and the external field $H_u$.

Plainly, then, the external field can be dispensed with entirely if $H_a$ has exactly the required value. Although this is attractive from the design point of view, it is not without its disadvantages. In the first place, if there is no external field the material

[1]) H. G. Beljers, The application of ferroxcube in unidirectional waveguides and its bearing on the principle of reciprocity, Philips tech. Rev. 18, 158-166, 1956/57.

is not always completely saturated, that is a greater number of Weiss domains have a "wrong" orientation, and as a result the damping of the microwaves in the forward direction is increased, which is obviously undesirable. In the second place, unless precautions are taken, the operation of such an isolator can be ruined by an interfering external magnetic field, which reduces the magnetization or may even cause it to disappear altogether if the interfering field is stronger than the coercivity of the material (approximately $2 \times 10^4$ A/m for the materials at present in use).

In this article we shall describe isolators with and without an external field.

A most suitable material in resonance isolators for wavelengths in the 8.6 mm region (the Q band) is *topotactically* oriented material [2]) of composition $\mathrm{Ba(Zn_{0.35}Mn^{II}_{0.15}Ti_{0.5})(Fe_{0.95}Mn^{III}_{0.05})_{11}O_{19}}$. This has an anisotropy field of more than $85 \times 10^4$ A/m, which means that the external field need not exceed about $18 \times 10^4$ A/m. For wavelengths in the region of 4.3 mm (the V band) a material having a much higher anisotropy field is needed. Ferrites of this kind have recently been developed in the Philips Irvington Laboratory [3]), and one of them, which has an anisotropy field of about $188 \times 10^4$ A/m, is eminently suited for use in a resonance isolator for 4 mm waves.

The anisotropy field is given the value required for a particular application by substituting other atoms for a certain fraction of the Fe atoms in the base material — a barium-ferrite for the 8 mm band and a strontium-ferrite for the 4 mm band. Because the magnitude of the anisotropy field is thus, within certain limits, controllable, it is proposed to designate this type of material by the collective name *controlled uniaxial anisotropy ferrites*, abbreviated to *c.u.a.f.* The resistivity of the two materials is very high ($> 10^7$ Ωcm); the dielectric losses are negligible.

It should be noted that the permissible power transmission of resonance isolators is limited by the temperature increase caused by the energy absorbed in the ferrite. If the temperature exceeds a certain value, the damping in the inverse direction begins to drop appreciably. In the isolators for 8 mm waves the energy dissipated should not exceed an average of 1 W.

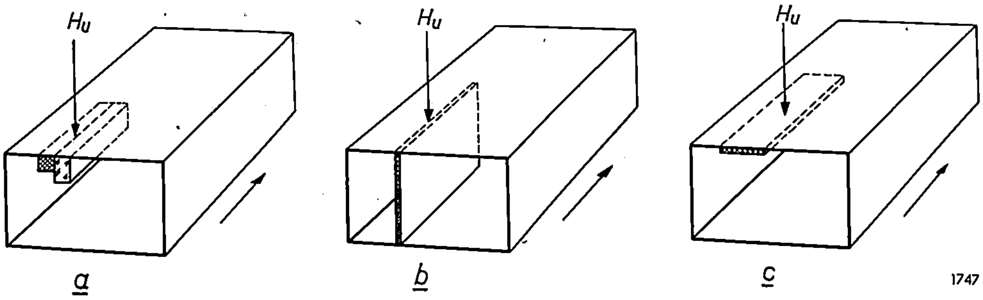


Fig. 1. Schematic representation of microwave resonance isolators in which the damping is produced by gyromagnetic resonance in a ferrite.
*a*) Actual arrangement
*b*) Extreme case, with thin ferrite sliver parallel to the plane of the electric lines of force (*E* plane).
*c*) Other extreme, with ferrite parallel to plane of magnetic lines of force (*H* plane).
In all three figures $H_u$ is the external magnetic field, and the arrow at the right indicates the forward direction of propagation through the waveguide.

## Isolators with weak external magnetic field

The construction of resonance isolators operating with an external field is illustrated schematically in *fig. 1a*. Fitted at a suitable place, side by side, on the broad wall of a rectangular waveguide are a square bar of c.u.a.f. material and a fused quartz strip. This form and the location of the ferrite, lie between two extreme cases where a very thin ferrite sliver is located either in a plane parallel to the electric lines of force (see *fig. 2*) or in a plane parallel to the magnetic lines of force (fig. 1*b* and *c*). In the first of these extreme cases the field in the waveguide is considerably distorted owing to the presence of the ferrite sliver, and measures are needed to minimize the resultant reflection. This is often done by making the ferrite sliver trapezium-shaped. Furthermore, the ratio of the attenuations (measured in decibels) undergone by the waves in the forward and inverse directions — this ratio may be regarded as a kind of figure of merit — is usually not so favourable as in an isolator in which the ferrite sliver

[2]) See F. K. Lotgering, Topotactically crystal-oriented ferromagnetics, Philips tech. Rev. **20**, 354-356, 1958/59.

[3]) F. K. du Pré, D. J. de Bitetto and F. G. Brockman, Magnetic materials for use at high microwave frequencies (50-90 Gc/s), J. appl. Phys. **29**, 1127-1128, 1958.

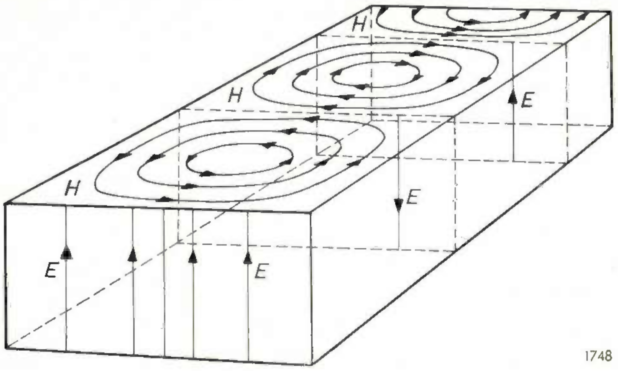


Fig. 2. In the $TE_{10}$ mode of vibration of a rectangular waveguide the electric lines of force ($E$) are perpendicular to the broad side faces and the magnetic lines of force ($H$) are parallel to these faces.

is parallel to the magnetic lines of force. Isolators of the latter type, however (fig. 1*c*), demand a somewhat stronger magnetic field — which is disadvantageous only when a soft ferrite is used — and the width of the sliver, that determines the maximum attenuation that can be achieved per unit length [4]), is rather limited.

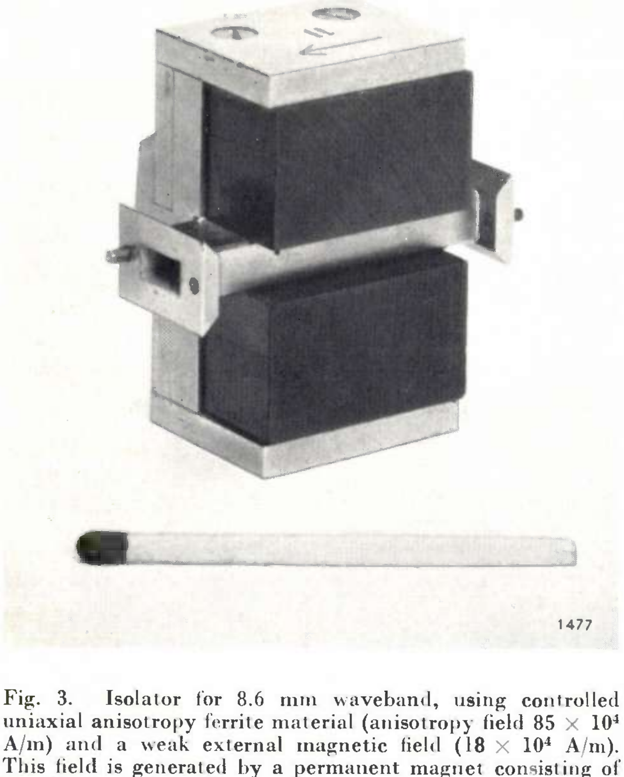


Fig. 3. Isolator for 8.6 mm waveband, using controlled uniaxial anisotropy ferrite material (anisotropy field 85 × $10^4$ A/m) and a weak external magnetic field (18 × $10^4$ A/m). This field is generated by a permanent magnet consisting of two blocks of ferroxdure in an iron yoke.

The intermediate form which we have chosen, and which has proved entirely satisfactory in practice, does not require oblique shaping of the ferrite. Because of the fairly considerable thickness of the ferrite bar, the attenuation per cm length is high. So, too, is the ratio of the attenuations in the forward and inverse directions. This favourable property is partly due to the presence of the silica strip. Since much of the microwave energy traverses the waveguide via and close to the silica strip (dielectric constant ≈ 4, dielectric losses minimal), the attenuation in the inverse direction is substantially greater than in an isolator without dielectric, but the attenuation in the forward direction is not. The explanation of these effects is complicated and not yet wholly clear.

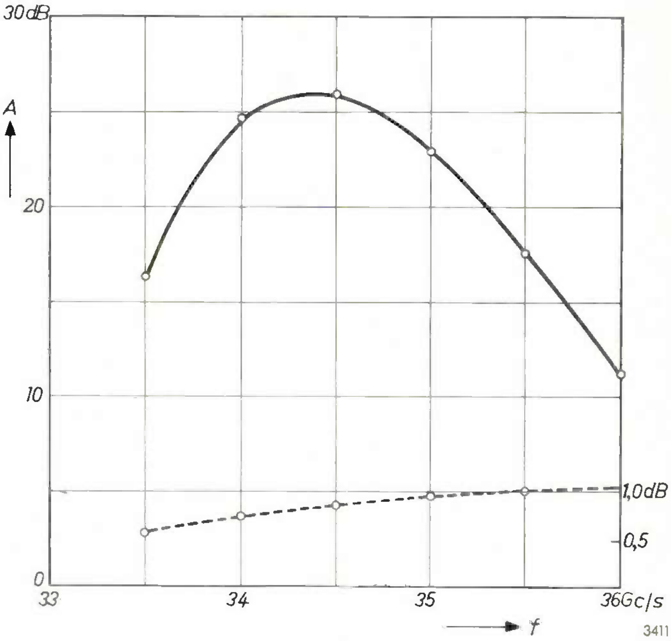


Fig. 4. Transmission characteristics (attenuation $A$ versus frequency $f$) of the resonance isolator with external magnetic field, for the 8.6 mm wave band. The scale values on the left relate to the curve for the inverse direction (solid line), those on the right to the curve for the forward direction (broken line). The maximum ratio between the attenuations (in decibels) occurring in the two directions is approximately 30, and is obtained at a frequency of about 34.3 Gc/s ($\lambda$ = 8.7 mm).

In the isolator for the 8-9 mm wave band [5]) (*fig. 3*) two bars of c.u.a.f. material, having an anisotropy field of 85 × $10^4$ A/m, are mounted end to end. Their dimensions are 12 × 0.80 × 0.40 mm and 12 × 0.60 × 0.36 mm. The slight difference in their widths makes it possible to obtain a broader characteristic (*fig. 4*), inasmuch as the resonance

[4]) The fact that a stronger magnetic field is needed may be deduced directly from formula (2). In the case of the $H$ plane strip we have $N_y \approx 1$ and $N_x \approx N_z \approx 0$, whereas for the $E$ plane: $N_x \approx 1$ and $N_y \approx N_z \approx 0$.

[5]) See also H. G. Beljers, Ferrite isolators in the 8-9 mm waveband, Commun. Congrès int. Circuits et Antennes Hyperfréquences, Paris 21-26 Oct. 1957, Part II (Suppl. Onde électrique **38**, No. 376 ter), pp. 647-648, 1958.

frequencies differ somewhat for the two bars owing to the slight disparity between their demagnetizing factors. The external magnetic field is roughly $18 \times 10^4$ A/m. It is provided by a permanent magnet consisting of two blocks of ferroxdure held in an iron yoke. An air gap, which can be bridged by a shunt allows fine adjustment of the external field strength; the latter is so adjusted that, together with the anisotropy field, it yields exactly the value required to produce gyromagnetic resonance at 35 Gc/s.

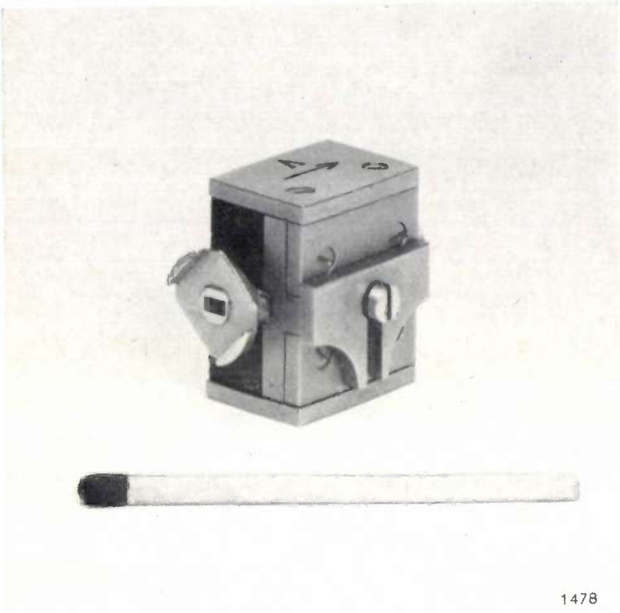


Fig. 5. Resonance isolator for 4.3 mm waves. Anisotropy field of magnetic material approx. $188 \times 10^4$ A/m. External magnetic field approx. $16 \times 10^4$ A/m. The movable shunt on the front serves for adjusting the external field to the exact value required.

The resonance isolator for the 4-5 mm waveband (*fig.* 5) uses bars of c.u.a.f. material which are half as large as in the 8.6 mm isolator and have an anisotropy field of $188 \times 10^4$ A/m.

## Resonance isolator without an external magnetic field

The construction of the isolator about to be described, which operates without an external field, differs considerably from that of the other. This isolator uses two c.u.a.f. ferrite strips one on each side of a thin sliver of dielectric material (in this case aluminium oxide with a dielectric constant $\varepsilon$ of 9). The two ferrite strips, which, like the dielectric, take up the whole height of the waveguide, are magnetized in opposite directions. Together with the walls of the waveguide, which are of iron to give magnetic screening, they form a closed magnetic circuit, that is to say there is no demagnetization. The plate with strips is mounted centrally in the waveguide (*figs.* 6 and 7).

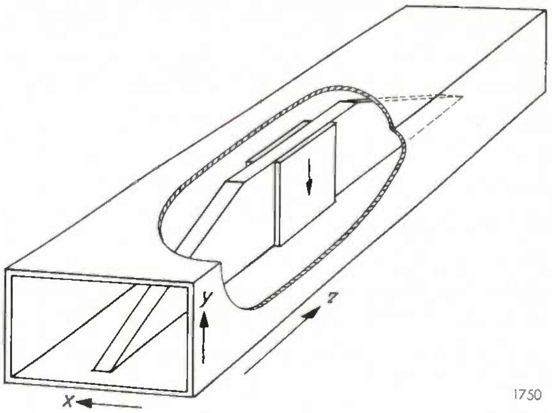


Fig. 6. Schematic representation of a resonance isolator for 8.6 mm waves, which needs no external field. Two thin ferrite strips (2.0 × 3.5 × 0.15 mm; anisotropy field $85 \times 10^4$ A/m) are fixed to the sides of a plate of aluminium oxide (thickness 1.1 mm) mounted centrally in the waveguide. The ends of the plate are cut obliquely to avoid reflections. The visible ferrite strip is magnetized in the direction of the arrow, the other in the opposite direction. The forward direction is that of the positive $z$ axis.

Here, too, a large part of the microwave energy passes through and near the dielectric. Calculations show that the effect of the side walls of the waveguide is of secondary significance, and further that the magnetic field of the microwaves in the side faces of the dielectric plate is elliptically polarized. The ellipticity (by which is meant the ratio $H_x/H_z$,

Fig. 7. Resonance isolator for 8.6 mm waves, with no external field, as schematically illustrated in fig. 6. In front of it can be seen the aluminium-oxide plate with ferrite strips, used in isolators of this type.

cf. fig. 6) decreases asymptotically with increasing plate thickness to the value $\sqrt{\varepsilon/(\varepsilon-1)}$, that is in our case to about 1.06. As mentioned earlier, to obtain minimum damping in the forward direction, elliptical polarization is precisely what is wanted. Calculation shows that the ellipticity of the rotating field must have the value

$$\sqrt{\frac{H_a + M(N_x - N_y)}{H_a + M(N_z - N_y)}}\,.$$

For very thin ferrite strips this expression approximates to

$$\sqrt{(H_a + M)/H_a}\,,$$

which in our case comes to about 1.12.

From the above we may infer that the thickness of the aluminium-oxide plate is here of great importance, and also that it is possible to choose this thickness such that the imposed requirements are fulfilled. Owing to the symmetrical arrangement of the whole assembly it is obviously not possible, as it was in the other two isolators discussed, to choose the dimensions of the materials more or less freely and then to minimize the damping in the forward direction by determining the most favourable position in the waveguide for the plate with the ferrite strips.

The best result is obtained when the dielectric and ferrite are given the dimensions indicated in the caption to fig. 6, which are in good agreement with the calculated values. To avoid reflections the plate is again cut obliquely at the ends. The same ferrite material can be used as in the isolator operating with an external field, the correct resonance frequency being obtained because of the entirely different shape of the ferrite strips. *Fig. 8* shows the characteristic of one of the first isolators of this type made during the development stage.

The follwing calculation will serve to demonstrate the fact that, in this isolator with no external field, the required resonance frequency is nevertheless obtained with the same c.u.a.f. material. For very thin ferrite samples $N_y$ and $N_z$ are negligible and $N_x \approx 1$, so that formula (2) can be written:

$$f = 0.035\,\sqrt{(H + M)H}\,. \quad . \; . \; . \; . \; . \quad (3)$$

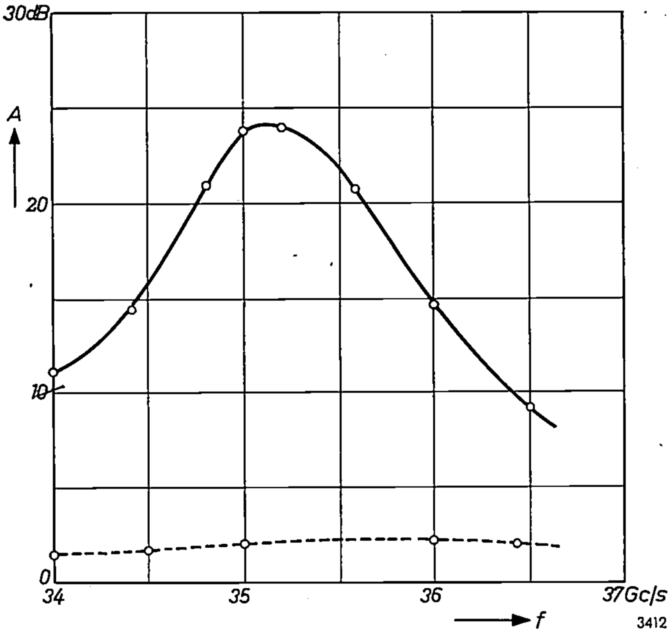


Fig. 8. Transmission characteristics of the resonance isolator without external field. The attenuation ratio here reaches a maximum of 12.

For a material with $M = 28\times10^4$ A/m (3500 gauss) and for a resonance frequency of 35 Gc/s, we obtain an $H$ value of $86\times10^4$ A/m. This is more or less the strength of the anisotropy field. For the isolator with external field, where $N_x \approx 4/5$, $N_y \approx 1/3$ and $N_z \approx 0$, we have:

$$f = 0.035\,\sqrt{(H_a + H_u + {}^7/_{15}M)(H_a + H_u - {}^1/_3 M)}\,, \quad (4)$$

which, with $H_u = 18\times10^4$ A/m, yields an $H_a$ value of more than $82\times10^4$ A/m. (The fact that $N_x$ is here $^4/_5$ and not $^2/_3$, as might be expected, is due to the fact that the image of the ferrite in the wall must be taken into account.) In view of the effect of the dielectric on the resonance frequency, the agreement may be described as reasonably good.

**Summary.** The gyromagnetic resonance effect, which occurs in magnetic materials in the presence of a suitable magnetic field, can be utilized for making nonreciprocal microwave transmission devices, such as directional isolators. In the millimetre wave bands the magnetic field strength required is extremely high ($10^6$ A/m for 8 mm waves and $2\times10^6$ A/m for 4 mm waves) and could not normally be generated with a permanent magnet of manageable proportions. Crystal-oriented anisotropic ferrites have now been developed, however, which possess in one direction a very high anisotropy field, and since the strength of this field can be deducted from the total magnetic field required, it is possible to use these materials for constructing isolators that need only a weak external field or none at all. The article describes resonance isolators of this type for wavelengths in the region of 8.6 mm and 4.3 mm (35 and 70 Gc/s, respectively).

# APPLICATIONS OF MICROWAVE TRIODES

by J. P. M. GIELES.

621.385.3.029.6

*In recent years various disc-seal triodes have been developed at Philips which are capable of operating at wavelengths of 7.5 cm, 5 cm and less. The article below discusses a number of applications of these triodes that have already been realized, and others that are still in the development stage. Most of the applications are in microwave radio links for telephony and television.*

Microwave triodes possess several attractive features. In addition to favourable phase characteristics, long life and easy replacement, they are easy to construct compared with other microwave tubes, they require no very high voltages and they can give a high $G \times B$ product, i.e. a considerable gain $G$ even when the bandwidth $B$ is large.

Another advantage of microwave triodes over other tubes used in the centimetre wavebands is their flexibility: triodes are relatively simple circuit elements and their simplicity makes them suitable for many and various functions. In complicated equipment containing many tubes, such as relay stations in microwave radio links, it is important for technical and economic reasons to limit the number of tube types to a minimum. It is therefore

**Table I.** Summary of data on microwave triodes manufactured or in development at Philips. The last two tubes are designed for a frequency of 6000 Mc/s, the others for 4000 Mc/s. For various purposes, however, the tubes can be operated up to frequencies about a factor of 2 higher.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | EC 157 | EC 59 | OZ 92 | 49 AL | 22 EC | 5 cm, 1 W | 5 cm, 10 W |
| Literature references | [1][2][7][8][9]<br>[10][11][12] | [3][4][7][9][12] | [5] | [12] | — | [6][12] | [12] |
| Cathode diameter (mm) | 3.2 | 4.5 | 4.5 | 4.5 | 4 | 3.1 | 4.5 |
| Cathode-grid spacing (μ) | 40 | 60 | 60 | 40 | 40 | 25 | 40 |
| Grid-anode spacing (μ) | 240 | 300 | 300 | 300 | 240 | 300 | 500 |
| Grid-wire diameter (μ) | 7.5 | 30 | 30 | 15 | 12 | 7.5 | 15 |
| Grid pitch (μ) | 50 | 130 | 130 | 90 | 75 | 40 | 90 |
| Anode voltage (V) | 180 | 500 | 500 | 220 | 180 | 300 | 600 |
| Anode current (mA) | 60 | 250 | 250 | 200 | 140 | 60 | 300 |
| Current density (A/cm$^2$) | 0.8 | 1.6 | 1.6 | 1.3 | 1.2 | 0.8 | 1.9 |
| Transconductance (mA/V) | 19 | 20 | 20 | 30 | 25 | 27 | 20 |
| Low-level gain at 100 Mc/s bandwidth (3 dB below peak) (dB) | 13 | 10 | 10 | 12 | 12 | 10 | 10 |
| Output power at 8 dB gain (W) | 1.5 | 12 | 12 | 6 | 5 | 1.2 | — |
| AM-PM conversion at 8 dB gain (°/dB) | —0.8 | —1.2 | —1.2 | —1.0 | — | — | — |
| Cooling | air *) | water | water or air | air | air *) | air | water |

*) An amplifier is in course of development in which cooling is effected by natural convection and radiation.

not surprising that the triode really comes into its own in such equipment, and has already found wide application in various installations built by Philips.

*Table I* gives a survey of the microwave triodes developed in recent years at Philips, or still in course of development. In this article we shall discuss in turn their application as:
amplifiers,
frequency multipliers,
oscillators,
mixers,
limiters,
amplitude modulators,
frequency modulators, and
high-level detectors.
Most applications have been studied on the EC 157. It is to be expected that the other types of microwave triode will behave similarly in equivalent circuits.

## Amplification

Microwave triodes used for amplification purposes can be classified, according to signal level, as pre-amplifiers or as power amplifiers. Articles have appeared in this journal on the EC 157 and the 5 cm, 1 W tube as pre-amplifiers, and on the EC 59 and OZ 92 triodes as power amplifiers. For particulars, see the articles referred to in Table I. It should be noted that, in view of the high $G \times B$ product, the gain can be raised far beyond the tabulated values by reducing the bandwidth. For example, the bandwidth of an amplifier fitted with an EC 157 was reduced from 100 Mc/s to about 40 Mc/s, giving an increase in gain from 12 dB (= 16×) to 17 dB (= 50×). At a bandwidth of 40 Mc/s a cascade arrangement of three amplifiers with EC 157 triodes gave a total gain of 50 dB.

Not so well known is the application of the EC 157 as a *low-noise input stage* at frequencies below about 1000 Mc/s, used primarily in equipment for radio astronomy. Although its disc-seal construction (resulting in higher capacitances and limited switching possibilities) makes this tube less suitable as a pre-amplifier at frequencies far below the design frequency (4000 Mc/s), it is precisely there that the noise properties of the electrode system are favourable compared with those of a crystal mixer used as an input stage. The minimum noise figure, which is a tube characteristic, can be determined at any frequency by slightly varying the input matching. *Fig. 1* shows the average of the minimum noise figures, expressed in dB, of 13 type EC 157 triodes, measured as a function of frequency [13]).

The strongest point of microwave triodes, which is to produce simply and efficiently a high output power at a large bandwidth, makes them especially suited for use as *power amplifiers*. When used as such in microwave radio links it is most important that amplitude variations should not give rise to disturbing phase variations, and hence to distortion. The extent to which this occurs is called the

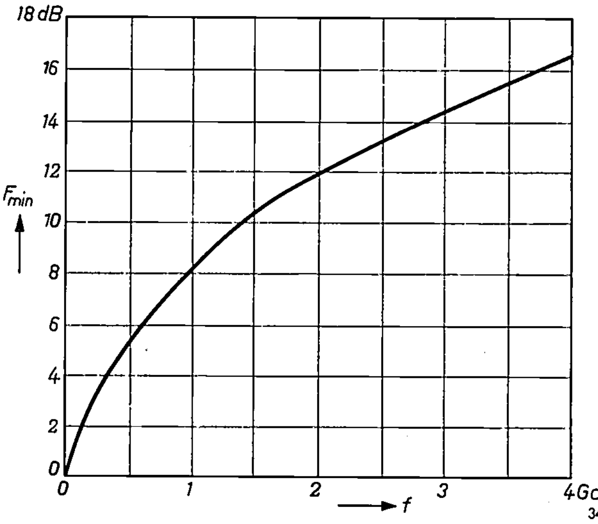


Fig. 1. Average $F_{\min}$ of the minimum noise figures in dB of 13 type EC 157 triodes, versus frequency.

[1]) G. Diemer, K. Rodenhuis and J. G. van Wijngaarden, The EC 57, a disc-seal microwave triode with L cathode, Philips tech. Rev. **18**, 317-324, 1956/57. The EC 157 differs from the older type, EC 57, in having a cathode of longer life.

[2]) J. P. M. Gieles, A 4000 Mc/s wide-band amplifier using a disc-seal triode, Philips tech. Rev. **19**, 145-156, 1957/58.

[3]) V. V. Schwab and J. G. van Wijngaarden, The EC 59, a transmitting triode with 10 W output at 4000 Mc/s, Philips tech. Rev. **20**, 225-233, 1958/59.

[4]) J. P. M. Gieles and G. Andrieux, A wide-band triode amplifier with an output of 10 W at 4000 Mc/s, Philips tech. Rev. **21**, 41-46, 1959/60 (No. 2).

[5]) E. Mentzel and H. Stietzel, A metallic-ceramic disc-seal triode for frequencies up to 6000 Mc/s, Philips tech. Rev. **21**, 104-109, 1959/60 (No. 3).

[6]) M. T. Vlaardingerbroek, An experimental disc-seal triode for 6000 Mc/s, Philips tech. Rev. **21**, 167-171, 1959/60 (No. 6).

[7]) J. G. van Wijngaarden, Possibilities with disc-seal triodes, Onde électr. **36**, 888-892, 1956.

[8]) K. Rodenhuis, A 4000 Mc/s triode with L-cathode construction and circuit, Le Vide **12**, 23-31, 1957.

[9]) G. Andrieux, Amplificateurs de puissance à triodes pour 4000 Mc/s, Onde électr. **37**, 777-780, 1957.

[10]) J. P. M. Gieles, The measurement of group delay in triode amplifiers at 4000 Mc/s, Onde électr. **37**, 781-788, 1957.

[11]) M. T. Vlaardingerbroek, Measurement of the active admittances of a triode at 4 Gc/s, Proc. Instn. Electr. Engrs. **105 B**, Suppl. No. 10, 563-566, 1958.

[12]) H. Groendijk, Microwave triodes, Proc. Instn. Electr. Engrs. **105 B**, Suppl. No. 10, 577-582, 1958.

[13]) These measurements were done by G. A. W. J. Spanhoff and N. van Hurck. For the application of the EC 157 in an amplifier for radio astronomy, see C. L. Seeger, F. L. H. M. Stumpers and N. van Hurck, A 75 cm receiver for radio astronomy and some observational results, Philips tech. Rev. **21**, 317-333, 1959/60 (No. 11).

amplitude-modulation/phase-modulation (AM-PM) conversion of the tube; it is expressed in degrees of phase variation of the output voltage due to a variation of 1 dB in the input signal. The AM-PM conversion of triodes is of the order of —1°/dB, which is much less than that of most other short-wave tubes.

Special attention may be paid to the *parallel amplifier*. It is possible to connect two amplifiers in parallel such that, for the same gain, they deliver twice the output power. *Fig. 2* shows a parallel output stage of this kind for operation at 4000 Mc/s, equipped with two type 49 AL triodes. This stage is capable of delivering an output of 10 W at an anode voltage of 220 V, the gain being more than 8 dB. Both amplifiers, like a single amplifier, are equipped with ferrite isolators. (Because of this arrangement the data for fig. 2 differ somewhat from those given in Table I.)

In this way outputs up to 50 W at 4000 Mc/s have been obtained from EC 59 triodes in the laboratory. Microwave radio links generally require no more than 10 or 20 W, which is ample to establish a dependable link. Apart from doubling the output power, the parallel output stage is more reliable than a single stage since, with proper design, the failure of one tube does not necessarily involve interruption of the link.

The use of triodes as microwave amplifiers may best be illustrated, perhaps, by their application in a "*straight-through*" *relay station* in a microwave radio link, i.e. a relay station in which the gain is obtained without demodulating to an intermediate frequency. This method of amplification stems from the fact that the triode as an amplifier of centimetre waves is superior to the tubes normally used for amplification in the IF wavebands. In view of switching and modulation problems a straight-through relay station does not lend itself directly to general use, but for long-distance links it has the great advantage of causing much less phase-distortion than conventional IF amplifiers. A laboratory version of such a straight-through relay station is shown in *fig. 3*. The block diagram is given in *fig. 4*.

A six-cavity input filter $F_1$ is followed by four pre-amplifier stages fitted with EC 157 triodes. All amplifiers are coupled by isolators $D$. By means of a frequency-shifting stage $M$, which will be discussed below under the head "Mixing", the frequencies of the input and output signal are separated sufficiently to exclude the possibility of spurious oscillations due to aerial feedback. On the recommendation of the Comité Consultatif International des Radiocommunications, the frequency shift is set at 213 Mc/s. After a three-cavity filter $F_2$ for suppressing the unwanted frequencies of the frequency-shifting stage, and three further amplifying stages, a small portion of the signal is taken off for automatic gain control. This feeds back through a DC amplifier $AVC$ on the field current in the ferrite attenuator $At$, whose attenuation depends on the magnetic field produced by this current. This system ensures that the signal level at the input of the last stage remains constant within 0.5 dB for a variation of up to 25 dB in the input signal of the first stage. A single output stage is used, fitted with an EC 157, so that the output power of the whole installation is about 1.5 W. The overall bandwidth is about 40 Mc/s at 0.1 dB below peak, and the group-delay variation over 20 Mc/s amounts, without compensation, to about 1 millimicrosecond, which is less than 1/10th of the value in normal IF amplifiers. The intermodulation noise, which is due to the non-linearity of phase characteristics, is accordingly very low. The weak point of this relay station is the input noise factor, which, owing to the attenuation caused by the input isolator and the input filter, amounts to as much as 18 dB. This drawback can be overcome, however, by increasing the input signal, that is to say by raising the transmitted power of each relay station in the chain. If, for example, the parallel output stage shown in fig. 2, or a single output stage with an EC 59, be connected at the end of the arrangement described, an output power of 10 W can be achieved.

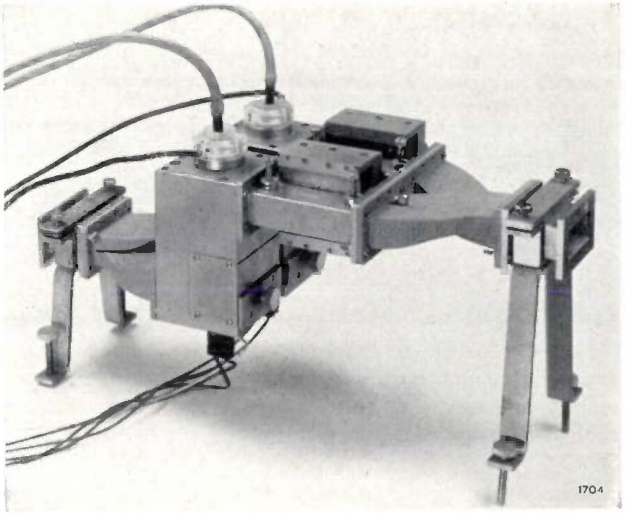


Fig. 2. Parallel output stage, consisting of two amplifiers each equipped with a type 49 AL triode.

| | | | |
|---|---|---|---|
| Frequency | 4000 Mc/s | Bandwidth | 100 Mc/s |
| Anode voltage | 220 V | Gain | 8.2 dB |
| Anode current | 2 × 200 mA | Output power | 10 W |
| Cooling air: | 27 l/min under pressure of 10 cm water. | | |

Finally, microwave triodes can also be used successfully as output amplifiers *at lower frequencies*. Wide use is made of the EC 157 in microwave links at frequencies in the 2000 Mc/s band. An experimental push-pull output stage with two EC 59 triodes in class C delivered an output of 76 W at 900 Mc/s [14]), which roughly corresponds to the performance of other tubes in that frequency band.

### Frequency multiplication

The usual practice in microwave radio links is for the incoming radio-frequency signal to be transposed to the intermediate frequency by mixing it in a crystal mixer stage with the signal from a local oscillator. After amplification the IF signal is returned to the radio-frequency band by means of a second local oscillator (see section on mixing). The two RF local oscillators needed must be

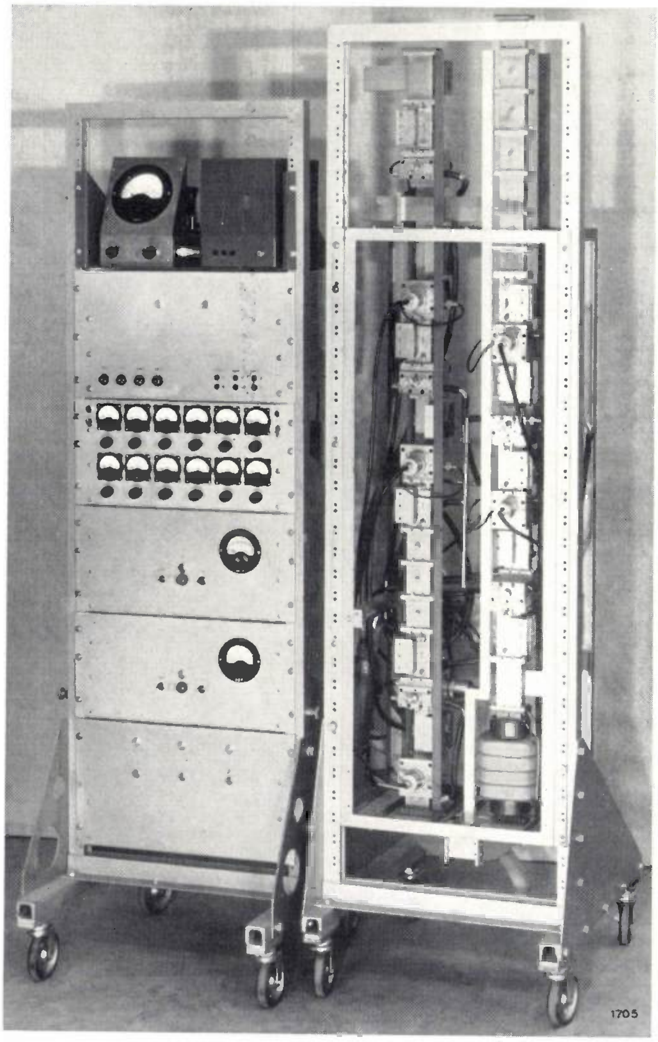


Fig. 3

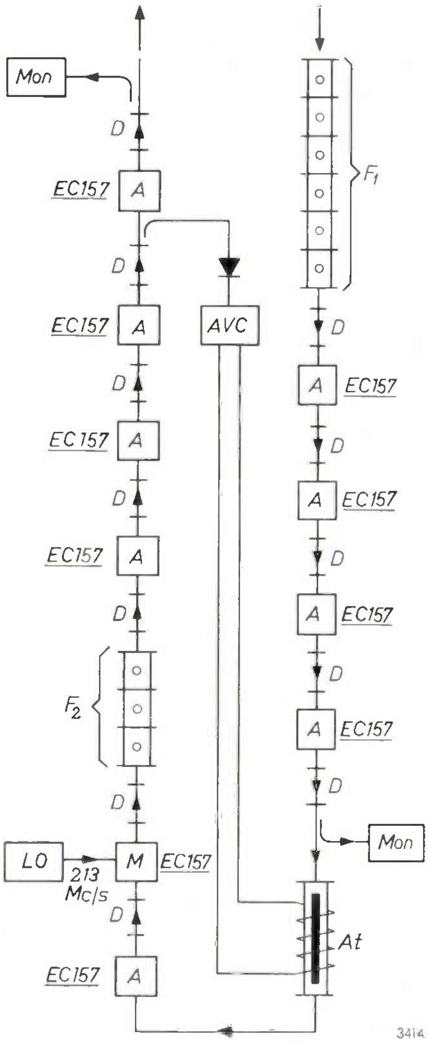


Fig. 4

Fig. 3. A complete "straight-through" relay station for a microwave radio link operating on 7.5 cm.

Fig. 4. Block diagram of "straight-through" relay station, in which all amplification is obtained in the 4000 Mc/s frequency band. $F_1$ six-cavity input filter. *D* ferrite isolators. *A* amplifier stages equipped with EC 157 triodes. *Mon* monitors. *At* ferrite attenuator. *M* mixer stage producing frequency shift between input and output signals. This frequency shift (213 Mc/s) prevents spurious oscillation due to aerial feedback. *LO* local oscillator. $F_2$ three-cavity filter for suppressing unwanted frequencies. *AVC* DC amplifier supplying the signal for the automatic gain control. Overall output power, ~ 1.5 W; bandwidth at 0.1 dB below peak, ~ 40 Mc/s; group-delay variation over 20 Mc/s, without compensation, ~ 1 mμsec.

[14]) Experiment done by W. J. Smulders.

extremely stable in operation. *Fig. 5* shows the block diagram of a typical local oscillator of this kind [15]). The first stage is a crystal oscillator (*Cr*) whose frequency is in the region of 25 Mc/s, and the required frequency, which may be, for example, in the region of 4000 Mc/s, is obtained by means of a series of frequency multipliers (*M*). The last stages use EC 157 tubes, the final multiplication being six-fold. A complete local oscillator designed on this principle for use in Philips microwave radio links can be seen in *fig. 6*. The output power is of the order of 200 mW, which is sufficient both for the receiving and transmitting ends of a relay station.

The above-mentioned frequency of 4000 Mc/s is by no means an upper limit; the EC 157 can be used as a frequency multiplier at frequencies far higher. In an experimental set-up an EC 157 was driven at 4000 Mc/s by an input power of approximately 1 W. With the anode surrounded by a circuit tuned at the required frequency, all harmonics were obtained up to 24 000 Mc/s. One of the multiplier stages used is shown in *fig. 7*. The available power, particularly at the highest frequencies, is not large but is more than enough to be used, for example, in frequency standards. With the 5 cm, 1 W triode, excited at say 8000 Mc/s, it may be possible to reach the millimetre waveband.

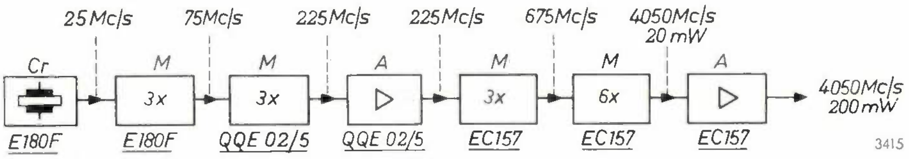


Fig. 5. Block diagram of a local oscillator [15]) for microwave radio-link relay station. *Cr* crystal oscillator. *M* frequency multipliers, the last two fitted with EC 157 triodes. *A* amplifiers.

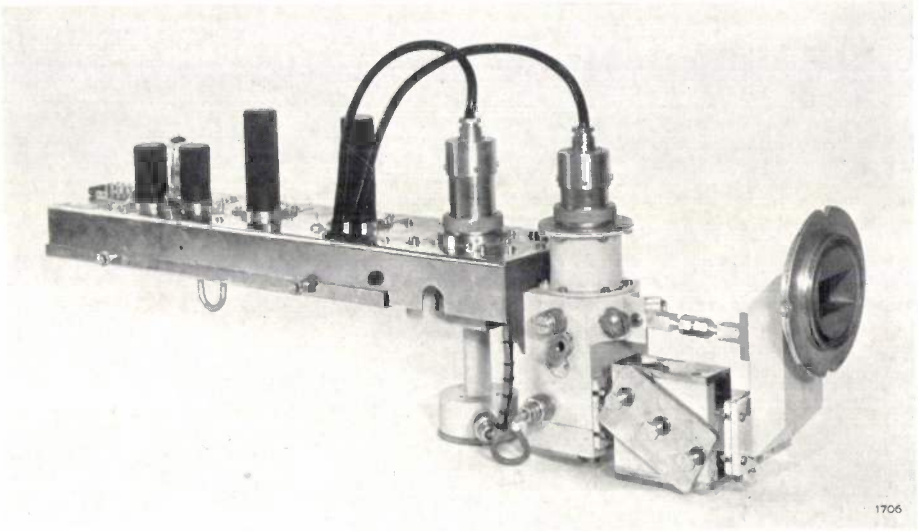


Fig. 6. A complete local oscillator, as in fig. 5, used in Philips microwave radio-link equipment.

## Applications in oscillators

The local oscillator described above contains, as frequency multipliers, many electron tubes, which may be regarded as so many potential sources of

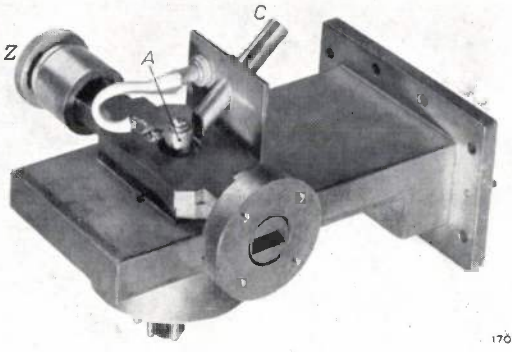


Fig. 7. Frequency multiplier stage equipped with an EC 157 triode, capable of producing 20 000 Mc/s by multiplying 4000 Mc/s. *A* anode. *C* nozzle for cooling-air supply. *Z* knob for adjusting tuning plunger. Driven by about 1 W at 4000 Mc/s the available power at the output is approx. 0.1 mW at 20 000 Mc/s.

[15]) This design, like the mixer circuit in fig. 11, was developed by H. J. Kramer in the microwave radio-link laboratory of the N.V. Philips' Telecommunicatie-Industrie at Huizen.

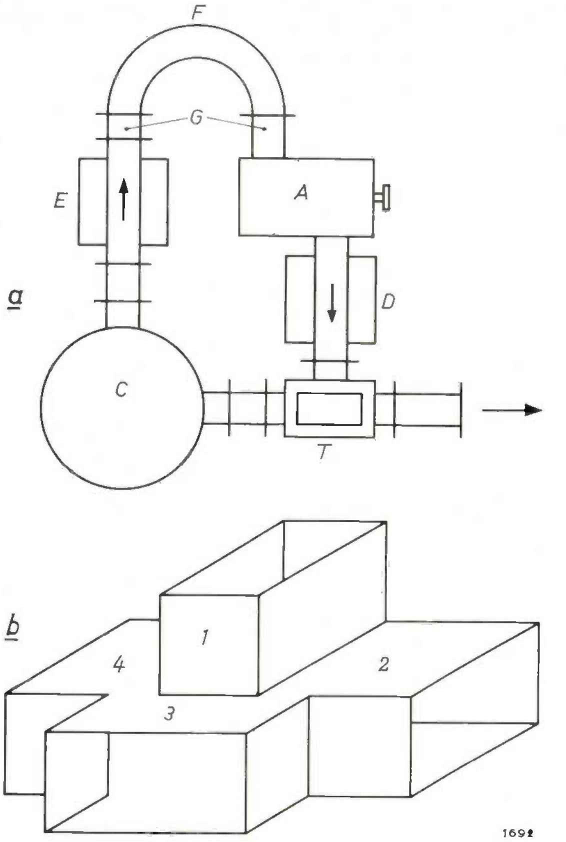
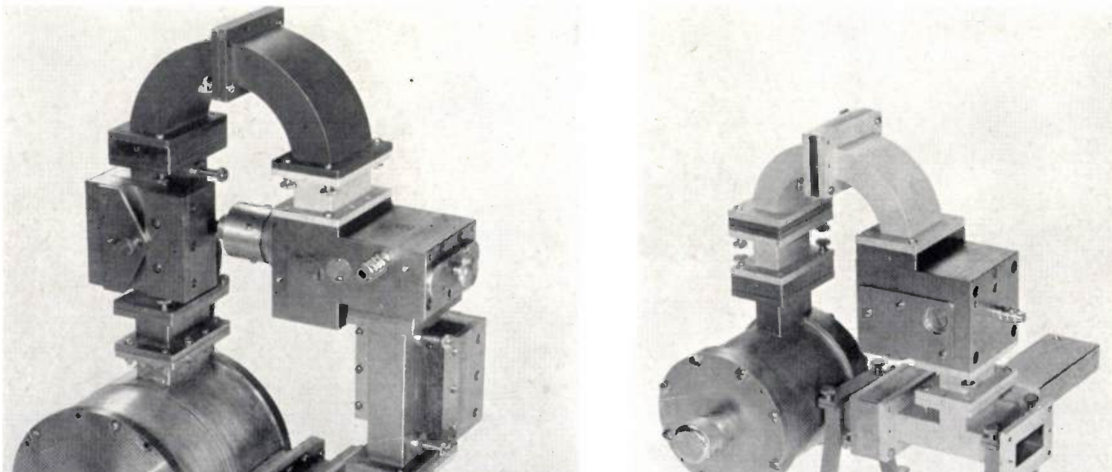


Fig. 8. *a*) Diagram of a 4000 Mc/s oscillator with an EC 157 triode. *A* amplifier. *C* cylindrical cavity resonator of "Invar", filled with dry air. *T* magic-tee (see *b*) as output coupler. *D*, *E* isolators. The phase of the signal returning via the curved section *F* to *A* is adjusted by means of spacer sections *G*. *b*) Magic-tee form of hybrid junction. Branches *1* and *3* have no corresponding dimensions parallel. In a magic-tee an electromagnetic wave entering one of the four branches *1*, *2*, *3* or *4* will divide equally between only *two* of the other three branches: paths *1*→*3*, *3*→*1*, *2*→*4* and *4*→*2* are not possible. (See for example G. C. Southworth, Principles and applications of waveguide transmission, Van Nostrand, New York 1950, pp. 339-340.)

breakdown. Attempts have therefore been made to design an oscillator which, with only one tube, will oscillate directly at the required frequency just as stably as a crystal oscillator. The components for such a local oscillator were available: the EC 157 triode can compete in every respect with tubes for lower frequencies, and with the aid of a cavity resonator the same high $Q$ for high frequencies can be achieved as with the crystal. As the material for this resonator we can use "Invar" which, like quartz, possesses a very low thermal coefficient of expansion (both about $10^{-6}$ per °C). If the same care is paid to the filling of the resonator as to the envelope around the crystal, similar results may be expected.

*Fig. 8a* shows a diagram, and *fig. 9a* a photograph, of a 4000 Mc/s oscillator built on these lines. The amplifier $A$ and the cylindrical cavity $C$ are incorporated in a closed loop of waveguides. They are both terminated by an isolator ($D$ and $E$, respectively). The circuit further contains a magic-tee $T$ as the output coupler. In this junction the output power of the amplifier is split in two (see caption to fig. 8*b*): one half is fed to the output waveguide,



*a* *b*

Fig. 9. *a*) Oscillator for 4000 Mc/s fitted with an EC 157, built on the principle of fig. 8*a*. *b*) Similar oscillator, but for 6000 Mc/s and fitted with a 5 cm, 1 W triode. Both oscillators, *a* and *b*, are reproduced on the same scale.

the other half returns via the cavity to the input of the amplifier. A matched termination of the fourth branch of $T$ absorbs any waves reflected from the cavity or the load.

Calculation has shown that the oscillator reaches optimum stability when the anode bandwidth of the amplifier is about 130 Mc/s and the attenuation of the resonant cavity 4.6 dB. The attenuation introduced by the output coupler amounts to 3 dB, and that due to each of the two ferrite isolators is about 1 dB. Since the loop gain must be unity, the operating point of the tube should be such that the gain is 9.6 dB; at this gain, and at the bandwidth of 130 Mc/s, its output power is 950 mW. The isolator $D$ reduces this to 750 mW, half of which, i.e. 375 mW, is the available power at the output.

The Invar cavity has very smoothly finished interior walls. It is excited in the $TE_{012}$ mode and is so designed as to give a loaded $Q$ of 22 000. The cavity contains dry air, and is sealed by thin sheets of mica at the waveguide flanges and a synthetic-resin seal along the edges of the lids.

The oscillator is aligned as follows. First of all the cathode side of the amplifier is matched at the required frequency, and at the anode side the bandwidth is adjusted to about 130 Mc/s. The anode voltage is 200 V, the anode current 60 mA. After the complete oscillator loop has been assembled, the bent waveguide $F$ (fig. 8*a*) is removed and a signal of the desired frequency is applied to the amplifier input. A detector is connected to the isolator $E$, and the amplifier and resonant cavity are then tuned exactly. Finally the curved section $F$ is replaced and the phase of the returning signal is regulated by varying the loop-length by means of spacers $G$ until the output power is maximum. If necessary the frequency can be finely adjusted with a trimming screw in the resonator.

The stability of the frequency during variation of the various parameters is measured by a beat-frequency method. Some results at 4000 Mc/s are given in *Table II*. The heater voltage proves to be the most critical of the parameters; the anode voltage has only a very minor influence. After exchanging the tube, only the anode side being retuned to maximum output power, it was found that the frequency had moved only 10 kc/s from the original value.

Fig. *9b* shows a similar oscillator for the frequency band around 6000 Mc/s, equipped with a 5 cm, 1 W type triode. Since the internal feedback in this tube is much lower than in the EC 157, isolators are not needed here.

**Table II.** Stability at 4000 Mc/s of the oscillator in fig. 9*a*.

| Varying parameter | Variation of frequency with increasing parameter | |
|---|---|---|
| | Absolute | Relative |
| Temperature (25-65 °C) | +1.4 kc/s per °C | $+0.35\times10^{-6}$ per °C |
| Heater voltage | ~ −0.06 kc/s per mV | ~ $-1\times10^{-6}$ per % |
| Anode voltage | +0.05 kc/s per V | $+0.025\times10^{-6}$ per % |
| Anode current | +1 kc/s per mA | $+0.15\times10^{-6}$ per % |

A triode can be used successfully as an oscillator not only at the design frequency and below, but also at far higher frequencies. Up to the design frequency the $\frac{1}{4}\lambda$ mode of the anode resonant cavity will be used; at much higher frequencies this is no longer possible because the resonant cavity would have to be smaller than is compatible with the tube dimensions. Nevertheless, the tube can be tuned to thees high frequencies by shifting the short-circuiting wall of the resonator half a wavelength outwards, thus using the $\frac{3}{4}\lambda$ mode. *Fig. 10* shows an experimental oscillator with an EC 157, which, on the principle described, is capable of generating frequencies up to about 8000 Mc/s.

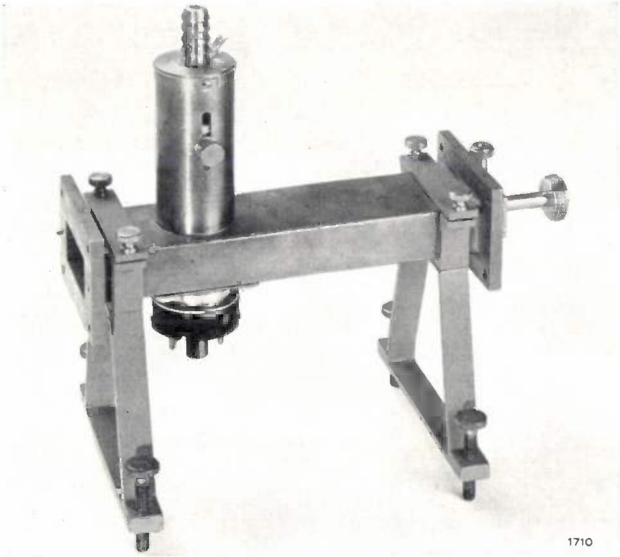


Fig. 10. Experimental oscillator with an EC 157 triode, with which frequencies up to about 8000 Mc/s are reached by using the $\frac{3}{4}\lambda$ mode of oscillation of the anode resonant cavity.

### Mixing

If we apply to the input waveguide of an EC 157 amplifier the high-frequency signal from a local oscillator (LO signal) and apply to the cathode an IF signal, sum and difference frequencies appear at the anode. By tuning the anode circuit to one of these, e.g. the sum frequency, we obtain a mixer stage. A mixer of this kind is used in Philips microwave radio-link equipment for the purpose of transposing the signal from the intermediate frequency (70 Mc/s) to one of the radio-frequency bands at 4000 Mc/s. The circuit diagram is shown in *fig. 11* [15]), and the data are given in the caption. The anode circuit is followed by a three-cavity filter, which suppresses the LO signal and the unwanted side-band (difference frequency).

The anode current of the EC 157 is fairly small, being only 25 mA, at which value considerable AM-PM conversion (see p. 18) occurs. This is almost entirely compensated, however, by the AM-PM conversion of the driving tube. Since AM-PM conversion is mainly due to variation of the input capacitance with the magnitude of the driving signal, its sign in an earthed-grid arrangement will be opposite to that in an earthed-cathode arrangement. The AM-PM conversion of the driving tube is positive and that of the triode negative, and both are about 2°/dB, so that by suitable design of the circuit we can build a mixer stage capable of handling fairly strong signals and nevertheless virtually free of AM-PM conversion (< 0.2°/dB).

The triode as a mixer is also used in a straight-through relay station in microwave radio links for imparting to the signal the earlier-mentioned frequency shift of 213 Mc/s (see fig. 4). The RF signal is applied to the input of an EC 157 amplifier, and a sufficiently strong LO signal of the desired shift frequency is applied to the cathode (*fig. 12*). The amplitude-modulated RF signal is thus mixed with the shift frequency, which again gives rise to two sidebands. The sideband required is then selected by suitable filters.

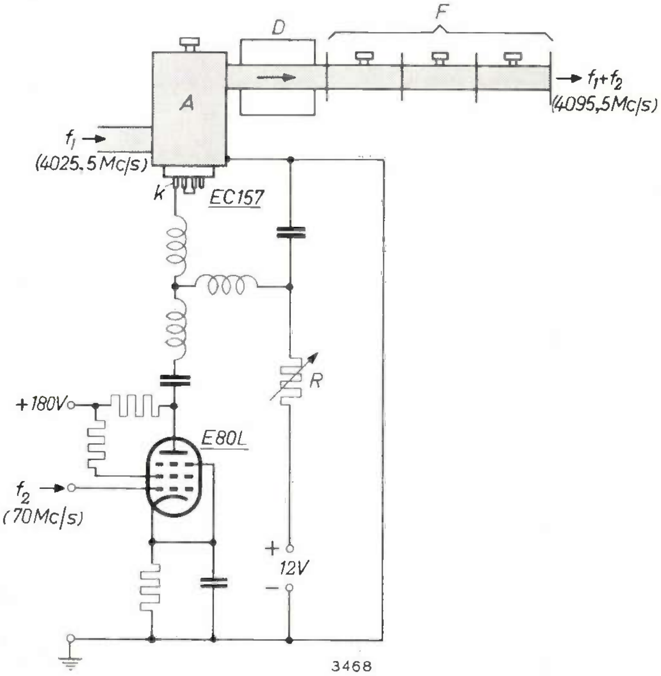


Fig. 11. Circuit [15]) for an EC 157 used as a mixer for transferring the modulation of an IF signal ($f_2$) to an RF signal ($f_1$). *A* amplifier. *D* isolator. *F* three-stage filter. *k* cathode lead of EC 157.

| | |
|---|---|
| Conversion gain | 3 dB |
| Bandwidth at 0.1 dB below peak | 25 Mc/s |
| Input power of EC 157 | 200 mW |
| IF signal on grid of E 80 L (r.m.s.) | 0.35 V |
| AM-PM conversion | <0.2 °/dB |
| Anode voltage of EC 157 | 180 V |
| Anode current of EC 157 (variable by means of *R*) | 25 mA |

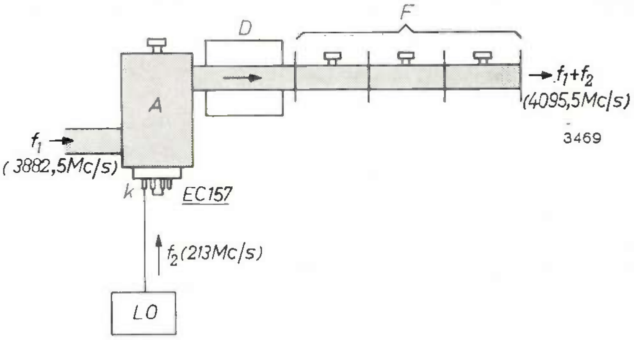


Fig. 12. Circuit for an EC 157 used as a mixer to produce a frequency shift of the RF signal in the straight-through relay station shown in fig. 3. *LO* local oscillator. Other symbols as in fig. 11.

| | |
|---|---|
| Conversion gain | 1 dB |
| Bandwidth at 0.1 dB below peak | 50 Mc/s |
| LO power | 500 mW |
| Anode voltage of EC 157 | 180 V |
| Anode current of EC 157 | 20 mA |

### Limiting

If a microwave triode is biased to operate at a low anode current and the input power increases, a situation is soon reached where the output power can rise no further. In this saturation region the tube may thus be expected to function as a limiter.

We can put this to the test by measuring the output power $P_o$ as a function of the input power $P_i$; for the EC 157 this results in curve *A* in *fig. 13*. With increasing $P_i$ this curve does not become flat, as we should wish for a limiter, but, at a larger driving signal, exhibits a very sharp dip. This effect is due to the passive feedback admittance between the cathode and anode circuits [16]). What in fact happens? The output power may be regarded as consisting of two parts. One part comes from the

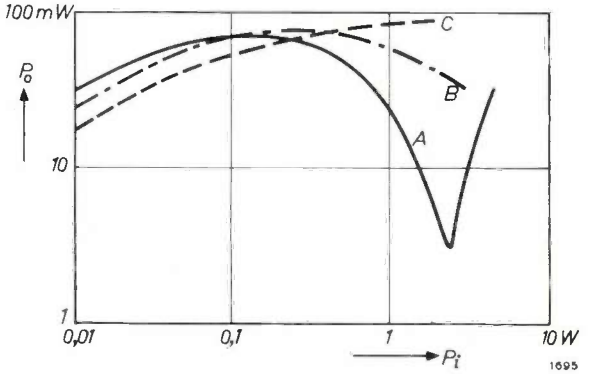


Fig. 13. The EC 157 as a limiter: output power $P_o$ as a function of input power $P_i$. Anode voltage 200 V, anode current 10 mA. Curve *A*: amplifier not neutralized, curve *B*: partly neutralized, and curve *C*: completely neutralized.

[16]) G. Diemer, Passive feedback admittance of disc-seal triodes, Philips Res. Repts. **5**, 423-434, 1950. See also the articles referred to under footnote [1]), p. 323, and [2]) pp. 152-154.

amplifying action of the triode. This part, which predominates in the case of small driving signals, approaches a constant value as the signal increases. The other part arises from the passive feedback admittance and has no saturation value, but continues to increase linearly with the input power. A point is reached, therefore, where these two parts are equal, and what happens then depends only on the phase relationship. Evidently in our case the two parts are virtually in antiphase, resulting in almost complete cancellation.

To avoid this undesirable effect we have tried to introduce between the anode and cathode cavities additional feedback, just sufficient to neutralize the existing feedback (internal neutralization). This is done, as shown in *fig. 14*, by passing a double coupling loop $L$ of suitable dimensions through the partition between the two resonant cavities. In this way it proved possible to neutralize an amplifier completely. The result is curve $C$ in fig. 13. It can be seen that the neutralization also results in a drop in gain for small $P_i$. The effect of partial neutralization will be to shift the dip towards a higher $P_i$. We then obtain curve $B$.

For use as a limiter the optimum curve lies somewhere between $B$ and $C$. A drawback is that, with stronger neutralization, the flat region also shifts towards a higher $P_i$. In that region, then, the triode functions well as a limiter. The output power can remain constant within 0.4 dB for an input-power variation of 10 dB [17]).

### Amplitude modulation

Even in microwave telecommunications, where amplitude modulation is not usual, it may be necessary in certain circuits to modulate a signal in amplitude, as for example in the circuit described in the next section. An amplitude modulator is also often indispensable in laboratory test equipment. Its operation is actually the same as that of the mixer discussed above, but in accordance with general usage we shall confine the term "modulation" to those cases where the modulating signal is an audio or video signal.

Depending on the strength of the RF signal, different circuits have to be used to minimize distortion. For weak signals it is sufficient to introduce the signal as a voltage source of low internal resistance in series with the cathode resistor, as shown in *fig. 15*. As long as the driving signal is not too strong, this voltage modulaton can be applied up to a modulation depth of 80%.

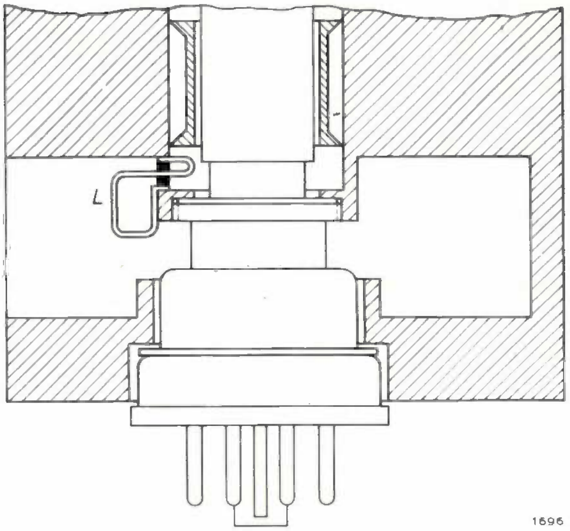


Fig. 14. Neutralization of an EC 157 amplifier by means of a double coupling loop $L$ through the partition between anode and cathode resonant cavities.

Where a tube is operated close to the saturation level, however, current modulation is found to give the best results. For this purpose the signal source should be a current source of high internal resistance connected in series with the cathode resistor. The circuit adopted in practice is shown in *fig. 16*. The modulating voltage is applied between grid and cathode of a power pentode, type EL 86, connected in series with the amplifying tube. With this circuit an EC 157 almost driven to saturation can be modulated to a depth of about 90%.

In both cases the variation of the output *power*, and not of the output *voltage*, is proportional to the modulating voltage. This means that the detector must have a square-law characteristic.

### Frequency modulation

Frequency-modulated microwaves are difficult to obtain with only one microwave triode used as an

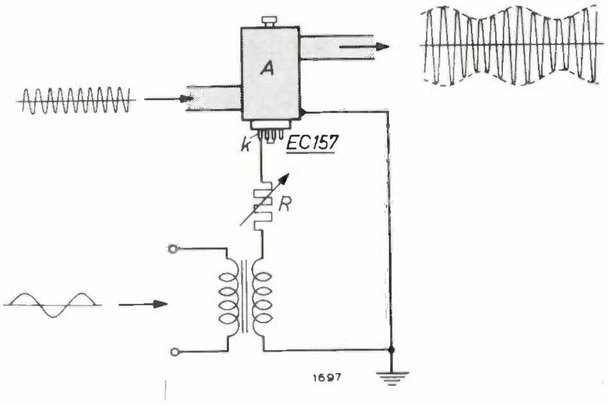


Fig. 15. Amplitude modulator with an EC 157 for modulating small signals. $A$ amplifier. $k$ cathode lead. $V_a = 200$ V, $I_a = 20$ mA (variable by means of $R$).

[17]) The AM-PM conversion of these and similar limiters has not yet been investigated.

oscillator. The oscillator frequency is then almost entirely determined by the tuned circuits, and although the frequency can be varied to some extent by varying the anode voltage and anode current, the circuit is not attractive for frequency modulation because the frequency deviation attainable is small and because of the large amplitude modulation which it involves.

The situation is different, however, if *two* tubes are used in a special circuit, the principle of which is illustrated in *fig. 17*. Two EC 157 amplifiers $A$ and $B$ are interconnected by waveguides (bold lines in the figure) via two magic tees $T_1$ and $T_2$. The path $T_1AT_2$ is made a quarter wavelength longer than the path $T_1BT_2$. Furthermore, $T_1$ and $T_2$ are directly interconnected via $PQ$.

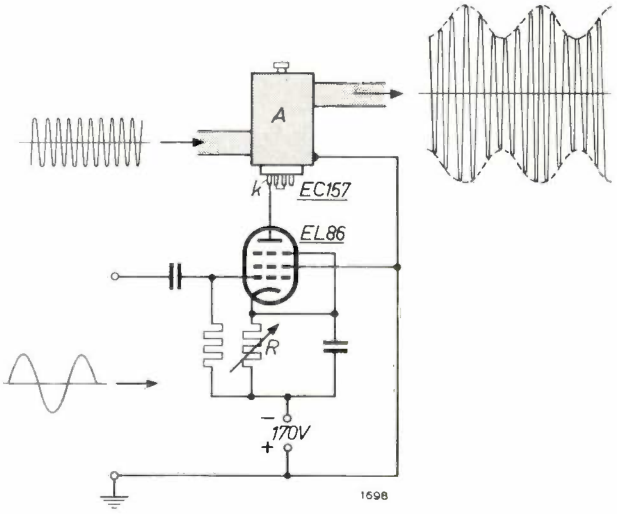


Fig. 16. Amplitude modulator with an EC 157 for modulating strong signals. Symbols as in fig. 15. $V_a = 200$, $I_a = 60$ mA.

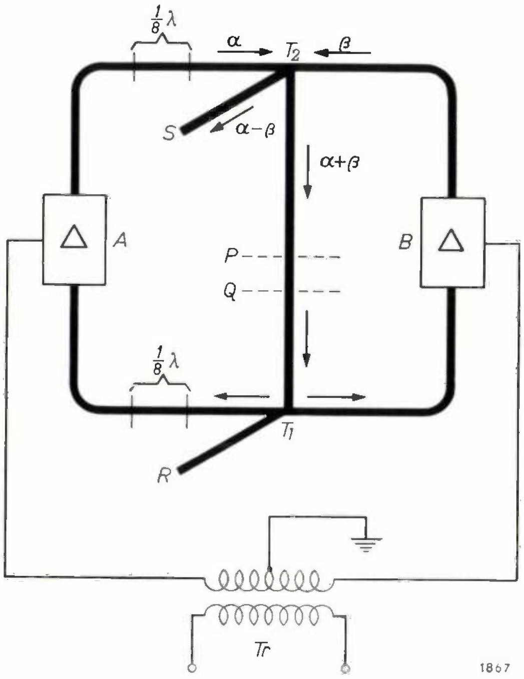


Fig. 17. Illustrating the principle of a frequency modulator with two EC 157 triodes. The thick lines represent waveguides. $A$, $B$ amplifiers. $T_1$, $T_2$ magic-tees. The branch $R$ has a matched termination, the branch $S$ functions as output. $Tr$ centre-tapped transformer (input).

To explain the operation of the circuit we shall imagine that the connection between $T_1$ and $T_2$ is broken for a moment at $P$ and $Q$. An electromagnetic wave entering $Q$ divides in $T_1$ into two waves of the same power and the same phase, one travelling to the left and the other to the right. These two waves are equally amplified in identical amplifiers $A$ and $B$, and thus arrive in $T_2$ with equal amplitudes and, owing to the $\frac{1}{4}$-wavelength extra path-length via $A$, with a phase difference of 90°. They can therefore be represented by the complex quantities $\alpha$ and $\beta$ in *fig. 18*. In $T_2$ the waves $\alpha$ and $\beta$ combine in such a way that a wave $\alpha + \beta$ travels towards $P$ and a wave $\alpha - \beta$ along $S$.

Now suppose the waves $\alpha$ and $\beta$ are not only amplified in $A$ and $B$ but also modulated in amplitude in push-pull, as represented diagrammatically by the centre-tapped transformer $Tr$ in fig. 17. In practice this amplitude modulation can be effected by one of the methods described under the previous heading; the instantaneous output powers of the amplifiers then become varied in the ratio $(1 + m \sin pt) : (1 - m \sin pt)$, where $m$ is the modulation depth and $p$ the angular frequency of the modu-

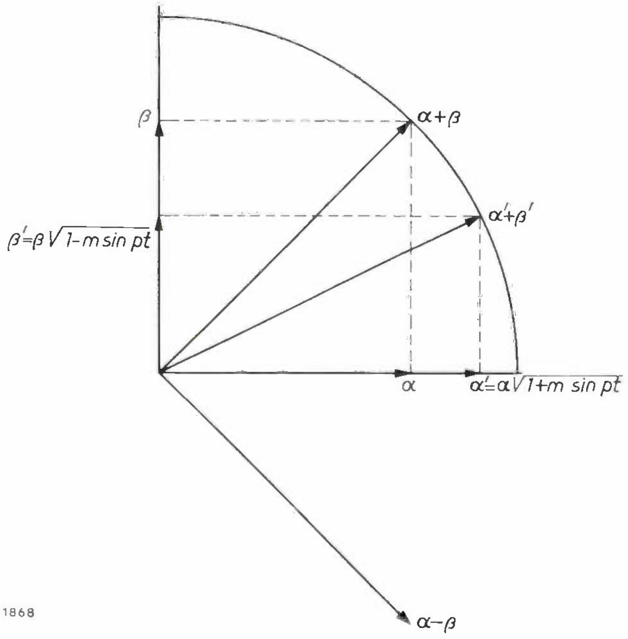

β
α+β
α'+β'
$\beta' = \beta\sqrt{1-m\sin pt}$
α
$\alpha' = \alpha\sqrt{1+m\sin pt}$
α−β
1868

Fig. 18. Vector diagram illustrating the occurrence of phase modulation when the amplifiers $A$ and $B$ in fig. 17 are push-pull amplitude-modulated.

lating signal. The amplitudes of the waves arriving in $T_2$ are thus given by $\alpha' = \alpha\sqrt{1 + m \sin pt}$ and $\beta' = \beta\sqrt{1 - m \sin pt}$, and those at the output of $T_2$ by $\alpha' + \beta'$ and $\alpha' - \beta'$.

It is easy to show that the moduli of both these output waves are equal and constant, and that the arguments vary with time. The output waves from $T_2$ are therefore not modulated in amplitude but only in phase: the vector points of $\alpha' + \beta'$ and $\alpha' - \beta'$ thus move, during a modulation period of the amplifiers, along arcs of the same circle in fig. 18.

If we now connect $P$ again with $Q$, the circuit will start to oscillate, and it will do so at a frequency such that the total phase shift in the feedback loop is exactly a whole multiple of $2\pi$. Since the modulating voltage causes the "phase length" of the circuit to vary, the frequency at any given moment will be such as to satisfy the just-mentioned phase condition. In this way, then, we have obtained a frequency-modulated oscillator. The branch $S$ of $T_2$ functions as the output, and the branch $R$ of $T_1$ has a matched termination.

To produce a large frequency deviation it is necessary that the phase variation introduced should constitute as large a part as possible of the total phase shift in the loop. This amounts to keeping the whole feedback path as short as possible. A compact assembly is achieved by combining the two magic-tees and their connection $PQ$ in a single block (*fig. 19*). The two amplifiers are normal amplifier units, except that input and output are at the same side. The amplifiers are fitted to the left and right of the block in fig. 19 (see *fig. 20*), and the difference in path length is obtained by means of spacer sections. The bottom opening at the front of the central block ($R$ in fig. 17) has a matched termination, and the upper opening ($S$ in fig. 17) is the output of the modulator.

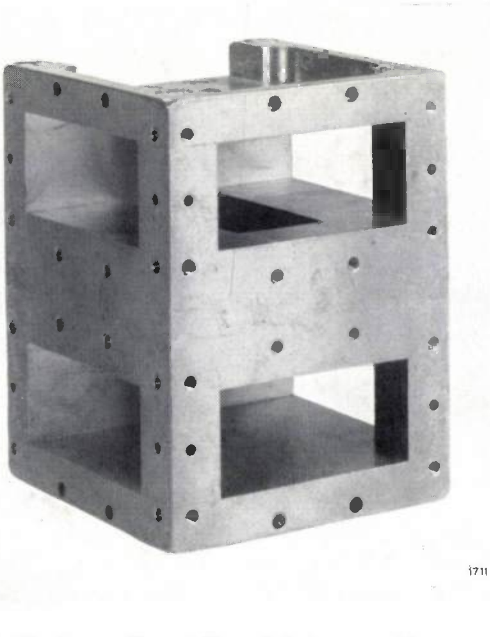


Fig. 19. Magic-tees $T_1$ and $T_2$ and their connecting waveguide (see fig. 17) combined in a single block, giving a compact construction of the frequency modulator.

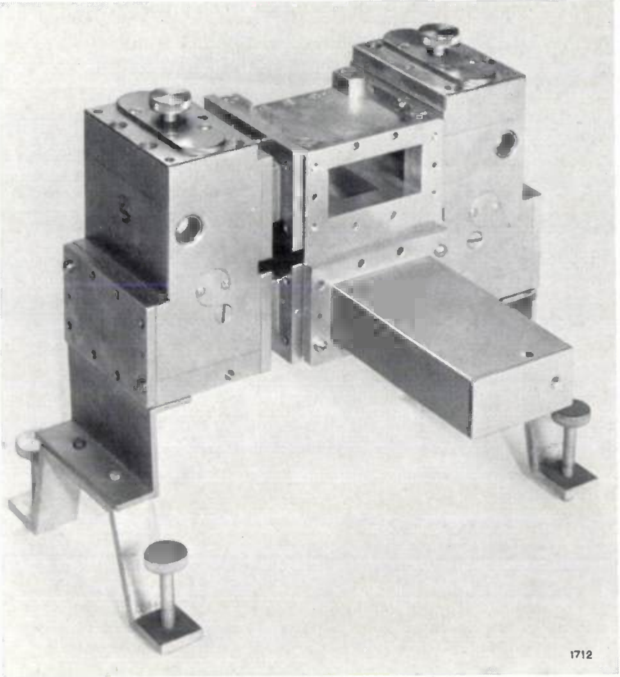


Fig. 20. Complete frequency modulator, built on the principle of fig. 17, using the block shown in fig. 19. The feedback paths of the two triodes are unequal in length and in part common. When the waves in the triodes are amplitude-modulated in push-pull, the output signal will be frequency-modulated.

To help shorten the feedback path, no isolators are fitted at the anode side of the amplifiers. This makes it necessary to neutralize the amplifier blocks, which is done by the method described above (fig. 14).

As a result of this special circuit arrangement, the output power of the frequency modulator is fairly high. As we have seen, the moduli of the output waves of $T_2$ are identical, so that the sum of the output powers of the amplifiers divides into two equal parts in $T_2$. Of these, one part is again split in $T_1$ into two equal parts, and the other is immediately available as output power. We may therefore expect the output power to be equal to that of a single tube driven by a signal strong enough to produce a gain of 3 dB. What the actual power will be depends on the setting of the modulator. The above reasoning was based on a phase difference of 90° between $\alpha$ and $\beta$. Plainly, by varying this angle we can play off the output power against the frequency deviation. If we make the angle larger we immediately increase the deviation, but from fig. 18 we see that

$\alpha + \beta$ then becomes smaller, that is to say, less power is fed back and the tubes will oscillate at a lower level. Conversely, a smaller angle gives a smaller deviation, but $\alpha + \beta$ is then larger, more power is fed back and the amplifiers oscillate more strongly.

We can make this process more effective if, at the same time, we vary the anode-circuit bandwidth of the amplifiers, depending on whether we want a large deviation or a large output power. *Table III* gives some results of measurements made at two settings representing roughly the extremes of deviation and output power.

Table III. Data on the frequency modulator in fig. 20, equipped with two EC 157 triodes, in two different settings.

| | For large frequency deviation | For high output power |
|---|---|---|
| Anode voltage | 200 V | 200 V |
| Anode current | 60 mA | 60 mA |
| Anode-circuit bandwidth | 150 Mc/s | 60 Mc/s |
| Centre frequency | 4000 Mc/s | 4000 Mc/s |
| Maximum frequency deviation | 60 Mc/s | 15 Mc/s |
| Linear frequency deviation | 20 Mc/s | 5 Mc/s |
| Output power | 0.6 W | 3.6 W |

## High-level detection

In millimetre radar, where pulses of extremely short duration are used (about 5 mμsec), a very large bandwidth (about 200 Mc/s) is required in the IF section and video amplifiers in the receiver [18]). A bandwidth as large as this is very difficult, if not impossible, to achieve with conventional tubes. As far as the IF amplifier is concerned, the difficulty is solved by using a travelling-wave tube, operating at a frequency of 4000 Mc/s. A normal crystal detector cannot, however, handle anything like the maximum output power of the travelling-wave tube without becoming overloaded. Considerable video amplification is therefore needed behind the detector in order to drive the cathode-ray tube.

This video amplification, which is very difficult to realize for such a large bandwidth, can be dispensed with entirely if the detection is done by an EC 157 triode, used as an anode-bend detector. The travelling-wave amplifier can boost the IF signal to about 1 W, and the triode detector, driven by this signal, delivers over the whole bandwidth video signals of about 10 V, which can be used for driving the cathode-ray tube directly. This arrangement is shown schematically in *fig. 21*.

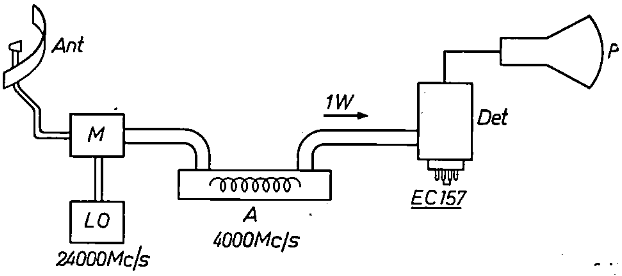


Fig. 21. The EC 157 used as an anode-bend detector for strong signals in millimetre radar. The IF amplifier *A* is a travelling-wave tube, which supplies a signal of about 1 W at 4000 Mc/s to the triode detector *Det*. The detector output signal is strong enough, at the required bandwidth, to drive the cathode-ray tube *P* directly, i.e. without video amplification. *Ant* antenna. *M* mixer. *LO* local oscillator (24 000 Mc/s).

An ordinary amplifier block can be used for such a detector; the top cover is removed and also the inner conductor and plunger, in order to minimize stray capacitances. The circuit used is shown in *fig. 22*. The tube is biased by a fixed positive

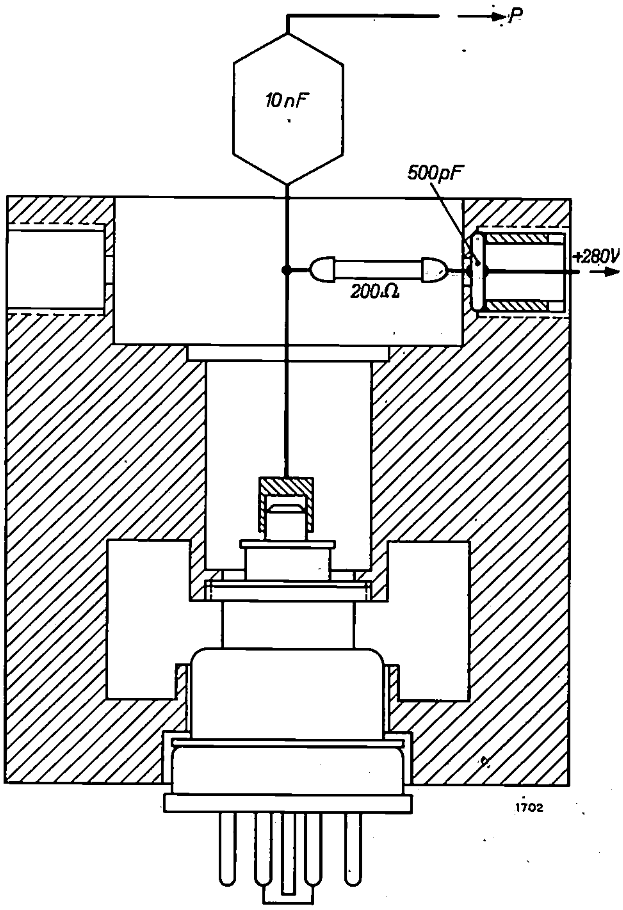


Fig. 22. Arrangement of EC 157 as anode-bend detector at 4000 Mc/s in a slightly modified amplifier block. The tube is biased to cut-off by a fixed voltage of +7 V on the cathode. The cathode-ray tube is connected at *P* (see fig. 21).

[18]) See R. J. Heijboer, Millimeterradar, Ingenieur 71, E. 41-E. 48, 1959 (No. 14) (in Dutch).

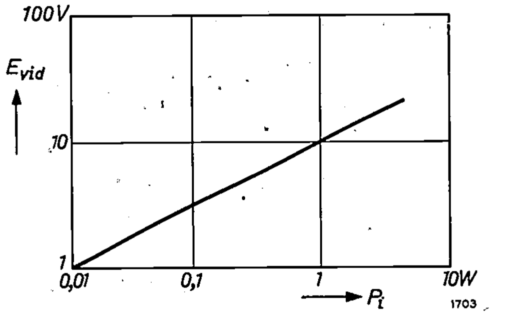


Fig. 23. Detection characteristic of the EC 157 used as an anode-bend detector in the arrangement shown in fig. 21: the video voltage $E_{\rm vid}$ on the anode is plotted as a function of the IF power $P_{\rm i}$ at the input.

potential of 7 V on the cathode, which corresponds roughly to the cut-off point. *Fig. 23* shows the video voltage on the anode as a function of the IF power at the input.

**Summary.** Because of such features as simplicity, favourable phase characteristics and large product of gain and bandwidth, disc-seal triodes for centimetre waves are suitable and attractive for many and various applications, especially in microwave radio links. Tabulated details are given of the microwave triodes developed or in development at Philips, and their uses are discussed as amplifiers, frequency multipliers, local oscillators of extremely stable frequency (4000 and 6000 Mc/s), mixers, limiters, amplitude and frequency modulators and high-level detectors for millimetre radar.

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

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

**2736:** W. Kwestroo and H. A. M. Paping: The systems BaO-SrO-$TiO_2$, BaO-CaO-$TiO_2$ and SrO-CaO-$TiO_2$ (J. Amer. Ceram. Soc. 42, 292-299, 1959, No. 6).

The solid phases formed at 1400 °C in air in the three-component systems BaO-SrO-$TiO_2$, BaO-CaO-$TiO_2$ and SrO-CaO-$TiO_2$ are described. Besides solid solutions of components with known structures, some new ternary compounds have been studied. The dielectric constants and loss factors of a number of specimens are given. Crystallographic data of the compounds $BaCaTiO_4$, $Ba_3Ca_2Ti_2O_9$ and $Ca_3Ti_2O_7$ and of the solid solution series $(Ba,Sr)_2TiO_4$ are presented. The preparation of the new compounds is described in detail.

**2737:** K. van Duuren, A. J. M. Jaspers and J. Hermsen: G-M counters (Nucleonics 17, No. 6, 86-94, 1959).

The article describes some modern designs of Geiger-Müller tubes. Diversity of design provides G-M tubes for many uses: hollow anodes for $4\pi$ and liquid-flow counting, compact coincidence units and integrator tubes for inexpensive, reliable monitors and alarm devices.

VOLUME 22, 1960/61, No. 2 pp. 29-60 Published 9th November 1960

Philips Technical Review

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

# A HIGH-SPEED SCANNING RADAR ANTENNA

by F. VALSTER.

621.396.677.7:621.396.965.

*The considerable mass of conventional radar antennae and the wind forces acting on their large surface area, are a hindrance to the rapid scanning desirable for certain purposes. In the design discussed below, based on a principle due to Rinehart, only one small part of the antenna system rotates, making it possible to achieve very high speeds of revolution.*

In plan-position radar it is frequently desirable to have a rapidly rotating beam. With conventional antenna constructions, however, this meets with difficulties due to the high moment of incrtia of the rotating mass, which consists of a primary radiator (dipole or horn) combined with a reflector or lens. Further difficulties are occasioned by the varying wind forces on the large revolving surfaces.

Attempts have therefore been madc to design radar antenna which dispense with the need for large rotating systems. The various methods devised to overcome this problem are all based on the Luneburg lens, which we shall presently discuss. An experimental antenna of this kind has also been built at Philips, with the object of studying the effect of the scanning speed on the radar picture.

In the following we shall first discuss some theoretical considerations relating to the speed of revolution of a radar beam, after which we shall describe the design of the present antenna and some results achieved.

## Reasons for a high scanning speed

A high scanning speed in radar may be desirable for two reasons: firstly, to obtain a clear display of fast-moving objects, and secondly, for the sake of a bright radar picture. We shall examine these reasons in more detail.

### *Clear display of fast-moving objects*

If the beam rotates slowly, a fast-moving object travels so far in the time taken by the antenna to complete one revolution that the radar echo on the P.P.I. screen is seen to move in a series of jumps. A jerky echo is difficult to distinguish from the irregular brightness variations due to noise, which are always present on the screen, and is also difficult to follow if there are a lot of permanent echoes. The latter was the case, for example, with one of our short-range radar sets, where the echoes of cars moving along a street were very hard to follow between the numerous permanent echoes from the street itself.

The optimum speed of revolution depends on the speed $v$ of the observed object, on the radius $R$ of the area covered on the radar screen, and on the radius $r$ of the screen (possibly also on the diameter of the luminous spot). If $d$ be the distance over which the echo of the object moves during one revolution of the beam and $N$ the speed of revolution, the following relation holds:

$$N = \frac{r}{R}\frac{v}{d}. \qquad \ldots\ldots (1)$$

The maximum value of $d$ at which the irregular displacement of the echo is not yet troublesome is about 0.25 mm. If the screen has the usual diameter of 30 cm, it follows from (1) that:

$$N = \frac{v}{6R}\ \text{r.p.s.}, \qquad \ldots\ldots (2)$$

where $R$ is expressed in km and $v$ in km/h. This relation is shown graphically in *fig. 1* for various values of $R$.

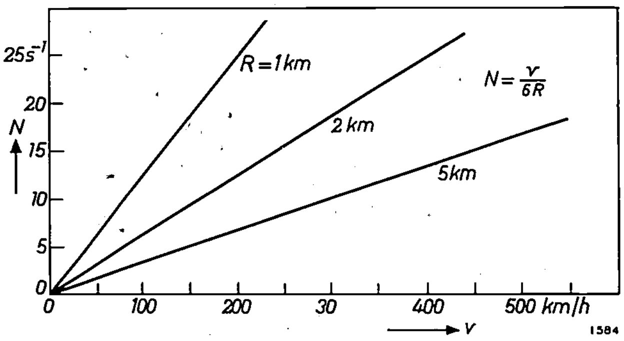


Fig. 1. The speed of revolution $N$ of the beam for which the echo on a radar screen of 30 cm diameter moves 0.25 mm per revolution, is plotted as a function of the speed $v$ of the object, for various values of the radius $R$ of the area covered.

### *Bright radar picture*

In some cases the radar picture is required to be bright enough to make it unnecessary to darken the room. Phosphors giving this brightness — and which are used in television picture tubes, for example — have a very short afterglow, however, and therefore the radar pattern must be traced many more times per second, i.e. fast scanning is necessary. The use of a television picture tube for radar would call for a minimum scanning speed of 20 r.p.s.

### *Operational consequences of high scanning speeds*

A high scanning speed is not without its effects on the operation of the radar system. The faster the beam is rotated the more pulses must strike the objects per second if weak echoes are to be clearly distinguished from random noise fluctuations. In other words, a high scanning speed implies a high pulse-repetition frequency, and this involves difficulties with the modulator. Moreover, a high pulse repetition frequency $f$ limits the radius $R$ of the area covered, since a fresh pulse must not be sent out before the echoes of the last pulse have been received from objects at a distance $R$. This restricts the radius $R$ to $c/2f$, where $c$ is the speed of light.

## The Luneburg lens

Scanning speeds of, say, 10 r.p.s. for small radar sets and 1 r.p.s. for large installations are evidently impracticable with conventional antenna systems — quite apart from the high power they would entail for the drive.

The solution of this problem is to be found — in principle at least — in a lens possessing rotational symmetry and which focusses the energy emitted by the primary radiator into the required beam. All that need then revolve (around the lens) is the primary radiator.

A familiar example of such a lens is the Luneburg lens [1]). This is spherical in shape and its refractive index $n$ depends only on the distance $r$ to the centre point, according to the following function:

$$\left.\begin{aligned} n(r) &= \sqrt{2-\left(\frac{r}{a}\right)^2} && \text{for} \quad r \leqq a \\ n(r) &= 1 && \text{for} \quad r \geqq a \end{aligned}\right\}, \quad . \quad (3)$$

where $a$ is the radius of the lens. The refractive index thus changes continuously from $\sqrt{2}$ in the centre of the lens to 1 at the surface ($r = a$). A property of this lens is that every point $O$ of its surface is imaged at infinity in the direction diametrically opposed to $O$ (*fig. 2*). Rays entering the lens from $O$ thus emerge as a parallel beam.

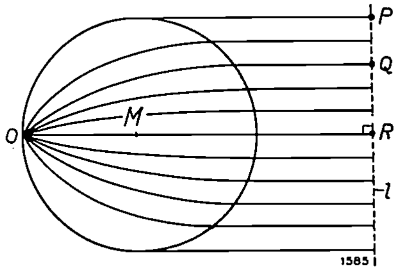


Fig. 2. The Luneburg lens has the property that every point $O$ on its surface is imaged at infinity in the direction diametrically opposite to $O$.

A lens having a refractive index in accordance with (3) is obviously very difficult to make and, as far as we know, never has been made. The difficulty can be circumvented, however, with a two-dimensional analogue of the lens (i.e. one which forms a beam in one plane) such as described by Rinehart [2]) and later calculated as one of a family of lenses of revolution by Toralda di Francia [3]). Our lens, too, is an analogue of this type.

Let us take another look at fig. 2. In order to produce plane wave fronts $l$ (perpendicular to the projection of the radius $OM$) from the spherical wave fronts issuing from $O$, the rays from $O$ to the

[1]) R.K. Luneburg, Mathematical theory of optics, Brown University Press, Providence (Rhode Island, U.S.A.) 1944, p. 213.

[2]) R. F. Rinehart, A solution of the problem of rapid scanning for radar antennae, J. appl. Phys. **19**, 860-862, 1948. J. S. Hollis and M. W. Long, A Luneburg lens scanning system, Trans. Inst. Radio Engrs. **AP-5**, 21-25, 1957 (No. 1). For another solution, see G. D. M. Peeler and D. H. Archer, A two-dimensional microwave Luneburg lens, Trans. Inst. Radio Engrs. **AP-1**, 12-23, 1953 (No. 1).

[3]) G. Toralda di Francia, A family of perfect configuration lenses of revolution, Optica Acta **1**, 157-163, 1954/55.

straight line $l$ must all, according to geometrical optics, have the same optical length. This is the case in the Luneburg lens, the rays passing through a medium of higher refractive index the shorter their geometrical length. The rays $OP$, $OQ$ and $OR$ thus have the same optical length, in other words the optical length

$$\int_{O}^{l} n \, \mathrm{d}s$$

is independent of the path followed.

Paths of equal optical length between the point $O$ and the line $l$ are also produced if the rays can be given equal geometrical length and made to pass entirely through a medium with $n = 1$; in that case a medium with a refractive index varying according to (3) is not necessary. Thus, instead of letting the rays $OQ$ and $OR$, for example, travel partly through an optically denser medium than $OP$, we let each of them make a detour such that $OQ$ and $OR$ get the same geometrical length as $OP$. If $n = 1$ along all rays, they have the same optical as well as the same geometrical length, and therefore $l$ again forms a wave front. This is only possible, however, for beaming in two dimensions, the third dimension being required for the detour.

## Form of the antenna

In the foregoing we have tacitly assumed that the energy can indeed be propagated along a curved surface. Assuming that this is possible (we shall return to this presently), we see that the lens from which we started becomes a sort of waveguide, which has the special property of converting circular wave fronts from a point source into linear wave fronts and has the advantage above the lens that it does not require a medium of varying refractive index.

The surface of the waveguide must meet the following requirements:

*a*) It must possess rotational symmetry.

*b*) All geodesics from a point on the circumference must have the same length from that point to a straight line perpendicular to the projection of the radius vector of the point; we can regard this line as the aperture of the antenna.

(A geodesic is the shortest superficial line between two points on any specified surface, and, according to Fermat's principle of least-time, this is the path along which the energy will propagate between those two points.)

The surfaces that satisfy these conditions can be found by the methods of differential geometry. *Fig. 3a* gives a sketch of such a surface, and fig. 3*b* a meridian cross-section of the surface. The equation for the meridian curve is expressed most simply by the coordinates $\varrho$ and $s$, where $\varrho$ is the direct distance from a point $P$ to the axis of symmetry, and $s$ is its distance *along* the meridian curve to the point $A$ (fig. 3*b*). The equation is then (see the article by Rinehart, footnote [2]):

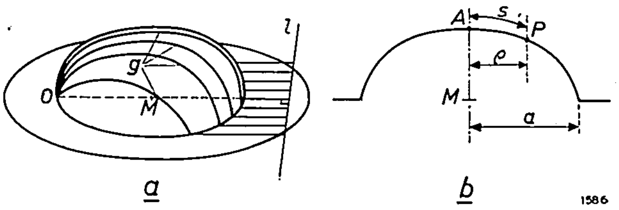


Fig. 3. *a*) In the hat-shaped surface illustrated, all geodesics $g$ connecting the point $O$ to the straight line $l$ (perpendicular to $OM$) are of equal length. The surface also possesses rotational symmetry.
*b*) Meridian cross-section of the surface shown in (*a*).

$$s(\varrho) = \tfrac{1}{2}(\varrho + a \arcsin \varrho/a) \text{ for } 0 \leqq \varrho \leqq a. \qquad (4)$$

For $\varrho > a$ the surface defined by (4) gives way to a contiguous horizontal edge.

## A model demonstrating the generation of the surface

Before going into the practical version of the antenna we shall discuss a model which demonstrates how a surface fulfilling the above-mentioned conditions can be generated. This model also helps to elucidate the points discussed above.

*Fig. 4a* shows a rectangular board *1* in which a circular hole is cut, of radius $a$. Fitted in this hole is an inflatable balloon *2*. A number of cords *3*, of equal length, are knotted together at one side, and the knot is fixed to a point $O$ on the circumference of the hole. The free ends are then secured to the board in such a way that they lie on a straight line $l$ which is perpendicular to the radius vector of point $O$ (while this is done the balloon is not yet inflated and the cords are still slack). Fitted around the hole is a ring *4*, mounted in such a way that the cords can slide freely under it and at the same time are made to pass along the board from the balloon to the straight line $l$.

When the balloon is inflated, the cords are tensioned, and they slip into positions over the surface of the balloon such as to allow the balloon to expand as much as possible; in other words, each delineates its shortest path (the geodesic path) over the balloon. When this state has been reached, the balloon and the board together form a surface on which all geodesics from the point $O$ to the straight line $l$ are of equal length.

The circular hole by itself is insufficient to guarantee that the surface thus produced will possess rotational symmetry. To make it fulfil this condition, more than simply the one set of cords issuing from point $O$ is required; other sets must be applied at points distributed around the circumference of the hole. Fig. 4*b* shows an arrangement using three sets of cords.

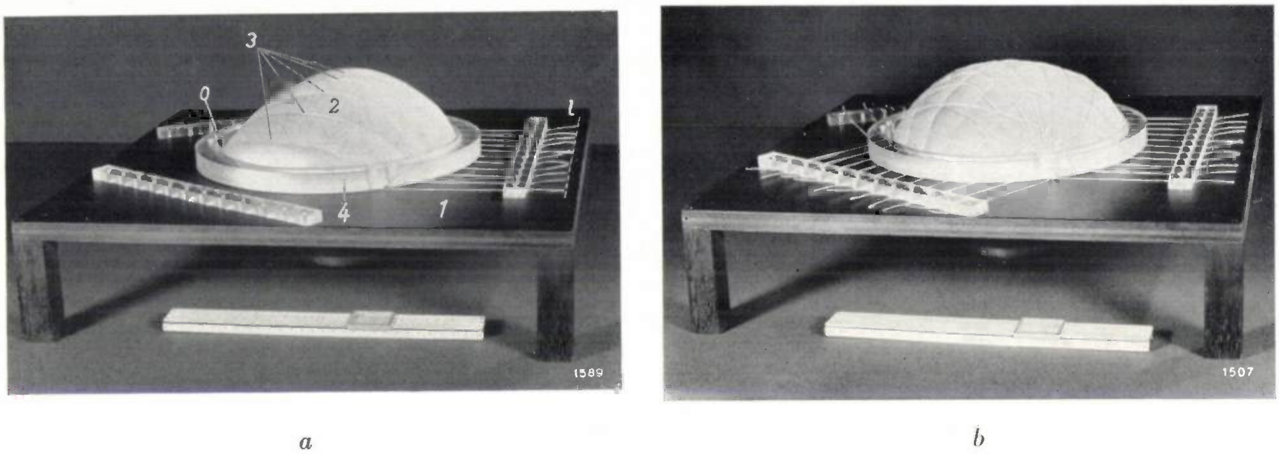


Fig. 4. *a*) A model illustrating the generation of a surface on which all geodesics between a point $O$ and a straight line $l$ (perpendicular to the radius vector of $O$) are of equal length. *1* board with circular hole. *2* inflated balloon. *3* cords passing partly over the balloon and partly along the board, and which all have the same length from the point $O$ to the straight line $l$. *4* ring, under which the cords pass freely and which holds the cords to the board surface between the balloon periphery and line $l$.
*b*) Three sets of cords are used here, producing a surface that better approaches rotational symmetry.

## Construction of the antenna

We have still been working on the assumption that the energy *can* be propagated along a curved surface such as that in fig. 3. We shall now consider how this can be achieved, in other words, how a curved waveguide with the required profile can be constructed.

We shall consider first the familiar case of wave propagation between two parallel, flat plates of a material that is a good conductor. If we choose the distance $2d$ between the plates smaller than $\frac{1}{2}\lambda$, where $\lambda$ represents the wavelength in free space, the energy has only one possible mode of propagation, viz. the TEM mode, in which the electrical field strength is perpendicular to the plates, and the velocity of propagation is equal to that in free space ($n = 1$, which is what we wanted).

We now apply the same principle to the curved surface. In order to compel the energy to propagate along the required lens surface, we use two metal plates which have virtually the same form as this surface, but are separated from it at either side by a distance $d$ (measured perpendicular to the surface). If we again choose the distance $2d$ between the plates smaller than $\frac{1}{2}\lambda$, the deviation in our curved waveguide from the mode of vibration occurring in the flat case is negligible, provided the radius of curvature is everywhere large in relation to $\lambda$. This condition is easily fulfilled.

The question now is how the two metal plates can be given the correct shape. The first step is to calculate exactly the meridian curve of the geometrical surface possessing the above-mentioned properties (rotational symmetry, geodesics of equal length). For this purpose it is convenient to change from the coordinates $\varrho$ and $s$ in eq. (4) to rectangular coordinates. On the outside of the meridian thus found we now construct the curve that remains a constant distance $d$ away from it (perpendicular to the meridian), and on the inside a curve which remains a constant distance $d + \delta$ therefrom, $\delta$ being the thickness of the sheet metal to be used, e.g. 1.5 mm. The curves produced are then the meridians (profiles) of the two dies in which two sheet metal plates can be forced into the correct shape (*fig. 5*).

To achieve reasonable accuracy with this graphic construction of the die profiles (deviation from the ideal profile nowhere more than e.g. 1% of the diameter $2a$) it is necessary to use meridian curves several times larger than the required dies. From a number of points on this enlarged meridian we then draw perpendiculars to the curve, and mark on each a distance $d$ on the one side and a distance $d + \delta$ on the other. In this way we find the profiles of the dies. The coordinates of these profiles can also be computed by carrying out the above process analytically. This involves complicated calculations, which, without the help of an electronic computer, would be far too time-consuming .

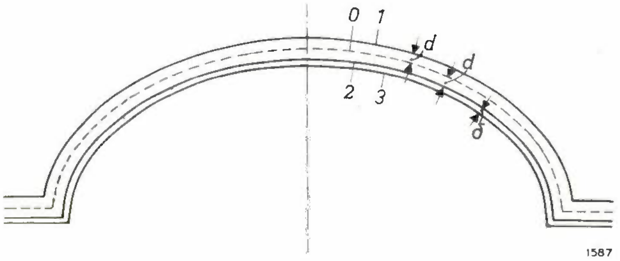


Fig. 5. Curve *0* is the calculated meridian curve of the geometrical surface possessing the required properties. Curves *1* and *2*, at a distance $d$ ($< \frac{1}{4}\lambda$) at either side of *0*, are meridians of the metal surfaces to be made. Curve *3* lies at a further distance $\delta$, equal to the thickness of the sheet metal used. *1* and *3* are the meridians of the jigs in which the metal sheeting will be stamped into the appropriate shape. (For clarity, the distances $d$ and $\delta$ are exaggerated in comparison with the other dimensions.)

The two plates, shaped like hats, can be seen in *fig. 6*; the upper plate is marked with geodesics and wave fronts.

Further constructional details are illustrated schematically in *fig 7*. The two aluminium "hats" are kept at the right separation $2d$ by spacer blocks of "Polyfoam", a material whose dielectric constant differs very little from unity and thus behaves almost like air for the purposes of wave propagation. The brims of the two hats are not flat and parallel, but slightly tapered, giving a certain beaming effect in the vertical plane.

The whole assembly is supported from below by a central bushing *8* and from above by a tubular structure (*fig. 8a* and *b*). The rectangular waveguide *4* (fig. 7) is rotated by a small motor; the mouth of the waveguide constitutes the point source and revolves in the gap between the brims of the two hats. The fixed waveguide *7* connects the waveguide *4*

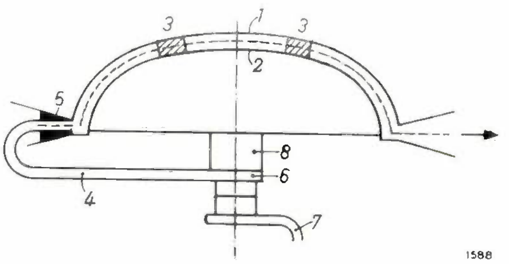


Fig. 7. Cross-section of the antenna system. *1* and *2* are the two hat-shaped plates, with spacers *3* of "Polyfoam". *4* rotating waveguide. *5* wedge-shaped block, preventing back radiation. *6* rotary joint between the rotating waveguide *4* and the fixed waveguide *7*, which connects the system to the transmitter and receiver. *8* bushing supporting the construction.

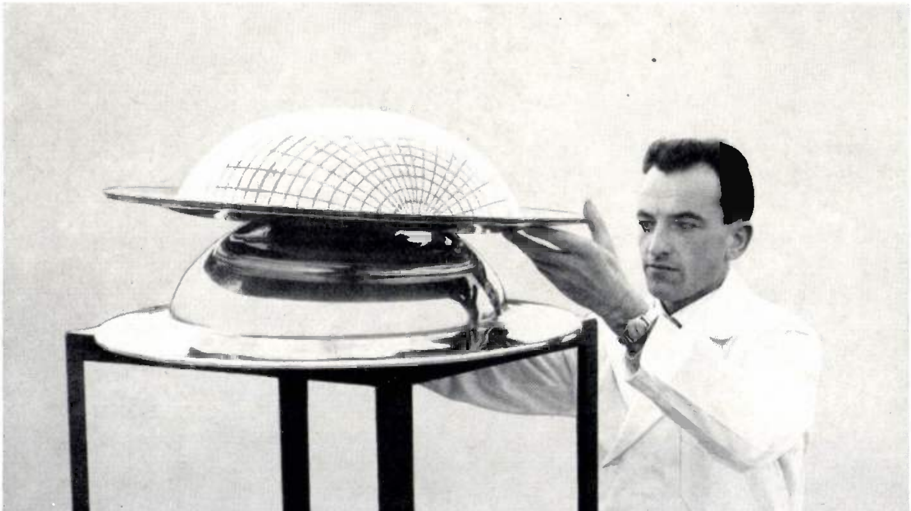


Fig. 6. The two hat-shaped surfaces, made from sheet aluminium. Geodesics and wave fronts are drawn on the upper surface.

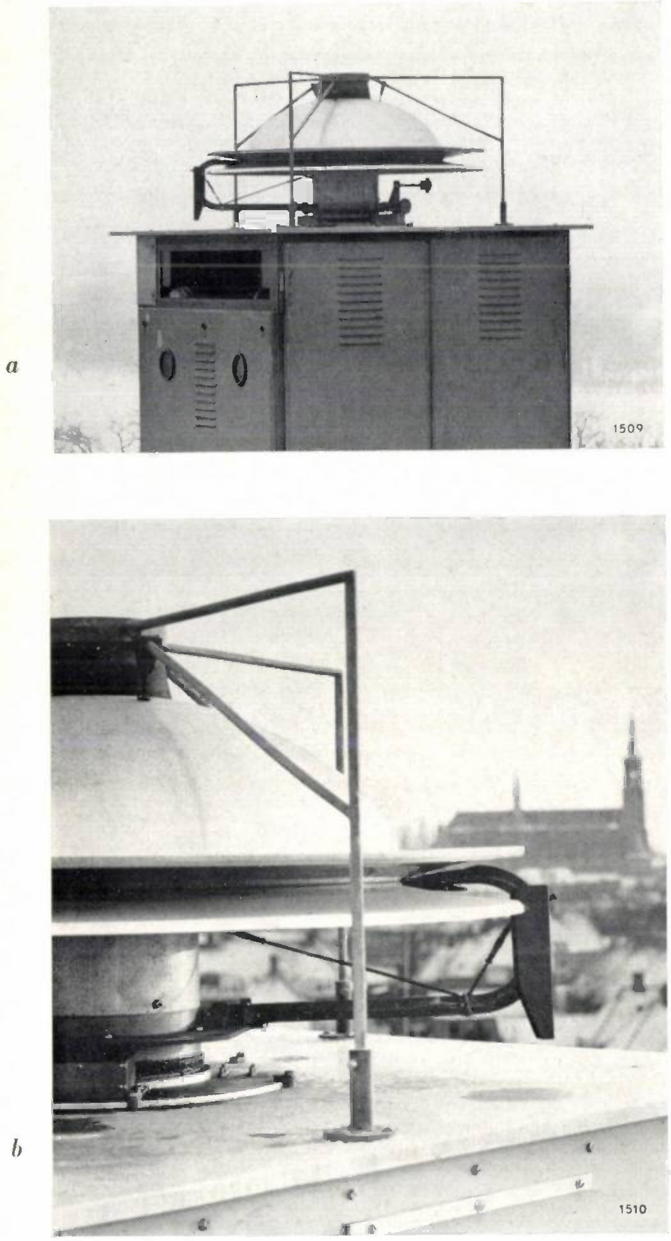


Fig. 8. *a*) Mounting of the antenna on the housing containing the radar transmitter and receiver.
*b*) Close-up view of antenna.

to the radar transmitter and receiver via the rotary coupling *6*. Back radiation is prevented by a wedge-shaped block *5* fixed around the mouth of waveguide *4*. It consists partly of aluminium with one or more grooves one quarter wavelength deep, functioning as a "choke", and partly of graphite-filled "Philite", which absorbs any residual radiation.

**Performance**

The antenna shown in fig. 8 forms part of an experimental radar installation operating on a wavelength of 1.25 cm. The pulses have a duration of 1/30 µs and a repetition frequency of 12 500 c/s. The diameter of the antenna (i.e. without the brim) is 50 cm; the distance $2d$ between the two "hats" is 0.5 cm, i.e. smaller than $\frac{1}{2}\lambda$.

The directional pattern in the horizontal plane shows a main lobe whose beam width $\Theta_\mathrm{h}$ between the 3 dB directions, according to a rule of thumb, should be $\lambda/D$ radians ($D$ is the length of the antenna, in this case the length $2a$ of the antenna aperture); this rule of thumb, then, gives $\Theta_\mathrm{h} = 1.25/50$ radians $= 1.43°$. Since the hats show slight deviations from rotational symmetry, the directional pattern was found in fact to be somewhat dependent on the position of the energy feed. The average value of $\Theta_\mathrm{h}$ was found by experiment to be 1.5°, which is thus in good agreement with the expected result.

The size of the side lobes depends on the radiation pattern of the primary radiator, which determines the amplitude distribution of the field in the antenna aperture. With the design chosen the level of the largest side lobes proved to be 18 dB below that of the main lobe, which is a sufficient difference for the purpose in view.

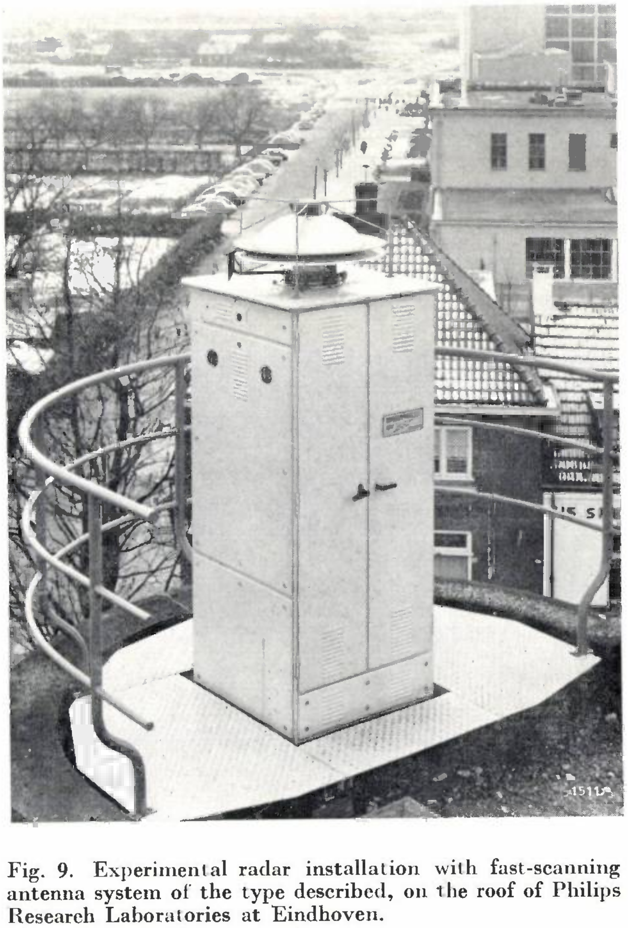


Fig. 9. Experimental radar installation with fast-scanning antenna system of the type described, on the roof of Philips Research Laboratories at Eindhoven.

The angle between the tapered brims of the hats is so chosen as to give the beam in the vertical plane an angular width $\Theta_V \approx 25°$.

The scanning speed of the antenna is variable up to a maximum of 10 revolutions per second. At the maximum speed the motor draws a power of only 200 W.

The radar system (*fig. 9*) using this antenna gives a display of an area in which streets with fairly heavy traffic can be seen. When the beam rotates at 10 r.p.s. the moving vehicles are indeed much more clearly distinguishable between the numerous permanent echoes than when the beam is rotated at the more usual speed of 1 or 2 r.p.s.

**Summary.** A high scanning speed (e.g. 10 r.p.s.) is desirable in plan-position radar when 1) the echo of fast-moving objects is to be followed clearly against a background of noise and numerous permanent echoes and 2) when a very bright radar picture is wanted, which is only realizable with short-persistence screens. Conventional radar antennae are not adapted to high-speed scanning. A rapidly revolving beam is obtainable, however, by using an analogue of the two-dimensional Luneburg lens, as described by Rinehart. This analogue is based on a geometrical property of a rotationally-symmetric hat-shaped surface whereby circular wave fronts in a given plane emanating from a point source are converted into linear wave fronts. The practical form of this idea consists of two metallic "hats", situated at a distance of less than one quarter wavelength on either side of the surface in question. The two hats constitute a waveguide. The point source is the mouth of a rectangular waveguide which revolves in the gap between the two brims; at the opposite side the radiation emerges in the form of a narrow beam. Some constructional details of an antenna of this type are given. In one version, for a wavelength of 1.25 cm, the main lobe has an average 3 dB beam width in the horizontal plane of 1.5°, and in the vertical plane of about 25°. A motor power of 200 W is sufficient for a speed of revolution of 10 r.p.s. The expectation that fast-moving objects are better observable at this speed than at e.g. 1 r.p.s., is confirmed.

*In October 1959, Professor Dr. Balthasar van der Pol died at Wassenaar at the age of 70. In the years from 1922 to 1949, during which he acquired international repute in the field of radio science and engineering, he directed the work done in that field in Philips Research Laboratory at Eindhoven. In the same period he also played a prominent part in international organizational activities concerned with radio communications. After his retirement from industry he continued in these activities, principally as Director of the C.C.I.R. (Comité Consultatif International des Radiocommunications) at Geneva, an office he held until a few years before his death. Amongst the numerous distinctions conferred upon him were the Medal of Honor of the Institute of Radio Engineers, of which he was for some time vice-president, the Valdemar Poulsen Gold Medal, several honorary doctorates, and the honorary presidency of the U.R.S.I. (Union Radio Scientifique Internationale).*

*The obituary notices that appeared in journals both in the Netherlands and abroad were obviously unable to go far into the details of his scientific work. It gives the editors pleasure to be able to publish here a more comprehensive appreciation, written by Prof. Bremmer, one of Van der Pol's former and closest collaborators and probably the most familiar with his work. In our view such an appreciation is a fitting way of paying tribute to the memory of this remarkable man, and at the same time it gives the reader a glimpse of many important aspects of the evolution of radio science.*

*The title photograph shows Van der Pol (right) in conversation with Prof. Holst, the founder and first director of this laboratory.*

# THE SCIENTIFIC WORK OF BALTHASAR VAN DER POL

by H. BREMMER.

538.56.001.5:621.37.001.5

In this article we shall try to review the many-sided scientific work done by the late Prof. Dr. Balthasar van der Pol, former member of the Philips Research Laboratories, Eindhoven, who died on the 6th October 1959. His numerous investigations were concerned primarily with problems of radio science and associated fields. The international name he soon acquired was due no doubt in the first place to his pioneering work on the *propagation of radio waves* and on *non-linear oscillations*. We shall see later how his study of these subjects led him to take an interest in many other problems, including problems of pure mathematics.

Van der Pol was one of those rare men who are equally at home in physics, engineering and mathematics. This enabled him to deal personally with

*In October 1959, Professor Dr. Balthasar van der Pol died at Wassenaar at the age of 70. In the years from 1922 to 1949, during which he acquired international repute in the field of radio science and engineering, he directed the work done in that field in Philips Research Laboratory at Eindhoven. In the same period he also played a prominent part in international organizational activities concerned with radio communications. After his retirement from industry he continued in these activities, principally as Director of the C.C.I.R. (Comité Consultatif International des Radiocommunications) at Geneva, an office he held until a few years before his death. Amongst the numerous distinctions conferred upon him were the Medal of Honor of the Institute of Radio Engineers, of which he was for some time vice-president, the Valdemar Poulsen Gold Medal, several honorary doctorates, and the honorary presidency of the U.R.S.I. (Union Radio Scientifique Internationale).*

*The obituary notices that appeared in journals both in the Netherlands and abroad were obviously unable to go far into the details of his scientific work. It gives the editors pleasure to be able to publish here a more comprehensive appreciation, written by Prof. Bremmer, one of Van der Pol's former and closest collaborators and probably the most familiar with his work. In our view such an appreciation is a fitting way of paying tribute to the memory of this remarkable man, and at the same time it gives the reader a glimpse of many important aspects of the evolution of radio science.*

*The title photograph shows Van der Pol (right) in conversation with Prof. Holst, the founder and first director of this laboratory.*

# THE SCIENTIFIC WORK OF BALTHASAR VAN DER POL

by H. BREMMER.

538.56.001.5:621.37.001.5

In this article we shall try to review the many-sided scientific work done by the late Prof. Dr. Balthasar van der Pol, former member of the Philips Research Laboratories, Eindhoven, who died on the 6th October 1959. His numerous investigations were concerned primarily with problems of radio science and associated fields. The international name he soon acquired was due no doubt in the first place to his pioneering work on the *propagation of radio waves* and on *non-linear oscillations*. We shall see later how his study of these subjects led him to take an interest in many other problems, including problems of pure mathematics.

Van der Pol was one of those rare men who are equally at home in physics, engineering and mathematics. This enabled him to deal personally with the theoretical-mathematical treatment of technical problems, and also to meet professional mathematicians on their own ground and persuade them of the importance of a rigorous mathematical approach to such problems. He saw how difficult it often is to bring together the practitioners in these three fields. His success in this respect was a valuable facet of his life's work.

As a mathematician, Van der Pol showed a strong preference for heuristic methods, by which, along partly intuitive lines, results are readily arrived at that can only be verified later. To this extent he resembled Heaviside. This brilliant man, who lived from 1850 to 1925, used mathematical methods which proved to be very fruitful, but generally left it to others to demonstrate their validity. Heaviside was nevertheless the first to get to the bottom of such technical problems as the role of inductances in long-distance telephone cables. Van der Pol was a great admirer of this self-made scientific genius, as testified by his inaugural address on 8th December 1938 as extra-mural professor at Delft [1]). The address was entirely devoted to Heaviside, and underlined the significance of his work at a time when the application of Maxwell's equations was not yet a commonplace. The following sentence from Van der Pol's address will serve to illustrate the part played by Heaviside in the development of electrotechnical concepts that are now thoroughly familiar:

> "Electrotechnology and physics are also indebted to Heaviside for the concept and the word "impedance", not merely as understood in present-day alternating-current theory, but in a very much wider sense which includes on-off switching effects, a form which even today has still not been dealt with exhaustively."

In all his work Van der Pol owed much to the inspiration of Heaviside. This appears directly from his preoccupation with the methods of *operational calculus*, which were created by Heaviside for application to electrical networks. No one has explained better than Van der Pol the merits of operator methods. In particular he demonstrated their value for the discovery of numerous relations in widely diverse fields of mathematics. In one respect, however, there was a marked constrast between Heaviside and Van der Pol. The former suffered all his life from the difficulty of getting his ideas accepted, not least because of the almost illegible form of his writings. The latter's publications, on the other hand, were models of careful exposition and clarity of thought.

Van der Pol's lucidity was also much in evidence in the lectures he gave and in the innumerable meetings and discussions over which he presided. In this connection, mention must be made of his considerable talents as an organizer. The ease with which he could present a matter, and his complete familiarity with the whole field of radio, soon made him a leading figure in international organizations concerned with radio science and engineering, especially in the U.R.S.I. and the C.C.I.R. It was no surprise when, after retiring from Philips Research Laboratory, he was invited to be the first director of the C.C.I.R. upon the establishment of its permanent secretariat. In the time that he held this office (from 1949 to 1956) he worked with devotion to ensure that justice was done to pure scientific research in the advice issued on the preparation of technical regulations for international radio communications. It is scarcely believable that, with all his activities on international committees, he should still have found the time for abstract investigations. He succeeded in this only because his zest for work was quite out of the ordinary, and remained with him to the end of his life.

Van der Pol maintained that only a few great men have had a decisive bearing on the development of physics. He used to say that, faced with the enormous number of scientific investigations, it was easy to forget that the really fundamental work is contained in a mere handful of publications. He himself was therefore strongly drawn towards personal contact with the leading figures in the world of science. In that respect he was fortunate at the very beginning of his career. After taking his degree *cum laude* at Utrecht in 1916, he did experimental research in England until 1919. At first he worked in London under Fleming, the man who, in 1904, was granted the first patent on a diode. There followed a period of two years at Cambridge, where he worked under Sir J. J. Thomson. On his return to the Netherlands he was attached for three years to the Teyler Foundation at Haarlem, where he worked under Lorentz. Thus, Van der Pol was in close association with two scientists, one of whom he himself wittily described as the "discoverer" of the electron (in 1897 Thomson had determined the ratio $e/m$ from electron orbits in a combined electrical and magnetic field) and the other as the "inventor" of the electron (in 1896 Lorentz had calculated $e/m$ from the Zeeman effect). In the years that followed, Van der Pol was in close contact with other distinguished scientists, including Appleton, who was awarded the Nobel prize in 1947 for his pioneering research on the ionosphere.

[1]) B. van der Pol, Oliver Heaviside (1850-1925), Ned. T. Natuurk. 5, 269-285, 1938.

### Propagation of radio waves

In England Van der Pol soon came up against a problem which was to remain one of his great lifelong interests. We refer to the effect of the earth on the propagation of radio waves. In 1909 Sommerfeld [2] had put forward a theory which explained the observed phenomena on the assumption that the earth's surface could be regarded as flat. As early as 1901, however, Marconi had succeeded in sending radio signals across the Atlantic Ocean. In 1915, after experience had been gained in the use of undamped waves, it had even proved possible to establish telephonic communication by radio across the Atlantic. It was evident that the curvature of the earth must have some effect on the propagation of the waves, because linear propagation straight through the highly absorbent earth would attenuate the signal beyond possibility of detection. It was conceivable that, as a result of diffraction, the waves might follow the surface of the earth far beyond the horizon as optically observed from the transmitting aerial. The existing theory on this idea was criticized by Poincaré, who showed in 1910 [3] that, in the case of diffraction, the field over great distances must decrease proportionally to exp $(-\beta D/\lambda^{\frac{1}{3}})$, where $\lambda$ is the wavelength and $D$ the distance from the transmitter to the receiver measured over the earth's surface. Unfortunately, it was not at that time possible to test this theory, since there was no way of deriving a reliable numerical value for the constant $\beta$.

This was roughly the situation when Van der Pol first came into touch with the propagation problem. Unless the value of $\beta$ was extremely small, the field upon diffraction would be so strongly attenuated that there could only be one other explanation for the long range of the radio waves: the presence of the ionosphere. The existance of conducting layers at high altitudes in the atmosphere, known as the ionosphere, was postulated in 1902 by Kennelly and Heaviside to account for the propagation of waves over great distances. The radio waves would then follow a zig-zag path, being reflected successively from the ionosphere and from the surface of the earth.

The mathematical difficulties of the pure diffraction problem, assuming the absence of the ionosphere, were bound up with the numerous Bessel functions occurring in the solution. This induced Van der Pol to encourage the mathematician Watson to study this problem, at a time when the latter was working on his well-known book on Bessel functions, which appeared in 1922. Watson's results, which were published in 1918 [4], proved beyond doubt that the value of $\beta$ was too large (it would be 0.00376 km$^{-\frac{2}{3}}$ if the earth were a perfect conductor and the atmosphere completely homogeneous) to explain the long range of radio waves without invoking the help of the ionosphere. The implications of Watson's calculations were discussed by Van der Pol in a paper published in 1919 [5].

This did not mean, however, that further study of the pure diffraction problem, leaving the ionosphere out of account, was no longer necessary. In the first place the wave propagated along the earth, the "ground wave", makes an important contribution over short distances, particularly during the day, compared with the contribution from the "sky wave" which is propagated via the ionosphere. Moreover, the ionosphere has little effect on the extremely short waves used in television, frequency-modulated transmissions and radar. The further study of the ground wave was therefore still of importance. In cooperation with K. F. Niessen, Van der Pol elaborated on Sommerfeld's theory [6], which could still be used for short distances, and made a special study of the influence of the height of the transmitting and receiving aerials above the ground. In later years the ground-wave theory was worked out by Van der Pol and the present author for distances at which the earth's curvature plays an essential part [7]. A mathematical method was developed for computing the fields of the ground wave for all topographical conditions. It was found that, even in the case of very short waves, the decrease of the field strength beyond the transmitter horizon was very much more gradual than had formerly been thought. This is illustrated in *fig. 1*, where the ratio of the actual field strength $E$ on the earth's surface to the field strength $E_{\mathrm{pr}}$ in free space is plotted as a function of the distance from transmitter to receiver for a given case, viz. a transmitter at a height of 100 metres above the earth's surface and surface conditions corresponding to dry ground. The various curves show this ratio for four different wavelengths (differing successively from the other by a factor of 10) and for the limiting case $\lambda \rightarrow 0$. A distinct shadow effect, whereby the field strength is reduced by a factor

[2] A. Sommerfeld, Ann. Physik **28**, 665, 1909.

[3] H. Poincaré, Jahrb. drahtl. Telegr. Teleph. **3**, 445, 1910.

[4] G. N. Watson, Proc. Roy. Soc. A **95**, 83, 1918.

[5] B. van der Pol, Phil. Mag. **38**, 365, 1919.

[6] B. van der Pol and K. F. Niessen, T. Ned. Radiogen. **7**, 1, 1935.

[7] B. van der Pol and H. Bremmer, Phil. Mag. **24**, 141 and 825, 1937.

of at least 4 within a distance of 5 km beyond the horizon, does not appear until the wavelength has dropped to about 7 mm.

The numerical results of the ground-wave propagation theory are of particular importance at wavelengths smaller than about 10 m, in which case the heights of the transmitting and receiving aerials above the earth's surface have a considerable nfluence. During Van der Pol's term of office with the C.C.I.R. a wide programme of computations was decided upon in order to get numerical data on ground-wave propagation, including the effect of refraction in the lowest layer of the atmosphere. The results were published in the form of graphs [8]), an example of which is shown in *fig. 2*; this refers to a wavelength of 5 m, again for representative ground conditions (conductivity of 0.01 mho/m and a relative dielectric constant of 10). The graph indicates, for example, the extent to which the field increases with increasing height $h_2$ of the receiver for a fixed transmitter height $h_1$ of 10 m. The results are the same if the transmitter and receiver heights are interchanged.

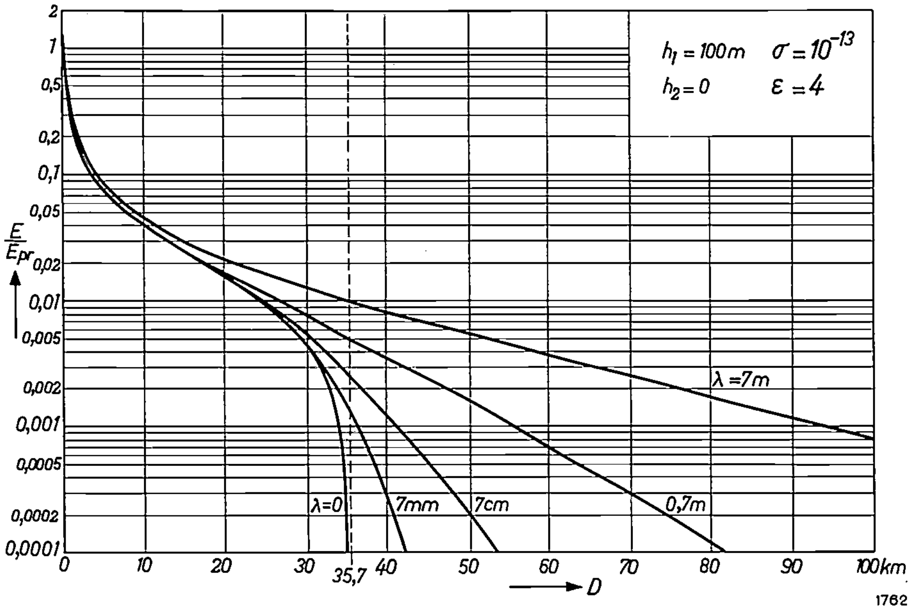


Fig. 1. Attenuation of a radio wave propagated along the curved surface of the earth (ground wave) for various wavelengths $\lambda$, calculated for dry ground and transmitter height $h_1 = 100$ m, receiver height $h_2 = 0$ m. The ratio of the field strength $E$ on the earth to the field strength $E_{pr}$ at the same distance in free space is plotted as a function of the distance $D$. A distinct shadow effect at the optical horizon (in this case at $D = 35.7$ km) appears only in the case of millimetre waves.

The rigorous theory on which such data are based is applicable to all problems of waves meeting a spherical obstacle. This situation also arises when light rays refracted in spherical drops of water give rise to rainbow effects. These considerations accordingly led to a new treatment of the rainbow in terms of wave theory [7]) *).

We shall now return to those cases in which the sky wave conducted via the ionosphere is more important than the ground wave. As shown by Watson's calculations, mentioned above, this is the normal state in short-wave transmissions over great distances. The important role played by the ionosphere, established beyond all doubt by Watson's results, was later emphasized by Van der Pol in the following sentence from his above-mentioned inaugural speech:

> "You know of that conductive layer, high in the atmosphere, which makes radio communication over long distances possible, and whose existence is just as important to radio technology as the existence of iron in the earth is to electrical engineering as a whole".

[8]) Atlas of ground-wave propagation curves for frequencies between 30 Mc/s and 300 Mc/s, C.C.I.R. Resolution No. 11, Geneva 1955.

*) *Editorial note:* By coincidence, a part of this wave treatment of the rainbow is to be seen scribbled on the blackboard in the title photograph.

Although the existence of the ionosphere was not definitely established until 1925, when direct reflection measurements placed it beyond all doubt, Eccles and Larmor had worked out a theory as early as 1912 for wave propagation through a conductive gas, which could be immediately applied to the ionosphere. It was found that a conductive gas behaves formally like a medium possessing a certain conductivity and a relative dielectric constant smaller than unity. This means that the phase velocity of radio waves in the ionosphere exceeds the speed of light $c$ in a vacuum. However, the corresponding propagation velocity of discontinuities (group velocity) is smaller than $c$, as it must be to accord with the principles of the theory of relativity. By means of resonance measurements on lecher lines, Van der Pol determined, during his period in the Cavendish Laboratory at Cambridge, the conductivity and dielectric constant of the plasma in glow discharges; these experiments closely simulated the conditions in the ionosphere. After experimental difficulties had been overcome, it was possible in this way to measure for the first time an apparent dielectric constant smaller than unity. This represented experimental confirmation of the now generally accepted propagation mechanism for radio waves that are refracted in the ionosphere. This whole subject was clearly dealt with in Van der Pol's thesis in 1920, entitled "The influence of an ionized gas on the propagation of electromagnetic waves, and its applications in the field of wireless telegraphy and in measurements on glow discharges".

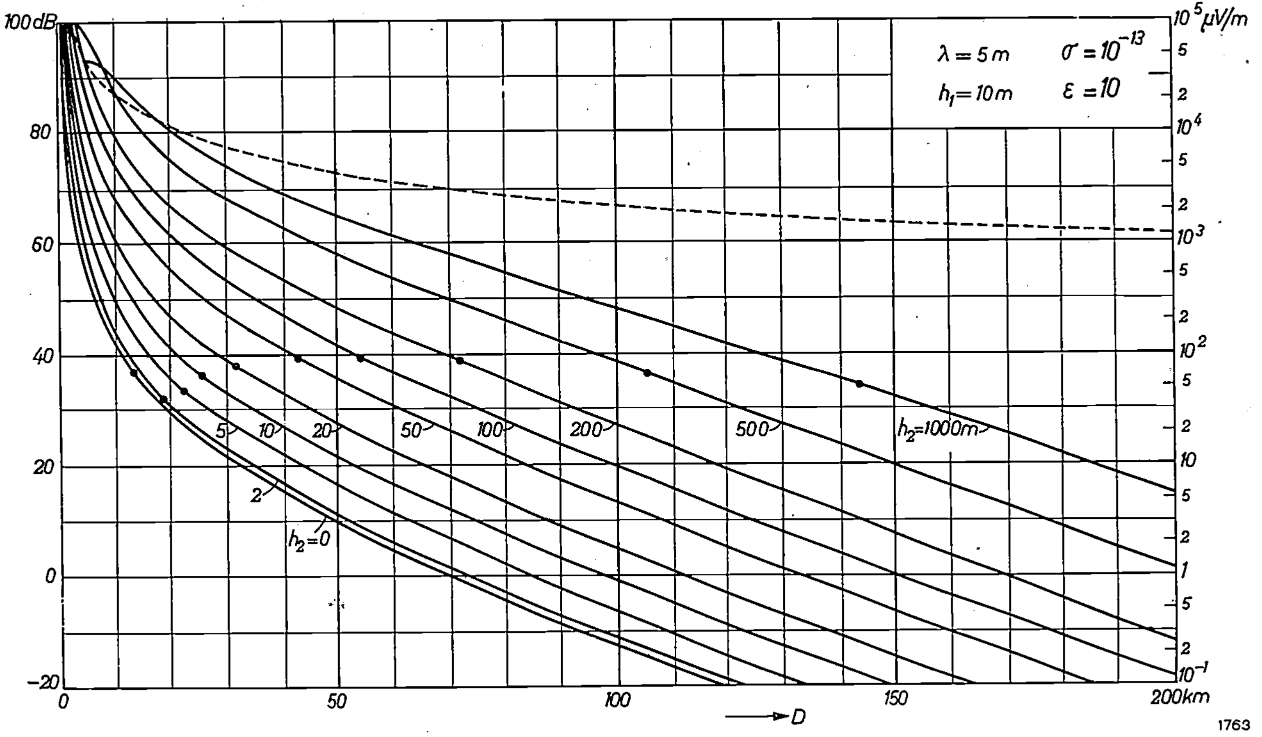


Fig. 2. Vertical component of the field strength of the ground wave (in dB and in μV/m) of a transmitter at height $h_1$ = 10 m, computed for various distances $D$ and receiver heights $h_2$, for $\lambda$ = 5 m and over slightly moist ground. The dashed curve would be obtained in the absence of the earth. The point on each curve indicates the horizon distance when the line connecting the transmitter and receiver touches the earth.

## Non-linear circuits: relaxation oscillations

We shall now consider another field of research in which Van der Pol was particularly active, that of non-linear oscillations. It was understandable that in the 'twenties his attention should be drawn to this subject, since the advances already made in the application of triode circuits made necessary a deeper understanding of oscillation phenomena in order to round off the theory and to discern further possibilities. The simple linear theory of oscillatory effects is arrived at when the $i_a$-$v_g$ characteristic of the triode (anode current versus grid voltage) is approximated by a straight line. The unsatisfactory point is that the amplitude of the excited oscillations then remains undetermined. The limitation of the amplitude is due to the decrease of the slope of the characteristic at either side of the operating point.

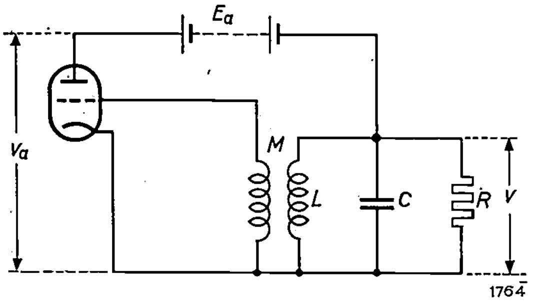


Fig. 3. Circuit for investigating oscillations of a triode circuit, in particular for studying the limitation of the amplitude (leading to Van der Pol's differential equation).

*Fig. 3* shows a representative circuit for investigating these effects, in which $M$ is the mutual inductance between the grid circuit and the $LC$ circuit (with shunt resistance $R$) in the anode lead. We take the triode characteristic to be given by:

$$i_a = \Phi(v_a + g v_g), \quad \ldots\ldots \quad (1)$$

where $g$ is the amplification factor of the triode. The non-linear function $\Phi$ of the variable $u = v_a + g v_g$ can be represented by a Taylor series in the neighbourhood of the point $u = E_a$:

$$\Phi(u) = \Phi(E_a) + \alpha\left(\frac{u - E_a}{k}\right) + \\ + \beta\left(\frac{u - E_a}{k}\right)^2 - \gamma\left(\frac{u - E_a}{k}\right)^3 + \ldots. \quad \ldots \quad (2)$$

The coefficients are normalized by the introduction of the parameter $k = g(M/L) - 1$. This circuit leads to the following differential equation for the time variation of the voltage $v$ (see figure):

$$C\frac{d^2v}{dt^2} + \left\{\left(\frac{1}{R} - \alpha\right) + 2\beta v + 3\gamma v^2 + \ldots\right\}\frac{dv}{dt} + \\ + \frac{1}{L}v = 0. \quad \ldots \quad (3)$$

Van der Pol showed in the first place that the coefficient $\beta$, which determines the first non-linear term of the characteristic, so important for *detection*, does not in itself establish a finite value of the amplitude. The effect of this term will merely be to cause the angular frequency of the $LC$ circuit to differ slightly from its value $(LC)^{-\frac{1}{2}}$ in the linear theory. The term with the coefficient $\gamma$ is the first term to influence the amplitude. The following terms determine less essential properties. For example the fifth-order term in (2) is necessary in order to understand certain hysteresis effects, also studied by Van der Pol [9]). In his amplitude investigations he therefore confined himself mainly to study the influence of the $\gamma$ term, taking $\beta = 0$. By neglecting all higher terms in the expansion of the characteristic, eq. (2), he obtained from (3) a differential equation which, after introducing reduced variables both for the voltage $v$ and the time $t$, assumed the form:

$$\frac{d^2y}{dx^2} - \varepsilon(1 - y^2)\frac{dy}{dx} + y = 0, \quad \ldots \quad (4)$$

where $\varepsilon = \sqrt{L/C}\,(\alpha - 1/R)$.

This differential equation is known as "Van der Pol's equation". Initial investigation of this equation for *small* values of the dimensionless parameter $\varepsilon$ showed that the solution, after a preliminary "time" $x_0$ of the order of $1/\varepsilon$, approximates to a normal sinusoidal oscillation of amplitude equal to 2. Van der Pol wondered what the corresponding solution would look like for larger values of $\varepsilon$, at least of the order of unity. The mathematical difficulties involved led him to apply a graphical procedure, the method of isoclines [10]). In the case under consideration this amounts to reducing equation (4) with the aid of the new variable $dy/dx = z$ to the following first-order differential equation:

$$\frac{dz}{dy} = \varepsilon(1 - y^2) - \frac{y}{z}. \quad \ldots\ldots \quad (5)$$

Plotting $z$ against $y$, this equation determines at every point a slope $dz/dy$, which can be indicated by a short dash. Drawing a continuous curve through a series of these dashes produces a curve which represents a possible solution in the variables $y$ and $z$. A start can be made at various points, and in this way solutions constructed that satisfy the relevant initial conditions. Further numerical integration produces from each curve a solution for $y$ as a function of $x$.

This procedure is illustrated in *fig. 4*, where various solutions are given for the relation between $y$ and $dy/dx$, when $\varepsilon = 1$. *Fig. 5* shows the solution for $y$ itself, corresponding to one of these, and also solutions for $y$ for the cases $\varepsilon = 0.1$ and $\varepsilon = 10$, similarly

[9]) B. van der Pol, Phil. Mag. 43, 700, 1922.
[10]) B. van der Pol, Phil. Mag. 2, 978, 1926.

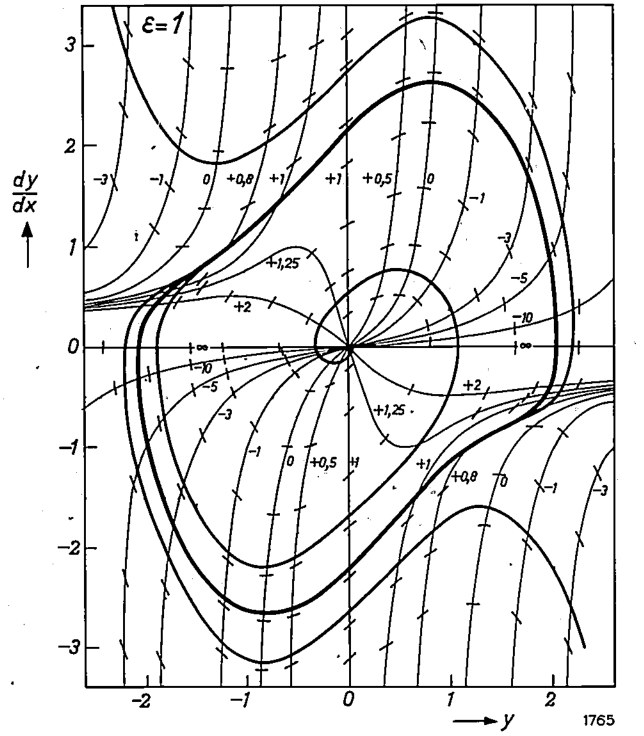


Fig. 4. Illustrating the isocline method of solving the differential equation (5), for $\varepsilon = 1$.

obtained. It follows from fig. 5 that, as the time increases, the solution approaches asymptotically to a non-sinusoidal periodic function, which differs more from a sinusoidal oscillation the larger $\varepsilon$ is; the larger the value of $\varepsilon$ the sooner is the asymptotic end state reached. Furthermore, the period in this end state for large values of $\varepsilon$ is seen to approach a value $\Delta x$ which is of the order of $\varepsilon$; it thus differs appreciably from the corresponding period for very small values of $\varepsilon$, which, in terms of the reduced time unit $x$, approaches $2\pi$. Returning to the original variables $v$ and $t$ in eq. (3), this means that the period for large values of $\varepsilon$ is of the order of magnitude of $L(a - 1/R)$. This is a *relaxation time*, i.e. a time that determines the decay of an aperiodic process. (Better known is the case occurring in other circuits where a relaxation time is determined by an $RC$ product, see e.g. fig. 6.) For this reason, Van der Pol gave the name *relaxation oscillations* to these oscillatory effects which, at large values of $\varepsilon$, constitute the end state of processes defined by the differential equation (4).

Van der Pol first described these oscillations in 1926, after which he continued to investigate them, theoretically and experimentally, in cooperation with J. van der Mark [11]). A particular study was made of *forced* oscillations [12]), produced when an external alternating voltage is applied to a system like that in fig. 3. In the simplest cases the situation can be described by substituting for the right-hand side of Van der Pol's equation a term proportional to cos $\omega t$. For small values of $\varepsilon$ this provided a better understanding of already known properties of oscillator circuits. This applied particularly to the suppression of the free oscillation of a system that occurs when its frequency is close to that of the imposed external oscillation. The resultant synchronism with the latter oscillation was found to be particularly pronounced when, with increasing $\varepsilon$, the first, approximately sinusoidal, free oscillation gradually changes into a free relaxation oscillation. Another way of putting this is that the relaxation oscillations very readily assume the frequency of a superposed alternating voltage if the latter does

[11]) B. van der Pol and J. van der Mark, Onde électrique 6, 461, 1927.

[12]) B. van der Pol, Phil. Mag. 3, 65, 1927.

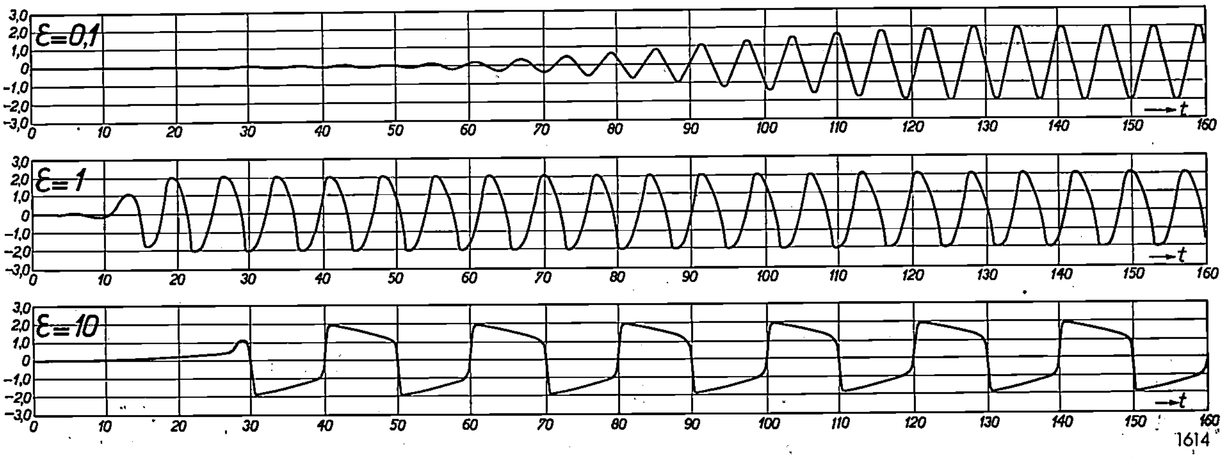


Fig. 5. Solutions of the differential equation (4), obtained by integration from solutions of eq. (5), for $\varepsilon = 0.1$, $\varepsilon = 1$ and $\varepsilon = 10$. (The variable along the abscissa, wrongly denoted by $t$, is the reduced quantity $x$.)

not differ too much from the frequency, determined by a relaxation time, of the natural free relaxation oscillations. If the circuit be modified so as to reduce this relaxation time to about *half* its original value, fairly abrupt synchronization will occur with half the frequency of the external oscillation, and after a further change with a *third* of this frequency, and so on. In other words, it is possible to generate relaxation oscillations that can successively be synchronized with an external oscillation and its sub-harmonics. These synchronization effects occur the more readily the larger the value of $\varepsilon$, or more generally, the more numerically important are the higher terms in eq. (2). Whilst the frequency of a relaxation oscillation can thus very easily be influenced, the amplitude is not nearly so readily controlled.

This synchronization with sub-harmonics later proved to be of great practical importance in the early development of television [13]). A relaxation oscillation is used in television for the scanning of successive lines of the picture. With suitable circuitry a succession of sub-harmonics can be formed, finally producing a relaxation oscillation with a synchronized frequency of 25 (or 30) c/s. This is then used to effect the return to the first line after the complete scanning of each frame. A simple circuit for demonstrating the frequency division of relaxation oscillations is shown in *fig. 6*, where the battery voltage $E$ must be higher than the ignition potential $V_1$ of the discharge tube $B$. In the absence of the alternating voltage $E_0 \sin 2\pi f_0 t$, the capacitor $C$ charges up via the high resistance $R$ until its voltage $V$ reaches $V_1$; it then discharges through the low resistance $r$ of the ignited gas discharge. The discharge is extinguished when the capacitor voltage has dropped to a value $V_2$, after which the charging and discharging process is repeated. Applying now the alternating voltage $E_0 \sin 2\pi f_0 t$, and continuously reducing the frequency of the free relaxation oscillations (which is proportional to $1/RC$) by gradually increasing the capacitance $C$, it is easy to demonstrate the sudden occurrence of synchronization with the successive subharmonics having frequencies $f_0/2$, $f_0/3$, $f_0/4$ and so on. This is further illustrated in *fig. 7*, where the variable capacitance $C$ is set out along the abscissa. The stepwise changes in the period of the forced oscillation can be read from the ordinate.

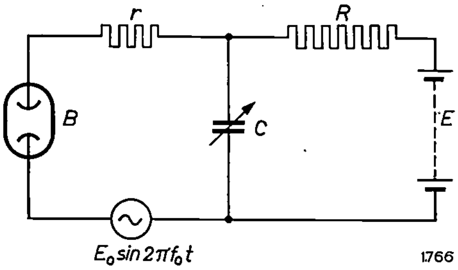


Fig. 6. Circuit for demonstrating the frequency division produced by the synchronization of relaxation oscillations with the sub-harmonics of an imposed voltage $E_0 \sin 2\pi f_0 t$.

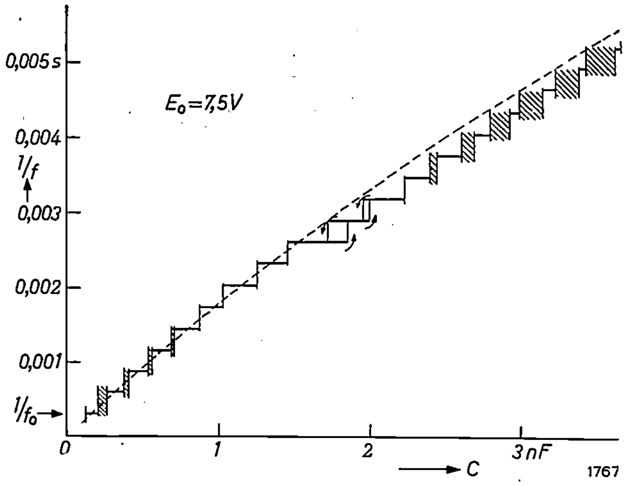


Fig. 7. Synchronization with sub-harmonics. The graph represents the period $1/f$ of the forced relaxation oscillation of the circuit in fig. 6 excited at a constant frequency $f_0$ while the capacitance $C$ is gradually varied.

Subsequently, in collaboration with C. C. J. Addink, Van der Pol entered upon the study of more general synchronization phenomena. It proved possible to obtain synchronization not only with frequencies $f_0/2$, $f_0/3$, $f_0/4$ etc., but also with frequencies $(n/m)f_0$, where $n$ and $m$ are arbitrary integers ($n$ may even be $> m$). For these experiments the frequency of the imposed oscillation was varied instead of one of the parameters of the relaxation oscillator [14]).

Characteristic of the operation of the circuit in fig. 6 is the occurrence of an essentially aperiodic process (in this case the exponential charge and discharge of the capacitor) which is periodically interrupted and restarted at certain critical values of a relevant parameter. Such situations are frequently found in nature, and they always give rise to relaxation oscillations. Van der Pol was struck by their very number and variety. He drew up a long list of examples, which included: the aeolian harp, the scratching of a knife on a plate, the periodic recurrence of epidemics and economic crises, the periodic density variation of two (or another even number) types of animals that live together

[13]) J. van der Mark, An experimental television transmitter and receiver, Philips tech. Rev. 1, 16-21, 1936.

[14]) B. van der Pol, Actualités scientifiques et industrielles No. 718, published by Hermann, Paris 1938, p. 69-80.

and of which one serves as food for the other, the sleep of flowers, phenomena associated with periodic rain showers after the passing of a meteorological depression, and finally the beat of the heart.

The latter example led Van der Pol to a very interesting and instructive practical application of relaxation oscillations, namely an electrical model for simulating all the rhythmic movements of the human heart [15]). With this system it was a simple matter to record "electrocardiograms", and to study the normal heart beat as well as cardiac disorders. The importance of the system to medical science will be obvious. The heart may be regarded as consisting of three systems: the sinus venosus, the two auricles and the two ventricles. The operation of each system can be simulated by a circuit of the type in fig. 6. The unilateral effect of the sinus on the auricles, and of the auricles on the ventricles, was represented by non-amplifying triodes. After other details of the operation of the heart had been taken into account the model illustrated in *figs. 8* and *9* was finally produced. The ignition of the gas discharges in the circuits simulating the auricles and ventricles represents respectively an auricular and a ventricular systole. The synchronizations in this model with subharmonics of the resultant dominant heart beat can be recognized in the human heart in certain cases.

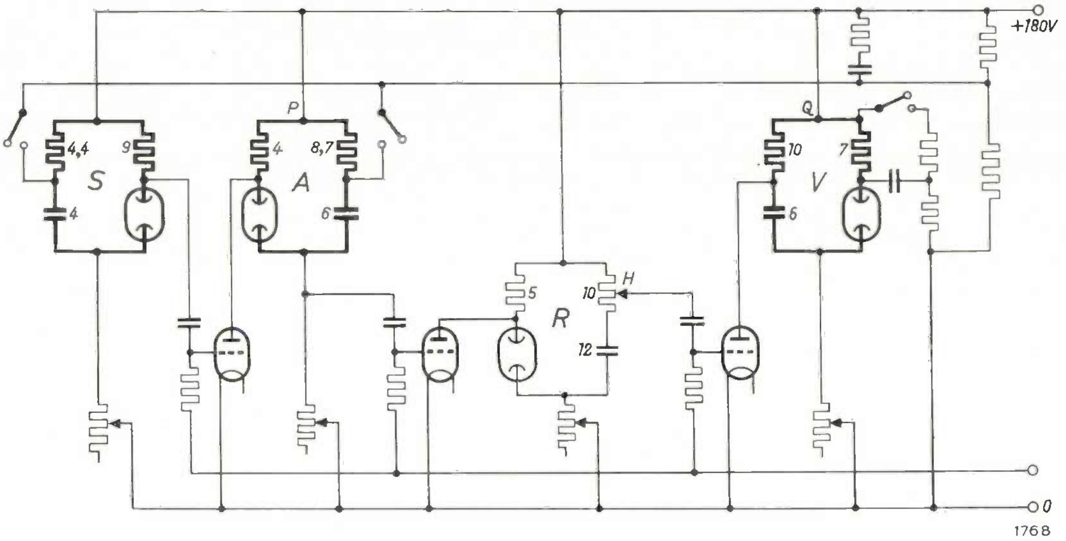


Fig. 8. Circuit simulating the functioning of the heart. Resistance values in kΩ, capacitances in μF. The three circuits *S*, *A* and *V*, each of which can enter into relaxation oscillations and which are coupled in a special way, represent respectively the sinus venosus, the auricles and the ventricles. *R* is a delay system which simulates the transit time of a stimulus through the bundle of His. The ignition of each of the neon lamps in *S*, *A* and *V* corresponds to a systole of that part of the heart. The coupling between auricles and ventricles can be varied with the potentiometer *H* (Erlanger's experiment). Electrocardiograms can be made by recording the current in one of the leads from point *P* or *Q*.

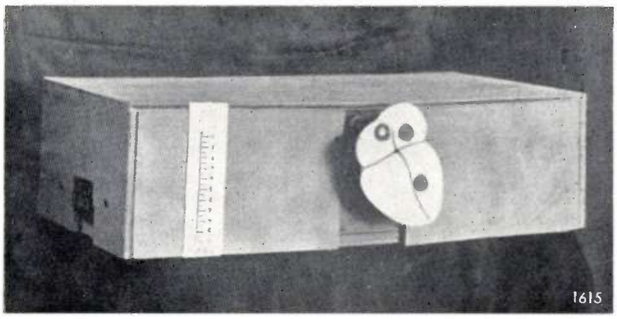


Fig. 9. Photograph of the heart simulator. The three neon lamps of the circuits *S*, *A* and *V* (see fig. 8) are mounted on the three corresponding parts of the anatomical representation of the heart.

The development of the heart model out of the work on relaxation oscillators was a side-track similar to that which led from considerations of radio-wave propagation to a new wave-theoretical treatment of the rainbow. Van der Pol liked to point out how radio problems were apt to draw attention to fields which, at first sight, seem to have very little connection with radio [16]).

### Transient phenomena and operational calculus

Operational calculus was introduced by Heaviside with the original object of providing a simple means

[15]) B. van der Pol and J. van der Mark, Phil. Mag. 6, 763, 1928.

[16]) The examples mentioned and various others will be found in: B. van der Pol, Proc. World Radio Convention, Sydney 1938.

of studying switching transients in electrical networks. The methods devised were later developed into a system of calculus which has proved to be extremely useful in mathematics as a whole. No one saw this more clearly than Van der Pol, who demonstrated by numerous examples the possible applications of operational methods to widely diverse fields of study.

To illustrate the starting point of opcrational calculus we shall briefly discuss an aspect of the historical development of alternating current circuit theory, often examined by Van der Pol. At first the aim was to determine how a certain system responds to an externally applied harmonic vibration cos $\omega t$, or, more generally, exp j$\omega t$, where cos $\omega t$ is the real component. Investigations were made to ascertain in what way the properties of the system depend on the frequency $f = \omega/2\pi$. As a case in point we shall consider an amplifier. The ratio $G(\mathrm{j}\omega)$ between the output voltage and input voltage were studied in the case where both were proportional to exp j$\omega t$. If the function $G(\mathrm{j}\omega)$ is known, we can calculate with the aid of Fourier analysis the output voltage for any arbitrary time function of the input voltage. The case of the response to a discontinuous input voltage can also be treated in this way. In applying this procedure it was often overlooked, however, that the complex function $G(\mathrm{j}\omega)$ depends on *two* real (though mutually related) functions, and that it is consequently not enough to take account only, for example, of the modulus $|G(\mathrm{j}\omega)|$. This was discussed some considerable time ago in this journal by J. Haantjes [17]). On the other hand, all properties can equally be derived from the knowledge of a single real function, such as the unit step function response $G_s(t)$. This is defined as the output voltage rcsulting from the application at a given moment $t = 0$ of a direct voltage of unit magnitude to the input side, when no voltage was present for $t < 0$. (By giving the input function *unit* voltage, $G_s$ is a dimensionless function of time.) All characteristics of the amplifier can be described in terms of this unit function. It can be shown that for a given arbitrary primary voltage $v_{\mathrm{pr}}(t)$ at the input, the output voltage of the amplifier will be

$$v(t) = \int_{-\infty}^{+\infty} G_s(t-\tau)\,\frac{\mathrm{d}}{\mathrm{d}\tau}\,v_{\mathrm{pr}}(\tau)\,\mathrm{d}\tau = \int_{-\infty}^{+\infty} G_s(\tau)\,\frac{\mathrm{d}}{\mathrm{d}t}\,v_{\mathrm{pr}}\,(t-\tau)\mathrm{d}\tau. \quad . \; . \; (6)$$

This representation is simpler than that where the system is described in terms of the above-mentioned complex function $G(\mathrm{j}\omega)$, which defines the behaviour of the system in the case of a continuous, harmonic input voltage. If the latter function is used, the effect of an arbitrary input voltage $v_{\mathrm{pr}}(t)$ is given by the expression:

$$v(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} G(\mathrm{j}\omega)\,\left[\int_{-\infty}^{+\infty} v_{\mathrm{pr}}(\tau)\,\mathrm{e}^{\mathrm{j}\omega(t-\tau)}\,\mathrm{d}\tau\right]\mathrm{d}\omega\,, \quad (7)$$

which is generally more difficult to handle for a nonperiodic $v_{\mathrm{pr}}(t)$ than eq. (6).

The above shows that it is sometimes simpler to study the behaviour of a system with reference to a discontinuous process (here the sudden application of a unit step voltage, giving the step-function response $G_s(t)$), than with reference to a continuous process (the continuous application of an alternating voltage exp j$\omega t$, from $t = -\infty$ to $t = +\infty$). An objection to working with the unit-step function might be that, although mathematically easier to handle in such cases, true discontinuity cannot in reality be achieved. It should be remembered, however, that any discontinuous process can be regarded as the limit of a physically realizable process, and that, strictly speaking, there are in fact no perfectly continuous processes either. Van der Pol once remarked that there seemed to have been a long-standing aversion to the study of discontinuous phenomena. He described this aversion wittily as a "horror discontinuitatis", resembling the "horror vacui" assumed by 17th century physicists.

We have discussed these preliminaries to show how useful Heaviside's operational methods can be for the study of electrical systems. This appears from the fact that eq. (6) — a so-called composition product or convolution integral, in this case of the two functions $G_s(t)$ and $v_{\mathrm{pr}}'(t)$ — is particularly suitable for treatment by the methods of operational calculus. In the version of operational calculus adopted by Van der Pol and the author [18]), to any arbitrary function $h(t)$ (the "original") one may allot a new function, the operational "image" $f(p)$, defined by the following Laplace integral:

$$f(p) = p\int_{-\infty}^{+\infty} \mathrm{e}^{-pt}\,h(t)\,\mathrm{d}t\,. \quad . \; . \; . \; . \quad (8)$$

[17]) J. Haantjes, Philips tech. Rev. 6, 193, 1941.

[18]) B. van der Pol and H. Bremmer, Operational calculus, based on the two-sided Laplace integral, Cambridge Univ. Press 1950. — The version of the operational calculus developed in this book is characterized by the fact that the integral in eq. (8) has the lower limit $-\infty$, as opposed to the more conventional definition used by Carson, Doetsch, Wagner *et al.*, where the lower limit is always 0. The advantages of using $-\infty$ as the lower limit are explained in the book.

The image of a convolution integral is simply the product of the images of the individual factors. In the case of eq. (6) we therefore find the operational image of the function $v(t)$ as the product of the images of $G_s(t)$ and $v_{pr}'(t)$. By finding the original of this operational image we then have a complete solution of the problem.

The other expression for $v(t)$, eq. (7), which must obviously be equivalent to eq. (6), has a form which does not lend itself so well to an operational solution. Since eq. (7) is precisely the form most suited for the treatment of periodic processes, we can see why the operational calculus is less useful for dealing with such processes than with transients. Nevertheless, eq. (7) must evidently also have some relation to operational calculus; this relation becomes clear if we recall that (7) represents a Fourier integral which, after appropriate substitutions, can be put in the form of the Laplace integral (8). It is precisely the linking of operational methods to an exponential integral of this kind (which is related to the familiar Fourier integrals) that has made this calculus into a rigourously grounded mathematical tool.

In order to apply operational methods to the solution of all kinds of concrete problems, such as determining the above-mentioned integral (6) with given explicit functions, it is evidently useful to have a list of images available that correspond to originals given by elementary functions. Van der Pol called such a list his "dictionary", whilst the rules of operational calculus were the "grammar". One of these rules, for example, is the foregoing theorem concerning the convolution integral.

Armed with such a grammar and dictionary Van der Pol studied numerous properties of electrical networks, and of filters in particular. He arrived at theorems, often of very general application, which concerned for instance the relation between the unit function response of a given filter and that of a new filter derived from it; the latter was produced from the original by replacing all inductances by capacitances, and vice versa. General theorems of this kind sometimes led to interesting special cases. One notable example was the discovery that the charging of a system of capacitors from a D.C. source always takes place with an efficiency of 50%, irrespective of the number of resistances and inductances in the circuit [19]). This is to say that the final energy of the charged capacitors amounts to half the energy which the source, after being suddenly switched on, must deliver in order to sustain the currents which must flow until the system reaches its final state.

Many of the mathematical relations discovered by Van der Pol were arrived at quickly by the skilful manipulation of the methods of operational calculus. An example of such a relation in the case of continuous functions is the remarkable and previously unknown identity;

$$a \int_0^\infty e^{-at} \sin^{2n} t \, dt = \frac{(2n)!}{(a^2+2^2)\,(a^2+4^2)\ldots(a^2+4n^2)}\,,$$
$$\text{for} \quad \mathrm{Re}\, a > 0. \quad . \quad . \quad . \quad . \quad (9)$$

Van der Pol liked in particular to work on relations for *discontinuous* functions. A representative example is the function $A_2(t)$, which indicates the number of pairs of integers $(m,n)$ for which $m^2 + n^2 < t$; operational methods led at once to the surprising relation:

$$\tfrac{1}{4}\{A_2(t)-1\} = \left[\frac{t}{1}\right] - \left[\frac{t}{3}\right] + \left[\frac{t}{5}\right] - \left[\frac{t}{7}\right] + \ldots, \quad (10)$$

where the symbol [] here represents the largest integer $\leqslant$ the number within the brackets.

As a last typical example of the application of operational calculus we mention the "moving average" or "sliding mean" theory, which was extensively studied by Van der Pol. Starting from an arbitrary function $g(t)$, the sliding mean is defined by the new function

$$g^*(t) = \int_{t-\Delta}^{t+\Delta} g(\tau)\, d\tau. \quad . \quad . \quad . \quad (11)$$

In many cases we want to know the function $g(t)$ when $g^*(t)$ has been measured. Take, for example, the intensity $g(\omega)$ as a function of frequency $\omega$ in a spectrogram recorded with a spectrograph. Owing to the necessarily finite width of the slit in the spectrograph, the actual intensity $g(\omega)$ is smeared out into a generally less-rapidly varying function $g^*(\omega)$. The operational solution of this problem leads to many ways in which $g$ can be expressed by a series in $g^*$, suitable for numerical use.

## Other subjects

In the foregoing we have discussed the most important general subjects on which Van der Pol was engaged. We shall now touch briefly on various other, more specialized problems on which he worked for shorter or longer periods — on some all his life —

[19]) B. van der Pol, Physica 4, 585, 1937.

and which will be found in the bibliography of his numerous publications [20]).

Amongst the earliest researches of Van der Pol were special subjects concerned with the theory of electrical networks, including current distributions in an arbitrary number of coupled circuits. His interest in antenna theory was first aroused by investigations into the radiation and natural frequency of antennae possessing end capacitance, during which the inadequacy of the concept of radiation resistance was discussed. In England he carried out experimental research on the conductivity of seawater, in connection with its bearing on the propagation of radio waves. When he joined the Philips laboratory his interest extended to virtually all problems of radio technology, then still in its early stages. Partly in cooperation with K. Posthumus, Y. B. F. J. Groeneveld, T. J. Weijers, G. de Vries, C. J. Bakker, W. Nijenhuis, C. J. Bouwkamp and others (mentioned elsewhere in this article) he investigated the general properties of triode characteristics, the distribution of the electrical field and electron paths inside a triode, the theory of grid detection, general properties of oscillators and filters (including the equivalent circuit of a quartz oscillator, and the non-linear theory of hysteresis effects in two coupled circuits, one of which is part of a triode generator), noise in radio valves, radiation patterns of beam antennae, and many other subjects.

Van der Pol had an important share in the preparatory work leading to the introduction of radio broadcasting in the Netherlands. Together with R. Veldhuyzen, M. Ziegler and J. J. Zaalberg van Zelst he did extensive research into the field-strength distribution of broadcast waves over the Netherlands. The results were plotted on charts, showing contours of constant field strength. The deviation of these contours from a circular shape indicated the influence of the type of ground traversed between transmitter and receiver. In the chart for the Hilversum transmitter on 298.8 m wavelength, for example, the absorbent effect of the city of Amsterdam and of the dry sandy ground of the chain of hills east of Utrecht could be clearly seen. This chart led Van der Pol to point out that the field-strength distribution in the Northern provinces would change after the damming of the Zuiderzee, since the conductivity of this large expanse of water would decrease as it became progressively less salty. This was confirmed by later measurements.

When a new transmitter was planned to replace the old one at Hilversum, the reciprocity theorem was applied to determine its most favourable location: instead of starting with a transmitter in a central (though not yet determined) point, measurements were made of the fields due to two auxiliary transmitters in two distant points in the country, where reception from a centrally-situated transmitter would have been weakest. The most favourable site for the new transmitter was found by determining, in the central region of the country, at what point the received field was strongest on the line connecting the points of intersection of corresponding contours charted for both auxiliary transmitters (*fig. 10a, b* and *fig. 11*).

Long before there was any question of the practical exploitation of frequency modulation, which has since become so important in broadcasting, Van der Pol had investigated the theory of this system of modulation. Fundamental insight here requires much more mathematical knowledge than the simpler system of amplitude modulation, since the differential equations involved can no longer be solved by time functions proportional to exp $j\omega t$. In the first place, Van der Pol gave a suitable definition for the vague concept "instantaneous frequency", $\omega_{\text{inst}}$. Writing the general frequency-modulated signal as:

$$A \cos \varphi(t) = A \cos \left[\omega_0 t + m\omega_0 \int_0^t g(t)\mathrm{d}t + \Phi\right], \quad (12)$$

this definition reads:

$$\omega_{\text{inst}}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\varphi(t) = \omega_0 \{1 + m g(t)\}. \quad . \quad (13)$$

Together with F. L. H. M. Stumpers he then investigated the so-called quasi-stationary approximation whereby the currents and voltages assume values at any given instant that conform to the then existing value of the instantaneous frequency as defined by (13). It was found that the current produced in an impedance as a result of a frequency-modulated voltage is then solely determined by the phase characteristic of the impedance [21]). In the 'forties Van der Pol also carried out experimental research on frequency modulation, but lack of space must preclude its discussion here.

In the theoretical investigation of frequency modulation an important part was played by

[20]) A complete bibliography of Van der Pol's publications, compiled by C. J. Bouwkamp, will appear shortly in Philips Res. Repts.

[21]) B. van der Pol, J. Instn. Electr. Engrs. **93 III**, 153, 1946.

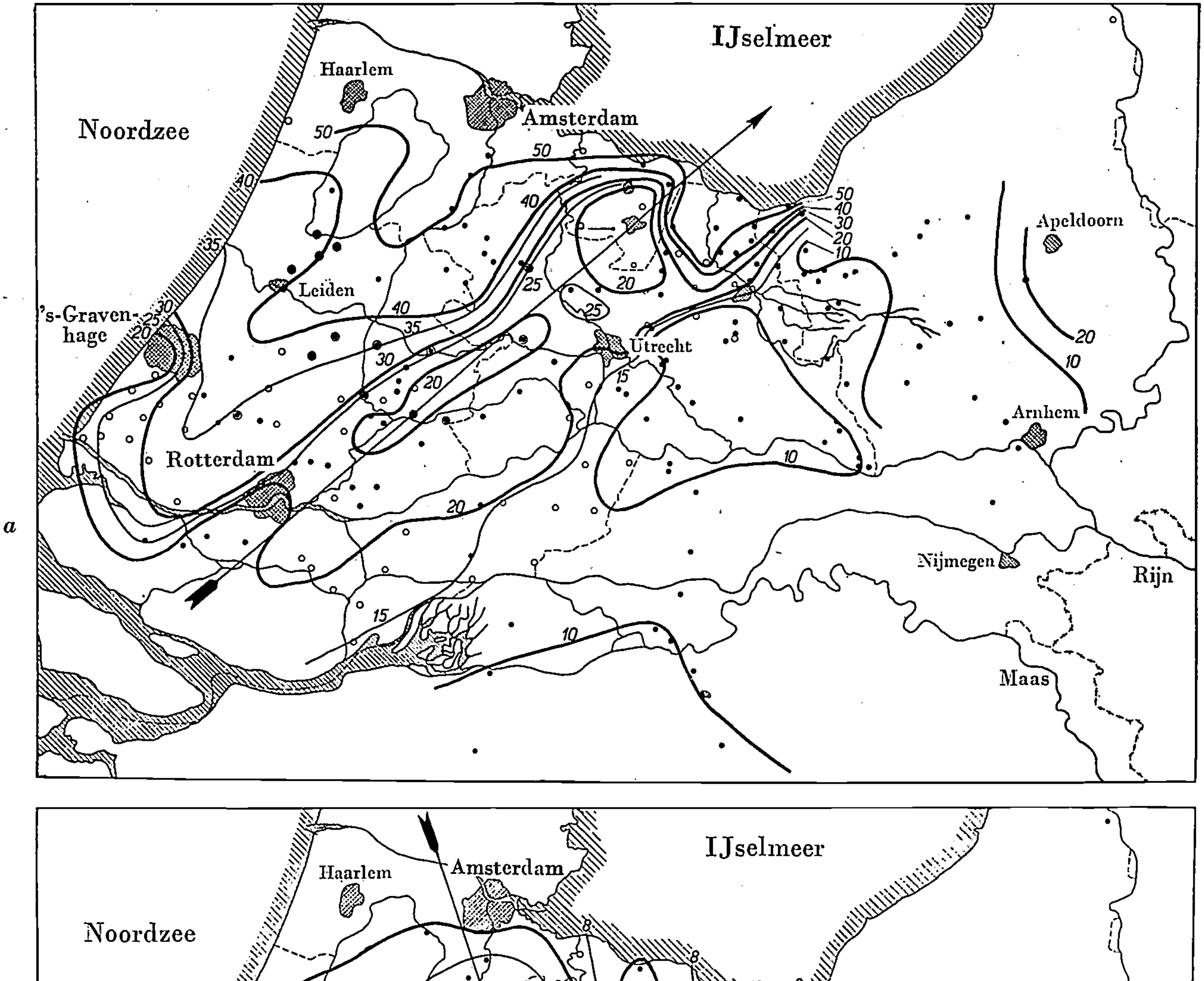


Fig. 10. Field-strength measurements carried out in 1934 in the central region of the Netherlands for determining the most favourable site for the new broadcasting station, planned at that time. In two remote parts of the country, where the poorest reception might be expected from a central transmitter, two identical auxiliary transmitters were erected, one at the Dollard estuary (wavelength 325 m), and the other at Maastricht (wavelength 317 m).

*a*) The field distribution of the Dollard transmitter represented by contours of constant field strength. The long arrow points to the transmitter.

*b*) The same for the Maastricht transmitter. (Note the shadow effect due to Rotterdam.)

Mathieu's differential equation. More generally, Van der Pol applied himself to the study of differential and difference equations related to radio problems. These studies again led to many and various investigations that had a less direct connection with physical or technical questions, or were even of a purely methematical nature. They included theta functions, elliptic functions, the gamma function, the theory of numbers, the properties of prime numbers, and so on. Nevertheless, these problems cannot be seen as entirely distinct from radio technology. Theta functions, for example, are related to the theory of potential functions, and Van der Pol's interest in the theory of numbers was inspired by the part it plays, for instance, in the above-mentioned synchronization effects in relaxation oscillators. Incidentally, these abstract researches also led to a quite unexpected "industrial by-product", as a result of Van der Pol's study of the distribution of Gaussian prime numbers. In the system of all complex numbers with integral real and imaginary parts, these are the numbers that cannot be obtained as the product of two of them. When all these two-dimensional prime numbers are drawn in the complex plane, a curious pattern is produced (*fig. 12*). This was used as a tablecloth pattern which had particular success in America.

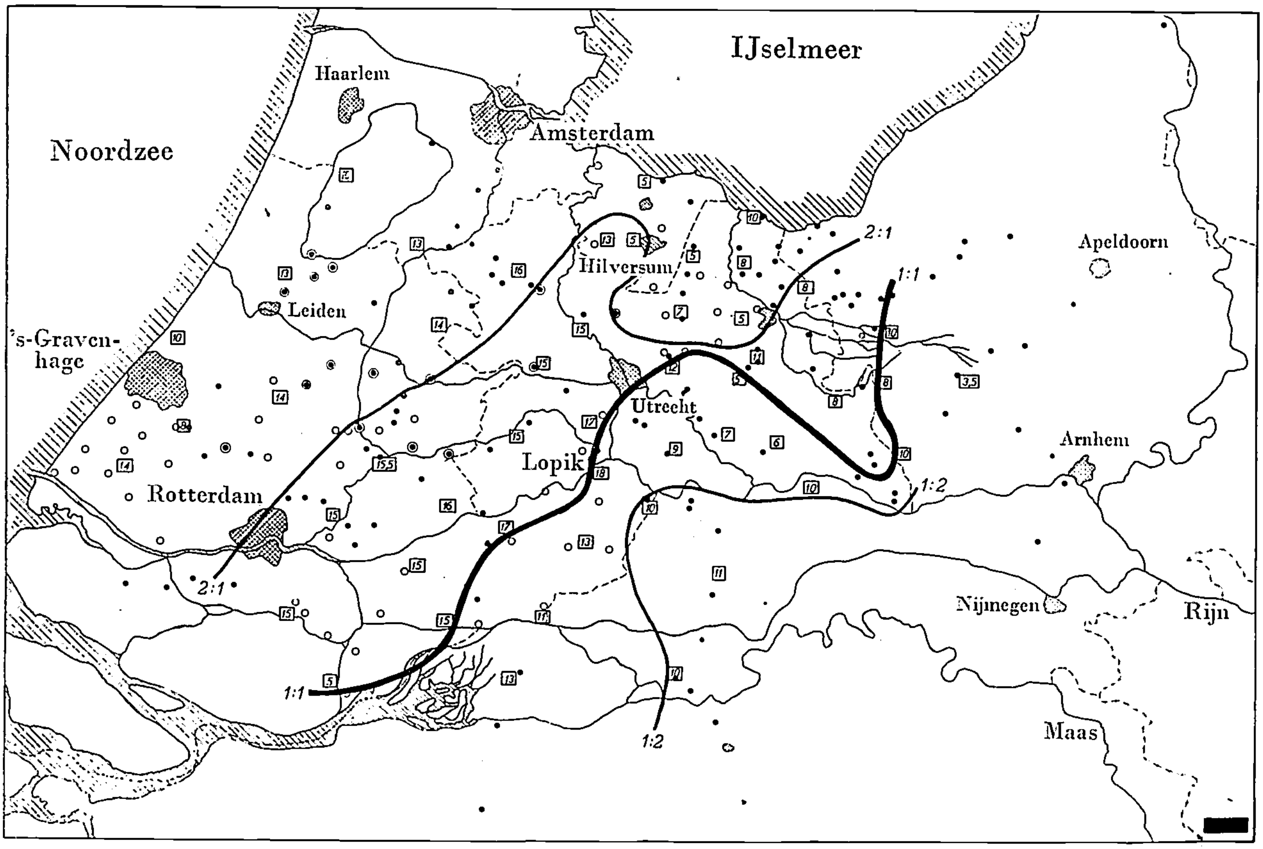


Fig. 11. The bold lines connect the places where the two auxiliary transmitters are received with equal strength (thick central line, 1 : 1) or with an intensity ratio of 2 : 1 and 1 : 2. As can be seen, the highest field strength is found on the thick line at the point 18 near the village of Lopik (where in fact the station was subsequently built). The numbers at other places indicate the relative value of the field which, if the transmitter of the station had been sited there, would have been obtained in the more unfavourable of the two poor-reception areas (Dollard or Maastricht). (Fig. 10*a*, *b* and fig. 11 are taken from B. van der Pol, Rapport van de veldmetingen van twee bij de Dollard en bij Maastricht opgestelde proefzenders, T. Ned. Radiogen. 7, 173-195, 1935.)

The association of mathematics and radio technology sometimes worked in the opposite direction. We refer to an investigation whereby Van der Pol, in cooperation with C. C. J. Addink, used electrical techniques for studying a mathematical function [22]). The investigation concerned the determination of the zero points of Riemann's $\zeta$ function, normally defined by the series:

$$\zeta(z) = \frac{1}{1^z} + \frac{1}{2^z} + \frac{1}{3^z} + \dots \qquad (14)$$

[22]) B. van der Pol, Bull. Amer. Math. Soc. **53**, 976, 1947.

Although this series can be used only for Re $z > 1$, the function can nevertheless be defined for the whole complex plane by means of an analytic continuation. Riemann had conjectured that the function thus defined would possess an infinite number of zero points, all of which would lie on the line Re $z = \frac{1}{2}$. By means of a relation derived by Van der Pol, the functional values along this line can be associated with the simple function:

$$e^{-x/2}\left[e^{x}\right] - e^{x/2} \quad . \ . \ . \ . \ . \ . \quad (15)$$

(the symbol [ ] has the same meaning as in eq. (10)). In fact, when the function (15) is cut off after a convenient large value of $x$, and then periodically continued, the coefficients of the Fourier series of the new function obtained in this way represent approximations to a series of values of the function $\zeta(z)/z$ along the line Re $z = \frac{1}{2}$. The problem of determining the zero points conjectured by Riemann was thus reduced to a Fourier analysis of the said function derived from (15). This Fourier analysis was carried out in an elegant experiment in which the sawtooth pattern of one period of the function in question was cut along the edge of a disc (*fig. 13*). The disc was rotated at a highly constant speed and light was passed through a radial slit and caused to fall on the serrated edge and on a photocell placed behind it. The photo-current thus obtained, varying with time, was mixed with a sinusoidal current of frequency $f$. The resultant signal, containing the difference tones between $f$ and all harmonics of the periodic function derived from (15), was used to excite a mechanical resonator of very sharp resonance. When the frequency $f$ was now slowly raised from zero, each successive harmonic of the above function excited the resonator in turn with its difference tone, and with an intensity corresponding to the relevant Fourier coefficient. By recording the excitation (as a function of $f$) a spectrum was obtained in which the required zero points of the $\zeta$ function could immediately be seen as minima; see *fig. 14*.

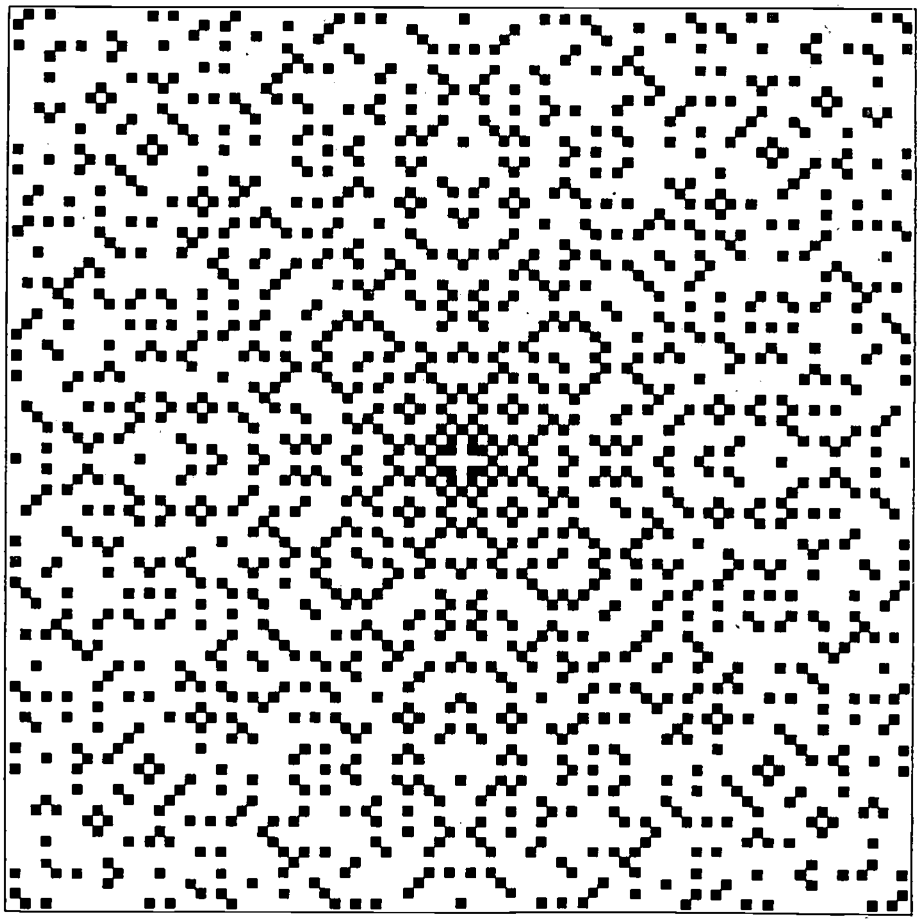


Fig. 12. The black squares represent in the complex plane (with origin in the centre) the complex prime numbers of Gauss, i.e. the numbers $(m + \mathrm{j}n)$ that cannot be written as $m + \mathrm{j}n = (a + \mathrm{j}b)(c + \mathrm{j}d)$, where $m, n, a, b, c, d$ are integers. This pattern was woven into a fabric marketed by an Eindhoven textile factory.

## Musical theory

We shall now turn from Van der Pol's scientific and technical investigations and in conclusion touch briefly on his interest in music. Although deeply interested in analysing the structure of a musical work, he realized that this could never be an approach to the creative aesthetic element —

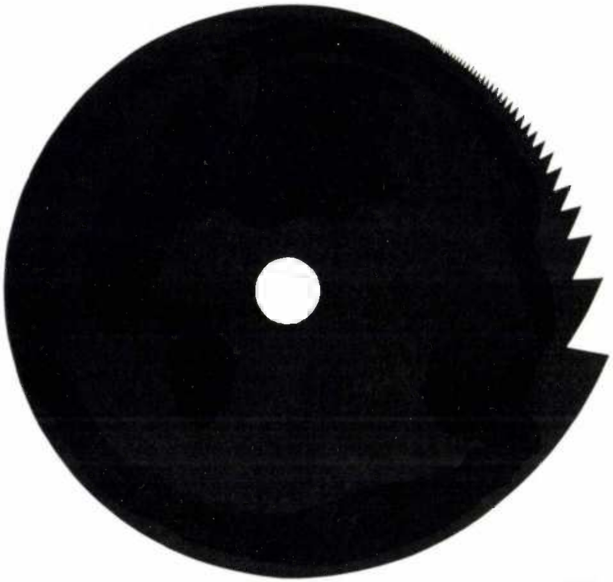


Fig. 13. Disc with edge serrations representing part of the function (15), for experimentally determining the zero points of Riemann's $\zeta$ function.

an element for which, as a competent musician himself, he had a deep understanding. It was his preoccupation with the theory of numbers that led him to look for mathematical laws in harmonic intervals [23]), or, for example, to find numerical definitions for concepts such as dissonance. In particular he discovered that the fractions presented by the successive intervals in the diatonic scale could be contained in a Farey series. In this case it is the series of all irreducible fractions whose denominators and numerators never exceed the number 5. Only two tones from the diatonic sequence (the *b* and the *d*, with *c* as the fundamental) do not fit into the Farey series, and the remarkable thing is that it is precisely these tones about which there is most uncertainty in tuning (if one is not bound to the equal-temperament scale).

It will not be surprising after the foregoing that Van der Pol emphasised that the sub-harmonics of a given note can be produced with the aid of synchronization effects on relaxation oscillations. In this way it was possible to confirm experimentally a postulate put forward by the music theorist Hugo Riemann, namely that the minor triad can be regarded as the combination of the 4th, 5th and 6th sub-harmonics of a given tone, just as applies to the major triad in respect of the 4th, 5th and 6th upper harmonics of another tone.

[23]) B. van der Pol, Muziek en elementaire getallentheorie, Arch. Mus. Teyler **9**, 507-540, 1942 (in Dutch).

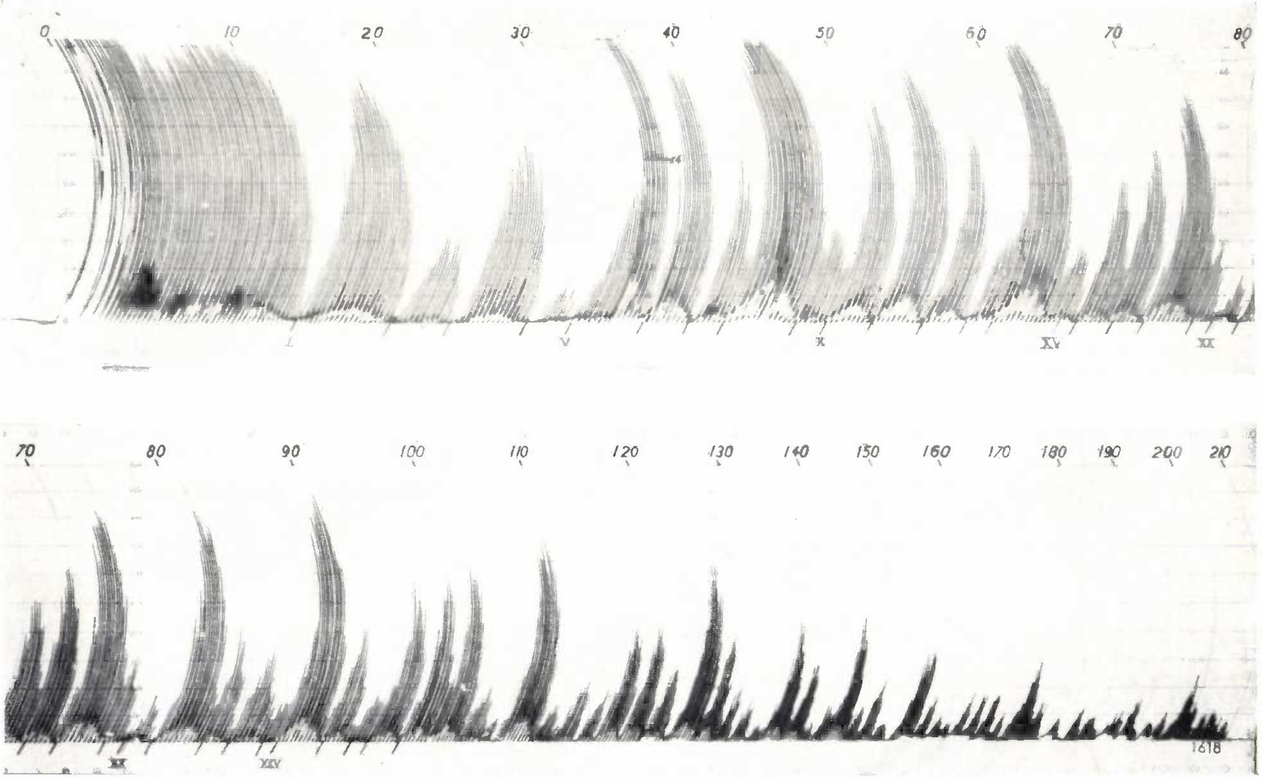


Fig. 14. Recording obtained with the aid of the disc in fig. 13, giving the Fourier coefficients of the function cut into the disc. The minima in the recording are the zero points of the $\zeta$ function on the line Re $z = \frac{1}{2}$. The first 29 zero points, denoted by dashes and Roman figures, were already known from calculations. (The imaginary part of $z$ is set out along the top edge.)

Van der Pol possessed absolute pitch, and was able, for example, to tune his violin to the right pitch without a tuning fork. He enjoyed exercising this gift and it was typical of the man that he should develop from it yet another interesting investigation. Together with Addink he devised an apparatus using an oscilloscope to check the tuning of an orchestra during a performance [24]. From hundreds of observations, both on permanently tuned instruments and on orchestras, they found that the middle *a* varied from 430 to 447 c/s, and also that there were quite distinct, more or less systematic variations of this frequency during a performance.

Van der Pol's enquiring mind led him to apply the principles of science to music. For him, however, there existed a much profounder relation between these fields. There is perhaps no more fitting way to end this review of the many-sided work of the great man of science that Van der Pol undoubtedly was, than to quote his own words, from a lecture given at the Teyler Foundation in Haarlem:

"Is there really any difference between the inspiration that leads to the conception of a beautifully flowing melody, a rich theme or a brilliant modulation, and that which leads to the conception of a new, elegant, unexpected mathematical relation or postulate? Are not both born in the same mysterious way and do they not both often demand laborious development? Art connotes skill, and skill too is the handmaid of science."

[24] B. van der Pol and C. C. J. Addink, Philips tech. Rev. 4, 217, 1939.

**Summary.** The review given deals principally with the work done by the late Van der Pol in the years from 1922 to 1949, when he was with the Philips Research Laboratories at Eindhoven. The main subjects discussed are the propagation of radio waves, especially the influence of the ground conditions on ground-wave propagation; non-linear circuit phenomena (in particular relaxation oscillations described by the "Van der Pol equation", and their synchronization); transients and operational calculus, which (following Heaviside, whom he greatly admired) Van der Pol used as a fruitful heuristic and later rigorously founded method. A brief discussion is devoted to various other subjects treated by Van der Pol in his numerous publications, including mathematical investigations prompted by problems of radio technology, such as the application of Mathieu's equation to frequency-modulation problems, and also studies relating to subjects of pure mathematics, such as elliptic functions and the theory of numbers (especially the properties of prime-numbers). Finally some examples are given of Van der Pol's interest in music and the theory of music.

# DEMOUNTABLE SEALS FOR GLASS HIGH-VACUUM EQUIPMENT

by B. JONAS *) and G. SEITZ *).

666.1.037.4

*An interesting variant is described of known methods of sealing glass components. The crux of the method is the inclusion of a metal ring in the seal. This makes it possible to disconnect the sealed parts in a matter of seconds, without causing damage.*

There are many industrial products in which replacements or modifications have to be made after assembly. For this purpose it must be possible to dismantle the product without causing significant damage. In the case of products made partly of glass, this reversal of the fabrication process may present difficulties. For conventional electron tubes and gas-discharge tubes the vacuum envelope and the seals are therefore constructed without a view to dismantling. Nevertheless, glass vacuum vessels that can be dismantled are quite important in certain products that are expensive and made only in small quantities and in laboratory work (experimental apparatus with interchangeable components) and development work [1]). The ways and means that have been devised to circumvent these difficulties are many and various, and include ground-glass joints of diverse kinds, rubber gaskets and wax seals, etc. The drawback of such devices is that they do not allow degassing at a sufficiently high temperature; as a rule, therefore, they can be used only if the glass tubes to which they are fitted are kept continuously connected to a pump during operation.

We shall describe here a method based on a technique which has been used for certain radio tubes, and which has now been developed to meet more stringent requirements. First a few words about the older technique.

In order to be able to dismantle a vacuum vessel (of hard or soft glass) the glass envelope can be made in two parts that meet in a circular seam. If this seam is filled with a glaze which has a low melting point (*fig. 1*) the joint can later be opened again by melting the glaze — without exceeding the softening point of the glass envelope — in exactly the same way as two metal parts soldered together are disconnected by melting the soldered joint. As stated, a "glazing" technique of this kind was used for a time in the manufacture of radio tubes. In the "Rimlock" range of tubes [2]), for example, the glass parts were not fused directly together, but bonded by means of a glaze seal, as in fig. 1. This technique was used here to prevent damage to the cathode, not with a view to the possibility of dismantling the tube.

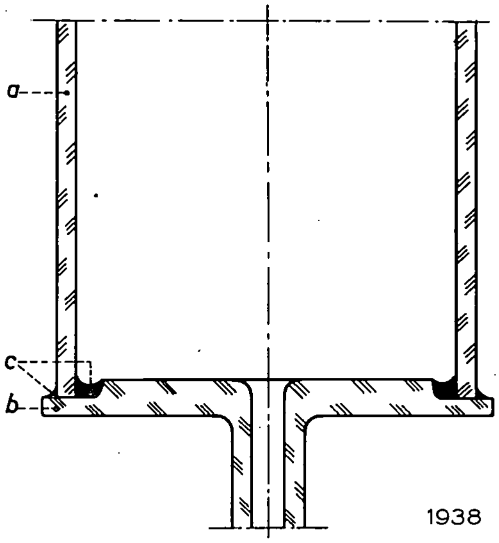


Fig. 1. Schematic representation of a glass seal using a glaze. *a* cylindrical glass envelope, *b* base, *c* melted glaze.

Having regard to the widely different properties of the types of glass used in tube manufacture (see *Table I*) the question arises whether such a method could find general application or would only be feasible in special cases. To effect a bond as described above, the following conditions must be fulfilled:

1) In the interval from room temperature to the temperature at which the tube glass begins to soften, the thermal expansion of the glaze must match that of the tube glass (to a first approximation they should be equal).
2) The flow point of the glaze must be sufficiently low so that softening and consequent deformation of the glass envelope cannot occur.

*) Zentrallaboratorium Allgemeine Deutsche Philips Industrie GmbH, Aachen Laboratory.

[1]) For details of vacuum-tight seals in *metal* vacuum equipment, see N. Warmoltz and E. Bouwmeester, Philips tech. Rev. **21**, 173, 1959/60 (No. 6).

[2]) See e.g. G. Alma and F. Prakke, Philips tech. Rev. **8**, 289, 1946.

Table I. Data relating to the sealing of glass parts using a glaze (and suitable metals for sealing in these glasses). $\alpha \times 10^7$ denotes the thermal expansion between room temperature and the annealing temperature of the glass; $T_w$ is the highest temperature which the various kinds of glass can withstand without deforming (corresponding to a viscosity of approx. $2.5 \times 10^{12}$ poise); $T_v$ is the flow point of the glazes (corresponding to approx. $10^4$ poise) [3]).

| | Glasses | | | | Glazes | |
|---|---|---|---|---|---|---|
| | Fused silica and glasses of extremely high $SiO_2$ (or $Al_2O_3$) content | Normal hard glass (borosilicate glasses). Suitable metals: tungsten, fernico and molybdenum | Normal soft glass (lime or lead glasses). Suitable metals: platinum, nickel-iron and chrome-iron | Soft glass types with high thermal expansion coefficient (lime and lead glass with high alkali content). Snitable metals: nickel and iron | Zinc-borate glazes | Lead-borate glazes |
| $\alpha \times 10^7$ (°C$^{-1}$) | 6 - 40 | 35 - 55 | 85 - 110 | 115 - 130 | 36 60 | 80 - 140 |
| $T_w$ (°C) | 1300 - 700 | 750 - 520 | 600 - 450 | 500 - 400 | | |
| $T_v$ (°C) | | | | | 750 - 550 | 500 - 300 |

Now it is possible to prepare glazes — on a basis of lead or zinc borate — possessing any desired coefficient of expansion, provided it is not unduly low. In most cases, then, the first condition can be met. A circumstance favourable to the fulfilment of the second condition is that the kinds of silicate glass used for tubes are "long", i.e. they all exhibit a relatively slow drop in viscosity with rising temperature, as required in the usual working procedures, whereas the glazes referred to, not only soften at relatively low temperature, but are "short", i.e. the drop in viscosity with rising temperature is fairly steep.

It becomes difficult to meet both the above conditions for the glaze only when the tube glass combines a low softening point with a low coefficient of expansion ($\alpha = 40$ to $45 \times 10^{-7}$), as required for sealing-in tungsten leads, for example. (No suitable glazes are known for types of glass whose thermal expansion coefficient is smaller than this.)

We have tacitly assumed in the foregoing that it is possible, during the sealing or unsealing process, to heat the whole envelope evenly. This is no problem if the tubes are small. The larger the objects, however, and with them the dimensions of the furnace, the more difficult it becomes to apply the process. Moreover, protracted heating at the temperature involved may be harmful to certain components inside the tube.

In such cases the aim will be to confine the heating to the glazed joint and to narrow zones of the glass envelope on either side. The axial temperature gradient of the glass wall in this region must be such that, with the given geometry, the glass will not crack. The temperature distribution in the wall is roughly as shown in *fig. 2*. The arrows indicate

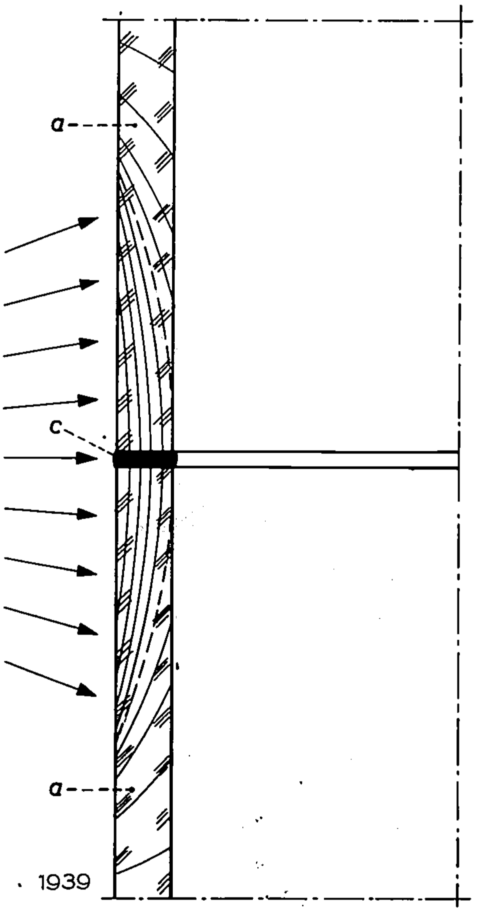


Fig. 2. Longitudinal section of the wall of two cylindrical glass tubes sealed by a glaze; *a* glass wall, *c* glaze. The drawing shows the isotherms inside the glass wall at the moment when the viscosity of the glaze has dropped sufficiently to permit completion of the joint. The isotherm at the flow point is shown by a dashed line. The arrows indicate the applied heat flow.

[3]) See e.g. G. Ch. Mönch, Neues und Bewährtes aus der Hochvakuumtechnik, Knapp, Halle a.d. Saale 1959;
W. H. Kohl, Materials technology for electron tubes, Reinhold, New York 1951;
J. H. Partridge, Glass-to-metal seals, Soc. of Glass Techn., Sheffield 1949.

the direction of the heat flow applied. The kind of heat source is immaterial here, what is important is that when heat is applied externally there must always be a fairly wide belt of high temperature if the flow point is to be reached over the whole width of the joint. Otherwise the joint will not fill out to the edge with molten glaze, resulting in poor bonding. Experiments have shown that the width of the zone subjected to temperature above the flow point should be nearly ten times the wall thickness.

At either side of this hottest zone there are then areas in which the temperature gradually drops to the general temperature of the whole tube, i.e. the temperature to which the tube as a whole may permissibly rise during the sealing or unsealing process. The higher this general temperature may be, the less danger there is of cracking. This is especially important if the tube has an awkward shape, e.g. if it is very compact or has flat surfaces perpendicular to the long axis.

As we have seen, the above principles makes it possible to seal glass parts together and to separate them again without damage and without causing undue softening of the glass. The method is not suitable, however, if there are components near the sealing zone that cannot tolerate either a high temperature gradient or high temperatures.

In this respect the "glazing" technique can be considerably improved if the heat source used for attaining the flow point in the joint is transferred to the joint itself. For this purpose a metal ring, coated on both faces with an appropriate glaze, can be introduced between the parts to be joined and can serve there as a heating element, either by inductive coupling or by the direct passage of current through it. The ring is bonded to the tube walls as the glaze melts, and forms part of the seal when the glass has cooled. The sealing technique evolved from this idea has been investigated on a number of cases in the Aachen laboratory, and will now be discussed.

First, some general observations. The energy supply is stopped as soon as the glaze is judged to be properly fluid, a state which is very soon reached once heating has begun. That is the moment for effecting the necessary displacements, i.e. final alignment when sealing the parts, and pulling apart when unsealing them.

Glazing the ring beforehand, as referred to above, appreciably shortens the period of high temperatures, and thus facilitates the sealing process. In simple cases, however, both operations can be combined.

Here, too, it is obviously an advantage, and indeed often essential, to preheat the complete tubes to a certain general temperature; this accelerates the sealing process and reduces the danger of cracking. Where the tubes are very long it is sufficient, of course, to preheat only the area surrounding the sealing zone. Preheating temperatures should not generally exceed 300 to 350 °C; as mentioned, temperatures higher than this may be damaging to components inside the tube.

The temperature distribution in the tube wall for the new method is roughly as shown in *fig. 3*. The differences compared with fig. 2 are immediately evident: the hottest zone (temperature above the flow point) has shrunk to a narrow strip on either side of the metal ring; the flow temperature can be attained everywhere in the joint at a much lower maximum temperature. An advantage not at once apparent from the figure is that the situation depicted is reached in a matter of a few seconds after switching on the power. This means that the heat supply needed to bring the glaze to the flow point is small, and thus that the dangerous temperature

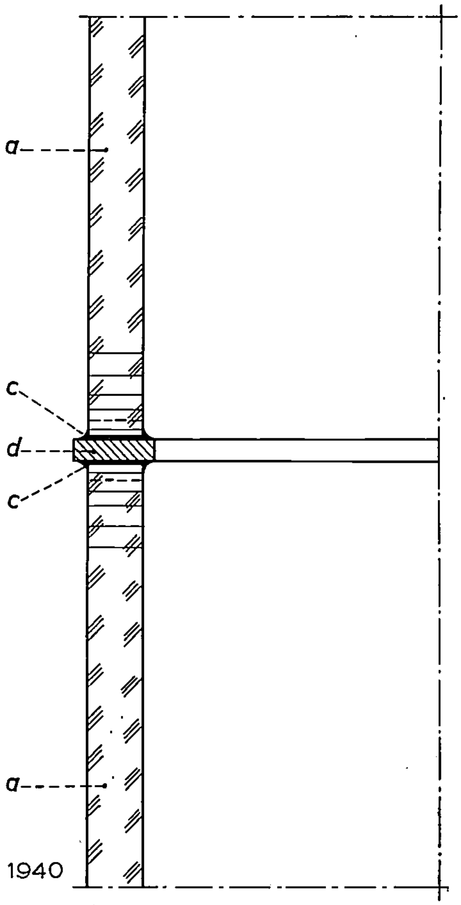


Fig. 3. As fig. 2, but a metal ring is now embodied in the joint. *a* glass wall, *c* glaze, *d* metal ring. The isotherm at the flow point is again a dashed line. The other isotherms are shown in the same steps as in fig. 2.

range is held only for a short time — a particularly important point where a glaze of relatively high flow point has to be used, as for example with types of glass having a low expansion coefficient (e.g. for use with tungsten-wire seals).

Having considered the broad outlines of the method, we can now touch on various particulars, including the choice of materials. As indicated in Table I, for almost every value of expansion coefficient of practical importance in tube manufacture there are several combinations of metal and glass available, each one of a pair matching its partner in expansion. A seal of two such materials, provided it is properly cooled, contains no dangerous stresses and hence has no tendency to break. Most of the sealing metals given in the table are of course also suitable for the heating ring in the method under consideration, the requirements in both cases being the same. Of course, the treatments of the various metals and alloys that may be necessary in practice, to prevent bubble formation, for example, or to improve the adhesion to the glass, are also needed here; they are carried out before the preparatory glazing of the ring.

It has been emphasized that the sealing of glass parts by means of a glaze can succeed only if the materials involved meet fairly stringent requirements as to their behaviour during expansion this being equally valid, as a rule, for glass-to-metal seals. In radio-tube manufacture, however, another technique of glass-to-metal sealing is widely used in which considerable differences in expansion are permissible. In these seals, cracking of the glass is precluded by using a metal part capable of plastic deformation, to compensate for the differences in expansion. This calls for metals of low yield point (see *Table II*) and also for a design that

Table II. Yield points at room temperature of some suitable sealing metals. At higher temperature these values are considerably lower.

| | |
|---|---|
| Sealing alloys | 30 - 50 kg/mm$^2$ |
| Copper (commercial) | approx. 15 kg/mm$^2$ |
| Copper (OFHC) | approx. 6 kg/mm$^2$ |
| Aluminium (99.99%) | approx. 2.7 kg/mm$^2$ |

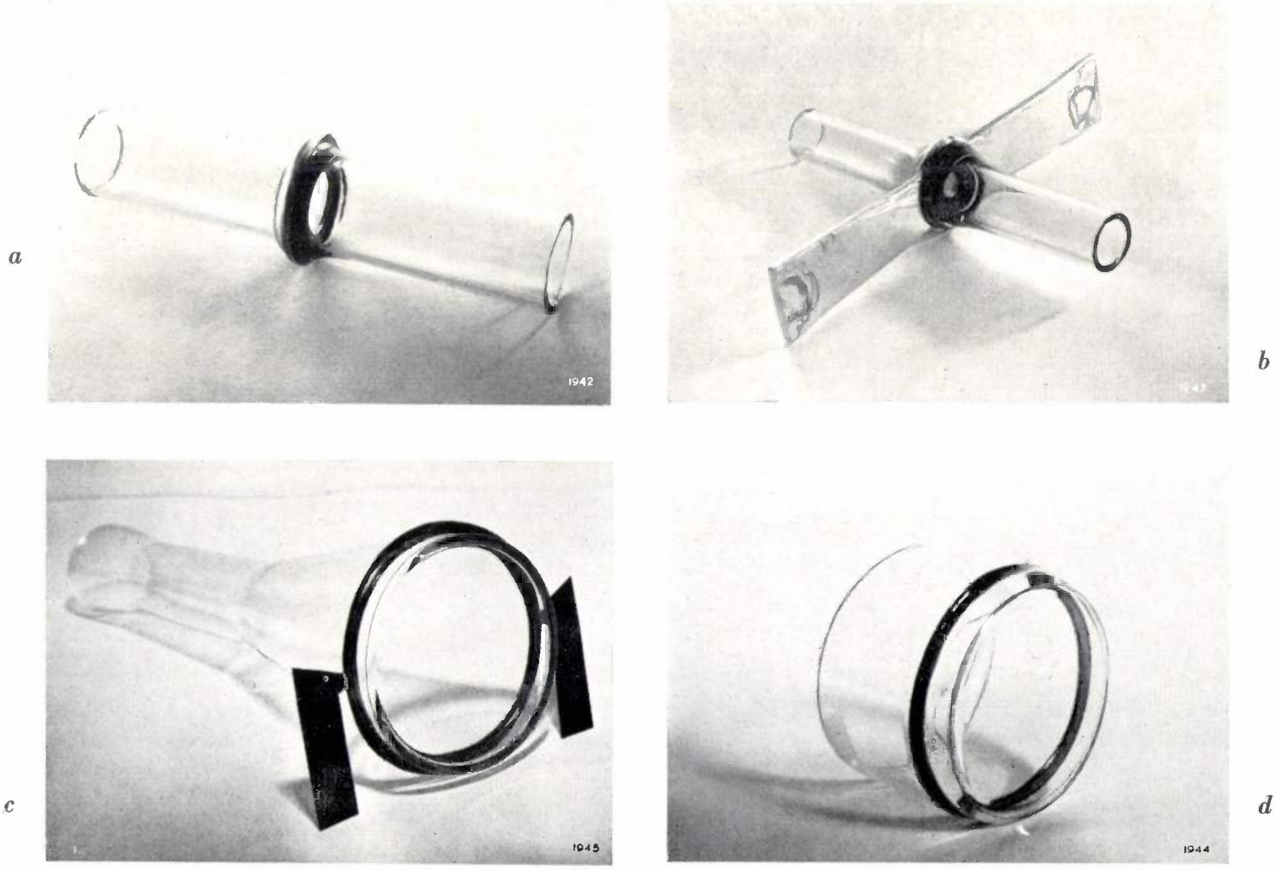


Fig. 4. Some glass high-vacuum seals, made by the method described.
*a*) Flanged tubes of soft glass, with copper ring.
*b*) Flanged tubes of soft glass, with aluminium ring. The aluminium "wings" serve as current leads for heating by direct conduction.
*c*) Seal of glass cap and cone with copper ring and projecting wings for current supply (soft glass).
*d*) Seal of glass cap and tube with fernico-type alloy (hard glass).

limits the magnitude of the forces produced (small cross-section of metal) and avoids tensile stresses in the glass [4]). Copper, gold and silver can be used with advantage here. If copper is suitably pretreated it bonds excellently with all hard and soft types of glass, provided only that the temperature during the sealing operation remains below the melting point of copper and that strong oxidation can be avoided.

The latter conditions are easily be met when applying the "flowing"-metal principle to the sealing method under consideration. This makes it possible to work at a relatively low temperature and moreover in any desired atmosphere. Indeed, the use of a ring of "flowing" metal in the joints discussed here offers significant practical advantages. To begin with, it adds aluminium to the small number of eligible metals. Aluminum bonds excellently to the type of glaze under consideration, and it also possesses outstanding virtues as a material for use in high-vacuum techniques. Because of its low melting point, however (657 °C), it has not been used hitherto for a direct glass-to-metal seal. This is again the limiting factor: aluminium can be used only for soft types of glass, since the flow points of all known glazes for hard glass are too close to the melting point of aluminium, or may even be higher.

Another advantage is that, if flowing metals are used and the rings are introduced as in fig. 3, the metal insert need not be so very thin as one might expect it to be. If the ring is of pure copper, it can have a thickness from 1 to 1.5 mm; if it is of pure aluminium, a thickness of 2 to 3 mm is permissible. This makes the rings much easier to handle, both mechanically and thermally. The dimensions are even more favourable where cylindrical seals are concerned (according to Housekeeper's method [5])). It will also be plain that the purer, i.e. the softer the metal, the thicker the ring may be. It should be added that rings that are not too thin have the further advantage of making a good vacuum seal possible between glass parts of very different thermal expansion coefficient.

*Fig. 4a-d* shows a number of glass joints made by the method described.

Soft metals possessing a high coefficient of expansion, differing widely from that of the glass, do have a certain drawback compared with metals and alloys whose expansion matches the glass. The latter allow considerable freedom as regards the shape of the joint, whereas the former demand a construction in which the outside and inside edges are left uncovered, as in fig. 3. Soft metals having a high expansion coefficient are therefore not suitable where — e.g. for reasons of insulation — a complete covering is necessary. Even a local overflow of glaze can easily cause a crack. The reason is evidently that the glass in such a case hinders the plastic flow of the metal. Quite small differences in expansion can then give rise to large forces which, at certain critical places — the glass or the metal itself, or the bond between them — may cause cracks or fracture.

If protruding edges are permitted, a seal as illustrated in *fig. 5* is advantageous, particularly in the case of relatively weak glass walls. The flanges make it possible to use metal rings of comparatively large cross-section (see also fig. 4*a* and *b*).

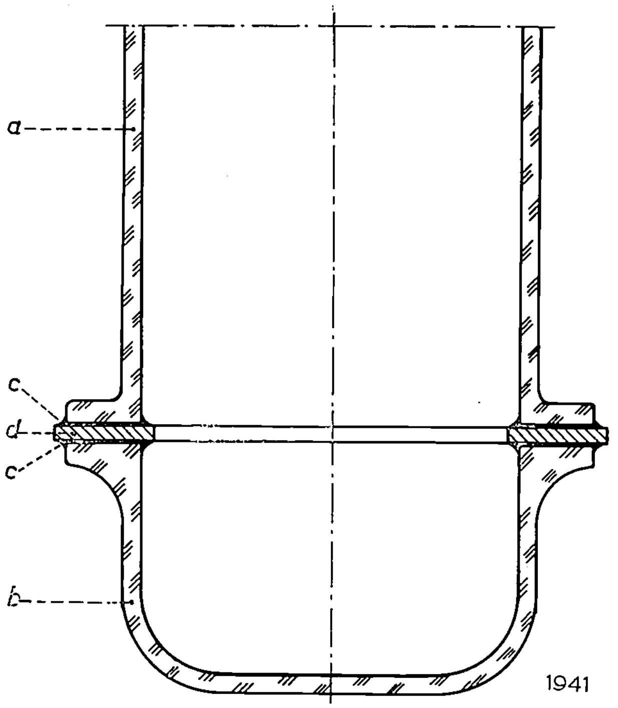


Fig. 5. Tube closed by glass cap. *a* flanged glass tube, *b* flanged glass cap, *c* glaze, *d* metal ring.

The considerations discussed in the foregoing were prompted by the question whether it was possible for the purposes of vacuum technique to make glass joints that would meet various conflicting requirements. They were to be capable of withstanding the necessary pumping temperatures, they should enable the tubes to perform their function after sealing-off, and they should be demountable without seriously damaging them or exposing them to high temperatures. In this brief report of our investigations, it is shown, in general terms, how far these objects can be achieved using glazed joints with the heat source embodied in the joint itself. Details of the method, for example the precise choice of glazes for hard and

[4]) See J. L. Ouweltjes, W. Elenbaas and K. R. Labberté, Philips tech. Rev. **13**, 109, 1951/52, in particular fig. 7, and B. Jonas, Philips tech. Rev. **3**, 119, 1938.

[5]) See the book by Partridge referred to under [3]), p. 160 *et seq.*, and that by Kohl, p. 60 *et seq.*

soft glass, and of alloys for glass-to-metal seals for the various types of hard glass, are, however, outside the scope of this article.

**Summary.** The method described here of making vacuum-tight demountable glass seals uses a metal ring coated on both sides with a suitable glaze. When the ring is heated, e.g. inductively, the glaze melts and the preheated glass parts are then pressed together. Embodying the heat source in the joint results in a favourable temperature distribution, the zone of highest temperature (and steep temperature gradient) being restricted to the immediate vicinity of the joint. Moreover the amount of heat supplied and the heating time are reduced to a minimum. As the metal ring remains in the joint, it can later serve again as local heat source if the joint is to be unsealed, without damaging the parts. A further advantage when the ring is of a soft metal is that, being capable of plastic deformation, it allows fairly wide disparities between the thermal expansion coefficients of the glass parts joined.

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

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, the Netherlands, where a limited number of repaints are available for distribution.

**2738:** H. B. G. Casimir, J. Smit, U. Enz, J. F. Fast, H. P. J. Wijn, E. W. Gorter, A. J. W. Duyvesteyn, J. D. Fast and J. J. de Jong: Rapport sur quelques recherches dans le domaine du magnétisme aux Laboratoires Philips (J. Phys. Radium **20**, 360-373, 1959, No. 2/3). (Report on various researches at Philips in the field of magnetism; in French.)

In the first part of the paper the crystalline anisotropy of a number of hexagonal oxidic compounds containing barium is discussed. In the absence of an external magnetic field the magnetization vector can point in an arbitrary direction with respect to the c-axis. This behaviour can be described with two anisotropy constants. Examples are given of materials with a preferential direction (c-axis), with a preferential plane (basal plane) as well as with a preferential cone for the magnetization vector. The latter case occurs at relatively low temperatures in crystals containing cobalt. There are also materials in which, at different temperatures, all three types of anisotropy occur. The relatively weak anisotropy in the basal plane, which has six-fold symmetry, has been measured. In crystals having only trivalent metal ions, two such ions can be replaced by one divalent and one quadrivalent ion. It appears that substitution of cobalt again promotes the occurrence of a preferential plane of the magnetization, as in the oxides which contain divalent metal ions. The classical dipole-dipole energy has been computed and it is shown that it can account for the observed anisotropy in the structure containing two successive barium layers, which, although not containing cobalt, shows a preferred plane for the magnetization vector. The anisotropy in the structure containing single barium layers, which has a preferred direction of the magnetization vector, is not explained by this mechanism, and presumably originates from spin-orbit interaction. The influence of controlled precipitation on the magnetic properties of alloys is discussed in the last section. With the aid of an electron microscope it is shown that a precipitate, consisting of long parallel needles in the optimal case, causes the high $(BH)_{\max}$ value (up to $12 \times 10^6$ gauss-oersteds) of single crystal "Ticonal" ("Alnico") containing 34% cobalt, that has undergone a special heat treatment in a magnetic field. It is further shown that a (110) [001] texture can be obtained in 3%-silicon iron only if the metal contains a precipitate of favourable composition (e.g. $Si_3N_4$ or MnS) and division.

**2739:** F. L. H. M. Stumpers: Interpretation and communication theory (Synthese **11**, 119-126, 1959, No. 2).

This article (contribution to a symposium held in 1954) describes in brief the analogy between the interpretation of translated texts and the interpretation of messages transmitted over noisy transmission channels. Also discussed are some of the statistical methods from communication theory that have been applied for the quantitative study of literature and language.

**2740:** P. M. Cupido: Some views on automatic control in glass factories (Glastech. Ber. **32 K**, I/1-I/5, 1959, No. 1).

A brief outline is given of some different modes of automatic control and the stability of the controlled system. The application of automatic control in glass processes is discussed. It is no remedy for bad furnace design or bad process conditions, but it will give a sound background for studying the process and will make it possible to evaluate improvements. If the optimum settings for the improved process have been found, automatic control will be the most reliable means of keeping the process conditions within the necessary limits.

2741: A. M. Kruithof and A. L. Zijlstra: Different breaking-strength phenomena of glass objects (Glastech. Ber. **32 K**, III/1-III/6, 1959, No. 3).

Survey of the mechanical strength of various glass objects. Not only thin glass fibres but also massive glass objects can have very high strengths (200 kg/mm$^2$) if care is taken that the surface is not damaged. From this it is concluded that the structure has only a secondary effect on the strength. A kind of "strength scale" is constructed showing how various combinations of structure and surface effects give rise to various levels of strength.

2742: J. Goorissen, F. Karstensen and B. Okkerse: Growing single crystals with constant resistivity by floating-crucible technique (Solid state physics in electronics and telecommunications, Proc. int. Conf., Brussels, June 2-7, 1958, edited by M. Désirant and J. L. Michiels, Vol. 1, pp. 23-27, Academic Press, London 1960).

See Philips tech. Rev. **21**, 185-195, 1959/60 (No. 7).

2743: J. H. Uhlenbroek and J. D. Bijloo: Investigations on nematicides, II. Structure of a second nematicidal principle isolated from Tagetes roots (Rec. Trav. chim. Pays-Bas 78, 382-390, 1959, No. 5).

A second nematicidal principle isolated from *Tagetes* roots has been identified as 5-(3-buten-1-ynyl)-2,2′-bithienyl. Upon catalytic hydrogenation 5-butyl-2,2′-bithienyl was obtained which by comparison of infrared spectra proved to be identical with a synthetically prepared sample. The 5-(3-buten-1-ynyl)-2,2′-bithienyl was further characterized by its infrared absorption.

2744: H. F. Hameka: Berechnung der magnetischen Eigenschaften des Wasserstoffmoleküls (Z. Naturf. **14a**, 599-602, 1959, No. 7). (Calculation of the magnetic properties of the hydrogen molecule; in German.)

To calculate the proton screening and the magnetic susceptibility of the hydrogen molecule, wave functions are introduced built up from calibration-invariant atomic wave functions. The calculation is done for two cases, the ground-state wave function being described either by the Wang or the Rosen approximation. In the first case the nucleus screening constant turns out to be $2.631 \times 10^{-5}$ and the susceptibility $-3.920 \times 10^{-6}$. In the second case the values are $2.732 \times 10^{-5}$ and $-4.045 \times 10^{-6}$, respectively.

2745: N. W. H. Addink: Note on the analysis of small quantities of material by X-ray fluorescence (Rev. univ. Mines **102**, 530-532, 1959, No. 5).

One point of particular importance in X-ray fluorescence analysis is the so-called inter-element effect: fluorescent radiation originating in one element A in a sample is partially absorbed by an element B, to an extent depending on the concentration of B. In the present investigation it is shown that this spurious effect can be eliminated by a) the use of only dilute solutions, b) the use of thin layers of powdered materials (sample required less than 1 mg). With method (b) systematic errors are held to within 4%.

2746: G. Klein and J. M. den Hertog: A sine-wave generator with periods of hours (Electronic Engng. **31**, 320-325, 1959, No. 376).

By means of an inverse-function generator it is possible to derive a triangular voltage accurately from a sinusoidal one. By applying negative feedback the reverse can also be achieved. Making use of this possibility an ultra-low frequency sine-wave generator was designed for maximal periods of 3½ hours. The distortion is then negligibly small. If a slight distortion is permissible, this period can be increased considerably. An important feature of this generator is the fact that no transient phenomena occur. The inverse-function generator can also be used for various other purposes, one of them being a logarithmic voltmeter covering the range from approximately 10 mV to some tens of volts.

2747: C. J. M. Rooymans: A new type of cation-vacancy ordering in the spinel lattice of $In_2S_3$ (J. inorg. nucl. Chem. **11**, 78-79, 1959, No. 1).

$In_2S_3$ has a crystal structure closely similar to the spinel structure: the formula can be written $In_{8/3}\square_{1/3}S_4$, analogous to spinel $AB_2O_4$ or $AB_2S_4$. In contrast to $\gamma$-$Fe_2O_3$, $In_2S_3$ has cation vacancies in certain tetrahedral positions; these positions are ordered, giving rise to a superlattice.

**2748:** D. J. Kroon, C. van de Stolpe and J. H. N. van Vucht: Etude de la résonance nucléaire magnétique de l'hydrogène inclus dans l'alliage $Th_2Al$ (Archives des Sciences **12**, fasc. spéc., 156-160, 1959). (Nuclear magnetic resonance study of hydrogen in $Th_2Al$; in French.)

Per molecule in the crystal of $Th_2Al$ there are four sites capable of accommodating hydrogen atoms. If only two of these sites are occupied by protons ($Th_2AlH_2$), the diffusion rate of hydrogen in the crystal is high at room temperature, resulting in a narrow resonance line. Below 100 °K this motion ceases and a broad resonance line results. From the temperature dependence of the line width it is found that the diffusion activation energy is 0.22 eV. Similar measurements have been made for $Th_2AlH_3$. From the shape of the resonance line in this case it is concluded that there is a certain equilibrium between "free" and "bound" hydrogen. If proton motion is hindered because all interstitial sites are filled ($Th_2AlH_4$ and $Th_2AlH_2D_2$), the line is broad even at room temperature. (See also Philips tech. Rev. **21**, 297-298, 1959/60, No. 10.)

**2749:** J. S. van Wieringen and A. Kats: Paramagnetic resonance of hydrogen in fused silica (Archives des Sciences **12**, fasc. spéc., 203-204, 1959).

Pure fused silica shows neither optical absorption nor paramagnetic resonance after irradiation with X-rays at room temperature. On the other hand, irradiation at the temperature of liquid nitrogen produces two absorption bands in the ultraviolet and a paramagnetic resonance spectrum whose intensity grows with the percentage water present. Paramagnetic resonance measurements suggest that the colour centres responsible for these effects are hydrogen atoms. They disappear after a few minutes at a temperature of 10-20 °C above that of liquid nitrogen.

**2750:** J. Davidse and B. T. J. Holman: A suppression filter with variable bandwidth (T. Ned. Radiogenootschap **24**, 199-209, 1959, No. 4).

This paper deals with the design of a notch filter with variable bandwidth. The loading capacitance is neutralized by means of a feedback circuit; in addition with this circuit negative load resistances can be realized. In this way very small bandwidths can be obtained. It is shown that bandwidth variation can be obtained by variation of the loading resistance. The transient response and the overshoot of the filter are calculated. Finally the practical circuit is given and discussed briefly.

**2751:** G. D. Rieck: Rekristallisation von Wolframdrähten (Hochschmelzende Metalle, 3rd Plansee-Seminar, Reutte/Tirol, June 22-26, 1958, edited by F. Benesovsky, pp. 108-119; published 1959). (Recrystallization of tungsten wires; in German.)

Tungsten for use in incandescent filaments is provided with a "dope" which promotes the growth of large crystals during recrystallization. These crystals show a fragmentation structure, in particular after bending, from which it can be concluded that the residues of the dope are present as filaments parallel to the wire axis. These large crystals have an orientation — the [531] direction lies in the wire axis — that differs from the texture of the small crystals of pure tungsten and which appears to depend on the action of the dope. The occurrence of this particular orientation should not be attributed to a deviation from the drawing texture, but can be explained by two facts. Firstly this orientation is able to survive the glide process occurring during deformation of the crystallites, whereas elsewhere a [110] texture arises. Secondly thc damaged walls of impurities inhibit the growth of these grains less than that of others.

This interpretation of crystal growth also provides an explanation of the observed fragmentation phenomena.

**2752:** J. M. Stevels: L'évolution de la technologie et de la recherche verrière depuis la guerre (Vetro e Silicati **3**, No. 14, 23-30, 1959). (Glass technology and research since the war; in French.)

The author demonstrates that the development of glass technology and research since the war has been extraordinary. The paper includes the following three sections: 1) basic research on glass, 2) manufacture of glass objects, 3) improvements in glass. It is interesting to note that there have been two different trends recently in the technology of glass: 1) realization that partially crystallized glass has particularly attractive properties, 2) realization that very often it is the finishing of the moulded glass that makes its excellent properties evident.

VOLUME 22, 1960/61, No. 3 pp. 61-104 Published 14th January 1960

Philips Technical Review

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

# EXPLORING THE ATMOSPHERE WITH RADIO WAVES

by H. BREMMER.

538.56:621.396:525.7

*In his inaugural address as extra-mural professor of the Eindhoven Technische Hogeschool on 12th February 1960, Dr. Bremmer spoke on the ways in which study of the behaviour of radio waves has enriched our knowledge of the atmosphere. In recent years important new discoveries have been made possible by the development of space research, enabling radio transmitters to be sent out beyond the ionosphere, and by the application of highly sensitive radar methods to observations from the earth.*

*With Professor Bremmer's kind consent and cooperation we print below the main contents of his address *), supplemented by some illustrations and a bibliography.*

In recent years space research has enabled us to enrich our knowledge of the physico-chemical structure of the atmosphere which, in its turn, helps to promote the advancement of space research itself. It is well known that rockets and artificial satellites are equipped with measuring instruments, and that the results of the measurements are sent back to earth in code form by a small radio transmitter in the vehicle. It is not so widely realized that the radio waves thus transmitted can also, during their travel, provide us with direct information on the space through which they pass. The behaviour of the waves is affected by this space, albeit very slightly. Radio waves transmitted from earth are similarly affected, since they too must cover a shorter or longer path through the atmosphere before reaching a receiver. In the latter case, however, it becomes extremely difficult to obtain information on the structure of the atmosphere above a height of about 400 kilometres. Unless very special installations are used, this is the maximum altitude reached by waves from a terrestrial transmitter, in so far at least as they return to earth at all after attaining a highest point and can be detected on earth. With rockets, on the other hand, radio investigations of the atmosphere can be continued beyond the 400 km limit. Projects of this nature were carried out on a limited scale during the recently concluded international geophysical year.

In the following we shall review the background of investigations in which radio waves are used to examine the structure of the atmosphere. In doing so we shall discuss both the results achieved with rockets and those obtained by measurements from the earth's surface.

## The ionosphere as a hypothesis to explain the range of radio waves

Physically, radio waves are related to visible light waves. Both are propagated at a constant speed and along straight paths only when the space through which they pass is either a vacuum or perfectly homogeneous in composition. Our atmosphere, however, is only an approximation to a homogeneous space. Variations in local weather conditions are evidence that the detailed structure of the atmosphere must differ from place to place. The paths of radio waves will therefore be bound to show deviations from straight lines. The fact that these deviations can be considerable was a conclusion reached when it proved possible with certain transmitters to achieve world-wide radio communication. If radio-wave propagation were essentially rectilinear the service area of a transmitter would not extend much beyond the horizon as seen from the

*) Published (in Dutch) by J. B. Wolters, Groningen 1960.

aerial. More distant receivers lie below this horizon and cannot be reached by straight connecting lines from the transmitter. Nevertheless, reasonable reception might still be expected for some distance beyond the optical horizon, i.e. up to the first part of the shadow region, into which radio waves are diffracted to some extent — more than in the case of light waves. The depth of penetration beyond the transmitter horizon can be found mathematically. Investigations by Watson [1]*) in 1918 proved beyond doubt, however, that it was not possible in this way to explain the reception all over the earth of stations working in the wave band between 100 m and 10 m, later so widely used.

As early as 1902, Heaviside in England and Kennelly in America had concluded intuitively that the explanation for the great range of radio waves was to be sought in a non-homogeneous structure of the atmosphere. They put forward the hypothesis of a conducting layer at high altitude. The radio waves would then be able to reach any point of the earth by zigzag paths, being alternately reflected from this layer and from the earth's surface. Heaviside moreover suggested that the layer might contain charged particles formed by the ionizing action of the sun, a supposition that was later shown to be correct. Incidentally, the possible existence of such a conducting layer had already been considered in 1878 by Balfour Stewart as a likely explanation of the daily variation in the earth's magnetism. This ionized region of the atmosphere, originally called the Kennelly-Heaviside layer, is now known as the *ionosphere.*

## Sounding the ionosphere

The ionosphere was at first merely a hypothesis to help explain disparate phenomena, and nothing at all was known about its height above the earth or about its other properties. It was not until 1925 that Appleton and Barnett estimated its height by an interpretation of the effect of fading, i.e. variations of the received strength of radio signals [2]. They showed that medium-wave fading in the hours of darkness could be understood by assuming that two waves reached the receiver simultaneously, one propagated along the surface of the earth, the "ground wave", and the other reflected from the ionosphere, the "sky wave". The observed fading could be explained by the interference of these two waves if the ionosphere were at a height of about 85 km. In the same year this height was first determined, more directly and accurately, by Breit and Tuve [3]. They calculated it from the observed difference in the times of arrival of the ground and sky waves. This is in fact the earliest known example of a radar experiment, since it involved an object that reflects radio waves (the ionosphere) which is not only detected but whose distance is determined. In their publication on the subject, Breit and Tuve remarked at the time: "We are hoping that such experiments will be performed by others as well as ourselves". Their hope has been fulfilled with a vengeance. More than a hundred observer stations are now daily carrying out numerous measurements on the principle indicated by Breit and Tuve. The simplest form of measurement consists in determining the time taken by a vertically transmitted signal to reach the ionosphere and return to earth. This is referred to as the echo time. Systematic "echo-sounding" of the ionosphere has become routine work, comparable with regular meteorological observations.

The echo effect mentioned depends on the wavelength or frequency used. The higher the frequency, the deeper the wave penetrates the ionosphere, and therefore the longer the echo time. Measurements showed that the ionosphere broadly consists of three successive layers extending from about 70 km to 400 km above the surface of the earth. By determining the echo time at many frequencies, and properly interpreting the results, it proved possible to lay bare the detailed structure of the ionosphere. The theory underlying the interpretation of such measurements had in fact been worked out in England by Eccles [4] as early as 1912; it concerned the propagation of electromagnetic waves through a gas containing charged particles, i.e. through a medium similar to the ionosphere. It had been found that where different kinds of charged particles are present at the same time, namely ions and electrons, only the latter affect the way in which a radio wave is propagated. The theory showed in particular that the propagation velocity of a radio wave entering the ionosphere, called the phase velocity, must increase if the wave on its way upwards encountered increasing concentrations of electrons. As a result the originally straight path would be bent downwards, given favourable conditions, and would then return to earth. Similarly curved paths, where waves are bent downwards after reaching a highest point in an atmospheric layer, were already known in the case of acoustic waves generated by explosions and artillery gunfire. The highest point in such cases was much lower, however, being in a layer at a height of about 30 km, this layer being characterized by a high ozone content.

*) References are given at the end of this article.

But let us return to the ionosphere. It is solely because the skyward radio waves undergo an increase in their velocity of propagation that radio communications are possible over great distances. Theoretically, then, it had been reasoned that such an increase must always occur when a wave enters a medium containing charged particles. It remained to verify this conclusion by a laboratory experiment. The first to do so was Van der Pol, who gave a full account of his methods in his thesis in 1920 [5]. A general understanding of the mechanism of radio transmission via the ionosphere thus existed five years before the separate ionospheric layers were first directly observed in 1925.

A further advance was made in 1930 when W. de Groot [6] gave a mathematical method for directly determining the electron concentrations in the layers from the observed frequency dependence of the ionospheric echo times. In this way it was found that in the best-known layers (called D, E and F layers, in ascending order) the number of electrons per cubic centimetre was of the order of $10^3$, $10^5$ and $10^6$, respectively. De Groot pointed out that it was only possible with this method to investigate the lowest part of each layer, i.e. the part below the level where the electron concentration is a maximum. Now, since the advent of rocket missiles, data can be collected on each layer through which the rocket passes [7]. This is done by measuring the frequency change — the so-called Doppler effect — of the radio signal sent back to earth by a small transmitter on board the rocket. Here the Doppler effect depends on the rocket's speed and on the local velocity of the radio wave; the latter depends in its turn on the electron concentrations near the rocket. Since the speed and course of the rocket are known, the electron density at all altitudes of the rocket can be determined directly from the variation of the Doppler effect. These new measurements broadly confirm the picture of the ionospheric layers earlier arrived at with the aid of echo sounding. Faith in the old results was so great that one commentator, discussing the confirmation provided by the rocket tests, said ironically: "This simply means to me that the rockets have in fact got through to the ionosphere". Nevertheless, the correspondence is not perfect; in particular it now appears that the layers of the ionosphere are not so clearly separated as they were formerly supposed to be. For example, the minimum of the electron density at the transition from the E to the F layer is extremely shallow or even imperceptible.

### The space above about 400 km

Until a year ago no detailed picture was known of the electron density above the middle of the F layer, which is the region above a height of about 400 km. More information has now been made available on this region by new radar experiments at a wavelength around 7 m, performed by Bowles [8]) of the National Bureau of Standards. A wave as short as this passes almost unhindered through the ionosphere, but reception is still possible, albeit with a great deal of trouble, of the extremely weak signal which is sent back to earth by the wave on its way up to very great altitudes. This weak signal results from the fact that the individual electrons in the upper atmosphere absorb energy from the oncoming radio wave and then scatter this energy in all directions.

Part of the scattered energy then returns to earth. The returning energy being proportional to the electron concentration, the latter can therefore be determined from the signal received. A measurement is made of the time variation of this signal shortly after the primary signal is sent out. In the first moments, only the contributions produced in the lowest regions of the ionosphere will be observed, these being the earliest to return. Thereafter the intensity is determined by the electron concentrations at increasingly higher levels.

To obtain a measurable signal strength in this experiment, a very powerful transmitter and a large aerial system are needed. These of course involve considerable costs, but the costs are very much lower than would be entailed if an artificial earth satellite were used to acquire the same data. The provisional results show that the electron density above the F layer decreases very gradually and that there are thus no further layers or high electron concentration. Future observations with the aid of artificial satellites will undoubtedly provide supplementary information.

Another important fact has been established by much simpler observations, viz. that up to very great heights a minimum concentration prevails of about 500 electrons per cm$^3$, or at least that there are always local regions present with this electron concentration. This has been inferred in particular from observations of the phenomena known as "whistlers". These are electrical disturbances which are generated by thunderstorms and are propagated from their terrestrial source along a line of force of the earth's magnetic field that extends far into the atmosphere; the path of propagation along this line of force shows a horseshoe bend, finally returning to earth somewhere in the opposite hemisphere (*fig. 1*).

From such observations Morgan and Allcock [9])

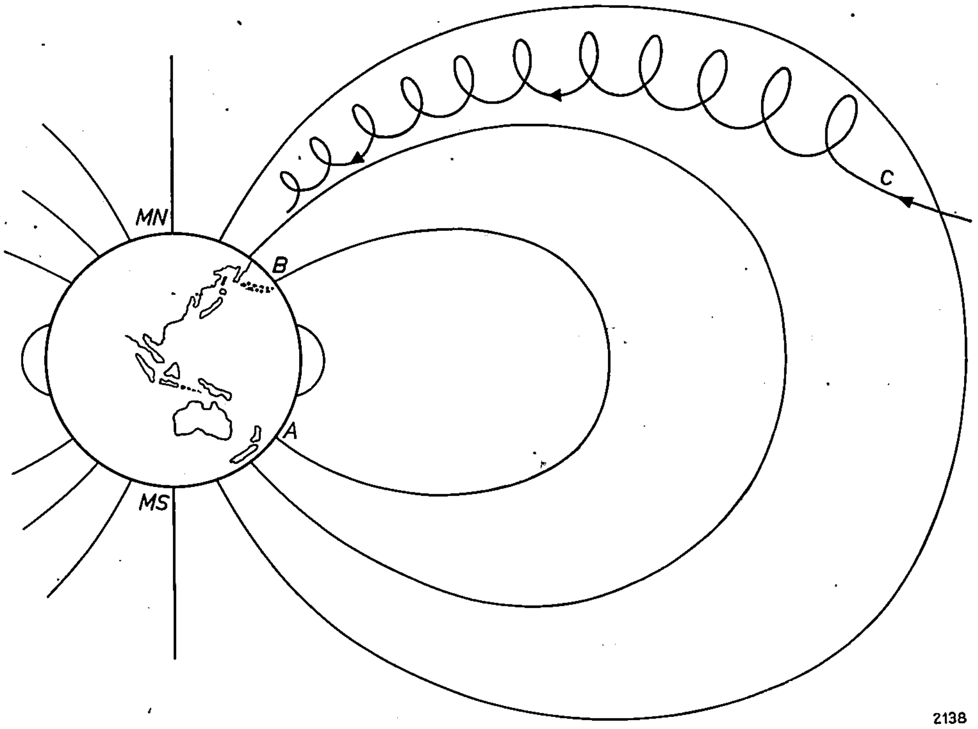


Fig. 1. The earth, with magnetic poles $MN$ and $MS$, showing a number of magnetic lines of force.

Electrical disturbances generated at $A$ (e.g. by a thunderstorm) are propagated along a line of force of the earth's magnetic field and reach a point $B$ in the other hemisphere. At $B$ they may give rise to a "whistler". The disturbance can travel to and fro many times along the line of force $AB$.

When a charged particle — emitted by the sun — approaches point $C$ in the transitional zone, where it first comes under an appreciable influence of the earth's magnetic field, it starts to describe a helical path around a line of force. Upon arrival in the lower atmosphere such particles may produce auroral effects. The paths of the lines of force indicate that the charged particles mainly enter in a region forming a ring around the magnetic north and south poles (aurora borealis and aurora australis).

recorded a case in 1955 where disturbances of this nature repeatedly made the long journey to and fro between Wellington in New Zealand and Unalaska on the Aleutian Islands. On their way they reached at their farthest point a distance of more than 20 000 km from the earth. Whistlers also contain frequencies in the audio range, which are heard in the receiver as a short fluting tone of descending pitch, hence their name. Another related kind of whistling atmospherics appears to originate in the upper atmosphere, probably as a result of fast-moving currents of ionized particles [10]): the sound heard resembles the twittering of birds, and has therefore been termed "dawn chorus". All these disturbances roughly follow a line of force of the earth's magnetic field. The saying that the traveller from afar has much to relate is certainly apt in their case. The properties of a whistler depend on the electrons it has encountered on its journey. Analysis of the incoming signal reveals in particular that the long journey is possible only if it is made through regions where the electronic concentration is at least of the order of the above-mentioned value of 500 electrons per cm$^3$. An important effect in this connection is that the electrons tend to arrange themselves in "filaments" along the lines of force of the earth's magnetic field.

The presence, formerly unsuspected, of relatively high electron densities up to very great heights above the earth (see *fig. 2a*) has been confirmed in other ways. In the first place, it agrees with the recordings of a positive-ion detector on board Sputnik III [11]). The concentration of positive ions is an indication of the presence at the same time of a concentration of negatively charged electrons. Further confirmation has come from a recent study by Siedentopf, Behr and Elsässer of the brightness and polarization of the zodiacal light [12]). In studying this phenomenon it is necessary to assume that the polarization effects are due solely to electrons, leaving out of consideration any effects that may be due to the dust particles also present. This means that the electron densities thus calculated are maximum values. Thus, from observations of three entirely distinct phenomena it has been made plausible that a minimum density of about 500 electrons per cm$^3$ exists up to very great distances from the earth.

The investigations mentioned hitherto enabled the electron concentrations to be inferred up to a certain height above the earth. It is also possible, however, to determine by direct means the total number of electrons contained in a column extending from the earth's surface far into cosmic space. For this purpose use is made of the recent radar experiments, where a radio signal is reflected from the moon. From the fading shown by the returned signal one can find the number of electrons contained in a narrow column reaching from the earth to the moon. Measurements by Evans [13]) and by Bauer and Daniels [14]) have shown that this number amounts to about $20 \times 10^{12}$ electrons in a column of 1 cm$^2$ cross-section. Suppose for a moment that the prevailing electron density was 500 per cm$^3$ over the whole distance from the earth to the moon (380 000 km); the column would then contain $19 \times 10^{12}$ electrons. This of course leaves hardly any margin for the much greater electron density that must in reality be present in the column at the position of the E and F layers: the necessary margin exists, however, if we assume an electron density outside these layers of, say, 400 per cm$^3$, which is of the same order of magnitude as the 500 per cm$^3$ mentioned above.

The results obtained indicate that the total number of electrons above the "middle" of the F layer is 3 to 5 times greater than the known number of electrons below that level (by "middle" we mean here the level of maximum electron concentration, denoted by $M$ in fig. 2*a*). With a similar method [15]) the total number of electrons can be determined between the earth and an artificial satellite in orbit at a specific height. This number can also be calculated from the recorded moment at which signals transmitted by such a satellite are last received after the satellite has disappeared beyond the horizon [16]). The same applies to the moment at which the signals are first picked up again when the satellite reappears above the horizon. In this way, with the aid of Sputnik II, it was established that the electron density in the upper half of the F layer declines much more gradually than it increases in the lower half upwards. This is in agreement with the measurements performed by Bowles [8]).

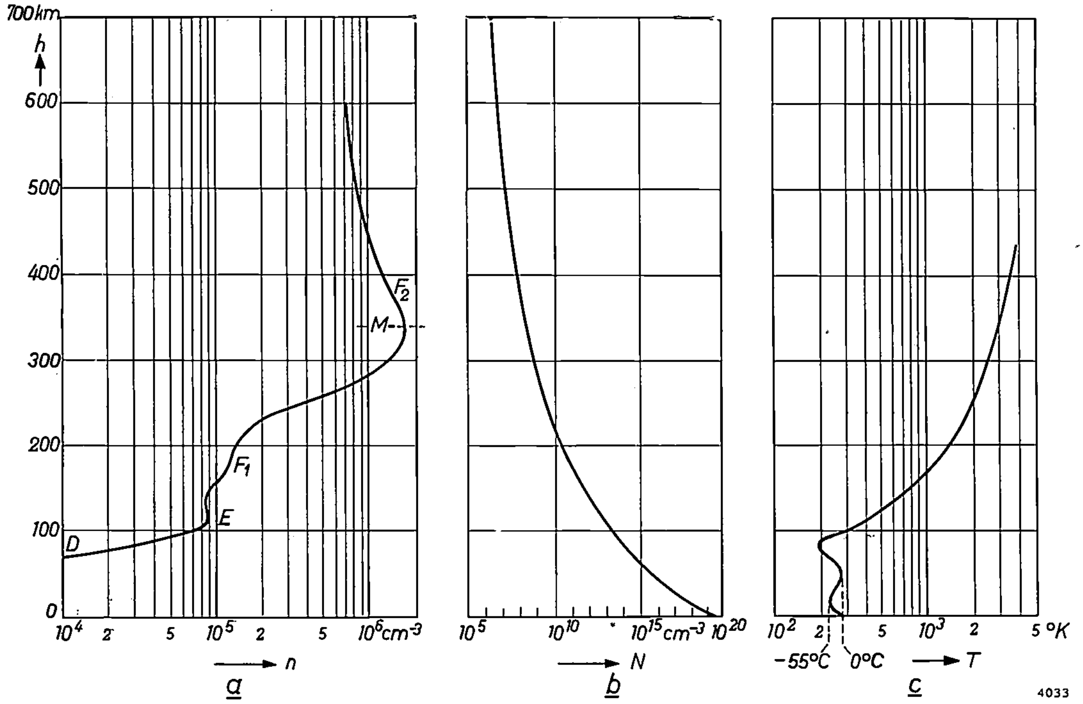


Fig. 2. The graphs show respectively as a function of height $h$ above the earth's surface:
*a*) the number of electrons $n$ per cm$^3$,
*b*) the number of particles $N$ per cm$^3$ (here mainly molecules in the lower regions and mainly atoms at higher altitudes),
*c*) the temperature $T$ (in °K, but also in °C at lower altitudes).
In (*a*) the letters $D$, $E$, $F_1$ and $F_2$ denote the correspondingly named layers of the ionosphere; the greatest electron concentration is found at the level $M$ (the "middle" of the F$_2$ layer). The change with height in the concentration of the electrons varies with the relative position of the sun and with the sun's activity. The curve shown is representative of a state during the day, when the F layer can be divided into an F$_1$ and an F$_2$ layer.

**Density and temperature of the atmosphere as a function of height**

There is thus a great deal of information available for studying the concentrations of electrons in the atmosphere. Being established with a relatively high degree of certainty, the data at our disposal represent an excellent starting point for building up a picture of the physical conditions existing in the space around our planet. In the first place, once we know the distribution of the electron concentrations in a particular layer, we can deduce from that distribution the molecular density and the temperature prevailing near the middle of the layer. This is possible because the formation of the layer partly depends on the absorption of ionizing radiation from the sun by atmospheric molecules or atoms. Factors thereby involved are the average concentration of these uncharged particles, and also the change which their concentration undergoes with changing height as a result of temperature. It has been found in this way that the gas density at a height of 100 km is roughly a million times smaller than near the earth's surface (fig. 2*b*), whilst the temperature at that height must be roughly equal to room temperature (fig. 2*c*). This does not imply that one would feel comfortably warm in this region, but simply that the molecular velocities of thermal agitation are just about the same there as in the air in which we live. Above the 100 km level, however, there must be an increase in temperature, otherwise the density of the uncharged gas atoms would decrease upwards much faster than it actually does.

Apart from the results obtained with radio waves, there are other indications that the decrease in the density of the air above about 100 km is so slow that it must necessarily be accompanied by a rise in temperature. These indications have come from air-pressure measurements aboard rockets and also from observations of auroral effects and of the light in the night sky[17]). As regards the latter, it should be recalled that on a clear, moonless night only about 30% of the faint light from the sky originates from the stars, directly and by scattering in the atmosphere, whereas some 40% is due directly to luminous gases in the upper atmosphere (the other 30% is scattered light due to other causes). From observations of this light we can thus estimate the density and temperature of the gases up to a height of about 1000 km.

The rise in temperature above 100 km is quite understandable, since the extremely thin atmospheric gas must finally make the transition to the so-called *interplanetary gas* of outer space. It is known that this gas has a very high temperature, and moreover it consists almost entirely of the simplest charged particles, namely protons and electrons. It is thought that the high velocities of these particles, corresponding to their high temperature, are such as to overcome the attraction of the planets, though not of the sun. The sun itself is probably the chief source of interplanetary gas, which might be regarded as a continuation of the rarefied and very hot gas of corona that envelops the sun.

Let us now consider the general picture of the atmosphere above about 500 km, as pieced together from theoretical insight and from the scarce data provided by rocket flights. The chemical composition of the air up to roughly 200 km differs only very slightly from that of the air we breathe, but above that height the atomically dissociated oxygen and nitrogen are gradually superseded by the much lighter atomic hydrogen. Above a height of the order of 1000 km the density has diminished to such an extent that the chance of gas particles meeting one another is very remote indeed and there are virtually no more interaction processes between the atoms, ions and electrons still present. Consequently the concentration of the uncharged atoms is henceforth entirely governed by the force of gravity, so that at very great heights only the lightest gas is found — hydrogen. At a distance of three earth radii, i.e. at a height of about 20 000 km, the (uncharged) hydrogen atoms have sufficient velocity, in view of the temperatures prevailing there (~1200 °C), to overcome the very weak gravitational pull of the earth. As far as the (electrically charged) ions and electrons are concerned, however, the effect of the earth's magnetic field, and the forces associated with it, are more important than the force of gravity. These particles remain much longer within the earth's sphere of influence, so that finally only the lightest, charged particles remain, i.e. protons and electrons. The latter are found there in the concentration referred to of the order of 500 particles per cm$^3$. For the sake of completeness it may be mentioned that Geiger counters carried by rockets have revealed the existence of at least two zones, called Van Allen zones, of highly intensive radioactive radiation [18]), whose maxima are situated at heights of about 3000 and 16 000 km above the earth's surface. The radiation is probably due mainly to high-speed electrons originating from the sun.

**Transition of the atmosphere to interplanetary gas**

The model described here seems to point to a very slow transition from the increasingly rarefied upper atmosphere to the region of interplanetary gas, which, according to Siedentopf and co-workers [12]), also possesses near the earth a density of the order of 500 electrons per cm$^3$. It should be remembered, however, that this gas does *not* take part in the daily rotation of the earth, whereas the air near the earth is carried around in its entirety. With increasing height, then, the air ought to be gradually less firmly bound to the earth. This, however, raises a theoretical difficulty. If the earth's magnetic field decreases with increasing distance at the same rate as it does near the earth, one can calculate the viscous forces due to electromagnetic effects that correspond to a minimum density of 500 electrons per cm$^3$. It is found that these forces are so strong as to suggest that the atmospheric air down to regions close to the earth is coupled more with the interplanetary gas, which does not move with the earth, than to the rotation of the earth. The movement of the air that does not entirely rotate with the earth should manifest itself in a prevailing east wind, and this should already be observable at a height of 100 km above the earth. This is certainly not the case, however. The difficulty disappears if there exists a transition zone in which the earth's magnetic field declines so rapidly as to be negligible beyond that zone. The form of the magnetic lines of force that fits this model is then such that the protons and electrons near the transition zone can only penetrate through it with great difficulty; the possibility of limited penetration from outside is then essential to explain auroral effects. The particles outside the transition zone are thus more or less isolated from those inside it. This makes it possible for all particles inside to rotate with the earth, whilst those outside it are coupled with the interplanetary gas. The transition zone, which may perhaps coincide with the central part of the outer Van Allen zone, at a distance of roughly 16 000 km, then acts as a natural boundary of the earth's atmosphere, at a level where it consists almost entirely of protons and electrons.

The transition zone thus screens the earth's magnetic field, and this implies theoretically that it must at the same time be the carrier of electric currents. It was for this reason that the existence of such a zone was first postulated, for these currents are the simplest explanation of phenomena connected with the disturbances of the earth's magnetism known as "magnetic storms". Since 1923 the relevant theory has been worked out in particular by Chapman, Ferraro and Martyn [19]). They have shown that a transitional layer must necessarily be formed whenever a stream of charged particles (thrown off by the sun) enters a magnetic field that initially decreases very gradually as a function of the distance to the earth. Any stream of particles will then tend to distort the latter field into a field of the type considered above. The regular production of such streams by solar eruptions accordingly maintains this type of field with a transitional layer. The most recent observations indicate that the transitional layer may be identical with a boundary region in which the earth's magnetic field loses its regular and slowly varying structure — in other words, the transitional layer represents the outermost zone in which the field still possesses a distinctly stable component.

### Ionospheric winds

We have just said that there are no indications of a prevailing east wind at a height of 100 km. This appears from a variety of investigations, again using radio waves reflected from the ionosphere [20]). The fading of these waves may be studied. Because of irregularities in the structure of the ionosphere, the moments at which the signal from a given transmitter is received most strongly by several receivers in each other's vicinity do not coincide. From the time differences recorded one can determine the direction and the strength of the wind prevailing at ionospheric altitudes. No prevailing east wind is observed, but winds of considerable velocities do occur at a height of 100 km. Wind velocitics of the order of 50 metres per second, i.e. 180 km/h, appear to be quite normal. Such hurricanes should not be imagined too dramatically, however, for the air density there is about a million times less than at the earth's surface. The mass of air displaced per unit time at such high velocities is therefore very small — too small, for example, to turn a rocket noticeably off course.

### Radar observation of meteors

In recent years the ionospheric winds in the E layer have been very systematically studied by radar observations of meteors [21]). In spite of its rarefication, the air at a height of 115 km is still dense enough to make meteors entering the atmosphere at that level white hot. The resultant vaporization is so intense that most meteors have completely evaporated before they can drop to an altitude of about 80 km. The heating process is accompanied by the ionization of atoms from the meteor, and a temporary trail of strongly ionized air is left behind. The trail may remain intact for as long as ten seconds, after which it dissipates as a result of diffusion. For a short time, then, there exists a cylindrical, expanding column in which electrons occur at a concentration often ten-thousand times greater than in the surrounding air of the ionosphere. Radar waves are reflected from a short-lived column of this kind, and meteors detected in this way occur at an average rate of one per second, against one every seven minutes observable by the naked eye. With a telescope almost as many meteors can be observed as with radar, but the latter has the advantage of being just as useful during the day as in the hours of darkness. These investigations can therefore be carried out both by day and night.

The study of meteors by radar has attracted considerable interest in the last ten years. Information can also be gathered in this way on air currents at high altitudes, since the ionized trails left behind by meteors are blown along during their brief existence by local winds. The component of this wind motion in the direction of observation can be derived from the Doppler shift in the frequency of the reflected waves. Statistical analysis of the wind components measured on large numbers of meteors makes it possible to calculate the force and direction of the prevailing wind in a given region of the ionosphere. These calculations confirm the above-mentioned wind velocities of the order of 50 m/s. The most direct indication of these wind velocities, however, is found from observations of the "luminous clouds" that are sometimes seen at high altitudes a few hours after sunset or before sunrise [22]).

### Atmospheric tides

The wind phenomena discussed here share to some extent the random nature of the winds near the earth's surface, which are governed by meteorological conditions. A large contribution to the ionospheric winds, however, is attributable to solar and lunar tidal forces. We are most familiar with tidal forces from the periodic alternation of ebb and flow in the seas. The same forces act on the air of our atmosphere, but they are much less noticeable. Our position as observers in relation to the atmosphere might be compared with that of someone trying to study sea tides from the bottom of the sea. This comparison suggests that the atmospheric tidal effects are perhaps much more pronounced at greater altitudes. A periodic vertical movement of air as a result of tides might, for example, manifest itself as periodic variations in the height of each ionospheric layer. In fact, echo-time measurements revealed for the first time in 1939 that these layers do indeed show a periodic rise and fall. In that year Appleton and Weekes [23]) found that the height of the E layer undergoes small variations of the order of 2 km, which are directly related to the phase of the moon.

Vertical tidal movements in the air, like those in the sea, are not conceivable without accompanying horizontal movements. One might therefore suppose tidal effects to be at the back of the wind phenomena detected by radar observations of meteors. Tidal winds do in fact appear to exist in the ionosphere. The effects due to the sun and moon separately can be kept distinct in this connection inasmuch as their respective contributions vary with the position of the sun and the moon. The tidal wind caused by the sun shows a highly regular pattern; at an altitude of 85 km in the northern hemisphere it may be described broadly as a wind, constantly changing in direction, and veering from the west in the morning and evening at half past eight local time; its maximum force is roughly 70 km/h.

Tidal winds as strong as this are out of the question near the earth's surface, where the prevailing winds are governed by meteorological conditions. Still, a tidal contribution does exist on the earth, albeit a very slight one. It can be found from the averages of barometric readings taken over a long period at times when either the sun or the moon is at the same position in the sky. By taking average readings the influences of incidental and constantly changing meteorological conditions are eliminated. In this way one finds a small tidal effect attributable to the sun, and further a 16 times smaller effect due to the moon. From the local distribution of these accurately determined statistical averages the associated, very slight tidal contribution to the wind on the earth's surface can later be calculated [24]). The results show that, on the equator, the tidal action of the *sun* superimposes on the meteorological winds an extra east or west wind which has a maximum force of less than one kilometre per hour. In higher latitudes this weak tidal component constantly changes direction, just as it does in the ionosphere. The even weaker atmospheric tidal effects due to the *moon* [25]) indicate that our faithful satellite has no significant influence on the distribution of the air in our atmosphere.

A comparison of the tidal winds blowing at velocities up to 70 km/h at a height of 85 km with the tidal wind of about 1 km/h near the earth might suggest that the tidal effects are generated primarily in the higher layers of the atmosphere. This, however, is by no means the case. If we take into account the rarefication of the upper air, we find that the energy transmitted by the tidal forces to unit volume of air is much greater near the earth's surface than in the ionosphere. The energy taken up near the earth, however, gradually moves upwards, thereby appreciably strengthening the tidal wind due to the sun, at least above a height of about 30 km.

It was for a long time puzzling that the sun's contribution to the atmospheric tides should be so much greater than the moon's. This seemed to conflict with the elementary theory according to which the moon's contribution should be 2½ times greater than that of the sun, a deduction based on the relative masses of the sun and moon and their distances from the earth. As regards the ocean tides, the moon is in fact the more effective of the two. We know that the times of high water along the coasts are almost entirely governed by the relative position of the moon; the weaker effect of the sun is responsible for the spring tides shortly after full moon and new moon. The fact that the sun plays the major part in the atmospheric tides was noticed by Laplace one-and-a-half centuries ago. In 1882 Kelvin suggested that a resonance effect might be involved. He conceived that the atmosphere might easily enter into an oscillatory movement whose period, for one reason or another, may well be close to the twelve-hour period governing the solar tides. On the other hand, the corresponding period of nearly thirteen hours for the lunar tides would not be close to a resonance period of the atmosphere. It was not until 1936 that it was first shown by Pekeris [26]) that among the resonance periods of the atmosphere there is in fact one of about twelve hours.

The duration of this resonance period, which has such an important bearing on the atmospheric tides, is very closely related to the temperature distribution in the upper levels of the atmosphere [27]). It would be quite different if the temperature of the air decreased upwards continuously. We know that the temperature has a minimum value of about —55 °C at an average height of 10 km, after which it rises and at about 50 km reaches an average value near 0 °C (fig. 2*c*). It then drops again until, at a height of roughly 85 km, it reaches a minimum value of about —80 °C. This is where the above-mentioned temperature rise begins, which continues right on up to the transitional zone bounding the atmosphere. It is the presence of two layers in which the temperature drops with increasing height that involves a resonance period of roughly twelve hours. The theory of resonance also indicates that the solar tidal wind at a height of 75 km must be about 100 times stronger than near the earth, and moreover that above 30 km it must blow in the opposite direction. This is entirely in agreement with observations of wind directions in the ionosphere, and serves to strengthen confidence in the theory.

In broad lines it can be said that theory and observation together have produced a satisfactory picture of the atmospheric tides in the lower atmosphere. According to the so-called dynamo theory [28]) the ionospheric tidal winds are partly responsible for the systems of electric eddy currents in the ionosphere; this ties up reasonably well with what can be deduced about these currents from the study of the earth's magnetism. Radar observations of meteors further indicate that above 100 km the tidal winds are rapidly attenuated, apparently because the tidal waves are strongly damped when they enter this region. This can be explained by the effects of viscosity and thermal conduction that first become effective there. Future investigations will undoubtedly deal in greater detail with the attenuation of tidal effects at high altitudes.

### Concluding remarks

The reader may now be wondering whether the investigations discussed are of any technical importance apart from their purely scientific interest. It should be remembered that research with the aid of radio waves can provide fresh insight into the uses of radio waves as a means of communication. In this connection it may be recalled how the familiar reflection and scattering of radio waves from ionized clouds prompted the American Thaler to turn this phenomenon to use for tracing guided missiles and nuclear explosions [29]), both of which give rise to such clouds.

I have tried to show in this survey how very valuable a tool radio has proved to be for exploring the mysteries of our atmosphere. In the skies above us there are many long-unsuspected phenomena at work, which are only now gradually yielding up their secrets. For the physicist it is a fascinating field of research, involving as it does such diverse branches of study as radiation theory and plasma physics, which are particularly important in the upper layers, the physico-chemical theories underlying the ionization and energy-exchange processes between the particles in the somewhat lower layers, aerodynamic and electrodynamic theories, which apply to the air currents above 100 km, and classical mechanics, which govern the tidal effects in the lowest layers. No one is entirely indifferent to the achievement of human ingenuity in establishing radio communication all over the earth, and it is natural that we strive to understand more of the mechanisms that make that communication possible. One final comment may not be out of place in this connection. It is remarkable how the subtlest-seeming phenomena play a fundamental role in radio telecommunications. For example, the highly rarefied gas of the ionospheric F layer has a lower density than the so-called vacuum in the best high-vacuum pumps; nevertheless, it is this gas, through the thinly distributed electrons it contains, that enables us to listen to a station at the other end of the earth.

## Bibliography

[1] G. N. Watson, Proc. Roy. Soc. A **95**, 83, 1918.

[2] E. V. Appleton and M. A. F. Barnett, Nature **115**, 333, 1925 and Proc. Roy. Soc. A **109**, 621, 1925.

[3] G. Breit and M. A. Tuve, Phys. Rev. **28**, 554, 1926.

[4] W. H. Eccles, Proc. Roy. Soc. A **87**, 79, 1912.

[5] B. van der Pol, The influence of an ionized gas on the propagation of electromagnetic waves and the applications thereof in the field of wireless telegraphy and in measurements on glow discharges, Dissertation, Utrecht 1920 (in Dutch).

[6] W. de Groot, Phil. Mag. **10**, 521, 1930.

[7] J. C. Seddon, J. geophys. Res: **58**, 323, 1953. H. Friedman, Proc. Inst. Radio Engrs. **47**, 272, 1959 (No. 2).

[8] K. L. Bowles, Phys. Rev. Letters **1**, 454, 1958.

[9] M. G. Morgan and G. McK. Allcock, Nature **177**, 30, 1956.

[10] See e.g. R. M. Gallet, Proc. Inst. Radio Engrs. **47**, 211, 1959 (No. 2).

[11] V. I. Krassovsky, Proc. Inst. Radio Engrs. **47**, 289, 1959 (No. 2).

[12] H. Siedentopf, A. Behr and H. Elsässer, Nature **171**, 1066, 1953.

[13] J. V. Evans, Proc. Phys. Soc. **B 69**, 953, 1956.

[14] S. J. Bauer and F. B. Daniels, J. geophys. Res. **64**, 1371, 1959 (No. 10).

[15] W. W. Berning, Proc. Inst. Radio Engrs. **47**, 280, 1959 (No. 2).

[16] I. L. Alpert, F. F. Dobriakova, E. F. Chudesenko and B. S. Shapiro, C. R. Acad. Sci. USSR **120**, 743, 1958.

[17] See e.g. S. K. Mitra, The upper atmosphere (Asiatic Soc., Calcutta, 2nd impression, 1952), Chapter X, Section 1.

[18] J. A. Van Allen, J. geophys. Res. **64**, 1683, 1959 (No. 11).

[19] See e.g. S. Chapman and V. C. A. Ferraro, Terr. Magn. atmos. Electr. **36**, 171, 1931.

[20] See e.g. C. O. Hines, Proc. Inst. Radio Engrs. **47**, 176, 1959 (No. 2).

[21] See e.g. L. A. Manning and V. R. Eshleman, Proc. Inst. Radio Engrs. **47**, 186, 1959 (No. 2).

[22] See Chapter VI, Section 14b, of book referred to under [17].

[23] E. V. Appleton and K. Weekes, Proc. Roy. Soc. **A 171**, 171, 1939.

[24] See e.g. J. Bartels, Handbuch der Experimentalphysik, **25**. Geophysik, Part 1 (Akad. Verlagsges., Leipzig 1928), p. 208.

[25] See p. 182 of book referred to under [24].

[26] C. L. Pekeris, Proc. Roy. Soc. **A 158**, 650, 1937.

[27] See e.g. K. Weekes and M. V. Wilkes, Proc. Roy. Soc. **A 192**, 80, 1947.

[28] See e.g. J. A. Fejer, J. atmos. terr. Phys. **4**, 184, 1953.

[29] Time (Atlantic edition), 17 Aug. 1959.

**Summary.** The substance is reproduced of the address delivered by the author on his inauguration as extra-mural professor at the Technische Hogeschool Eindhoven. The subject matter is that of investigations of the atmosphere with the aid of radio waves. After recalling the historic work on the ionosphere done by Appleton and Barnett and by Breit and Tuve, the author mentions the recent work of Bowles, who, by radar soundings from the earth, has obtained data on the upper region of the F layer. These data confirm observations made by instruments on board artificial satellites, namely that the electron density above the middle of the F layer changes much more slowly with height than below it. Amongst the important discoveries made possible by the development of astronautics are the Van Allen zones of intense radiation at distances of more than 2000 km from the earth. Other subjects discussed are the transition from the atmosphere to the region of interplanetary gas, the existence of winds and tides in the ionosphere, and the use of radar for observing meteors.

# MICROPHONY IN ELECTRON TUBES

by S. S. DAGPUNAR *), E. G. MEERBURG **) and A. STECKER ***).

621.391.816.2:621.385

*Microphony may be defined as the occurrence of an electrical interfering signal produced as a result of mechanical or acoustical vibrations of a circuit element, e.g. an amplifying tube. The effect is as old as the radio tube itself. At first it could be kept within bounds by mounting the tubes in resilient holders. At the levels of amplification common nowadays, however, this simple measure is far from sufficient. Theoretical and experimental investigations have shown what can be done in the design and construction of a tube to minimize microphonic effects. The article below, which embodies contributions from British, Dutch and German laboratories, gives some idea of these investigations and of the progress made in recent years in combatting microphony.*

## Introduction

Amongst the component parts of radio sets, amplifiers, etc., there are many that do not constitute a mechanically rigid assembly, but consist of parts capable of physical vibration at a frequency generally within the audio region. As the parts vibrate the distance between them alters, and this is accompanied by fluctuations in the electrical properties of the circuit element involved. Take, for example, a variable capacitor: if the plates vibrate with respect to one another, the result is a periodic variation in the capacitance. If the capacitor is part of the tuned circuit of an oscillator, the frequency of the generated voltage will also vary periodically, i.e. it will be subjected to frequency modulation, giving rise to interference in the output signal. This production of an interfering signal as a result of mechanically vibrating components is known as microphony.

Electron tubes are particularly subject to microphony, and in this article we shall be concerned solely with microphonic effects in electron tubes. Physical vibration of the electrode assembly not only causes variations in the capacitances between the electrodes but also fluctuations of the anode current and mutual inductance, and hence directly affects the gain of the tube.

There are many causes of vibration in an electron tube. Apart from incidental vibrations or shocks, there are those to which car radios, transceivers, radio equipment in aircraft, etc., are constantly subjected, there are the vibrations due to the motor in gramophones and tape recorders, the mechanical shocks caused by the operation of switches in various equipment, and above all the vibrations caused by the loudspeaker. Loudspeakers are often placed very close to amplifying tubes and can transmit vibrations to the latter both acoustically (via the air) and mechanically (through the cabinet, the chassis and the tube holders). This situation is particularly dangerous in that the loudspeaker itself reproduces the interfering microphony signal; if the gain is sufficiently high, this may give rise to acoustic feedback ("howling"), and if not, it may in any case produce troublesome reverberation. Microphony can also produce severe interference in television receivers. Loudspeaker vibrations here may be transmitted to amplifying tubes in the high frequency, intermediate frequency or video frequency part of the receiver, causing troublesome fluctuations in the brightness of the picture. Microphony in tubes in the deflection circuits may distort the picture as well as cause displacements in lines.

In recent years extensive investigations have been carried out in many laboratories both into the requirements to be met by tubes in modern equipment in order to minimize microphonic effects, and into the measures that can be adopted to make the tubes fulfil these requirements. This article will deal with the work done along these lines in various Philips laboratories and the results obtained [1]).

*) Mullard Radio Valve Co., Ltd., Mitcham, England.
**) Electron Tube Division, Philips, Eindhoven.
***) Development Laboratory of Valvo GmbH, Radioröhrenfabrik, Hamburg.

[1]) See also the following publications:
B. G. Dammers, On the microphony of the EF 86, Electronic Appl. **16**, 125-134, 1955/56;
B. G. Dammers, A. G. W. Uitjens, E. G. Meerburg and M. A. de Pijper, Reflections on microphony, Electronic Appl. **18**, 15-18, 1957/58;
B. G. Dammers, A. G. W. Uitjens, K. Hoefnagel, E. G. Meerburg and M. A. de Pijper, Causes and effects of microphony in the R.F. and I.F. stages of television receivers, Electronic Appl. **18**, 48-56, 1957/58;
A. Stecker, Die Mikrofonie der Elektronenröhre — Theorie und Analyse, Valvo Berichte **4**, 1-21, 1958 (also Electronic Appl. **18**, 99-117, 1957/58);
H. Hellmann, Die Prüffeldmessung der Mikrofonie von Elektronenröhren, Valvo Berichte **4**, 22-35, 1958;
D. Hoogmoed, Microphonic effects in electron tubes, Electronic Appl. **19**, 25-44, 1958/59.

### Factors determining the strength of the microphony

An electron tube subjected to acoustical and/or mechanical vibrations undergoes a periodically alternating acceleration. It is the magnitude of this acceleration that primarily determines the strength of the microphony. To give an idea of the accelerations involved, it may be mentioned that measurements with vibration pick-ups in radio and television receivers have shown [2]) that a loudspeaker fed with a power of 50 mW gives rise to tube accelerations from $0.1g$ to $0.25g$ ($g$ = acceleration of the force of gravity). A higher power evidently causes greater accelerations, the increase being proportional to the root of the power. In car radios the accelerations produced by engine vibrations are much greater than those caused by the loudspeaker. Of course, the type of car, the state of the engine and other conditions are important in this respect. Tests made on the instrument panels of numerous types of cars have shown that, under certain circumstances, accelerations up to $25g$ may occur.

Apart from the magnitude of thc vibrations to which the tube as a whole is subjected, the extent to which the vibrations are transmitted from the base or wall of the tubc to the electrodes also has an important bearing on the strength of the microphony. Further factors involved are the stiffness of the componcnts and the rigidity of their mountings.

A further point to be taken into account in this connection is the function of the tube in the apparatus concerned, since this function determines the parameter whose fluctuations may prove most troublesome. For instance, where a tube is to be used in a low-frequency amplifier, changes in the capacitances between the electrodes will seldom be important, whereas variations in the anode current as a result of electrode vibrations may be very important indeed, since these variations, after amplification, are usually applied to the loudspeaker and made audible. Capacitance variations, on the other hand, can be very troublesome in the oscillator tube in a superheterodyne receiver, particularly if the receiver is tuned to a high frequency. In that case the circuit capacitance is small, and as a result the tube capacitances have a considerable effect on the frequency of the voltage generated by the oscillator. Periodic variation of these capacitances thus gives rise to frequency modulation which, in an FM receiver, is heard through the loudspeaker. This may also be the case in an AM receiver if the set is not exactly tuned to the received signal. Variations in the frequency of an IF signal then give rise to amplitude modulation which, after detection, again results in an interfering low-frequency voltage. Amplitude and frequency modulation may also be caused by capacitance variations in one or more of the radio frequency or intermediate frequency circuits of a receiver, giving rise to fluctuations in the magnitude and phase of the output voltages of the amplifier stages involved.

In cases where microphony causes fluctuations of mutual conductance, the effect can be troublesome if the tube is used in the radio frequency or intermediate frequency circuits of an AM receiver, since a periodically varying mutual conductance results in a variable gain, and thus modulates the RF or IF signal voltage in amplitude.

Microphony in a tube is more troublesome the more amplifier stages are connected bchind the tube, in which case a correspondingly smaller variation in one of the parameters of the tube will be sufficient to produce an impermissibly large alternating current to the loudspeaker.

### Inconsistent nature of microphony

Because of the numerous factors governing its strength, microphony in practice is an irregular, inconsistent phenomenon. A tube fulfilling a certain function in a particular apparatus may give no difficulties, whereas in another function or another apparatus it may exhibit excessive microphony. The location of the tube and the position in which it is mounted may also have considerable influence. Moreover, individual tubes of the same type may show marked disparities. In spite of the extremely narrow tolerances used in the manufacture of components, it is impossible to avoid slight constructional differences from tube to tube. This has no significant effect on the purely electrical properties of the tube, but it may give rise to considerable differences as far as microphony is concerned. Consequently, certain practical methods of testing can only be carried out on a statistical basis; whether a particular modification introduced in a tube will improve the tube's microphonic behaviour in practice can only be established by investigating a fairly large number of individual tubes.

The inconsistent nature of microphony is accentuated by the fact that the frequency spectrum of the vibrations to which the tubes are subjected in practice is extremely irregular in shape. The reason for this is that the chassis, the cabinet and other structural elements of electrical apparatus exhibit many different resonance frequencies for mechanical and acoustical vibrations, so that the whole assembly behaves as if it consisted of large numbers of

[2]) See the article by Hellmann under [1]).

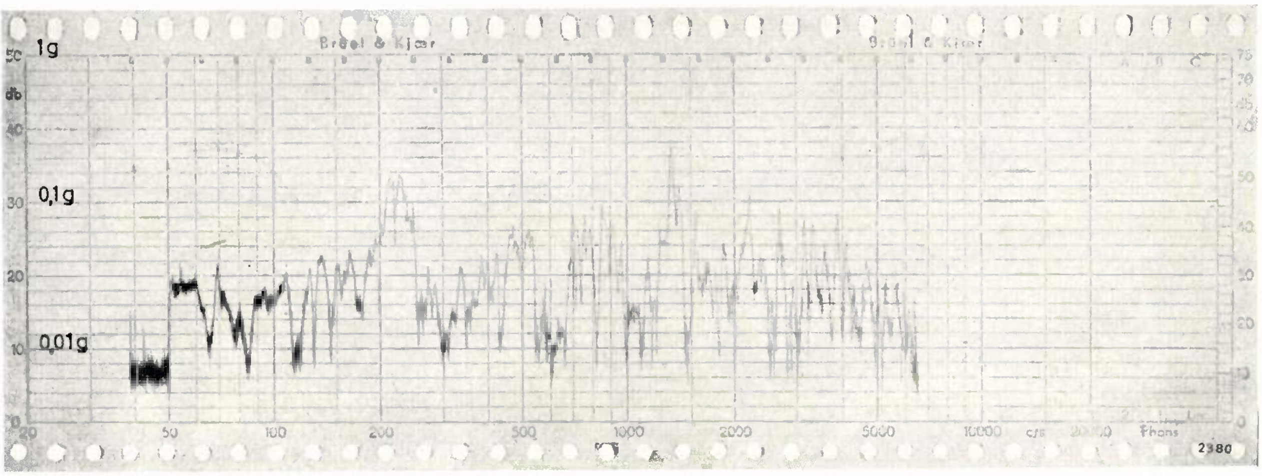


Fig. 1. Frequency spectrum of the acceleration to which a tube in a certain type of radio receiver is subjected when the loudspeaker is driven by a constant power of 50 mW at a varying frequency.

mutually coupled resonators. *Fig. 1* shows an example of a frequency spectrum of the acceleration undergone by a tube in a radio receiver when a constant electrical power of 50 mW is supplied to the loudspeaker at a variable frequency. A frequency spectrum of this kind is obtained by substituting for the tube a vibration pick-up, mounted in a container whose dimensions and weight correspond approximately to those of the tube. Where three pick-ups are used, mounted in directions perpendicular to each other, one can also determine the direction in which the accelerations occur. Such a combination of three vibration pick-ups is shown in *fig. 2*. The whole assembly is roughly as heavy as an electron tube and can be inserted in one of the tube holders in the apparatus under test.

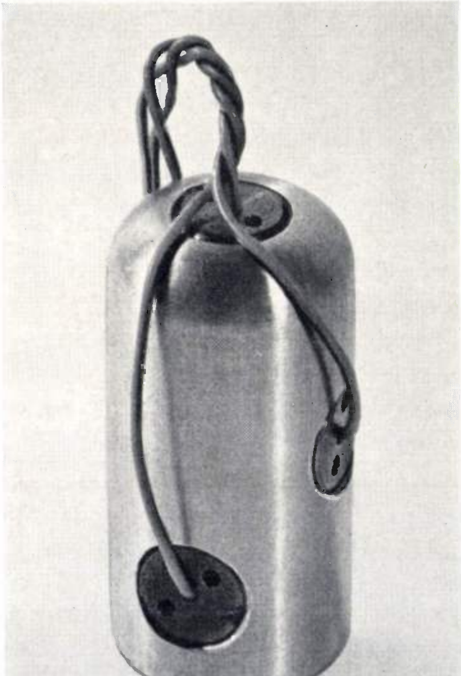


Fig. 2. Combination of three vibration pick-ups, used for measuring the vibrations to which tubes are subjected in electronic equipment. The whole assembly can be fitted in a tube holder in place of a tube.

## Methods of investigating microphony

There are various direct methods of investigating microphony that can easily be carried out without special equipment. For instance, the microphonic tendency of an audio amplifying tube can be ascertained by incorporating the tube in an amplifier circuit. The output voltage is applied via a variable amplifier to a loudspeaker set up near the tube. This arrangement is shown schematically in *fig. 3*, where *B* is the tube under test, *L* the loudspeaker and *A* the variable amplifier. The gain of *A* is adjusted until it is just sufficient to cause acoustic

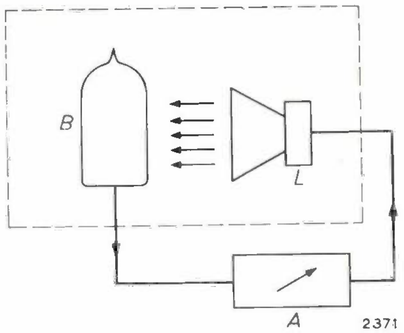


Fig. 3. Principle of a simple set-up for investigating microphony in an audio-frequency amplifying tube. *B* tube under test, *L* loudspeaker, *A* variable amplifier.

oscillation, after which the "sensitivity" of the combination of $B$ and $A$ at this setting is found. This is generally taken to mean the alternating voltage required on the control grid of $B$ in order to produce an output of 50 mW from $A$. The gain setting of $A$ so found is clearly too high when used with the tube $B$; it is thus possible to give a sensitivity of the combination that *is* permissible, to guide users of this type of tube.

Obviously, a specification of this kind is useful only in a circuit arrangement exactly corresponding to that with which the experiments were carried out. A small constructional change in the apparatus in which the tube is used can considerably alter the tendency to microphony. For this reason, and because of the above-mentioned spread between individual tubes of the same type, it is always necessary to allow a wide safety margin.

A radio-frequency tube can be tested in a similar way. The tube is incorporated in an RF amplifier stage and an unmodulated RF signal voltage is applied to the control grid (see *fig. 4*). A detector is connected to the output of this amplifier stage, which is again followed by a variable audio-frequency amplifier and a loudspeaker. Microphony now causes modulation of the RF voltage, and the detector delivers an audio signal which, via the AF amplifier and the loudspeaker, can produce acoustic oscillation.

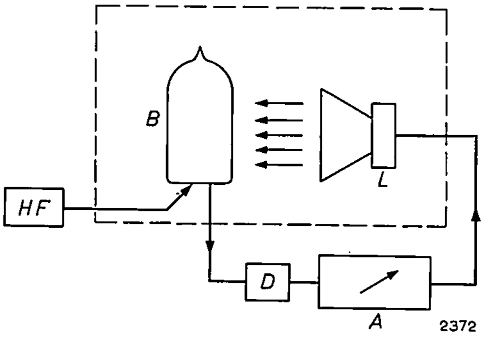


Fig. 4. Principle of a set-up for investigating microphony in a radio-frequency amplifying tube. $B$ tube under test, $L$ loudspeaker, $A$ variable amplifier, $D$ detector, $HF$ signal generator.

The methods described can be used for comparing different types of tubes or individual tubes of the same type, and also for checking the results of modifications made in a tube to reduce microphony. They give no indication, however, as to which components in a tube cause the microphony, and are therefore no help as regards the introduction of the necessary improvements.

In this respect another method is helpful. Instead of making the set-up howl, the loudspeaker is connected to a separate signal generator and amplifier, and the output voltage of the tube is measured with the aid of an amplifier and a vacuum-tube voltmeter (*fig. 5*). This makes it possible to choose and to vary the frequency of the vibrations to which the tube is subjected. The strength of the microphony is then found to vary quite irregularly with the frequency. This is partly due to the fact that the components of the electrode system have different resonance frequencies for mechanical vibrations, so that the tube behaves as if it consisted of a large number of mutually coupled resonators. A further factor, however, as mentioned above and illustrated in fig. 1, is that, even where the loudspeaker power is constant, the acceleration undergone by the tube shows a highly irregular spectrum. As a result, it is not easy to study the microphonic properties of electron tubes with a set-up as in fig. 5: there is always the possibility that the cause of strong microphony occurring at a particular frequency may lie outside the tube itself.

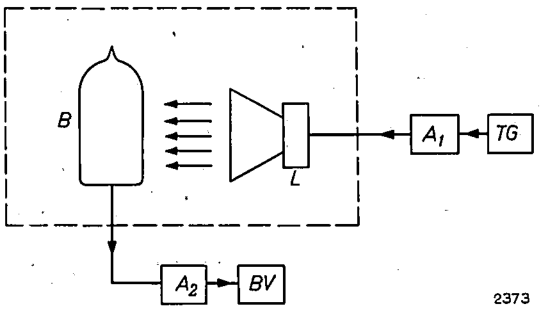


Fig. 5. Principle of a set-up for investigating microphony in electron tubes. The loudspeaker $L$ is fed via the amplifier $A_1$ with a voltage from the signal generator $TG$. The signal voltage due to microphony in tube $B$ is applied to a vacuum-tube voltmeter $BV$ via the amplifier $A_2$.

To arrive at results that are governed solely by the tube we must therefore set the tube in vibration directly and not via a loudspeaker, a cabinet and a chassis.

One method of achieving this is to subject the tube to an impact of given strength and to measure the resultant microphonic signal voltage. An apparatus designed for this purpose is shown in *fig. 6*. Even here, however, the results are not very satisfactory. A blow brings all components of the tube simultaneously into vibration, and only the total result can be measured from the signal voltage thereby generated. Consequently, this method too is really only suitable for comparing tubes one with the other, and not for tracing the causes of microphony.

A thorough study of the microphonic properties of tubes demands that the tubes be subjected to vibrations of constant acceleration and variable frequency. Only then is it possible to draw con-

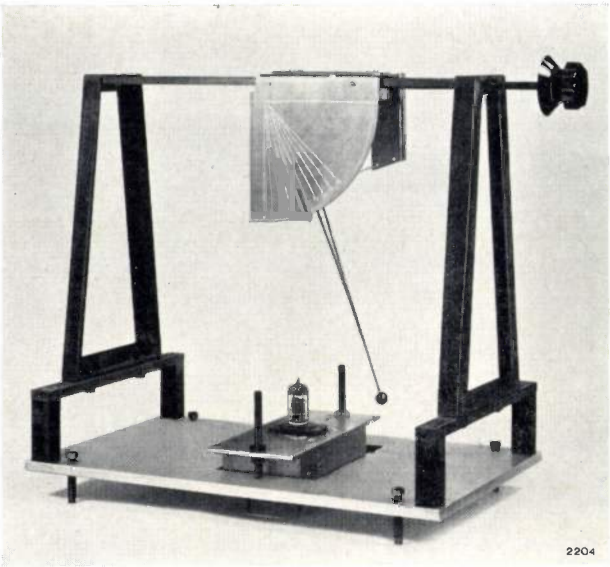


Fig. 6. Apparatus for studying the microphonic properties of electron tubes by subjecting the tube to a known impact.

clusions, or at least inferences, as to the cause of strong microphony at a particular frequency.

With this object in view, the constant acceleration has been achieved by testing the tubes in a specially designed vibrator. This method has for some time now been applied in several Philips laboratories to numerous types of tubes, and will now be discussed in some detail.

### A vibrator for the study of microphony

*Fig. 7* shows an axial cross-section through a vibrator designed for investigating microphonic effects in electron tubes. The construction of the vibrator closely resembles that of an electrodynamic loudspeaker. It consists primarily of a coil which can move in the air gap of a ring-shaped magnet. The coil is wound on an aluminium former, which is supported in sleeve bearings. The resonance frequency of the whole assembly is in the region of 30 kc/s, which is a great deal higher than the highest frequency of the range in which microphony tests are usually made (30 to 20 000 c/s). In this frequency range, then, where the alternating current in the coil is constant an alternating acceleration of almost constant peak value is obtained. (An alternating current of 100 mA was needed for a peak acceleration of 1g.) This can be quite easily checked by mounting a stationary metal plate a short distance from the upper surface of the coil and by measuring the variations occurring in the capacitance between this plate and the coil, this capacity being inversely proportional to the distance. For a constant charge, voltage variations appear across the capacitor thus formed, the magnitude of which is proportional to the deflection of the vibrator. These voltage variations can be amplified and measured. It follows from the theory of harmonic vibrations that, if the peak acceleration is to remain constant, the maximum deflection must be inversely proportional to the square of the frequency, i.e. with increasing frequency it must decrease by a factor of 4 per octave. With the vibrator described here this was indeed found to be the case in the required frequency range.

The way in which the tube under test is fixed to the vibrator calls for particular care. Strictly speaking, it should be perfectly rigid, otherwise the vibrations of the coil former will not be transmitted to the tube completely independent of the frequency. Now, some resilience in the mounting of the tube is unavoidable and consequently the tube resonates at the frequency of the mounting. To prevent

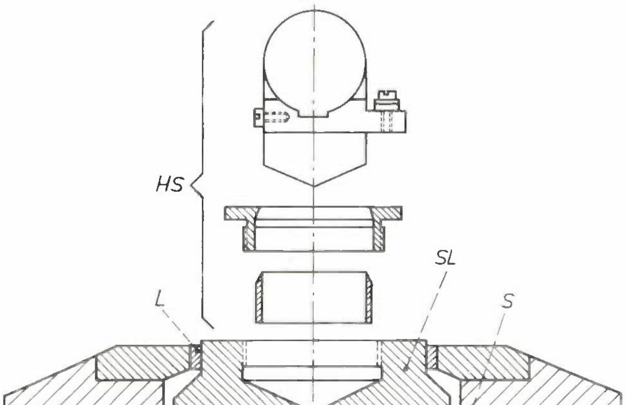
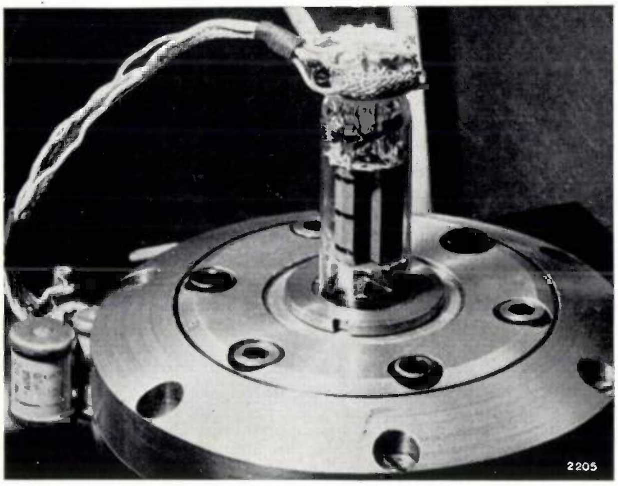


Fig. 7. Cross-section of a vibrator for studying microphonic effects in electron tubes. *S* coil, *SL* coil former, *L* sleeve bearings, *M* magnet, $P_1$ and $P_2$ pole pieces, *TO* vibration pick-up, *HS* adaptors for clamping the tube to the coil former.



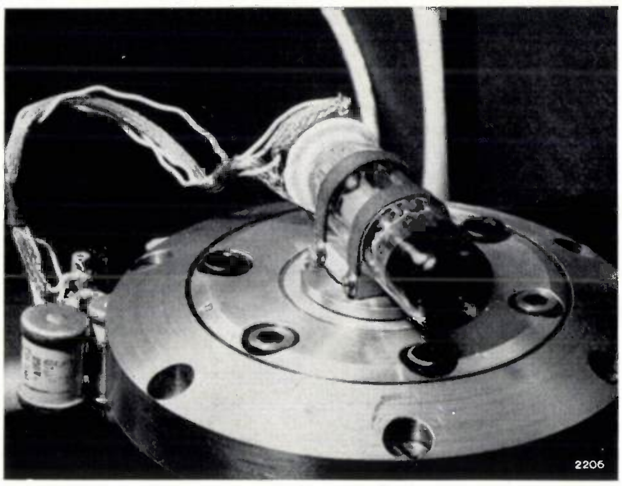


Fig. 8. Upper part of vibrator fitted with a tube to be subjected to vibrations: *a*) along the axis; *b*) perpendicular to the axis.

this affecting the measurements, the vibrator and mountings must be so designed that the resonance frequency is well above the frequency range under investigation. Some components used for this purpose are shown in the upper part of fig. 7.

To enable a constant check to be kept on the vibrations, a vibration pick-up of the piezo-electric type is fitted (with a piezo element of barium titanate); this is denoted by *TO* in fig. 7. This device also indicates whether insufficiently rigid mounting of the tube is causing a spurious resonance.

*Fig. 8* shows two photographs of the upper part of the vibrator, with a tube mounted in two positions, enabling it to be subjected to vibrations parallel or perpendicular to the axis.

## Measurements with the vibrator

With a vibrator as described above a tube can be subjected to a known acceleration which, as opposed to the methods illustrated in figs. 3, 4 and 5, is independent of the incidental resonance frequencies of other parts of the apparatus. If strong microphony appears at a particular frequency, it is now clear that this frequency corresponds to the resonance frequency of one of the tube components. Detection of these frequencies is accomplished by connecting the tube by flexible wires to an amplifying circuit: the signal voltage produced in this circuit as a result of microphony is measured whilst the vibrator frequency is slowly varied. The measurements are facilitated by using a recorder. Examples of spectrograms obtained in this way are shown in figs. 21-24.

In these and similar measurements the vibration frequency should be varied slowly and very evenly, the mechanical vibrations of the various components being very little damped. Because of the weak damping, the various resonances occur only very near to the exact resonance frequency: many peaks in the spectrogram are so sharp that they might easily be missed.

Once it has been found that a strong microphonic effect occurs at a particular frequency, the next thing to do is to trace the component responsible for it. There are various ways of setting about this. One obvious method is to calculate the resonance frequencies of components whose very slight movements can be expected, on theoretical grounds, to have a considerable effect on the electrical characteristics of the tube. One can then ascertain whether one of these frequencies coincides with a peak in the spectrogram. If this is so, it is reasonable to assume that the component in question must be the cause of this peak. Further experiments are then needed to show whether this assumption is correct or not.

In practice, this method turns out to be most unsatisfactory. The main reason is that calculations of the resonance frequencies of components in an electrode system can seldom be more than rough approximations. Exact formulae can be derived only for simple configurations, and the application of such formulae to practical cases calls for approximations and corrections; also, it is generally not accurately known just how the various electrodes are clamped or supported, or whether there is any play between them.

We shall illustrate the above method and its shortcomings by taking a grid as an example. The conventional grid construction is shown in *fig. 9*. Two uprights (or "backbones") $S_1$ and $S_2$ are mounted in holes in the mica discs $M_1$ and $M_2$. The grid wires $D$ are wound helically around the uprights. If we now regard $S_1$ and $S_2$ as freely vibrating rods, their resonance

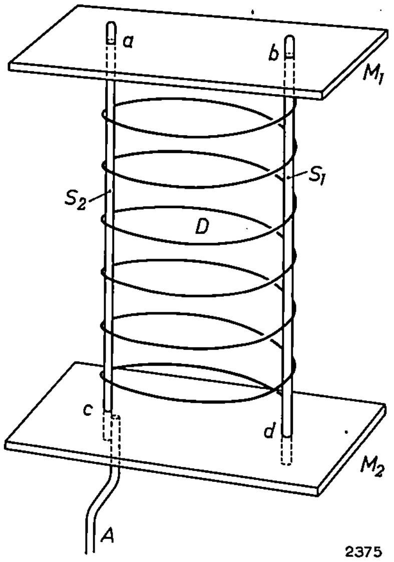


Fig. 9. Simplified construction of a grid. $S_1$ and $S_2$ uprights ("backbones"), $D$ grid wires, $M_1$ and $M_2$ mica strips on which the assembly is mounted at the points $a$, $b$, $c$ and $d$, $A$ lead-in wire.

frequency $f_r$ for mechanical vibrations can be calculated from the formula:

$$f_r = \frac{d}{4l^2}\sqrt{\frac{E}{\varrho}}\,K.$$

Here $d$ is the diameter and $l$ the length of the rod; $E$ is the modulus of elasticity and $\varrho$ the density of the material, and $K$ is a constant which depends on the way in which the upright is held. The magnitude of this constant is:

$K = 0.56$ if the vibrating rod is clamped at one end and free at the other,
$K = 3.56$ if the rod is clamped at both ends,
$K = 2.45$ if the rod is clamped at one end and held such that it can pivot at the other,
$K = 1.56$ if the rod is held such that it can pivot at both ends.

Owing to the unavoidable spread in the dimensions of the grid uprights and of the holes in the mica supports, it is never certain whether the uprights at positions $a$, $b$, $c$ and $d$ should be regarded as clamped, pivoted or free. Extremely small differences in dimensions, which may have no perceptible effect on the electrical properties of the tube, may have a marked effect, in view of the differences in $K$, on the resonance frequency of the grid uprights. A further inaccuracy in the calculation is due to the presence of the grid wires $D$. Their effect can be allowed for as an increase in the mass of the grid uprights, but this is obviously a rough approximation. Finally, the fact that a connection wire $A$ is attached to one of the uprights can also only be taken into account by very rough approximation.

For the grid wires two empirical formulae have been worked out [3]) which apply to wires bent in the form of an arc (*fig. 10a*) and in the form of a rectangle (fig. 10*b*). The formula for a grid wire as in fig. 10*a* is

$$f_r = \frac{0.217\,d}{2.78\,a^2 + 0.558\,R^2}\sqrt{\frac{E}{\varrho}}\,,$$

and for a wire as in fig. 10*b*:

$$f_r = \frac{0.217\,d}{2.9\,a^2 + 0.325\,R^2}\sqrt{\frac{E}{\varrho}}\,.$$

These formulae hold good only when $a$ and $R$ are roughly the same ($R/a < 2$). Consequently, and also because the actual shape of the grid wires never exactly satisfies fig. 10*a* or *b*, the result here too can never be more than a very rough approximation.

Another source of uncertainty is the fact that many elements capable of mechanical vibration in an electrode system are coupled to one another, resulting in resonance frequencies that do not correspond to those of the elements individually.

## Stroboscopic examination

The only way to point with certainty at one of the components of the tube as the source of strong microphony at a particular frequency is to observe directly that this component in fact resonates at that frequency, i.e. vibrates with a large amplitude. Since this "large" amplitude is not usually perceptible to the naked eye, it must be observed under a microscope. In order to make it possible to observe grids etc., it may further be necessary to make a number of tubes with special openings in the anode or in the screening.

As a rule the frequencies at which the investigations are carried out are too high for the eye to be able to follow the movement directly. The movements can be made visible, however, by illuminating the tube with a stroboscope. The vibrator and the stroboscope are then fed by two signal generators delivering alternating voltages whose frequencies differ by a few c/s. An arrangement designed for the purpose is shown schematically in

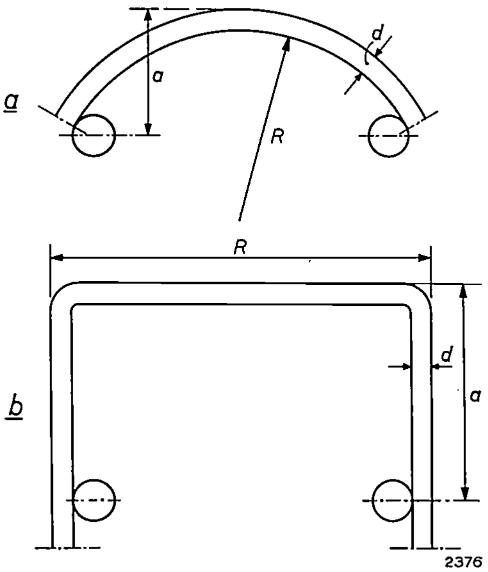


Fig. 10. *a*) Grid wire bent in the form of an arc, *b*) rectangularly bent grid wire.

[3]) See P. M. Handley and P. Welch, Valve noise produced by electrode movement, Proc. Inst. Radio Engrs. **42**, 565-573, 1954.

*fig. 11*. When the tube is set in vibration, the parts in question can be seen under the microscope to vibrate at a frequency equal to the difference between the frequency of vibration and the illumination frequency. The fact that one of the parts is resonating at the applied frequency is manifested not only by the increase in amplitude but also by the phase relationship between the impressed force and the deflection.

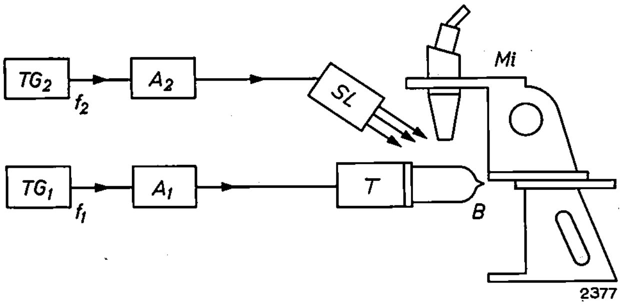


Fig. 11. Block diagram of an arrangement for stroboscopic investigation of microphony in electron tubes. $TG_1$ and $TG_2$ signal generators, $A_1$ and $A_2$ amplifiers, $T$ vibrator, $B$ tube under test, $SL$ stroboscope lamp, $Mi$ microscope.

When a vibrating system is driven by a force whose frequency is much lower than the resonance frequency of the system, the deflection is in phase with the force. If the frequency of the force is much higher than the resonance frequency, the deflection and the force are in antiphase. The transition from one of these states to the other takes place in a frequency range around the resonance frequency. The less the vibrations of the system are damped, the narrower is this frequency range. At the resonance frequency itself the phase shift between force and deflection is 90°.

To produce automatically a difference between the vibrator and the stroboscope of a few c/s at all frequencies, one might couple the tuning mechanisms of the two signal generators. It is very difficult, however, to make the coupling in such a way that the frequency difference remains sufficiently small over the whole frequency range of interest. If the difference is too great the eye can no longer follow the individual vibrations. When one of the components is then excited into resonance, the resonance will be scarcely perceptible.

An improvement in this respect is obtained if the stroboscope frequency is made exactly equal to the vibration frequency by connecting the vibrator and stroboscope to a common signal generator. Of course, the vibrating parts then appear to be stationary, the movement being frozen at all frequencies. When the frequency is slowly varied, however, and passes the resonance frequency of a component within the field of view of the microscope, the phase shift of 180°, mentioned above, can be seen to take place in the vibrations undergone by this component. As the frequency moves through a very small range, this part is then seen to make a single half-vibration and then stands still again. The concentrated attention required to observe this phenomenon is a serious drawback, however, to the application of this method in large-scale investigations. High demands are also made on the equipment; the frequency must be varied extremely slowly and continuously.

The stroboscopic method was not really a success until an apparatus had been designed with which it was possible, in the whole frequency range under investigation, to maintain a constant difference of 1 or 2 c/s between the vibration frequency and the frequency of the stroboscopic illumination. *Fig. 12* shows a block diagram of the equipment used for this purpose. As in fig. 11, the vibrator $T$ is driven by a signal generator $TG$ via an amplifier $A_1$. The output of the signal generator is again used to operate the stroboscope lamp, but only after first being applied to a frequency shifter $FS$ which delivers an output voltage whose frequency is a constant amount higher or lower than the frequency of the applied voltage. The output voltage of the frequency shifter is used to control the pulse generator $PG$, which delivers short voltage pulses to the stroboscope lamp $SL$.

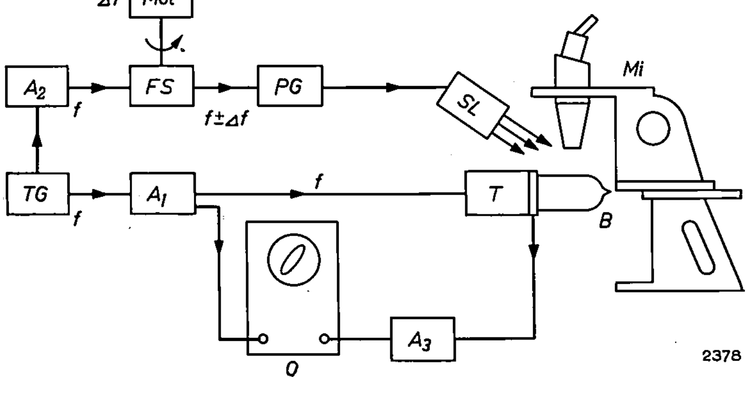


Fig. 12. Block diagram of equipment for stroboscopic investigation of microphony in electron tubes. By means of the frequency shifter $FS$ a constant difference of 1 to 2 c/s is maintained between the vibration frequency and the frequency of the stroboscopic illumination. $TG$ signal generator, $A_1$, $A_2$ and $A_3$ amplifiers, $Mot$ motor, $PG$ pulse generator, $SL$ stroboscope lamp, $Mi$ microscope, $T$ vibrator, $B$ tube under test, $O$ oscilloscope.

The frequency shifter consists of a rotor and a stator. The stator is provided with a three-phase winding. This is supplied with a three-phase voltage derived from a special amplifier $A_2$, fed with the signal-generator voltage (frequency $f$). The rotating field thus produced induces an alternating voltage in the single-phase winding of the rotor. When the rotor is stationary, the frequency of the latter voltage is equal to $f$, but when the rotor revolves at a speed of $\Delta f$ revolutions per second, the frequency of the e.m.f. induced in the rotor winding is an amount $\Delta f$ higher or lower than $f$, depending on the sense of rotation. Provided the rotor turns at a constant speed, a constant frequency difference is then maintained between the applied voltage and the output voltage.

The tube under test $B$ feeds an amplifier $A_3$, the output voltage of which is applied to one pair of plates of an oscilloscope $O$. To the other pair of plates a voltage is applied which is proportional to the current driving the vibrator. By observing at the same time the picture under the microscope and that on the oscilloscope screen it is now possible to ascertain with considerable certainty whether the vibrations of a particular component are responsible for the occurrence of strong microphony at a particular frequency. At that frequency the amplitude of the vibrations undergone by the component in question shows a maximum, and at the same time the alternating voltage produced by the microphony is seen on the oscillosope to reach a maximum value. The observer also sees the above-mentioned phase shift as the resonance frequency is passed. The phase shift also of course occurs between the current supplied to the vibrator and the voltage generated by microphony, and is thus displayed on the oscilloscope as a lissajous figure.

Owing to the extremely weak damping of the vibrations the effects referred to occur in such a very narrow frequency range that it is almost impossible for two or more components to resonate simultaneously, even when their resonance frequencies lie very close together.

*Fig. 13* shows a photograph of a set-up as here described. With his left hand the observer varies the frequency of the signal generator, whilst with his right hand he directs the microscope and the stroboscope lamp on to the component to be examined. Over the edge of the ocular he can see the screen of the oscilloscope (right in figure).

The equipment described can also be used in another way. Instead of the amplified voltage output of a signal generator we can apply to the vibrator the amplified alternating voltage produced by microphony in the tube under test. In many cases this will give rise to oscillation at a frequency corresponding to the resonance frequency of one of the components. This component will then vibrate with a large amplitude, and it will not generally be difficult to ascertain by means of the microscope and stroboscopic illumination exactly which component this is. Of course, this method can only trace the component that makes the major contribution to the microphony, since oscillation occurs at the resonance frequency of that component. It is also possible, however, to find the cause of strong microphonic effects at other frequencies if we include in the feedback path a filter that passes signals only in a limited frequency band. In that case, oscillation can occur only at a frequency within that band, and the component responsible for it can be traced with the microscope.

In the method using an oscillating circuit it is even more important than in the other methods described that the resonance frequency of the vibrator-plus-tube assembly should lie above the range of frequencies under investigation. If that is not the case, the circuit might start to oscillate at this resonance frequency, and the search for the "guilty" component would then be fruitless.

### Examples of microphony

It is unfortunately not possible in a photograph to give a good impression of the picture observed under the microscope when a component vibrates at its resonance frequency. Nevertheless, to give some idea of what is seen some photographs are shown that were obtained by double exposure at the extreme deflections of the vibrating component. With an arrangement as in figs. 12 and 13 it is a simple matter to freeze the observed picture of the vibrating component in any desired phase. All that is necessary is to stop the motor that drives the rotor in the frequency shifter. Obviously, the picture is then stationary too, and the required phase of the vibration can then be chosen by turning the rotor by hand.

Figs. 14 to 20 show various components of electron tubes and the picture seen under the microscope when the tube is made to vibrate at the resonance frequency of the respective component. The arrows indicate the direction of the vibrations.

*Fig. 14* shows a getter which, being relatively large and supported on one side only, has a low resonance frequency, namely 300 c/s. It is evident that a component as large as this, though not part of the actual electrode system, must have a noticeable effect on the operation of the tube if set in vibration.

*Fig. 15* shows the two filament leads of the tube, which have different resonance frequencies, viz. 570 and 600 c/s. The pictures observed under the

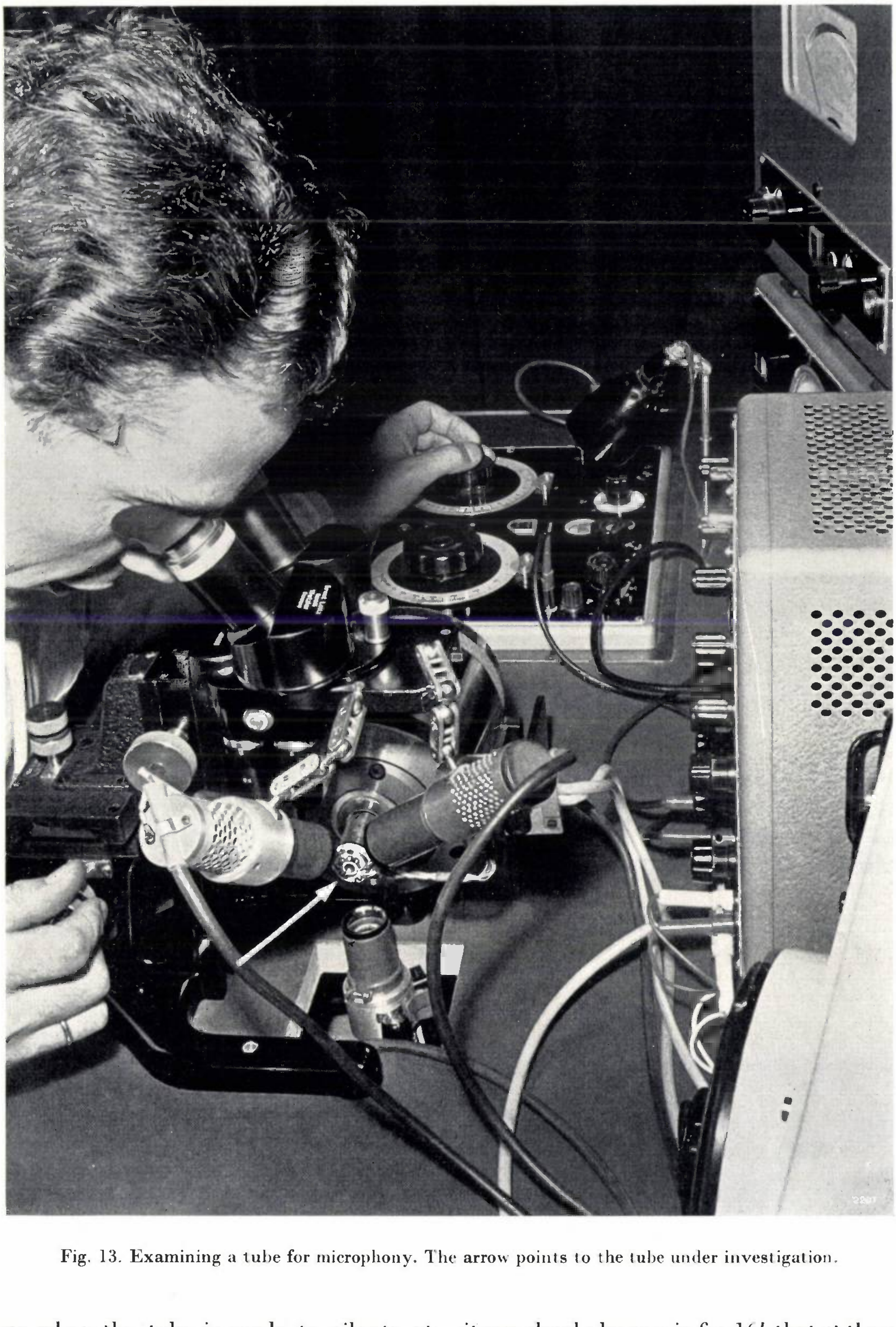


Fig. 13. Examining a tube for microphony. The arrow points to the tube under investigation.

microscope when the tube is made to vibrate at each of these frequencies appear in fig. 15*b* and *c*. It can be seen that, in each case, one of the two wires is virtually at rest whilst the other vibrates.

*Fig. 16* shows the suppressor grid of a pentode ɱade accessible to observation by an opening in the anode. Although the resonance frequencies of the turns of wire differ only slightly from one another, it can clearly be seen in fig. 16*b* that at the resonance frequency of one of them (approx. 2100 c/s) only that turn enters into vibration. This illustrates the fact that the mechanical vibrations are very little damped.

Grid-wire vibrations can cause impermissible microphony if they occur in the screen grid of a pentode in the RF or IF sections of a receiver. If the

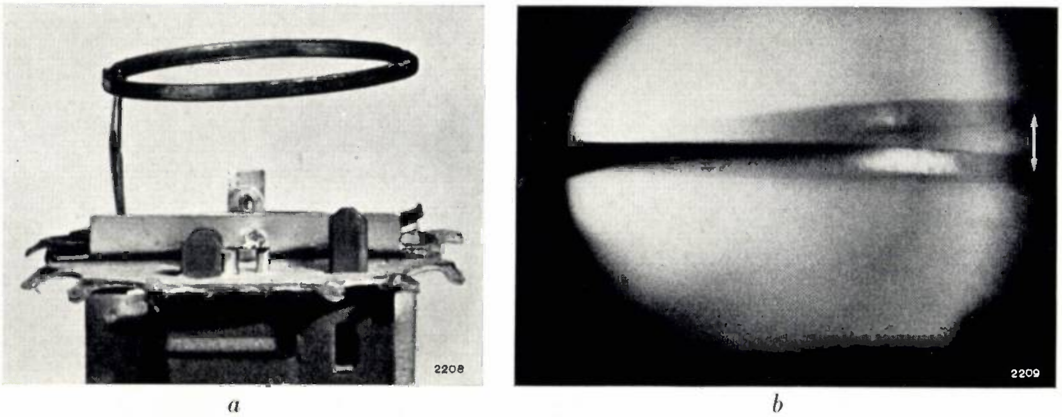


Fig. 14. *a*) Getter of an electron tube.
*b*) Picture seen in the microscope when the tube is made to vibrate at the resonance frequency of the getter (300 c/s).

mutual conductance in the pentode has been reduced to a low value by the automatic gain control, the electron current passes through only a few turns of the screen grid. A slight movement of one of these turns then has a considerable effect on the anode current and on the mutual conductance. The effect is less pronounced if the tube operates with a higher mutual conductance. More nuisance is then experienced from vibrations in the grid uprights, since this causes lateral movement of the whole grid. *Fig. 17* shows a grid undergoing vibrations of this kind in the triode portion of a triode-hexode. Here, too, it was necessary to cut an opening into the anode. The resonance frequency of this grid was 1900 c/s.

In *fig. 18* the end of a cathode can be seen that exhibited some play in the upper mica support of the electrode system, and therefore vibrated at a very low frequency (600 c/s). Cathode vibrations are usually damped more than those of other components, owing to the influence of the filament with

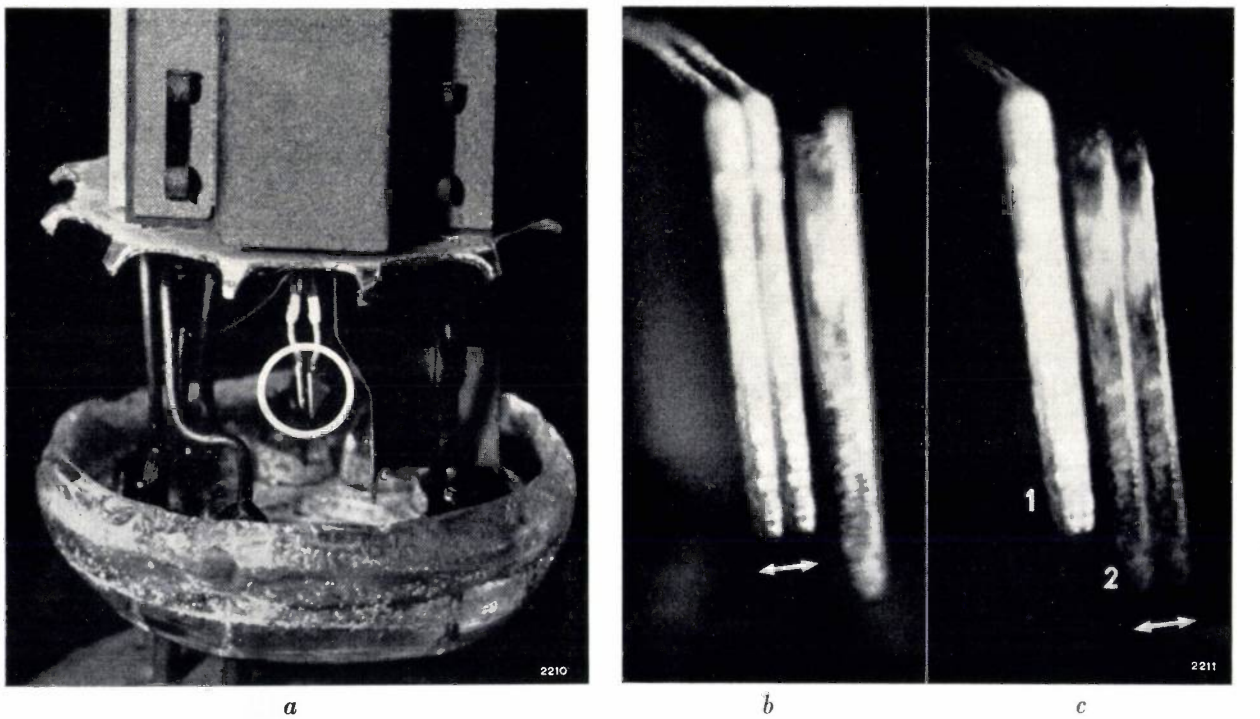


Fig. 15. *a*) Lower part of the electrode system of a vacuum tube. The circle marks the ends of the filament leads; *b*) and *c*) show the pictures of these leads seen under the microscope when the tube is successively made to vibrate at the resonance frequency of each lead (570 and 600 c/s).

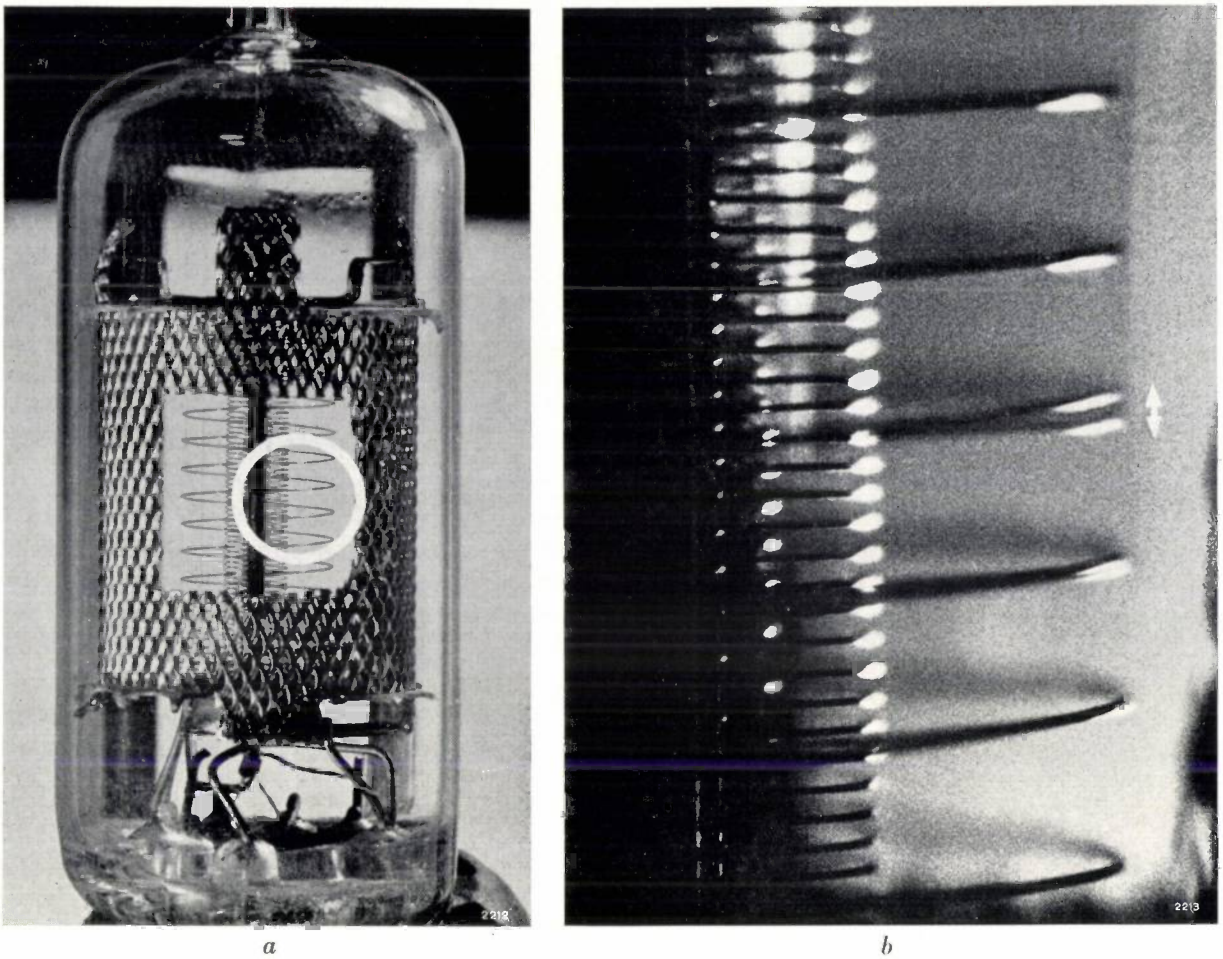


Fig. 16. *a*) Suppressor grid of a pentode, visible through an opening cut into the anode. The circle marks the part seen under the microscope, (*b*), when the pentode is made to vibrate at the resonance frequency of one of the grid wires (2100 c/s).

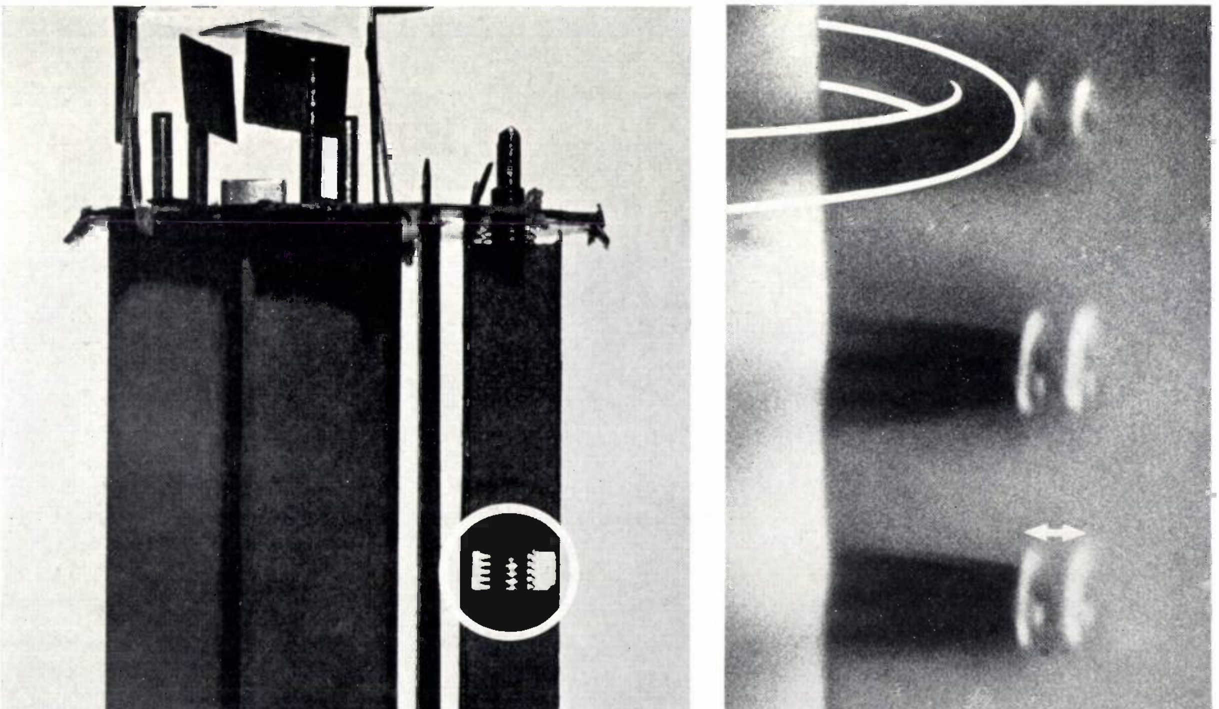


Fig. 17. *a*) Electrode system of a triode-hexode. The circle marks the grid of the triode portion, visible through a hole cut into the anode. *b*) Picture of the grid vibrating at its resonance frequency of 1900 c/s.

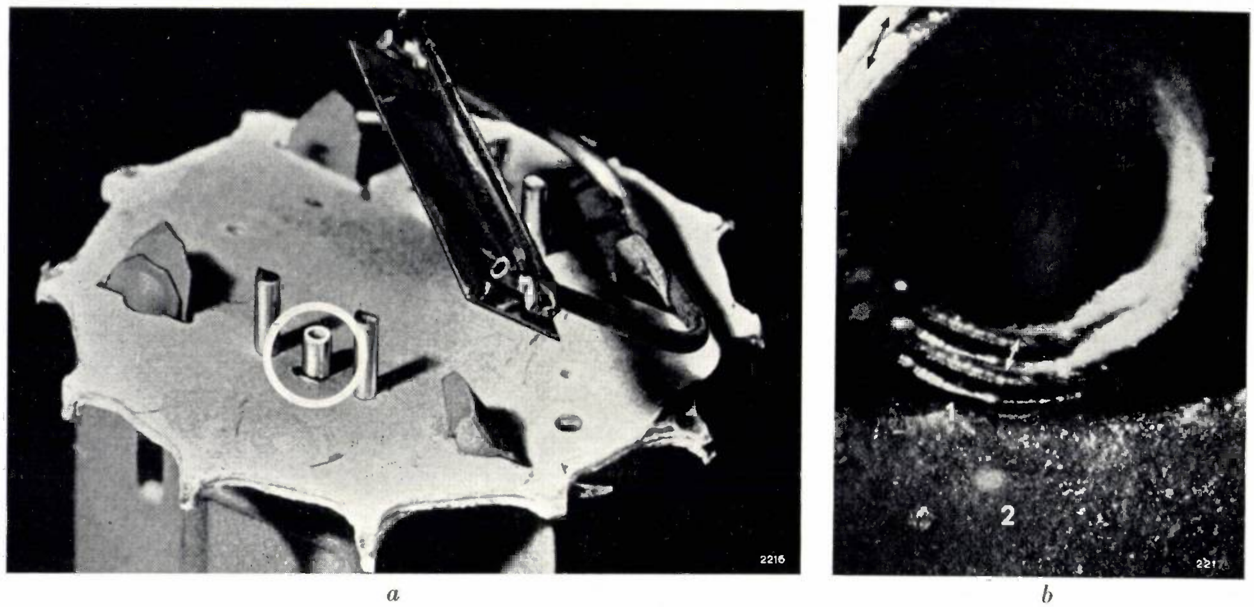


Fig. 18. *a*) Top view of the electrode system of a tube. The circle marks the end of the cathode, which showed some play in its hole in the mica disc.
*b*) Picture seen under the microscope when the tube was made to vibrate at the resonance frequency of the cathode (600 c/s).

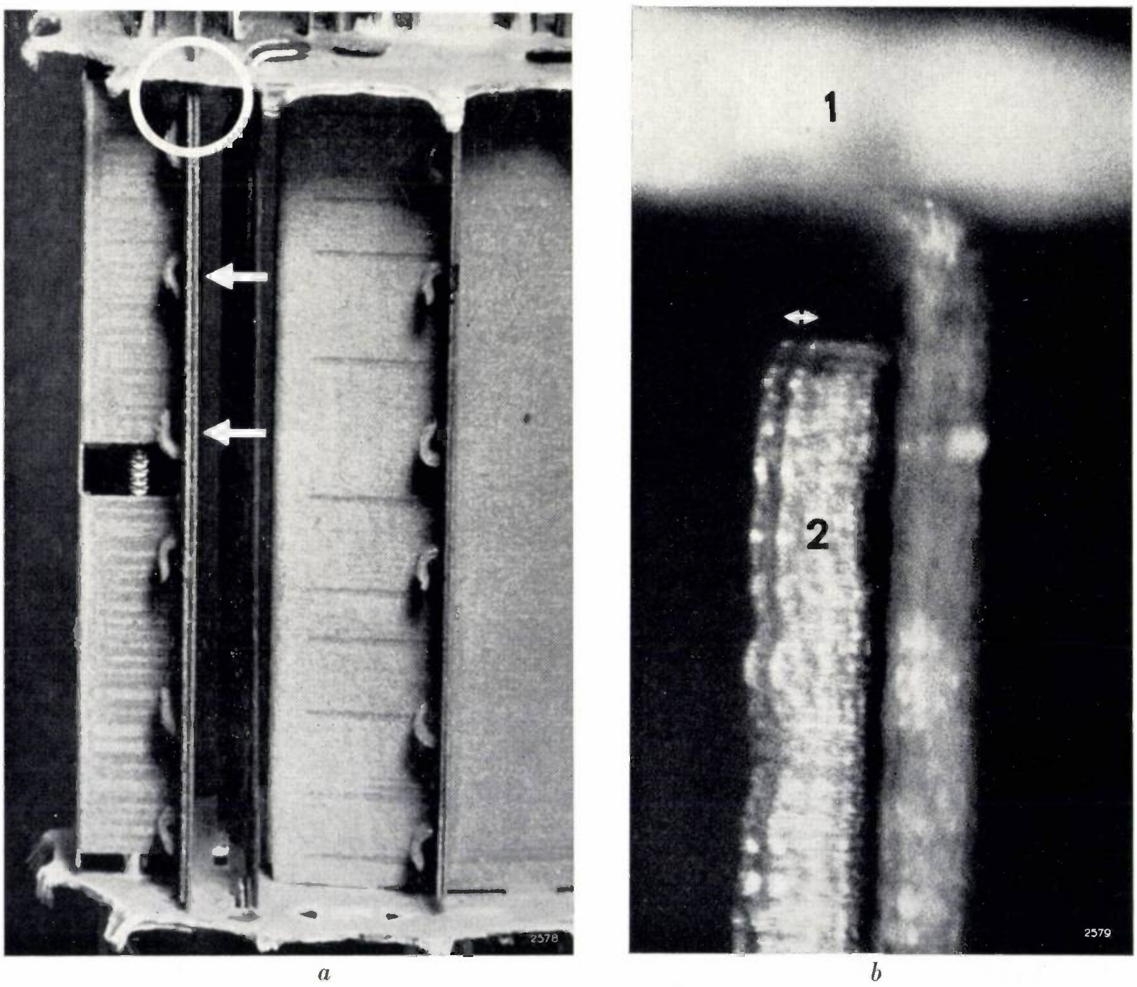


Fig. 19. *a*) Anode of an electron tube. The two parts were not fixed firmly enough at the positions indicated by the arrows, thus allowing free movement between them.
*b*) Picture seen under the microscope of the circled area when the tube was set in vibration at a frequency of 1300 c/s.

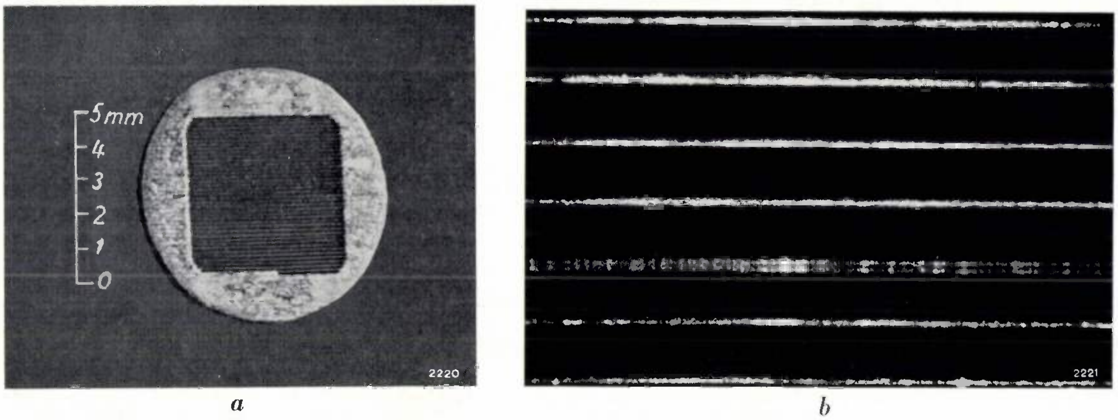


Fig. 20. *a*) Frame grid of a tube for very high frequencies.
*b*) Picture under the microscope when the grid was made to vibrate at the resonance frequency of one of the wires (37 000 c/s).

its insulation and to the emissive coating of the cathode.

Vibrations of one of the structural elements of an anode at a frequency of 1300 c/s can be seen in *fig. 19*. The reason for this vibration was that the parts of the anode at the position denoted by the arrows had not been properly fastened.

The fact that this method of investigation can also be used at higher frequencies than those mentioned above is illustrated in *fig. 20*, which shows

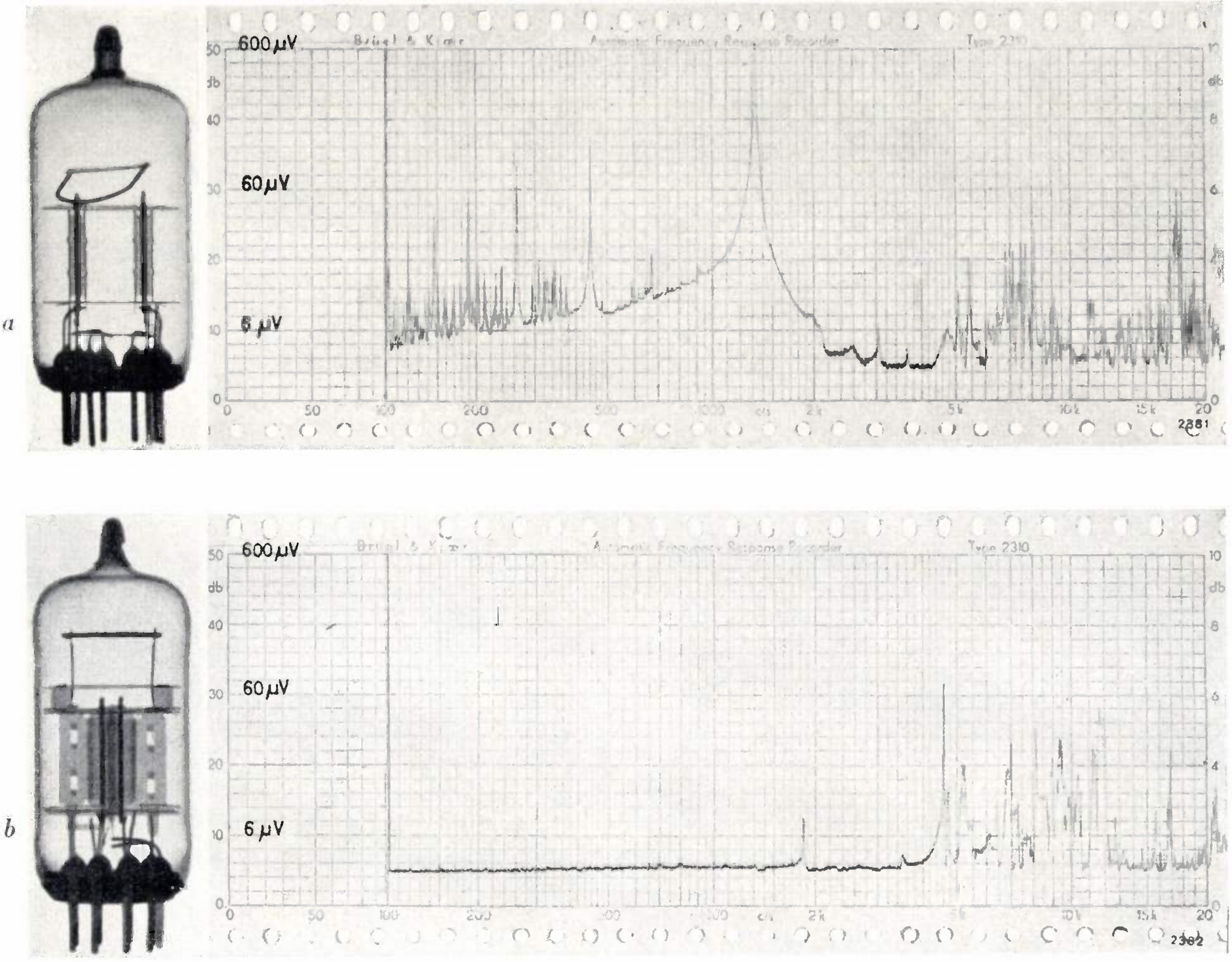


Fig. 21. Effect of getter construction on the microphony of an electron tube. The two constructions compared are shown on the left. The construction under *b* results in a considerable reduction of microphony.

the vibrations of one of the wires of a frame grid [4]). The frequency was 37 000 c/s. It need hardly be said that this imposes very high demands on the vibrator and on the rest of the circuit; the stroboscope lamp, for example, had to provide extremely short light pulses to produce a sufficiently sharp picture.

### The reduction of microphony

Once it has been established that a particular component makes a substantial contribution to the microphony of an electron tube, it is of course important to ascertain whether a structural modification designed to reduce the microphony really has the desired effect. This can best be checked by recording a spectrogram of the signal voltage due to microphony as a function of the vibration frequency. We shall illustrate this with some examples of improvements introduced. Figs. 21 to 24 show a number of spectrograms recorded whilst the electrode systems of the tubes under investigation were subjected to lateral vibrations at a constant peak acceleration of $\frac{1}{3}$ m/sec$^2$. The figures indicate the rms value of the alternating grid voltage producing the same interfering signal as caused by the microphony.

*Fig. 21* illustrates the result of modifying the support of a getter. The upper recording was made on a tube where the getter was fixed to a bracket which was welded to the anode at one point. This getter was found to be responsible for the strong microphony that occurred at a frequency of 1300 c/s. When the getter was secured at two places in the mica support, and thus no longer connected to one of the electrodes, a considerable improvement was obtained, as appears from the lower spectrogram. At frequencies below 1850 c/s the tube is now free from microphony.

*Fig. 22* shows the improvement achieved when a single mounting lug on the anode was replaced by a

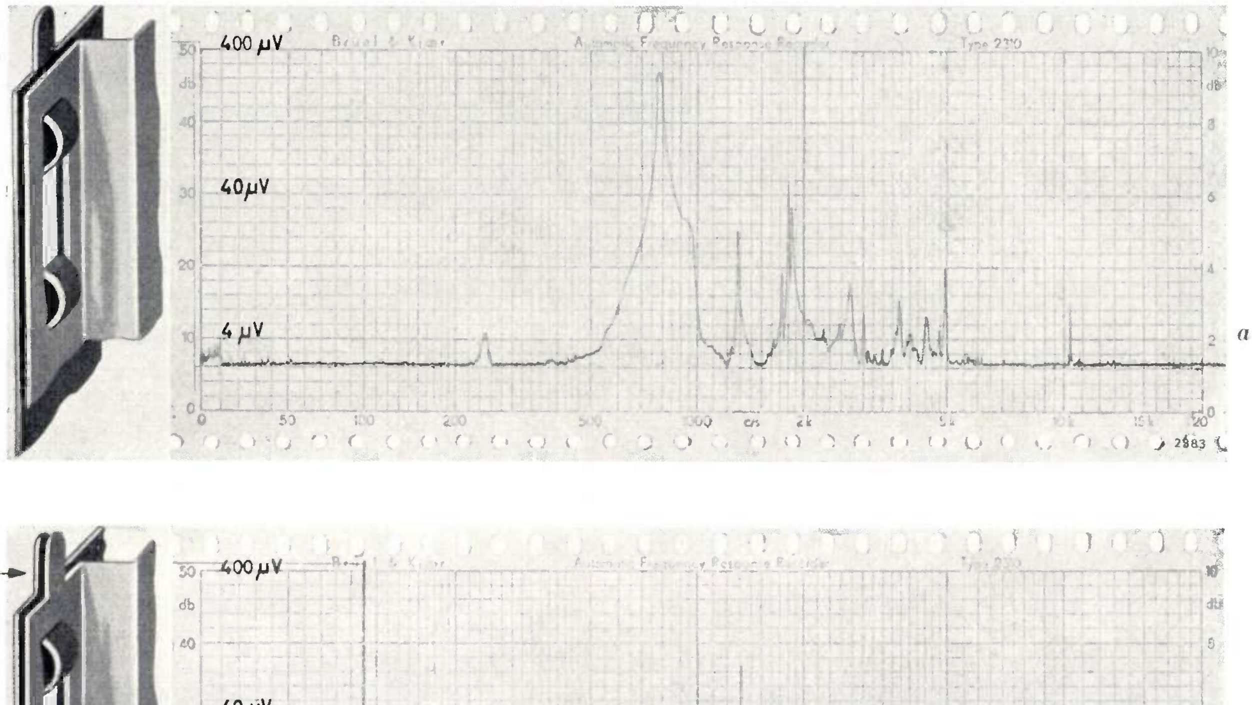


Fig. 22. Effect of anode fastening on the microphony of a tube. The method of fastening with a single lug (*a*) is inferior to that with a double lug (*b*).

[4]) The construction of a frame grid is described by G. Diemer, K. Rodenhuis and J. G. van Wijngaarden, The EC 57, a disc-seal microwave triode with L cathode, Philips tech. Rev. **18**, 317-324, 1956/57.

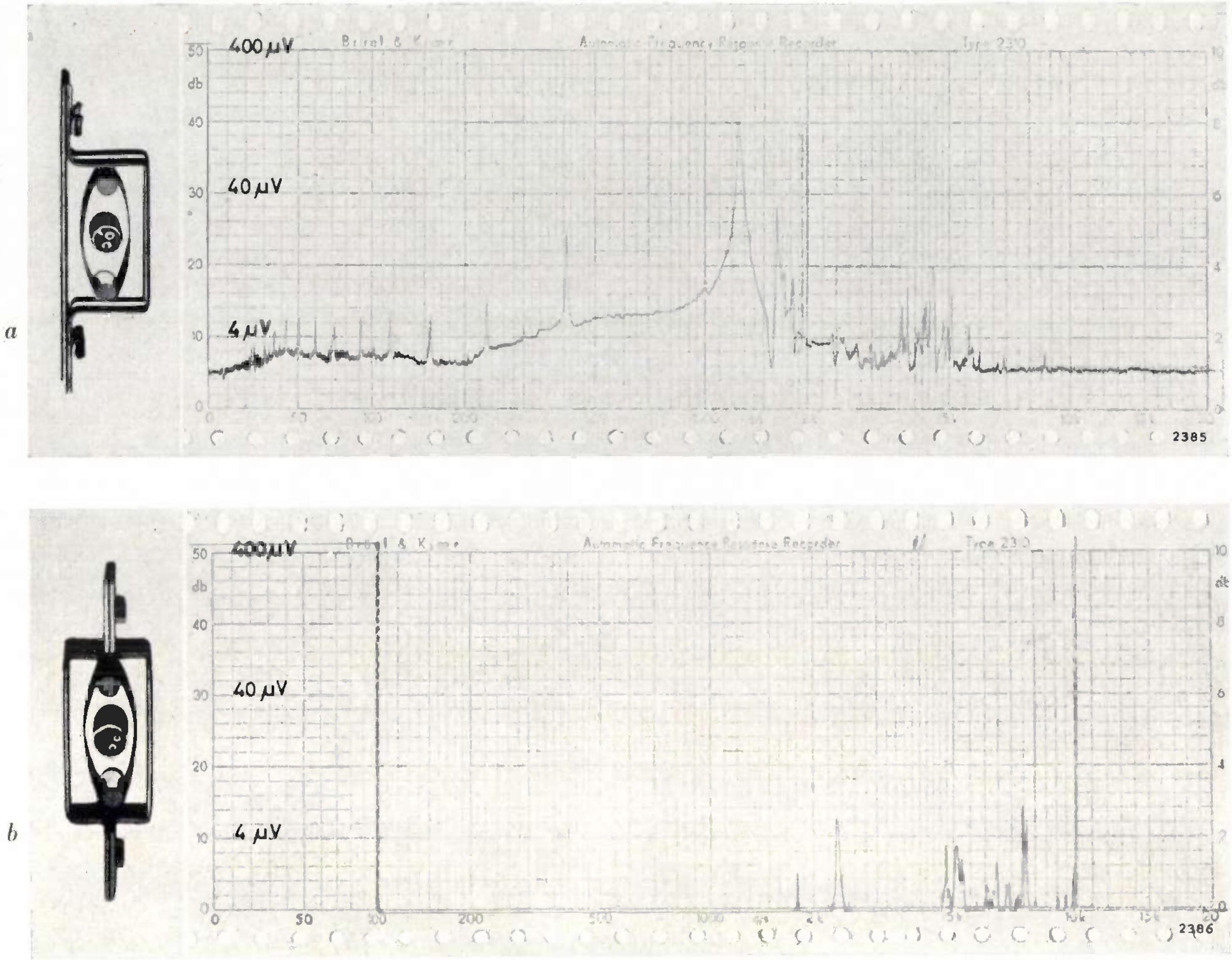


Fig. 23. The effect of anode design on microphony. The construction shown in (*a*) is greatly inferior to that in (*b*).

double lug, thereby eliminating the original play in the mica. In the construction shown in the upper figure, both parts of the anode were capable of relative vibration; as can be seen in the lower figure, the vibration is much less pronounced with the new construction. The high peak at about 780 c/s in the upper spectrogram is no longer to be seen in the lower recording.

The improvement obtained in another case, by modifying the construction of the anode, appears from *fig. 23*. An anode consisting of two rectangular sections, as in *b*, is far more rigid than an anode one of whose parts is flat, as in *a*. The marked improvement from the point of view of microphony is clearly to be seen in the spectrograms.

If the various components that give rise to microphony can be systematically traced and improved, the microphony can be almost entirely eliminated, as illustrated in *fig. 24*. The upper spectrogram relates to a tube which exhibited very troublesome microphony at various frequencies. The lower spectrogram, recorded after the necessary structural improvements had been made, shows that the tube is now virtually free from microphony.

It is not always possible in the series production of tubes to introduce all the improvements that would be desirable with an eye to microphony. Other considerations of quite a different nature may often be involved, such as the effect of these improvements on the electrical properties of the tube, on the cost price or on the production tools. However, if the causes of the microphony are sufficiently known — and they can nearly always be traced by the methods described above — a compromise can generally be found that satisfies these other requirements as well.

### Noise method of investigating microphony

For tracing the causes of microphony the foregoing method yields good results. In certain cases, however, a simpler and less time-consuming method may be sufficient. This may be the case when the purpose is not to investigate the causes of

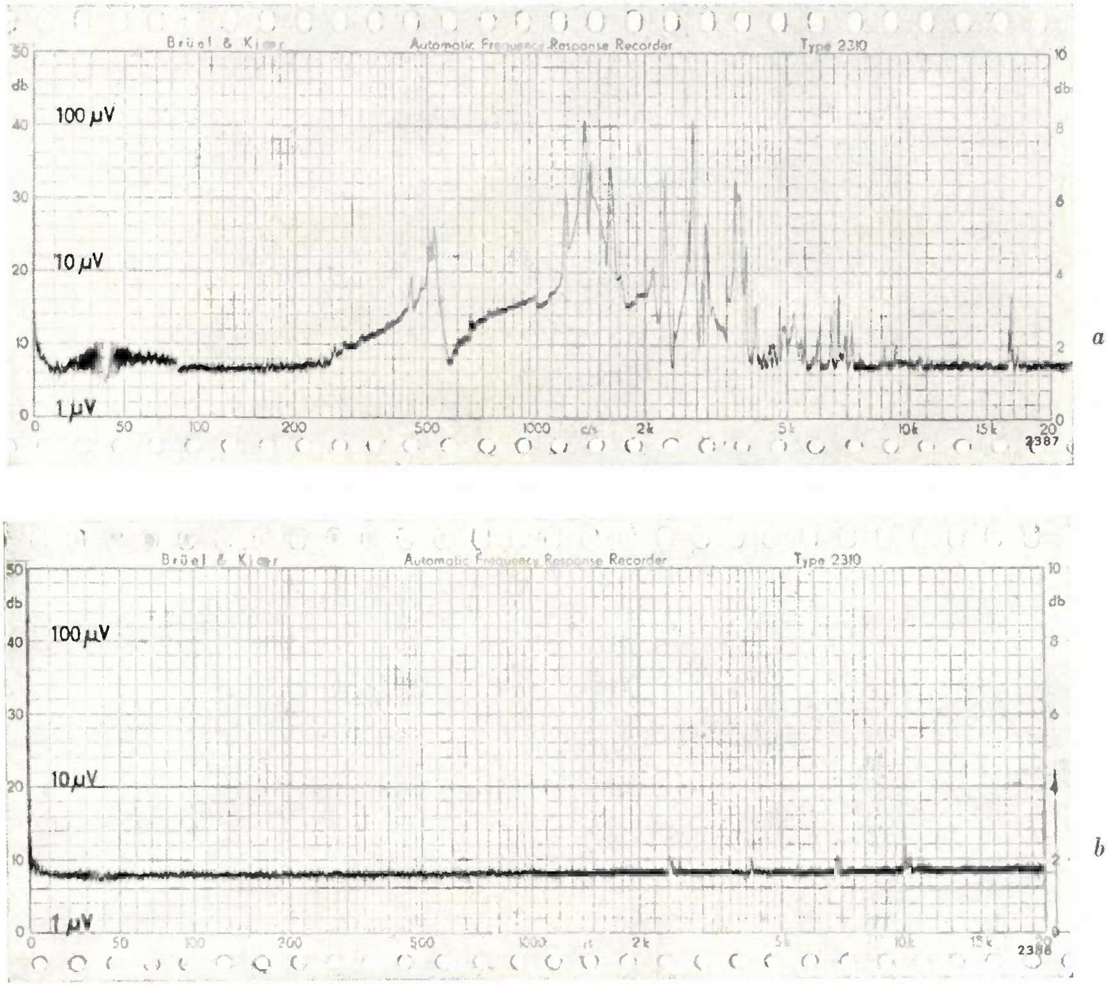


Fig. 24. Spectrograms of microphony in a tube, *a*) before and *b*) after certain structural modifications to reduce microphony.

microphony but simply to compare one tube with another in order, for example, to obtain statistical data on the effect of certain structural modifications. In such cases it is often enough to record a spectrogram. Using the method described earlier, viz. subjecting the tubes to sinusoidal vibrations with a variable frequency, several minutes will always be required to obtain a serviceable spectrogram. The frequency must not be varied too quickly because, as explained above, the mechanical vibrations of tube components are very little damped, and it is therefore likely that some peaks in the spectrogram will be missed if the test is done too quickly.

To conclude this account we shall briefly describe a method designed to produce a quicker result. The vibrator — and the tube under test — is excited not by a sinusoidal alternating current of variable frequency but by a current containing components with all frequencies at the same time, i.e. a current delivered by a noise source. In that case all components that have resonance frequencies in the frequency range under investigation will be excited into resonance simultaneously. If the tube is incorporated in an amplifier circuit, microphony gives rise to a signal composed of numerous alternating-voltage components. Measurement of the rms value of this signal gives in itself an idea of the extent to which the tube in question is "microphonic", but a better insight is obtained if the signal components are measured with a selective voltmeter which gives a reading in only a narrow frequency band. By shifting this small band over the whole investigated frequency range we can again obtain a spectrogram. This can be done in such a way as to display the spectrogram directly on an oscilloscope screen.

*Fig. 25* shows a spectrogram produced in this way. Reproducible graphs can be obtained with the selective voltmeter

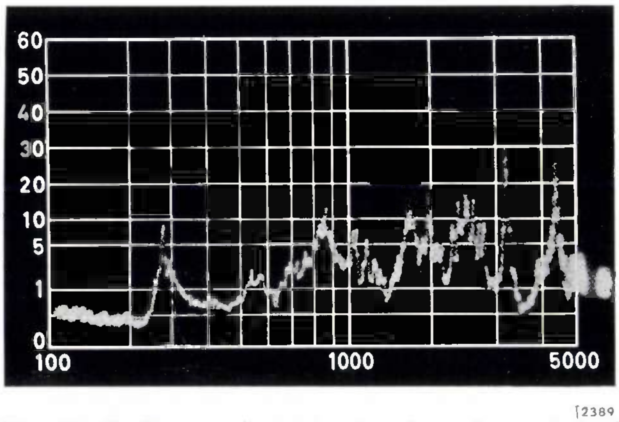


Fig. 25. Oscillogram obtained using the noise method of investigating microphony.

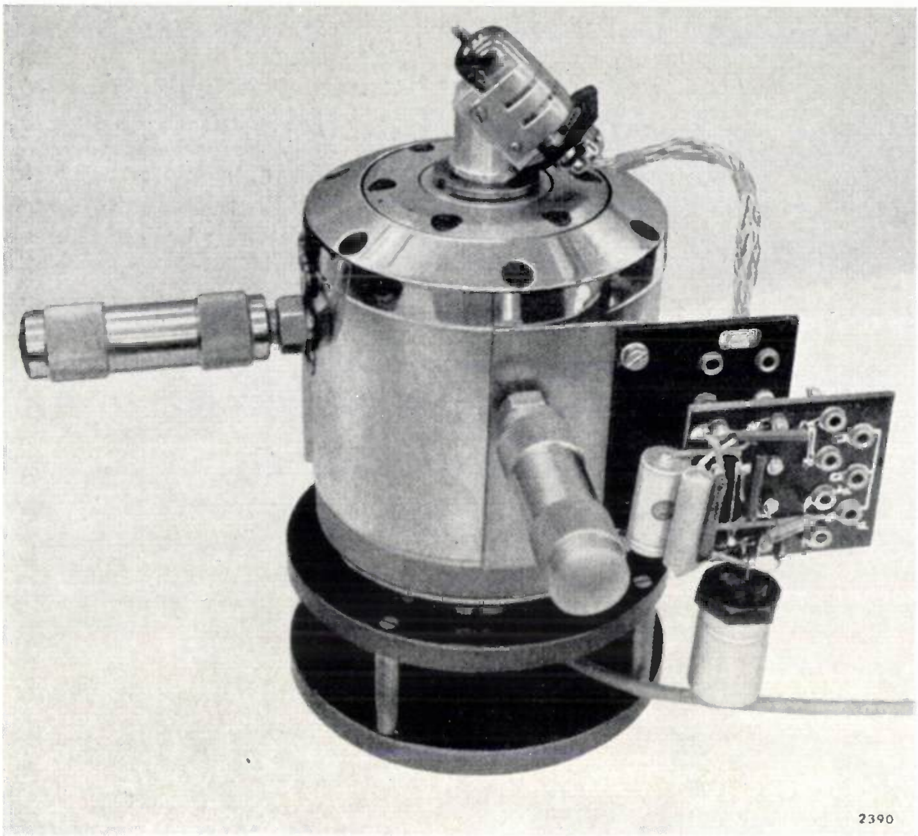


Fig. 26. Vibrator used for microphony investigations by the noise method. The tube under investigation is here mounted obliquely on the vibrator.

sweeping the whole frequency range in about 15 sec. Using an oscilloscope tube with a long-afterglow screen, the whole spectrum can be seen as a single display.

A vibrator used for investigating microphony by the noise method is shown in *fig. 26*. The tube is mounted obliquely in order to obtain a general picture of its microphonic properties, the tube thereby vibrating simultaneously in the lengthwise and lateral directions.

A drawback of the noise method is that the height of the peaks in the spectrum depends on their width. This is because the vertical deflection of the oscilloscope is proportional to the value of the microphony signal, averaged over the whole of the narrow frequency band passed by the selective amplifier. Consequently, a peak narrower than the bandwidth passed by the amplifier will appear to be shorter than a broader but otherwise equally high peak. The picture on the oscilloscope is therefore not an exact representation of the spectrum, and this must be taken into account when evaluating it.

Partly for this reason the noise method has not proved a great success. If one is prepared to accept this error, the results can be obtained just as quickly by the method using sinusoidal vibrations. When the whole frequency range is rapidly scanned, say in 15 seconds, here too the high peaks in the spectrum will not be reproduced in their true relationship. Even so, the resultant spectrogram is still better than that obtained by the noise method. Because of this, and the fact that the equipment for the noise method of investigating microphony is much more complicated, the system using sinusoidal vibrations has been given preference in our laboratories.

**Summary.** Various methods are described for investigating microphonic effects in electron tubes. Some direct methods requiring no special circuit arrangement can serve for comparing one tube with another, but they give no information on the cause of the microphony. For the latter purpose a vibrator has been designed by means of which a tube can be subjected to a vibration of constant peak acceleration and variable frequency. With the aid of a microscope and a stroboscope the components responsible for the microphony can then be traced by directly observing their vibration. Some results achieved are illustrated by spectrograms. Finally, a method using a noise generator is described, where the spectrogram is displayed on the screen of an oscilloscope.

# THE HEATING OF FOOD IN A MICROWAVE COOKER

by W. SCHMIDT *).

621.373.029.6:621.365.55

*Growing interest is being taken in a novel method of cooking food, i.e. by dielectric heating in a short-wave electromagnetic radiation field. For cooking raw food, heating pre-cooked meals and thawing frozen foods the method is very quick and hygienic. Suitable sources of power for this purpose are now available in the form of two types of magnetrons capable of continuous outputs of 2 kW and 5 kW. "Microwave cookers" have definite advantages in hotels and restaurants, where large numbers of meals have to be served in a short time, but it may well be that they will eventually also find their way into the home kitchen.*

In recent decades the heating of materials by high-frequency power has been applied on an ever-increasing scale in many branches of industry [1]). In the case of dielectric heating a limit is set to the delivered power by the electric breakdown strengths of the materials to be heated, which cannot withstand arbitrarily high voltages. The only way to increase the absorbed power is then to raise the frequency, hence the trend towards ever higher frequencies in dielectric heating.

In this respect the recent development of continuous-wave (CW) magnetrons giving an output of 2 and 5 kW represented an important advance [2]). These magnetrons operate at frequencies in the region of 2450 Mc/s, i.e. in the microwave range, and are particularly suitable for the dielectric heating of non-conducting materials. Besides their possible applications in industry, e.g. for drying wood and textile products, or for welding plastics, they have a promising application in the heating of foodstuffs, i.e. for preparing meals in a "microwave cooker". In this article we shall examine some of the problems involved in the construction of such a cooker using a 2 kW CW magnetron.

In the conventional methods of cookery (boiling, frying, roasting, baking, grilling) the heat is supplied by *convection* and *conduction* in water or fat, by *direct contact* with the heated pan or by *thermal radiation*. The heat in all these cases can only penetrate inside the food by conduction. The temperature gradient which this requires may not be too steep, as otherwise the surface of the food will suffer. The heating process therefore takes much longer than by the dielectric method, where the thermal conduction of the food is unimportant. In "microwave cookery" the food, with no water or fat added, is placed in a glass or earthenware dish, or even on paper or cardboard, and is heated through and through in a quarter or eighth of the normal time, without drying-out the surface of the food or scorching the paper. Pre-cooked or frozen foods can readily be warmed up again, the vitamins and the natural flavour, colour and other properties of the food being retained to a very high degree. There are various tricks by which the brown crust to which we are accustomed in conventional cookery can also be produced by microwave radiation; usually, however, a normal grill will be fitted to the oven for this purpose.

The use of a microwave cooker offers especial advantages for hotels and restaurants; the preparation of meals is very much quicker, and dishes can also be pre-cooked and warmed up when the time comes to serve them. This can be a great help in rush hours and where kitchen staff is short. In hospital kitchens experience has already proved that microwave cookers make it possible to provide a much more varied menu for patients on low-fat diets. Finally, a microwave cooker can also be a valuable asset in the home; a housewife who goes out to work will save a great deal of time preparing meals in this way, particularly if more pre-cooked meals in frozen form are made available by the foodstuffs industry.

In the following we shall first consider the physical principles underlying the method of dielectric heating in a microwave radiation field. After discussing the 2 kW CW magnetron marketed by Philips, we shall then examine the problems involved in the design of the oven proper, i.e. the space in which

*) Development Laboratory of Valvo GmbH, Radioröhrenfabrik Hamburg.

[1]) See e.g. the articles Heating by high-frequency fields, I. Induction heating, by E. C. Witsenburg, and II. Capacitive heating, by M. Stel and E. C. Witsenburg, Philips tech. Rev. **11**, 165-175 and 232-240, 1949/50.

[2]) W. Schmidt, Das Dauerstrichmagnetron Valvo 7091, Elektron. Rdsch. **12**, 309-314, 1958; or, Continuous-wave magnetrons types 7091 and 7292, Electron. Appl. **20**, 13-23, 1959/60 (No. 1).

the food is heated. Finally, an experimental model is discussed by way of illustrating the actual construction of a microwave cooker.

## Principles of dielectric heating in a microwave radiation field

At the frequencies commonly used for RF heating, the high-frequency power is fed into a coil or a capacitor. Conducting materials are heated in the magnetic field of a coil by the induction of eddy currents; non-conducting materials are placed in the alternating electric field of a capacitor, where the dielectric losses produce the desired heating. In the microwave range, i.e. at frequencies above 1000 Mc/s, the substance to be heated is placed in a resonant cavity, in which the electric and magnetic fields are so interwoven as to be practically indistinguishable. The "oven" of a microwave cooker accordingly consists of an appropriately dimensioned space bounded by metal walls. The microwave energy is conducted to the oven by waveguides. Multiple reflections from the walls fill the oven space with a radiation field, thus providing all-round irradiation of the food introduced. The method is suitable only for the dielectric heating of non-metallic objects, since the electromagnetic waves would be almost entirely reflected by good conductors.

If either pure dielectric or pure induction heating by microwaves is required this would be possible only in a small rectangular cavity whose ends run out into circular cylindrical spaces and whose length does not exceed $\lambda/4$. In the rectangular middle-section there will then be an alternating electric field between two walls which, at the frequency of 2450 Mc/s ($\lambda/4 = 31$ mm) are only 5 mm apart; in the cylindrical extensions there will be an alternating magnetic field of 10 mm diameter. Only very small objects could therefore be heated in these spaces, so that pure dielectric or pure induction heating at these frequencies has little practical significance.

Where an alternating electric field of amplitude $E$ prevails in a medium, the heat $P_{\rm w}$ generated in unit volume and unit time is given by [3]:

$$P_{\rm w} \propto E^2 f \varepsilon_{\rm r} \tan \delta , \quad \ldots \ldots \quad (1)$$

where $f$ is the frequency, $\varepsilon_{\rm r}$ the relative dielectric constant and $\delta$ the loss angle. This expression reveals the advantage of using microwave frequencies: even for small values of $E$ a considerable heating effect is obtained because of the high value of $f$. Whereas at low frequencies the heat generation is limited by the breakdown strength of the material, in the microwave range the limit is set by the maximum power which the generator is capable of delivering.

When an electromagnetic wave is propagated through a (non-magnetic) medium, it is attenuated as a result of the heat generation: the energy density of the wave decreases in the direction of propagation. For a vertically incident plane wave the "penetration depth" $z_{\rm i}$, which is conventionally defined as the distance at which the energy density has dropped to $1/e \approx 37\%$, is given by [3]:

$$z_{\rm i} \propto \frac{1}{f \sqrt{\varepsilon_{\rm r}} \tan \delta} . \quad \ldots \ldots \quad (2)$$

*Table I* gives the values of $z_{\rm i}$ calculated from measurements of $\varepsilon_{\rm r}$ and $\tan \delta$ on various foodstuffs. It can be seen that this penetration depth in some substances is rather small. Where fairly large volumes are involved there is consequently a danger that the substance will not be properly heated through. Equation (2) shows that the higher the frequency the less is the effective penetration of the heat. (In grilling this is, of course, turned to good advantage. Here, too, an electromagnetic radiation field is used — though of much higher frequency, i.e. in the infra-red — and this radiation is almost entirely absorbed in the surface layer, thus producing the familiar brown crust.) The limitation is not so serious as it seems, for in a resonant cavity the energy penetrates the substance from all sides. Moreover, experience has demonstrated that the results achieved in the cooking and thawing of

Table I. Values of $\varepsilon_{\rm r}$ and $\tan \delta$ for various foodstuffs, measured at various temperatures, and the calculated (theoretical) penetration depth $z_{\rm i}$ for microwaves of 2450 Mc/s.

| Foodstuff | Meas. temp. °C | $\varepsilon_{\rm r}$ | $\tan \delta$ | Penetration depth $z_{\rm i}$ (in cm) of 2450 Mc/s microwaves |
|---|---|---|---|---|
| Beef, raw | –15 | 5.0 | 0.15 | 5.8 |
| Beef, roasted | 23 | 28.0 | 0.2 | 2.46 |
| Peas, boiled | –15 | 2.5 | 0.2 | 7.9 |
| | 23 | 9.0 | 0.5 | 1.5 |
| Pork, raw | –15 | 6.8 | 1.2 | 0.66 |
| Pork, roasted | 35 | 23.0 | 2.4 | 0.18 |
| Potatoes, boiled | –15 | 4.5 | 0.2 | 6.1 |
| | 23 | 38.0 | 0.3 | 1.44 |
| Spinach, boiled | –15 | 13.0 | 0.5 | 1.42 |
| | 23 | 34.0 | 0.8 | 0.56 |
| Porridge | –15 | 5.0 | 0.3 | 3.7 |
| | 23 | 47.0 | 0.8 | 0.41 |

[3] Concerning the derivation of (1) and (2), see: W. Schmidt, Mikrowellengeneratoren mit abgeschlossenem Arbeitsraum zur dielektrischen Erwärmung von Nahrungsmitteln und Industrieprodukten, Elektron. Rdsch. **12**, 390 and 417, 1958, and **13**, 13, 1959; or, Microwave generators coupled to a loaded cavity for dielectric heating of foodstuffs and industrial products, Electron. Appl. **19**, 147-164, 1958/59 (No. 4).

foods are much better than might be inferred from the theoretical penetration depth given in table I.

## Magnetrons

Magnetrons operate with a high efficiency and are designed for a fixed frequency. They are complete generators in themselves. The designer wishing to use a magnetron as a microwave generator is therefore virtually unconcerned with problems of high-frequency engineering such as the construction and alignment of the frequency-determining oscillatory system or the feedback system.

Magnetrons are normally fitted with some sort of demountable output connection for taking-off the power. Provided there are no significant differences in individual characteristics, the replacement of a magnetron can be reduced to a simple mechanical operation.

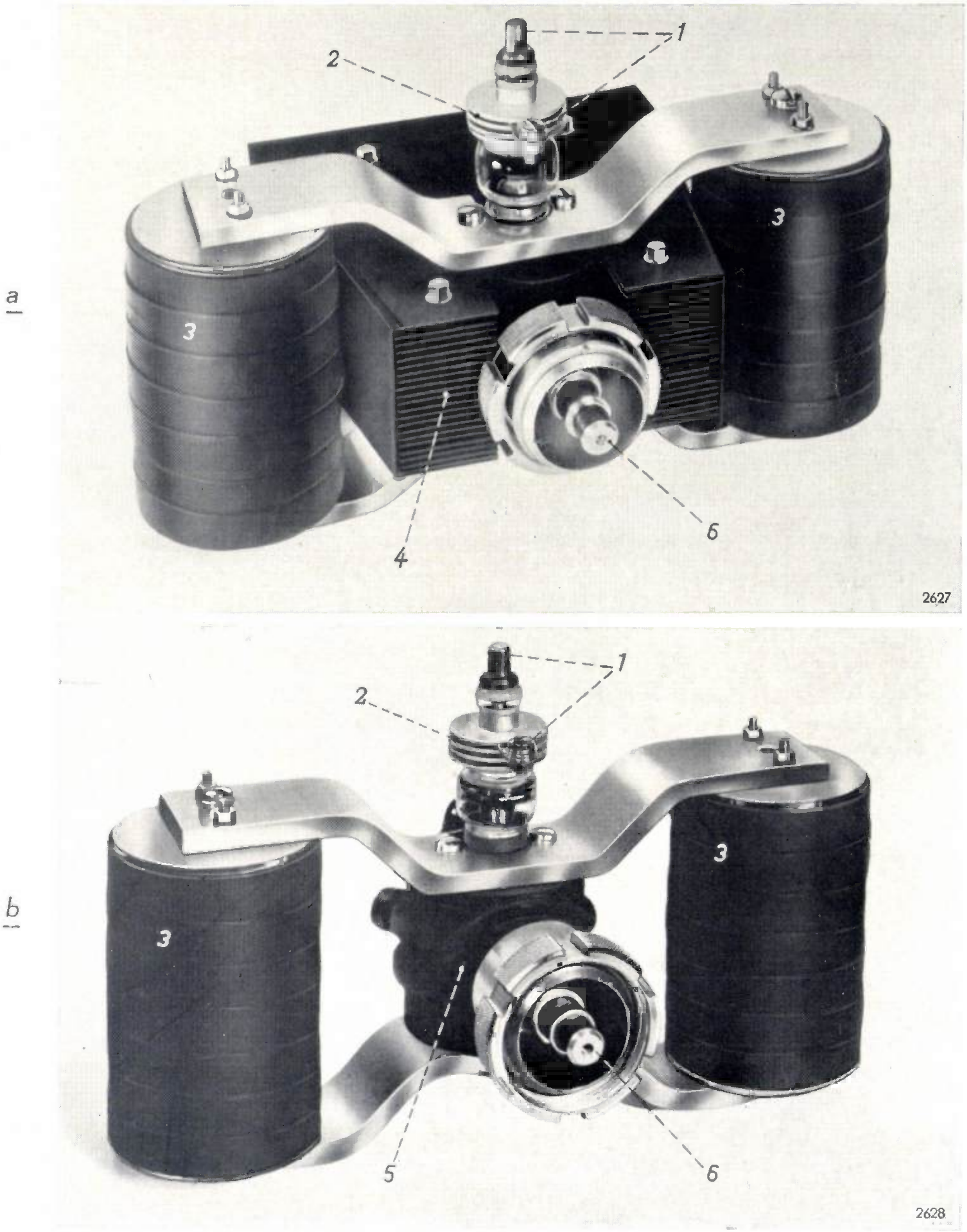


Fig. 1. Two CW magnetrons for 2 kW, 2450 Mc/s: *a*) type 7091, air-cooled and *b*) type 7292, water-cooled. *1* heater connections, the lower one also being the cathode connection. *2* cathode radiator. *3* ferroxdure magnets. *4* anode block with vanes for air cooling. *5* anode block with water cooling. *6* connection for coaxial output line (50 Ω). The inner conductor of the output connection is provided with a screw-thread, to ensure good contact with the output line even after long use at varying temperatures.

### *Construction of CW magnetrons types 7091 and 7292*

Pulsed magnetrons, which have been used for more than twenty years in radar and electronic navigation devices, are required to meet high demands as to frequency stability, pulse shape and reliability. With a CW magnetron for RF heating the emphasis is more on such demands as high efficiency, long life, low working voltage and insensitivity to load variations. The latter point is of particular importance, since the substances heated in a microwave oven differ widely in dielectric properties, shape and size and therefore subject the magnetron to widely different loads. Constancy of frequency is of less importance, the frequency bands available for industrial purposes (and which are regulated by law in the various countries) being fairly wide.

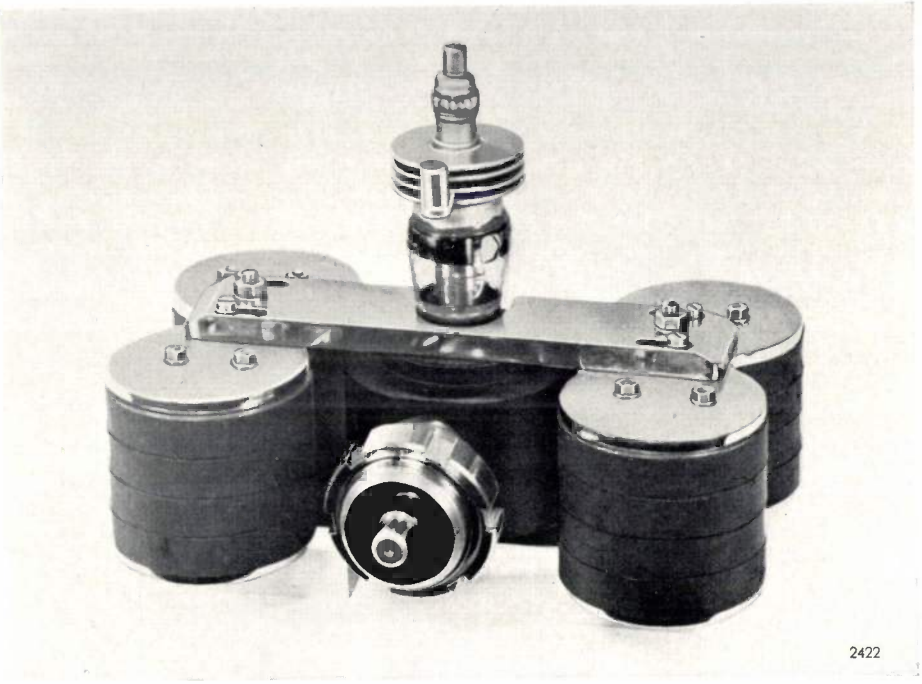


Fig. 2. CW magnetron type 55 125 for 5 kW. This magnetron also operates in the 2450 Mc/s band. The anode is water-cooled; the cathode radiator is cooled by a weak air current. The output power of 5 kW is obtained with an unsmoothed rectified voltage supply, the anode voltage being 6.5 kV and the average anode current 1.4 A. The magnetic field is provided by *four* columns of ceramic magnets. As in the 2 kW magnetrons, the power is extracted by a 50 Ω coaxial line.

For industrial purposes, i.e. where the high-frequency energy is used for heating, drying and sintering non-conducting substances or — as in the present case — for microwave cookery, Philips have developed two CW magnetrons, types 7091 and 7292. Microwave cookery, incidentally, was the first application of these magnetrons. Both types deliver a maximum power of 2500 watts at a frequency of about 2450 Mc/s [2]). Type 7091, which is air-cooled, was designed for ovens that may have to be moved from one place to another (*fig. 1a*). Type 7292, which is water-cooled, is intended for permanent installations (fig. 1*b*). Except for the method of cooling, the technical data for both types are identical.

Development work on a CW magnetron for an output power as high as 5 kW has recently been completed. This type, a picture of which is shown in *fig. 2*, and its possible applications, will not be dealt with here.

The oscillatory system of the magnetrons 7091 and 7292 consists of 20 sector resonant cavities in the anode [4]). A coaxial line takes the power off by means of two balanced coupling loops (*fig. 3*).

[4]) For a general description of the operation and design of magnetrons, see J. Verweel, Magnetrons, Philips tech. Rev. **14**, 44-58, 1952/53. See also G. A. Espersen and B. Arfin, A 3 cm magnetron for beacons, Philips tech. Rev. **14**, 87-94, 1952/53, and J. Verweel and G. H. Plantinga, A range of pulsed magnetrons for centimetre and millimetre waves, Philips tech. Rev. **21**, 1-9, 1959/60 (No. 1).

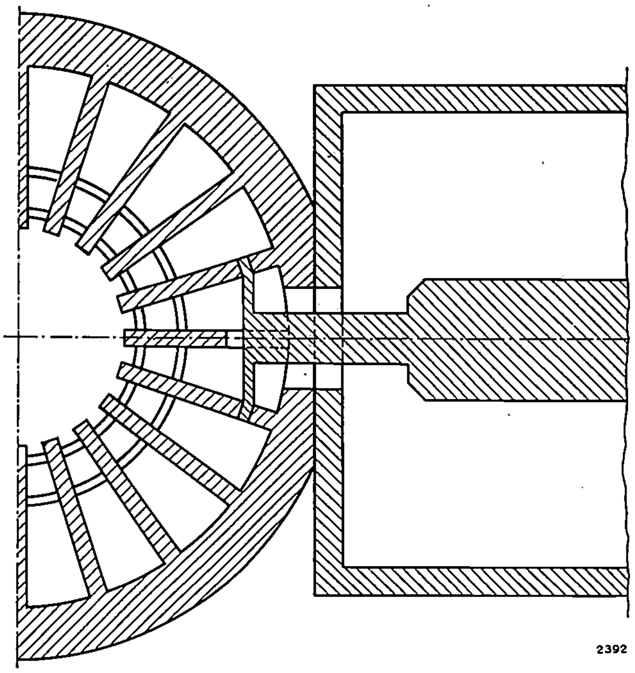


Fig. 3. Output system of types 7091 and 7292 magnetrons. The inner conductor of the coaxial output line is coupled to *two* resonant slots of the anode by a T-shaped "loop". The partitions between the resonant slots are alternately connected by two metal rings (ring strapping).

In the 2450 Mc/s frequency band it is possible to achieve high Q's for the unloaded cavity system, and therefore the coaxial output line can be permanently coupled to the magnetron. The circuit efficiency of the magnetron is then high, while the reserve of stability is still sufficient to deal with load reflections that may occur if the load is not exactly matched. To achieve this fixed coupling the system of coupling loops must link a large part of the alternating magnetic flux in a resonant cavity. However, in view of the fact that a high standing-wave ratio, as a result of load reflections, may give rise to high currents in the output line, it is necessary to use material of large cross-section for the coupling, and this makes the first requirement difficult to fulfil if a single coupling loop is used. The use of a double coupling loop provides the desired permanent coupling in the 7091 snd 7292 magnetrons in spite of the large cross-section of the loop material (which ensures good heat removal) and without any capacitive coupling with the resonant cavity, which would adversely affect operating stability.

The inner conductor of the output system of magnetrons 7091 and 7292, illustrated in fig. 3, forms part of a tapered vacuum seal. The insulation used is a ceramic, which is stronger than glass. The dielectric losses are low, even if load reflections considerably increase the standing-wave ratio in the line. The output system is fitted with a standard connecting flange.

The cathode in a magnetron is subjected to bombardment from electrons not captured by the anode [4]). These electrons absorb energy from the high-frequency field and convert this energy into heat when they return to the cathode. In CW magnetrons, then, the cathode load is not limited by the emission per unit area but by the energy density in the space between anode and cathode in relation to the surface area of the cathode. The thermionic emission is about 100 mA/cm$^2$, plus a considerable contribution from secondary emission. Since the energy transferred to the cathode in the form of heat by returning electrons depends on the input power and the magnitude of the load reflection, the cathode surface in CW magnetrons for industrial applications, where the load reflection may vary within wide limits, must be exceptionally strong mechanically and capable of withstanding large variations in temperature. Normal oxide cathodes are no longer adequate at high frequencies, which call for solid or sintered cathodes. Magnetrons 7091 and 7292 therefore have impregnated dispenser cathodes [5]), the emitting surface of which is a porous tungsten cylinder, impregnated with a substance that promotes thermionic emission. Such cathodes have a long life and stable emission even though subjected to high and varying temperatures. The mechanical strength of the tungsten jacket enables this kind of cathode to withstand back-bombardment as well as breakdown effects due to overloading.

A special feature is the use of a heater not in contact with the cathode proper: the cathode cylinder is thus heated only by radiation. The usual insulation of the heater — e.g. with aluminium oxide — is therefore dispensed with. In many applications, including microwave cookery in hotels and restaurants, the heater may remain switched on for hours on end, high-frequency power being taken off only now and then. Experience so far indicates that, under these conditions, this form of heater is very satisfactory. Mechanical strength is ensured by using a heater wire of considerable thickness, viz. 1.2 mm in diameter.

Operating costs are a decisive consideration where domestic and industrial apparatus are concerned. Since the magnetron represents a substantial part of the total costs, the expense entailed by its replacement must be kept as small as possible. For example, components outside the vacuum system of the magnetron should not require to be replaced when the magnetron is replaced. Moreover, in order that untrained personnel may effect replacements, the high-frequency system should require no regulation or adjustment.

In this connection the use of the ceramic ferroxdure for the permanent magnet has decided advantages. Possessing a much higher coercive force than

[5]) See R. Levi, Dispenser cathodes, III. The impregnated cathode, Philips tech. Rev. 19, 186-190, 1957/58.

magnetic alloys, it is virtually insensitive to stray magnetic fields and to changes in the resistance of the magnetic circuit [6]). As a result, a magnetron with built-in iron pole-pieces can be removed from the magnetic circuit without any fear that the temporary demagnetization will cause permanent weakening of the induction in the air gap. The pole-pieces used in magnetrons 7091 and 7292 keep the air gap to a minimum, thereby minimizing the amount of permanent-magnet material required. In this way the magnet system using ceramic material combines the "packaged" system's advantage of low costs with the "unpackaged" system's advantage of demountability.

Ceramic magnetic material also has a certain drawback, in that its magnetic properties are more sensitive to temperature variations than those of magnetic alloys. This is not, however, a serious objection. In fig. 1 it can be seen that the ferroxdure "pillars" are connected to the magnetron system by narrow iron yokes; the effect of this is to make the time constant of the rise in temperature of the magnetic material longer than half an hour. Moreover, the final temperature rise of the magnets is only 20% of that of the magnetron itself. In view of the short periods during which the magnetron is switched on for microwave cookery, its operation is not noticeably affected by the sensitivity of the magnets to temperature variations, particularly since, in a microwave oven, fluctuations of the mains voltage and of the load impedance can cause a relatively ten-times greater variation of the output power. To compensate for changes in output power, anode-voltage regulation is necessary. In the air-cooled magnetron, type 7091, the slight influence of the temperature variations can in any case be eliminated by fitting two additional air ducts to circulate air around the magnets. In the water-cooled magnetron, type 7292, the magnetron itself rises so little in temperature that the heating of the magnets may be discounted.

### *Operating data and performance charts*

Having considered the structural features of the CW magnetrons, we shall now examine the behaviour of the system magnetron — coaxial line — oven. The system is characterized by the way in which the frequency and the power delivered by the magnetron depend on the load impedance constituted by the coaxial line and the oven. It might be studied by drawing curves of constant frequency and constant delivered power in cartesian coordinates, the real and imaginary parts of the load impedance being plotted against one another. A more useful diagram, however, is obtained by drawing these contours in a polar diagram, with the modulus $\varrho$ and the argument $\varphi$ of the reflection coefficient plotted as radius vector and azimuth, respectively. The reflection coefficient is the ratio between the complex amplitudes of the reflected and incident waves, at any arbitrary cross-section of the coaxial line carrying the power from the magnetron. The position of this "reference plane" is usually chosen at the connecting flange of the magnetron. The modulus $\varrho$ is determined by measuring the standing-wave ratio $\sigma$ in the coaxial line; the relation between $\varrho$ and $\sigma$ is given by the formula $\sigma = (1 + \varrho)/(1 - \varrho)$. The argument $\varphi$ is related in a simple manner to the distance $\Delta l_{\min}$ from the reference plane to the nearest minimum in the standing wave, according to the expression: $\Delta l_{\min}/\lambda = \frac{1}{4}(1 - \varphi/\pi)$, where $\lambda$ is the wavelength in the coaxial line. To make the diagram simpler to use, it is convenient to plot, as in *fig. 4*, circles for constant values of $\sigma$ instead of for constant values of $\varrho$. Since $\sigma$ runs from 1 for $\varrho = 0$ to $\sigma = \infty$ for $\varrho = 1$, the value of $\sigma$ rapidly increases as the circle grows larger. Also, the diagram give values of $\Delta l_{\min}$, expressed in terms of $\lambda$, instead of values of $\varphi$ itself. At a constant magnetic field and constant anode current, the values of $\sigma$ and $\Delta l_{\min}/\lambda$ can be measured for arbitrary values of the load impedance (which need not itself be known). Each measurement produces a point in the diagram. The corresponding values of the frequency and output power are also measured. By doing this for various values of the load impedance, and by joining-up the points of equal power and also those of equal frequency, we obtain the Rieke diagram for the magnetron. Fig. 4 gives the diagram for types 7091 and 7292. The diagram relates to an operating point corresponding to 2000 W, that is to values of magnetic field and anode current such that, given ideal matching (centre point of diagram, $\sigma = 1$), the output power is 2000 W. The possible application and merits of a CW magnetron can now be assessed by noting the relation between the various quantities depicted in the diagram.

It is seen that as the power is increased, the power contour moves towards the upper right of the diagram. The operating point of the magnetron must not, however, enter the hatched area known as the "sink region", i.e. the region of instability, where the magnetron no longer oscillates properly. The frequency contours also converge upon this region,

[6]) See Philips tech. Rev. **13**, 194-208, 1951/52 and **16**, 141-147, 1954/55.

indicating that here the frequency, too, is highly unstable. Opposite this "electronic instability limit" can be seen the "thermal instability limit", *Th.* If this is exceeded, the cathode will be overheated due to back-bombardment by returning electrons. The anode, too, will be overheated because too little power is then withdrawn from the magnetron as a result of strong reflection from the load impedance. The aim is to design the whole magnetron system so as to keep these forbidden regions as far apart as possible and also to ensure that the midpoint of the diagram lies midway between the electronic and thermal instability limits. The maximum permissible standing-wave ratio, which is determined by the minimum distance between the centre point and the boundaries of the dangerous regions, is then as high as it can be. As regards the type 7091 and 7292 magnetrons, this is the situation at the 2 kW setting.

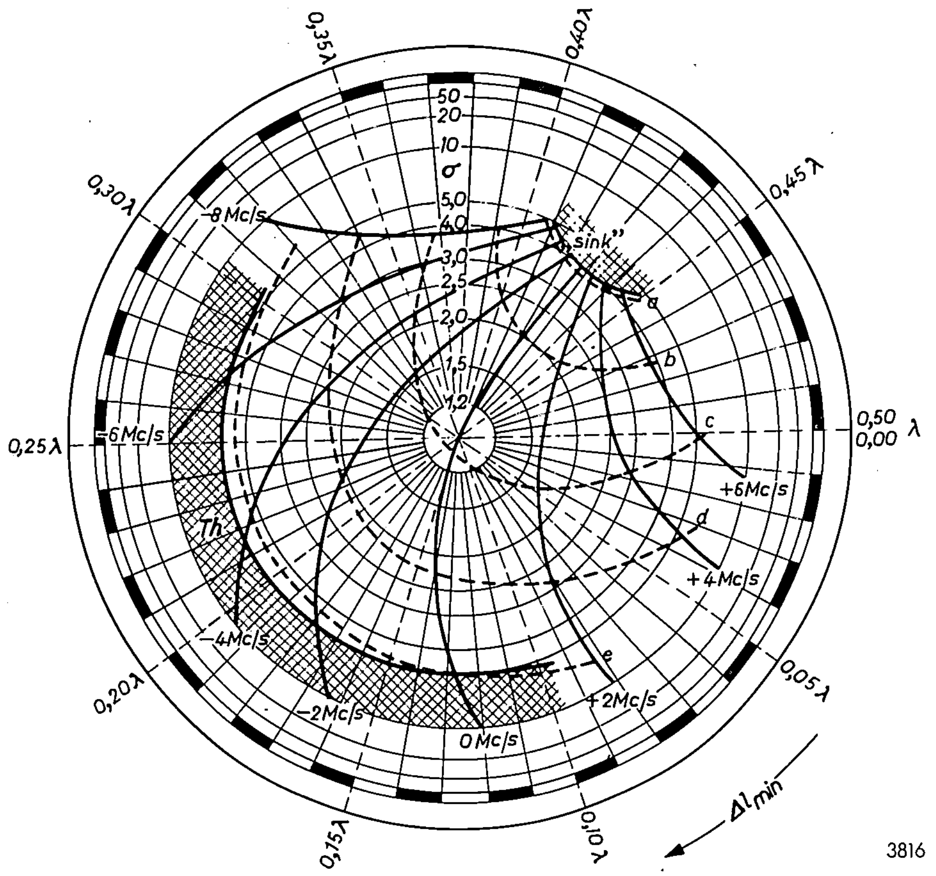


Fig. 4. Rieke diagram of a type 7091 or 7292 magnetron. The circles are contours of constant standing-wave ratio $\sigma$, the values being indicated. Around the periphery are set out the values of $\Delta l_{\min}$, being the distance from the reference plane (through the connecting flange of the magnetron) to the first minimum or node of the standing wave, expressed in terms of the wavelength $\lambda$ in the coaxial output line. The diagram contains contours of constant frequency (solid curves) and of constant output power (dashed curves). The letters *a*, *b*, *c*, *d* and *e* on the dashed lines refer to the table, where the corresponding values of output $P_o$ and anode voltage $U_a$ are given. The diagram holds for an operating point of 2000 W, i.e. the output power is 2000 W when the load impedance is a matched termination, giving a standing-wave ratio of $\sigma = 1$ (centre point of diagram). The hatched region top right is the region of electronic instability, called the "sink"; opposite to it is the region of thermal instability *Th.* When the magnetron is in operation, $\sigma$ must never be so high as to bring the operating point into one of these regions.

| | *a* | *b* | *c* | *d* | *e* |
|---|---|---|---|---|---|
| $P_o$ (W) | 2500 | 2250 | 2000 | 1500 | 1000 |
| $U_a$ (kV) | 4.7 | 4.6 | 4.5 | 4.4 | 4.3 |

Fig. 4 also indicates the extreme values between which the frequency will adjust itself if the standing-wave ratio is, say, 1.6. The circle for $\sigma = 1.6$ touches the contours for + 2 Mc/s and — 2 Mc/s, so that the operating frequency will be within the 2450 ± 2 Mc/s band. The microwave frequency band allocated for industrial applications is 2400-2500 Mc/s in most countries (2350-2450 in Germany). Owing to the spread in properties between magnetrons as manufactured, it is always possible to select individual specimens having frequencies suited to the local conditions.

It is most important that the operating point should not enter the sink region, as this can very quickly cause damage to the magnetron. The consequences are not so serious if the boundary of the thermal instability region is crossed. The position of the two danger zones depends, of course, on the anode current; they move inwards as the anode current increases, in which case the maximum permissible standing-wave ratio is reduced.

A direct voltage is commonly used for the anode of a CW magnetron. Magnetrons can also operate, however, with an alternating anode voltage, or with a rectified but unsmoothed alternating voltage, and the latter is in fact used in this case. With a full-wave rectified, unsmoothed voltage the ratio of the peak anode current to the mean value is about 2.5; the sink then lies outside the circle for $\sigma = 4$. With an AC supply the permissible value of $\sigma$ is smaller. If a smoothed rectified voltage were used — in which case the ratio of peak to mean current would be about 1 — the limit of electronic stability would lie further from the centre of the Rieke diagram, and on the face of it the permissible value of $\sigma$ might then be higher. The objection here, however, is that at values of $\sigma$ higher than 4 or 5 (in other words, with a reflection of more than 50% of the power), large concentrations of energy may occur at places of maximum field-strength, with consequent danger to the output line contacts and the vacuum seal of the output coupling. Standing-wave ratios higher than 4 or 5 are therefore ruled out, so that there is nothing to be gained from smoothing. This cuts out the expense of smoothing capacitors, and thus helps to keep down the price of the cooker.

Table II. Principal data for magnetrons 7091 and 7292, valid for ideal matching (operation at 2 kW) and provisional data for 5 kW magnetron type 55 125.

| Type of magnetron | 7091 and 7292 | 55 125 |
|---|---|---|
| Anode voltage | 4.5 kV | 6.5 kV |
| Mean anode current | 0.75 A | 1.4 A |
| Maximum peak anode current | 2.1 A | 2.4 A |
| Output power | 2000 W | 5000 W |
| Maximum permissible standing-wave ratio $\sigma_{\max}$: | | 2.5 |
| sink (electronic limit) | 4.0 | |
| thermal limit | 5.0 | |

In any case, smoothing would involve the extra danger that sparking and flashover effects might develop, as a result of capacitor discharges, into serious electrical breakdowns. The use of two-phase or three-phase rectifiers, with no smoothing, is therefore to be recommended for supplying a normal microwave cooker operating at 2 kW.

*Tables II* and *III* give some general data on the 7091 and 7292 magnetrons, and some performance data for the case of ideal matching. For comparison, the provisional data for the 5 kW magnetron are included.

Table III. General data for CW magnetrons 7091 and 7292, and provisional data for type 55 125.

| Type of magnetron | 7091 | 7292 | 55 125 |
|---|---|---|---|
| Power extraction system | 1⅝″ coaxial line; 50 Ω | | |
| Cooling of anode block | Air, approx. 1.7 m³/min | Water, at least 0.5 l/min, depending on inlet temp. | Water, approx. 2.5 l/min, depending on inlet temp. |
| Cooling of cathode radiator | Weak air current | | |
| Max. temperature of anode block | 125 °C | 125 °C | 125 °C |
| Max. temp. of cathode radiator | 180 °C | 180 °C | 180 °C |
| Heater voltage: | | | |
| upon switching-on | 5.0 V (+5%, —10%) | | 5.5V(+5%,—10%) |
| in normal operation | 2.0 V | | 3.5 V/1 V |
| Heater current: | | | |
| upon switching-on | 32 A | | 66 A |
| in normal operation | 18 A | | 52 A/28 A |
| Warm-up time | 120 s | | 240 s |

## The oven

To build a microwave cooker using the magnetron described, the first question to be decided concerns the shape of the oven, i.e. the space in which the food is exposed to the microwaves. The aim is to achieve in this space a uniformly distributed energy density, in order that the foodstuffs introduced shall be uniformly heated. If we choose a rectangular shape, the space will then act as a resonant cavity which, fed with energy through a coupling system, can be made to oscillate if the wavelength in vacuo $\lambda_0$, corresponding to the magnetron frequency $f$ (i.e. $\lambda_0 = c/f$, where $c$ is the velocity of light), and the side lengths $a_x$, $a_y$ and $a_z$ satisfy the following condition ($m$, $n$ and $p$ are integers):

$$\lambda_0 = 2\Big/\sqrt{\left(\frac{m}{a_x}\right)^2 + \left(\frac{n}{a_y}\right)^2 + \left(\frac{p}{a_z}\right)^2}. \quad . \quad . \quad . \quad (3)$$

This condition applies strictly only to an ideal resonant cavity with no damping and with no space charge [7]). In reality, however, all resonant cavities possess a certain damping, as a result of which their resonance curves show broad maxima, the more so if substances are introduced in them — as in the oven — which further increase the damping. Condition (3) need not, therefore, be exactly fulfilled; provided only the sides are longer than $\frac{1}{2}\lambda_0$, there will always be integers $m$, $n$ and $p$ in respect of which (3) is satisfied with sufficient accuracy. Such a set of $m$, $n$ and $p$ values is met by a field distribution consisting of a pattern of standing waves having $m$, $n$ and $p$ half-wavelengths (not to be confused with $\lambda_0$) in the three directions. If the sides are much longer than $\frac{1}{2}\lambda_0$, there will be many sets of values of $m$, $n$ and $p$ for which condition (3) is adequately satisfied. In such an oven, then, there can exist numerous modes of oscillation at the same time, and they are all simultaneously excited provided that the boundary conditions are fulfilled at the position of the coupling with the exciter waveguide. Thus, if the dimensions of the oven are large compared with $\frac{1}{2}\lambda_0$, superposition of the various oscillation modes leads to a more or less uniform distribution of energy in the oven. If no further measures are taken, however, non-uniformity will nevertheless appear over larger distances. The reason for this may be unbalanced coupling with the energy source, irregularities in the walls (e.g. the oven door), the presence of trays for putting the food on, and so on. When the foodstuffs are placed in the oven, they also affect the field distribution. Owing to the absorption of energy the resultant field is composed of wave trains which, depending on whether or not they have passed through the absorbent substance, transport energy of differing density. Added to this is the partial reflection of the waves from the surface of the substance; how large this reflection is depends on the dielectric constant of the substance. It must also be remembered that the excitation frequency may vary slightly because of the spread in characteristics between individual magnetrons. The desired uniformity in the distribution of energy in the oven must be assured over the whole range of excitation frequencies. We shall now consider the measures to be adopted to this end.

The broadening of the resonance peaks as a result of damping also tends to stabilize the field distribution, and hence the load impedance of the magnetron. The latter is of considerable importance, since one of the primary conditions to be met if the magnetron is to be interchangeable is that the empty oven shall represent the same load for all magnetrons of the type used. It should consequently yield approximately the same operating point in the Rieke diagram, irrespective of the frequency of the particular magnetron and without the need to adjust the plunger in the matched waveguide, which acts as a matching transformer, between the coaxial line from the magnetron and the oven. The damping in an oven measuring 44 × 40 × 36 cm should be at least equivalent to 100 cm$^3$ of water. Since all parts of the oven in which high-frequency currents flow give rise to losses, every oven possesses a certain amount of damping, but this is adequate only if the walls consist of a material whose resistance is not unduly low — e.g. V2A steel — and provided also that the oven contains such components as reflector plates and bars forming a hot grill. The damping can be increased if the oven trays for putting the food on are made of dielectric materials like glass or certain plastics, which give rise to greater losses than metal trays. They constitute a basic load which is sufficient to allow the cooker to be left switched on when the oven is empty. When there is food in the oven, only a few per cent of the consumed power is lost in these trays, their dielectric losses being much smaller than those of the food.

*Fig. 5* shows two sections through the oven of a microwave cooker. The oven contains a tray for the food, a zig-zag heating element for grilling and a reflector plate. The purpose of the reflector is to promote a homogeneous energy density in the oven, its principal action being to offset the irregularity

[7]) The derivation of (3) is given in the first article quoted in footnote [3]).

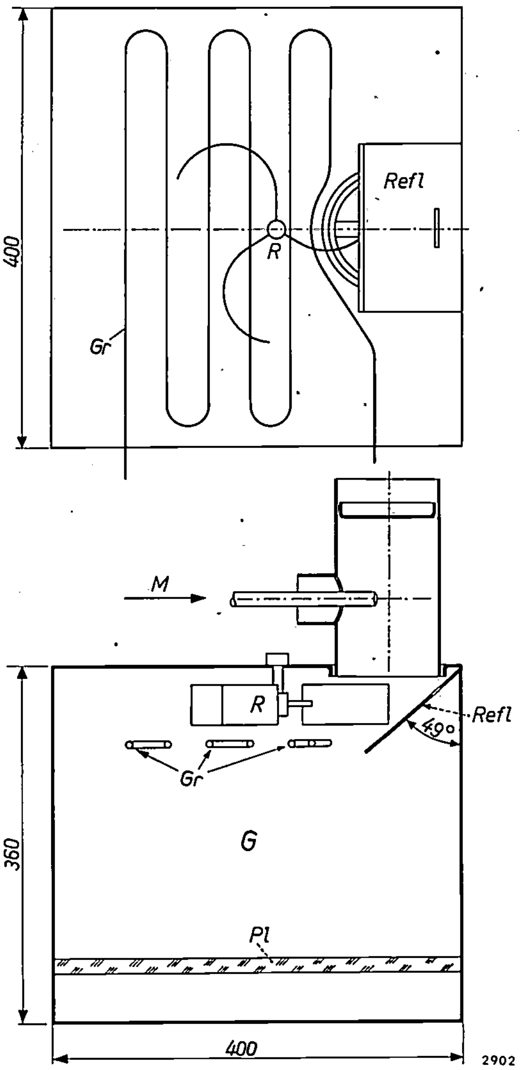


Fig. 5. Two sections through the oven of an experimental microwave cooker. *M* coaxial line from magnetron. *G* oven. *Refl* reflector. *R* field stirrer. *Gr* grill element. *Pl* tray for carrying the food. Dimensions in mm.

caused by the asymmetrical connection of the matched waveguide (see below). The same result can be obtained by fitting the internal edges of the oven with oblique metal strips.

The effect of all these measures and the best position of the reflector plates can only be found by experiment. For this purpose, 25 small dishes, each containing 50 cm$^3$ of water, were placed in an experimental oven. The rise in the temperature of the water in these dishes within a certain time gives qualitative and quantitative indications regarding the distribution of the energy density, averaged over time, in the plane where the dishes are situated. *Fig. 6a* shows the result of an initial experiment. The marked rise at the right of the oven is due to the fact that the matched waveguide is not connected symmetrically. There was a practical reason for this, namely that, in addition to this waveguide, room had to be found on the oven for the magnetron and a fan as well. Fig. 6*b* illustrates the improvement effected by introducing a reflector provided with a slit, as shown in fig. 5. The energy density can be made still more uniform — or at least its time average — by constantly varying its distribution. This may be done by varying certain quantities that affect the boundary conditions for the field distribution. Since the frequency of the magnetron can only be altered within narrow limits, the only possibility is to alter the geometry of the oven, e.g. by making the walls themselves, or special reflector plates, undergo periodic movement. The oven represented in fig. 5 contains a rotating blade reflector called a "field stirrer". Fig. 6*c* shows the density distribution of the dissipated energy after the stirrer has been given a suitable shape and position in the oven, again found by experiment. One might also, of course, homogenize to some extent the heat development in the food by making the oven tray rotate.

Once a uniform distribution of energy has been achieved by these measures, it is disturbed again by the absorption and reflection of energy that occurs as soon as food is introduced. It is therefore virtually impossible to achieve a homogeneously dense distribution of energy in the oven for any arbitrary content. In practice, however, it is found that sufficiently uniform heating is obtained without the radiant energy density being exactly homogeneous, the reason being that heat conduction in the food has an equalizing effect.

A door is required that gives easy access to the oven. When closed it must shut-in the microwave energy, and this raises problems concerning the door contacts. Chinks in the oven wall allow microwave energy to escape and also distort the field distribution inside. The first step is to try by mechanical means to ensure good electrical contact in the door joints, e.g. by arranging a series of contact springs around the opening. At high frequencies it is necessary to use very reliable contacts, but as they inevitably get dirty it is difficult to keep them functioning well over a long period. The second possibility of making the door "high-frequency tight" consists in using quarter-wave slots. Two such slots are used, opposite to each other and both $\lambda/4$ deep (in this case 31 mm deep at $f = 2450$ Mc/s, $\lambda = 12.5$ cm); together they form a waveguide $\lambda/2$ long, short-circuited at one end (*fig. 7a*). As a result of reflection a standing wave appears in these slots, the voltage and current distribution of which are shown in fig. 7*b*. As the current in the middle

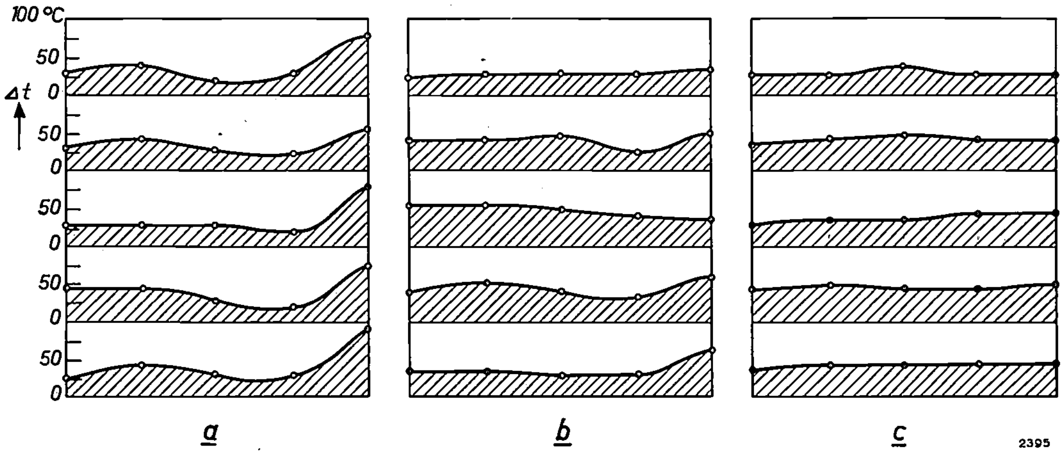


Fig. 6. Distribution of energy density in the oven of an experimental microwave cooker. The three diagrams *a*, *b* and *c* relate to the same horizontal section through the oven at the situation of the food. At five points along each of five straight lines in this cross-section is plotted the temperature rise in twenty-five dishes, each containing 50 cm$^3$ of water. In each case, *a*, *b*, *c*, the cooker was switched on for the same time.
*a*) Original distribution.
*b*) The distribution has been made more uniform by introducing a reflector plate containing a slit (see fig. 5).
*c*) The time-averaged energy distribution has been made very uniform by the use of a "field stirrer".

is zero, the contact at that position need not be perfect and that is where we can situate the join between oven wall and door. This produces the effect — at least in theory — of a space closed without joins, and one may therefore count on insensitivity to corrosion and dirt.

In practice the $\lambda/4$ slots are not in themselves sufficient, particularly not when the oven is empty and there are high RF currents flowing in the walls. The slots should therefore be combined with an effective mechanical contact, as illustrated in *fig. 8*. Here, however, the existing mechanical contact between door and housing at the right had to be improved by contact springs, which are not shown in the drawing.

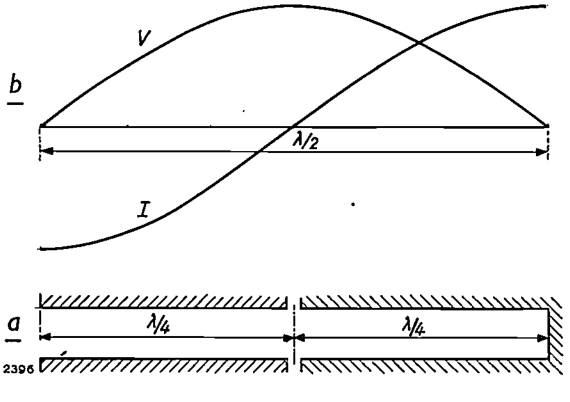


Fig. 7. Illustrating how the use of quarter-wave slots prevents the escape of high-frequency energy.
*a*) Two opposing slots, each $\lambda/4$ deep, constitute a $\lambda/2$ waveguide short-circuited at one end.
*b*) Voltage ($V$) and current ($I$) amplitudes along the $\lambda/2$ waveguide. At the position where $I = 0$, a poor electrical contact is of no consequence.

The wall currents being strongest when the oven is empty, the door contacts are then subjected to the severest load and there is then a greater chance of damage being done to the contacts and of microwave escaping. It should also be remembered that the microwave energy may not always be uniformly distributed over the whole circumference of the door but may be concentrated, radiating from corners or hinges if the construction is imperfect.

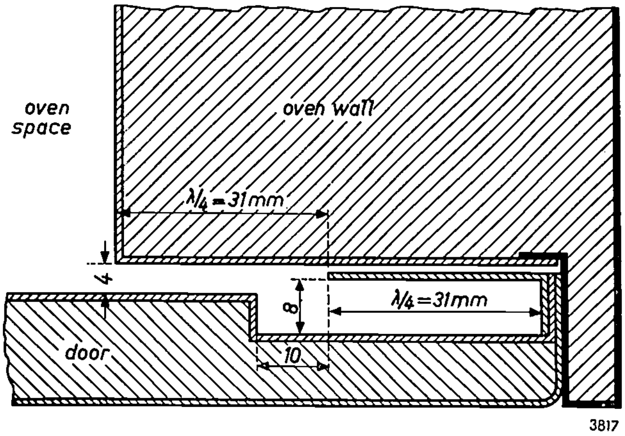


Fig. 8. The application of $\lambda/4$ slots to prevent the escape of microwave energy through the joins in the oven door. In practice it turns out that it is still advantageous to use contact-springs. These are not shown in the drawing.

### Construction and circuitry of a microwave cooker

To conclude we shall now touch on the actual construction of an experimental microwave cooker, and also discuss briefly the power supply circuit. A microwave cooker intended for a wide market

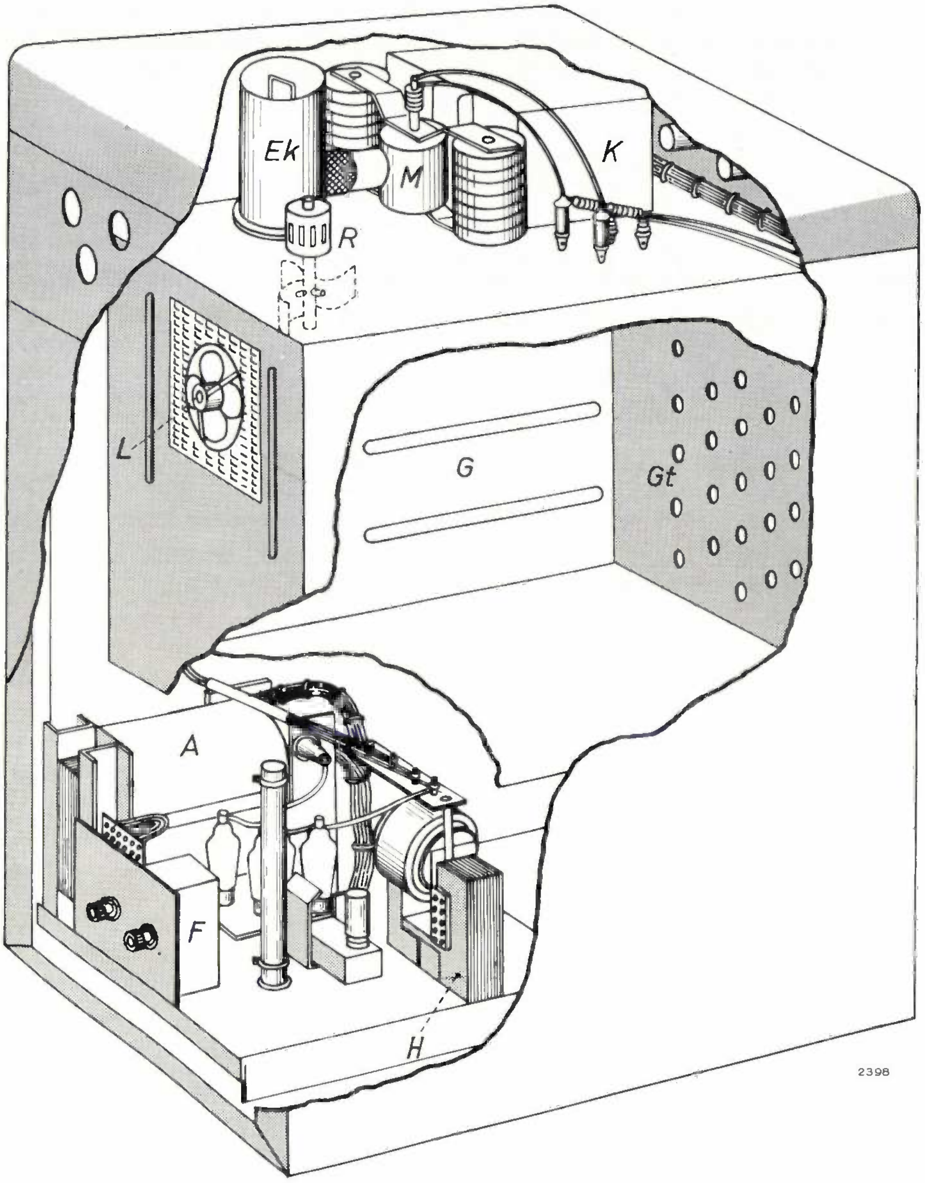


Fig. 9. Cutaway view of a microwave cooker. *M* magnetron. *K* cooling fan. *Ek* coupling waveguide. *R* field stirrer. *G* oven. *L* air extractor. *Gt* oven door. *A* anode-voltage transformer. *F* suppressor filter. *H* heater-current transformer (the secondary of which is at high tension and is correspondingly insulated).

and which will be handled by untrained users (cooks and housewives) must be reliable and easy to operate. It must therefore be provided with safety devices to prevent overloading and to exclude the risk of damage or accidents due to mistakes in operation.

*Fig. 9* gives a somewhat simplified cut-away view of a microwave cooker, seen from the back, that could be installed in a restaurant or private kitchen. As can be seen, the oven is about half-way up, at the same level as in an ordinary cooker. The magnetron *M*, with its coaxial output line and the matched waveguide *Ek*, is mounted on top of the oven. Here, too, are accommodated the cooling fan *K* and the motor for the field stirrer *R*. The rear wall of the oven is fitted with an extractor fan *L*. The oven door is provided with holes through which fresh air can enter and which make it possible to watch the food while it is cooking. Interior lighting, not shown in fig. 9, is then necessary. (There is no significant escape of microwave energy either through these holes or through the air extractor.) The oven may also be fitted with a grill element which, as mentioned, serves as part of the main load and also provides thermal radiative heating to give the customary brown crust. Food cooked in a microwave oven does not change much in appearance; in order, therefore, to ensure adequate cooking and to prevent overcooking it is desirable to provide the cooker with a built-in timer. This is mounted, together with various switches and pilot lamps, on a panel above the oven door.

The power pack is located under the oven. The position of the valves, transformers and other components can be seen in fig. 9. Further particulars of the circuit are given in *fig. 10.* A filter is inserted in the mains leads to suppress interference from transients that may be generated by the magnetron when it is switched on. When the main switch $S_1$ is turned on, current flows to the heaters of the magnetron and rectifier valves, and the fan $K$ for cooling the magnetron, the stirrer $R$ and the air extractor $L$ are started up. Block $H$ contains the heater-current transformer for the magnetron; the secondary is at high tension and is correspondingly insulated, and therefore the heater current is regulated at the primary side. Two minutes after $S_1$ has been switched on, the time relay $T$ closes the safety switch $V$, enabling the anode voltage for the magnetron to be switched on with $S_2$. The time the oven is to operate is preset with the timer $t$. In series with $S_2$ there are two other safety switches $Tk$ and $Lk$. The door switch $Tk$ cuts off the magnetron if the door is opened; if one were to put one's hand into the radiation field, damage could be done to internal tissues. Switch $Lk$ opens if the temperature of the magnetron is too high, indicating that the air or water cooling is not functioning properly.

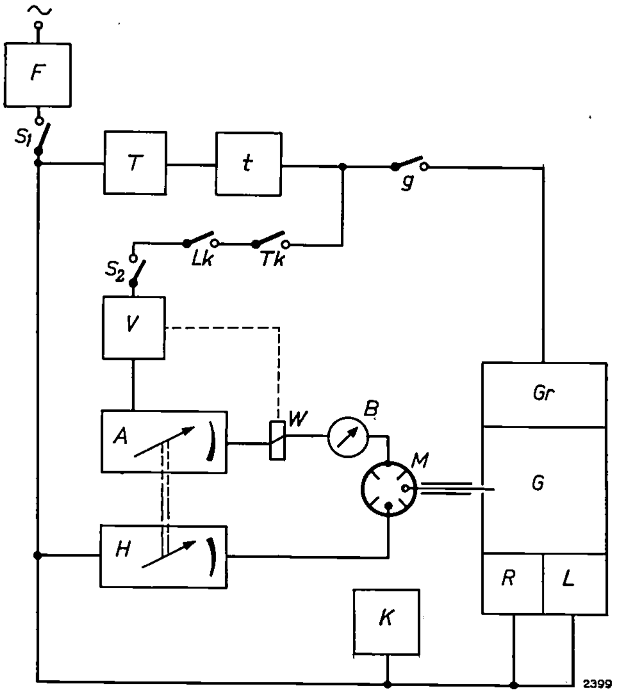


Fig. 10. Block diagram of a microwave cooker. $F$ suppressor filter. $S_1$ main switch. $T$ time relay. $t$ timer. $S_2$ switch for magnetron high tension. $Lk$ and $Tk$ safety switches, which can close only when the magnetron temperature is not too high and the oven door is closed. $V$ anode-voltage switch. $A$ power pack supplying anode voltage. $H$ filament-current transformer. $W$ overload safety device. $M$ magnetron. $B$ ammeter for anode current. $K$ cooling fan for magnetron. $G$ oven. $R$ field stirrer. $L$ air extractor. $Gr$ grill element. $g$ switch for $Gr$.

Block $A$ contains the anode voltage transformer, which is regulated at the primary side, and the rectifier valves. In connection with the Rieke diagram (fig. 4) it was mentioned that the magnetron may be fed with an unsmoothed, rectified voltage. The ratio between the peak and mean values of the anode current affects the permissible value of the standing-wave ratio $\sigma$, and this is taken into account in the design of the high-tension and rectifier section. The HT, after being switched on, should be turned up very gradually, otherwise the taste of delicate foodstuffs might be spoilt by sudden overheating. In blocks $A$ and $H$ (fig. 10) it is indicated symbolically that the heater voltage is gradually reduced as the anode current is turned up; if this were not so, back-bombardment would overheat the cathode. The meter $B$ in fig. 10 is an ammeter for the anode current; it provides a check on the power delivered by the magnetron and thus simplifies the operation of the cooker. For practical purposes all the dial need contain is a mark indicating the maximum permissible deflection of the needle. If the anode current becomes too high, e.g. as a result of mains fluctuations, the relay $W$ shuts off the high tension via the safety switch $V$.

The oven $G$ further contains a grill element $Gr$, which is operated by a separate switch $g$ but can also be brought into action by the timer.

The microwave cooker illustrated in fig. 9 is only an example. A continuous-feed oven might equally well be built, which could serve for heating or cooking large numbers of the same dishes in big kitchens or canteens. The 5 kW CW magnetron mentioned earlier (fig. 2) might be used with advantage here. A continuous oven lends itself well to automation, in which form microwave heating could also be used in industry, e.g. for drying wood, paper and textile products, for welding and sintering plastics and for numerous other purposes [8]). Compared with the traditional method of (dielectric) RF heating, the microwave method offers the advantage of a greater concentration of power in the workpiece. Compared with infra-red heating, microwaves have the advantage of great depth of penetration, giving effective heating from inside outwards.

**Summary.** After a review of the physical principles of dielectric heating in a microwave radiation field, a CW magnetron is described which is capable of delivering 2 kW continuous output at 2450 Mc/s. (Reference is also made to a recently-developed 5 kW CW magnetron.) The 2 kW magnetron is made in two

[8]) See e.g. W. Schmidt, Mikrowellengenerator zur dielektrischen Erwärmung und Trocknung nichtmetallischer Bahnen und Folien, Elektron. Rdsch. **13**, 359, 1959 (No. 10).

versions, air-cooled and water-cooled, and is especially suited for the heating of foodstuffs in a resonant cavity (microwave cookers). The magnetron has an impregnated dispenser cathode, delivers its power through a double coupling loop to a coaxial line, and is equipped with a ferroxdure magnet. The characteristics of the magnetron are discussed with reference to its Rieke diagram.

When used in a microwave cooker the magnetron is required to operate under a widely varying load impedance. The oven of the cooker must be so designed as to minimize load reflections, which give rise to standing waves in the magnetron output line. A uniform field distribution in the oven is achieved by means of reflector plates and a "field stirrer". It is important to provide for adequate damping in the oven, particularly when there is no food in it. Quarter-wave slots are used to prevent microwave energy escaping through the door joints. Finally the construction and circnitry of an experimental microwave cooker are discussed.

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

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, The Netherlands, where a limited number of reprints are available for distribution.

***2753:** J. Smit and H. P. J. Wijn: Ferrites: physical properties of ferromagnetic oxides in relation to their application (Philips Technical Library, 1959, XIV + 369 pp., 244 figures).

See the book notice printed below (p. 104).

***2754:** H. Bremmer: Mode expansion in the low-frequency range for propagation through a curved stratified atmosphere (J. Res. Nat. Bur. Standards **63 D**, 75-85, 1959, No. 1).

The expansion cited in the title is particularly useful when ionospheric propagation at low frequencies is considered. The complex problem dealing with two media, viz., a homogeneous earth and a surrounding stratified atmosphere, leads to intractable expressions. However, as the influence of the earth may be accounted for by an approximate boundary condition at the earth's surface, the problem is reduced to that of the outer medium only. The coefficients of the mode expansion for this simplified problem are derived while taking into account the earth's curvature; however, the latter proves to be negligible under very general conditions. The expansion derived is wanted in particular when the influence of a gradual transition in the electron density with height at the lower edge of the ionosphere is studied.

***2755:** W. J. Oosterkamp and J. Proper: The water equivalence of "Mix D" phantom material for soft X-rays (Brit. J. Radiol. **32**, 560, 1959, No. 380).

Correction to No. **2645**.

**2756:** H. F. Hameka: Calculation of the magnetic susceptibility of methane (Physica **25**, 626-630, 1959, No. 7).

The magnetic susceptibility of methane is calculated by employing molecular orbitals which are constructed from gauge invariant atomic orbitals. The result is $\chi = -13.7 \times 10^{-6}$; the agreement with the experimental value $\chi = -12.2 \times 10^{-6}$ is satisfactory.

**2757:** W. van Gool and A. P. D. M. Cleiren: Influence of hydrogen on the red ZnS-Cu fluorescence (J. Electrochem. Soc. **106**, 672-676, 1959, No. 8).

Self-coactivated ZnS-Cu phosphors were made by firing in different atmospheres. When $H_2S/H_2$ mixtures were used, the red fluorescence decreased with increasing amounts of hydrogen. With $Ar/S_2$ or with $N_2/S_2$ atmospheres no red fluorescence was obtained. These experimental results can be summarized by stating that, in order to obtain the red fluorescence, hydrogen must be incorporated into the phosphor and the sulphur pressure must be sufficiently high. The hydrogen either forms a part of the red center or it destroys or replaces a killer center that prevents the occurrence of red fluorescence when hydrogen is absent. In connection with the high sensitivity of the red fluorescence to small amounts of impurities it is suggested that the concentration of the red centers is much smaller than the amount of incorporated copper.

**2758:** H. Koelmans and C. M. C. Verhagen: The fluorescence of binary and ternary germanates of group II elements (J. Electrochem. Soc. **106**, 677-682, 1959, No. 8).

The fluorescence of binary and ternary germanates of Ca, Sr, Ba, Mg and Zn with different activators was investigated. Germanate phosphors activated with Pb, Ti and Mn are described. The ternary composition triangles are given together with the X-ray powder-diagrams of 23 hitherto unknown germanates.

**2759:** N. W. H. Addink: Determination of the transition probability (reciprocal of lifetime) of excited atoms and ions from spectroanalytical data and the importance of lifetime values in spectrochemistry (Spectrochim. Acta **15**, 349-359, 1959, No. 5).

The purpose of this investigation was to discover why some of the spectral lines used in the method of spectrum analysis by means of direct-current arc discharge do not meet the requirements of reproducibility. While studying the origin of these lines it was found necessary to calculate their transition probability .(reciprocal of lifetime), which proved that a relatively long lifetime is responsible for emission disturbances (collisions of the second kind).

**2760:** H. J. L. Trap and J. M. Stevels: Physical properties of invert glasses (Glastechn. Ber. **32 K,** VI/31-VI/52, 1959, No. 6).

The conventional silicate glasses are characterized by an irregular spatial Si-O network, in the interstices of which a number of network modifiers are situated. The physical properties of these glasses are mainly determined by the behaviour of this network. However, going to a rather low silica content, the coherence of these glasses and their physical properties can no longer be determined by the spatial Si-O network, since only rather short Si-O chains are present. The cations determine the behaviour of the glass (invert glasses). It is shown that in invert glasses certain properties (viz. those which are related to short-range phenomena, such as dielectric and mechanical losses, viscosity, coefficient of expansion) vary with composition in a direction opposite to that in conventional glasses. Properties reflecting an average overall situation (dielectric constant, refractive index) vary in the same direction as in conventional glasses. It is shown that the transition from conventional glasses to invert glasses takes place at compositions where the average number of non-bridging oxygen ions per $SiO_4$ tetrahedron is two. Consequently this criterion may be used to determine which fraction of "intermediates" is present in the form of network modifiers and network formers.

**2761:** J. M. Stevels: Netzwerkfehler in kristallinischem und glasigem $SiO_2$ (Glastechn. Ber. **32,** 307-313, 1959, No. 8). (Network imperfections in crystalline and vitreous $SiO_2$; in German.)

Crystalline and vitreous $SiO_2$ almost always contain network imperfections. Their concentration, however, is often so small that they are undetectable by the methods of chemical analysis. There are also network imperfections which by their very nature cannot be detected in this way. Modern physical methods are now available (dielectric loss measurement at low temperatures, paramagnetic resonance measurements and optical absorption methods) which can give not only some estimate of the concentration but also some idea of the kinds of imperfection present. Comparison of the network imperfections before and after irradiation with electromagnetic waves (short-wave U.V., X-rays or $\gamma$-rays) or neutrons, can lead to some idea of the "reactions" brought about by such radiations in crystalline and vitreous silica.

**2762:** C. J. Bouwkamp: Interaction of two crossed cylinders in the presence of Van der Waals forces (Nieuw Arch. Wisk. **7,** 66-69, 1959, No. 2).

Calculation of the Van der Waals interaction energy of two crossed infinite circular cylinders in terms of their radii and separation.

**2763:** H. C. Hamaker: A note on ANOVA in the transistor industry (Industr. Qual. Control **16,** 12-14, 1959, No. 1).

Discussion between the author and A. W. Wortham on an application of variance analysis to certain problems in transistor applications. The discussion centers around the problem of how to deal with one apparently discrepant observation, a so-called outlier.

**2764:** A. H. Boerdijk: Contribution to a general theory of thermocouples (J. appl. Phys. **30,** 1080-1083, 1959, No. 7).

Application of thermodynamics of irreversible processes to a thermocouple of which (*a*) the bars have an arbitrary shape, (*b*) the properties of the materials are arbitrary functions of temperature, and (*c*) the composition is, under certain restrictions, inhomogeneous and anisotropic, leads through introduction of a single place coordinate to two nonlinear differential equations describing the stationary distribution of temperature and electrical potential. Output powers and efficiencies are expressed in terms of the temperature gradients in the bars. The maximal values of the efficiencies obtained by variation of the shape of the bars are independent of the shape. Upper bounds of the efficiencies attainable by stationary thermoelectric conversion are derived. If the shape of the bars is restricted to general cylinders and truncated wedges or cones, the transient behaviour is described by two partial differential equations which contain two independent variables only. A periodic ripple in the electrical current has the same effect as a decrease of the electrical conductivities of the materials.

**2765:** J. S. C. Wessels: Studies on photosynthetic phosphorylation, III. Relation between photosynthetic phosphorylation and reduction of triphosphopyridine nucleotide by chloroplasts (Biochim. biophys. Acta **35**, 53-64, 1959, No. 1).

The photochemical reduction of TPN by isolated chloroplasts was investigated. A comparison of the rate of TPN reduction with that of photosynthetic phosphorylation provided evidence that the generation of ATP in the presence of vitamin $K_3$ or FMN is not coupled with the reoxidation of TPNH by the oxidized product of the photolysis of water. Photosynthetic phosphorylation could proceed unimpaired under conditions in which the chloroplasts had lost their ability to reduce TPN. On the other hand TPN reduction could be considerably stimulated by a chloroplast extract which did not affect photosynthetic phosphorylation. These results are discussed in relation to the recent finding that the reduction of TPN by chloroplasts is accompanied by ATP formation.

**2766:** M. J. Sparnaay: The interaction between two cylinder-shaped colloidal particles (Rec. Trav. chim. Pays-Bas **78**, 680-709, 1959, No. 8).

The theory of the stability of lyophobic colloids, as given by Derjaguin and by Verwey and Overbeek for the flat-plate model or the sphere model of colloidal particles, is applied to cylinder-shaped colloidal particles. Anisometry was taken into account by considering the interaction between two parallel particles and between two particles in a crossed position. Mathematical expressions are given for the repulsive and the attractive potential energy in these two cases. It can be inferred from these expressions that the behaviour of two parallel cylinders is intermediate between the behaviour of two flat plates and two spheres, whereas two crossed cylinders behave in much the same way as two spheres.

## Now available

J. Smit and H. P. J. Wijn: Ferrites: physical properties of ferromagnetic oxides in relation to their application (Philips Technical Library, 1959, XIV + 369 pp., 244 figures).

In recent years the most important developments in magnetic materials have been in the field of magnetic oxides. This book gives the reader an introduction to ferrites, that is to say, the magnetic oxides containing iron as their main component. In the many theoretical problems treated, the authors make use of simple physical models rather than rigorous mathematical methods. In view of the large and growing number of applications of ferrites in electronics and electrical engineering, this book is indispensable to those working in these fields. In addition there is much of direct interest to those concerned with metallurgy and inorganic chemistry.

The book is divided into four parts, with chapters as follows:

*Part A.* Theory: I. On the properties and the origin of magnetic fields in matter. II. Theory of ferromagnetism. III. Ferrimagnetism. IV. Magnetic anisotropies. V. Magnetization processes. VI. Dynamics of magnetization processes.

*Part B.* Measurements: VII. Methods of measuring ferromagnetic properties.

*Part C.* Intrinsic properties: VIII. Intrinsic properties of ferrites with spinel structure. IX. Intrinsic properties of ferrites with hexagonal crystal structure. X. Intrinsic properties of ferrites with garnet structure.

*Part D.* Polycrystalline ferrites: XI. Structure of polycrystalline ferrites. XII. Electrical properties. XIII. Static initial permeability. XIV. Frequency-dependence of the initial permeability. XV. Static hysteresis loops. XVI. Dynamic properties at high field strengths.

The book concludes with a list of references to the literature and an index.

French and German editions are in course of preparation.

VOLUME 22, 1960/61, No. 4 pp. 105-140 Published 16th February 1961

Philips Technical Review

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

# APPLIED STATISTICS

## AN IMPORTANT PHASE IN THE DEVELOPMENT OF EXPERIMENTAL SCIENCE

by H. C. HAMAKER.

31:519.2

*On 20th May 1960 Dr. H. C. Hamaker presented an address under the above title to mark his inauguration as extra-mural professor at the Technische Hogeschool, Eindhoven. As we have frequently done on similar occasions in the past, we print below the text of his discourse in extenso *). The author has introduced some changes here and there to suit a wider audience than that which attended the inaugural ceremony. Some years ago a series of articles on sampling systems from the pen of Prof. Hamaker were published in this journal **). These articles were largely confined to lot inspection. In recent years statistical methods have been more and more applied to research in general: the author traces here the broad lines of this development. A short bibliography is appended of articles on statistics contributed by the author in recent years to professional journals.*

Fifteen or so years ago the statistician in industry was still virtually unknown in this country and in many others. The notion that statistics might be usefully applied to technological problems was more often than not, as I myself discovered, dismissed with a shrug of the shoulders.

Since then, however, the situation has radically changed. The Philips Research Laboratories at Eindhoven now employ a group of eight or ten professional statisticians who are engaged solely on research in this field and on the application of statistical ideas and methods to industrial problems. Moreover, scattered throughout the company there are smaller groups of statisticians working on problems specific to a particular division and its products. Similar developments have taken place in other industries at home and abroad, and the market has not yet by any means reached saturation point. There is a constant demand for experienced statisticians, and for training and instruction at various levels that will enable industrial personnel to solve simple problems themselves and to judge when the services of a professional statistician can best be called upon. This is not a passing vogue. On the contrary, I believe that we are in the middle of an important phase in the evolution of scientific thought, which will have a lasting influence on almost every branch of science. I should like first of all to dwell on this aspect for a moment.

Technology, the arbiter and pacemaker of our modern society, is a product of the rapid advances made in the natural sciences. These in their turn sprang from man's innate urge to understand and explain the phenomena he perceives in the world around him. The first steps in this direction are to make objective observations, and to record these observations in terms of numerical quantities which can be compared with other similar quantities and used to perform mathematical operations.

Experience has shown, however, that numerical observations are hardly ever exactly reproducible; when an experiment is repeated the results invariably show a certain spread. Accurate results are obtained only when the observations are made under carefully chosen, controlled conditions. The scientist would therefore retire behind the walls of special laboratories where such conditions could be established. The utmost care was devoted to the construction of the equipment employed, and the objects to be studied were carefully selected; only those were admitted that appeared to allow accurate and reproducible observations. Although no observation is ever perfectly reproducible, no matter how elabo-

*) Published by Centrex, Eindhoven, 1960.

**) H. C. Hamaker, Lot inspection by sampling; The theory of sampling inspection plans; The practical application of sampling inspection plans and tables (with J. J. M. Taudin Chabot and F. G. Willemze), Philips tech. Rev. 11, 176-182, 260-270, 362-370, 1949/50.

rate the precautions taken, the accuracy achieved was such that the results could be interpreted without taking their spread into account. To facilitate the interpretation of the results the problems chosen were simple, involving only a limited number of factors. "Exact", reproducible observation, with only one factor varied at a time, became the ideal that typified the classic exact sciences. Hence the qualification *exact*.

It is in the nature of man to tend to overestimate the significance of methods and ways of thinking that have been successful, and to attribute to them a general validity beyond the confines of the field for which they were devised. The success of the exact sciences accordingly led to the belief that the path followed was the only true one, and those who followed it were inclined to look down on all investigations that had to be content with less accurate observations.

In this they were wrong, for there are many fields of research where exact observation is essentially impossible; for example, the branches of science concerned with living beings. We can, it is true, determine with some accuracy the metabolism of a single plant or animal, but the value of such observations depends on the extent to which the results may be interpreted as valid for the species as a whole. And such interpretations must inevitably take account of the variations from one individual to another, which are always considerable.

Again, it is often wrong to limit experiments to the variation of only one factor. Agricultural experiments are a case in point. Suppose we want to find out how the yield of different varieties of potatoes depends on the nature and quantity of manure. We must not investigate separately the influence of manure containing potassium and of manure containing nitrogen, because their combined effect cannot be predicted from their separate effects, i.e. there is an "interaction" between these variables. Nor can the experiment be divided into parts to be done in different years, for changeable weather conditions will make it impossible to compare the results. The only efficient experimental method is to deal with all the varieties and manures in one season. The elements of this experiment should be distributed over a limited area of land in such a way as to minimize unavoidable variations in the fertility of the soil. Even then, fairly important random fluctuations will continue to influence the results, and these must be taken into account by an appropriate mathematical method of interpretation.

Between 1920 and 1930, the well-known English statistician Sir Ronald Fisher, at the Agricultural Experimental Station at Rothamstead, was the first to recognize that the principles of agricultural experimentation should be entirely different from those of the "exact" experiments in the laboratory. In his book "The design of experiments" he showed the way which is still being pursued [1]).

According to these new principles there is nothing against varying several factors simultaneously in a single experiment, provided the layout is so designed that the effect of each factor can be separately established by an appropriate analysis of the observations. A further merit of such designs is that the extent to which the various factors influence one another can also be determined. This is of course impossible if only one factor is varied.

The random fluctuations impart an element of uncertainty to any conclusion drawn from the observations. This implies accepting that there is always a certain risk of drawing a wrong conclusion. This is no objection provided the design of the experiment permits an analysis of the magnitude of that risk: we are then in a position to keep it within reasonable bounds.

In these experimental designs statistical techniques of analysis and interpretation play a most important role. They are derived from the statistical principles evolved at the turn of the century by Galton and Pearson, which are based mainly on the theory of probability [2]). This explains why the new ideas have been developed by statisticians and are nowadays principally described in text books on statistics.

In my view, however, they should really be seen as a fundamental change in the principles of experimental science. The classic ideals of exactitude, and of varying only one factor at a time, forced the experimenter into a straitjacket, from which he was freed by the work of Fisher. Phenomena formerly excluded from the preserves of exact science, for being inaccessible to exact and reproducible observation, can now be subjected to a mathematically well-founded — and thus in a certain sense "exact" — treatment. This has considerably widened and deepened the field of applied science. No wonder, then, that statistics nowadays plays an active part in almost every branch of knowledge.

Although in the application of statistics the line of thought remains consistent, the methods adopted

[1]) R. A. Fisher, The design of experiments, Oliver and Boyd, Edinburgh 1935.

[2]) See e.g. K. Pearson, The life, letters and labours of Francis Galton, Cambridge Univ. Press 1914 (Part I), 1924 (Part II), 1930 (Part IIIa, b).
E. S. Pearson, Karl Pearson, an appreciation of some aspects of his life and work, Cambridge Univ. Press 1938.

are many and various and must always be adapted to the given problems and circumstances. To take the case of *agriculture* again, only one experiment a year can be carried out; elaborate and complicated experimental designs are. therefore justified. The experiment itself is done in spring and summer, the observations are harvested in the autumn, and the winter is spent in analysing them and in designing a fresh experiment.

In the *electrical engineering industry*, on the other hand, we are concerned with machines capable of turning out a thousand and more products an hour. There is no need here to skimp the number of observations, but the experiments must be kept as simple as possible to avoid organizational difficulties and laborious analysis. At such a rate of production the engineer is in a hurry, and he would really have preferred the answer yesterday to the experiments being done today.

Between these extremes we have the *chemical industry*, which may process one charge a day. Here the investigator must be sparing in his observations, because they are costly and because a comprehensive investigation would take too long.

In *medical research* we may want to compare two medicines: neither the patient nor his doctor must know which one has been administered, to preclude the possibility that their judgement of the results is coloured by preconceived opinions. Nor is it possible to carry out the experiment as and when it suits the investigator; its rate of progress is determined by the availability of patients. This has the advantage, however, that the results can be gradually accumulated and the experiment stopped as soon as a clear conclusion has been reached.

Again, in *psychological experiments* we are not primarily concerned with studying the effects of certain factors. The problems involved are entirely different. There are, for example, many possible methods of estimating human intelligence, but intelligence cannot be *measured*; it is only by comparing the various methods one with the other that we can try to discover in how far they answer their purpose.

The *sociologist* has to rely on surveys for his observations; statistical methods are important here for processing and interpreting the results, and also for devising an efficient survey technique.

I could go on like this for some time, but I shall turn now from the general to the particular, in order to illustrate some of the ideas and methods which form the statistician's stock-in-trade. I shall take the example that first springs to mind on this occasion, namely the address I am now delivering and inaugural addresses in general.

Let us imagine we want to know, for one reason or another, the average length of the words used in this discourse, length meaning the number of letters per word. How can we best find it?

My lecture consists of a collection of 6000 to 7000 words, and no-one will expect me to have measured them all. It is evidently possible to determine the average word-length with a fair accuracy by counting a limited number of words. We start by counting 200 words, picked at random from the text; in the terminology of statistics we take a "sample" of 200 words. The result is set forth in *fig. 1* in a so-called histogram. Of the 200 words, 34 were found to contain 2 letters, 40 had 3 letters, 34 had 4, and so on; the longest word, with 18 letters, was observed three times *). These 200 observations yield an average length

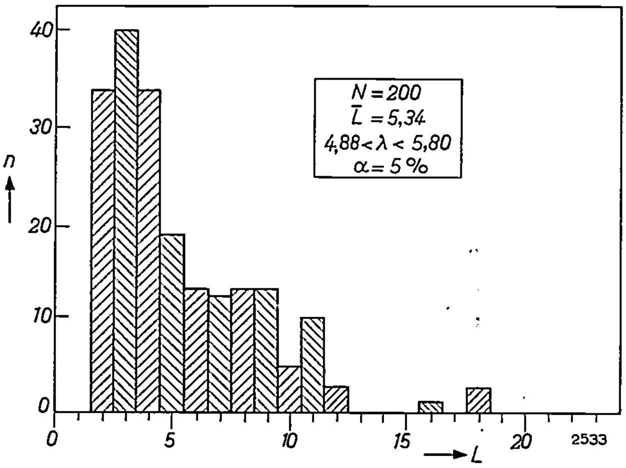


Fig. 1. Histogram of counted word-lengths in a random sample (drawn by lot) of 200 words. $n$ = number of times the word-length $L$ (= number of letters per word) is found.

$$\bar{L} = 5.34 \text{ letters per word.}$$

Of course, this value does not agree exactly with the "true" average length $\lambda$ which we should find if we counted all the words. Nevertheless, from the dispersion of the results, as appears in the figure, we can establish limits between which this unknown parameter ($\lambda$) can be said to lie with a given degree of confidence. This yields:

$$4.88 < \lambda < 5.80 \text{ letters per word}$$

with a *confidence* of 95%.

If I assert that $\lambda$ lies between these limits, I run a 5% risk of telling an untruth. The limits form a so-called *confidence interval*. For the confidence level we can choose any value we like; a higher level of, say, 99% will lead to a wider interval. The width of

*) The analysis refers to the original Dutch version of this lecture.

the interval is inversely proportional to the root of the number of observations, so that we must make four times as many observations if we wish to halve the width of the interval; i.e. 800 instead of 200.

Two important questions now arise. Firstly, how are we to pick out the words that will constitute the sample? And secondly, what degree of accuracy, or what width of confidence interval, should we aim at; in other words, how big must the sample be?

To the first question the statistician has a ready answer. The sample must be a random one; it may be obtained by numbering the words of this speech consecutively from beginning to end and determining the serial numbers of the words that are to make up the sample by a lottery.

There is nothing difficult about that, for every applied statistician worth his salt carries a lottery about with him. A simple example of such a device consists of a glass bulb with a hollow stem and filled with ten wooden beads, nine of them green and one red. When the beads are shaken from the bulb into the stem, the position of the red bead indicates a number between 0 and 9. By repeating this operation many times we obtain a set of *random numbers*, that is numbers from 0 to 9 in completely random order. Groups of four random numbers together then indicate the serial numbers of the words to be included in the sample; numbers higher than the total of words in the text are ignored.

Plainly, this method of sampling is extremely cumbersome, and in the present case unnecessary. In any language, long and short words succeed one another with a fair degree of regularity. Perhaps, in an inaugural address of this nature, one might expect to find rather more long words in the middle, when the speaker has warmed up to his specific technical jargon, than at the beginning or end, but this can have no pronounced effect. If, to be on the safe side, we choose one or two lines at random on each page and count the words on those lines, we shall certainly get a satisfactory estimate of the average word length. That in fact was how fig. 1 was constructed.

Another method would be to stick a pin into the text at random places, and afterwards to take the words thus pierced. That would lead to a biased result, however, for the long words have a greater chance of being pierced than the short ones. *Fig. 2*, which was produced in this way, shows this very clearly. The average length is now 7.67 letters per word, quite a lot more than the value of fig. 1. Even so, we can still arrive along these lines at a correct result. The chance of piercing a word is clearly proportional to the length of that word, and a statistical theory, making use of this datum, tells us that we must now compute the mean of the *reciprocal* word-length. The value found is 0.188 words per letter, and by taking the reciprocal of this mean we obtain a correct estimate for the actual average word-length:

$$\frac{1}{0.188} = 5.31 \text{ letters per word,}$$

which agrees very nicely with the value found from fig. 1.

This method finds practical application in traffic surveys for ascertaining the average distance travelled by motorists per journey. The system is to stop an arbitrary group of cars on the road and to ask the drivers where they come from and where they are going. The chance of a motorist being included in the sample is proportional to the length of his journey, and here again the above method of calculation must be applied.

Finally, there is a third method by which we can find the average word-length: we determine the average number of letters and the average number of words per full line of print, and then divide the first average by the second. The number of letters per line is found to be surprisingly constant, so that a few counts are sufficient to achieve reasonable accuracy. It is a simple matter to count the number of words in a line, and here an accurate result can be arrived at by making a large number of counts.

The three methods discussed are compared in *Table I*. Column B gives the average word-length, column C the 95% confidence interval, column D the width of this interval, and column E the total number of counts. The first two methods have roughly the same width of interval, but since the average word-length in the second method is greater,

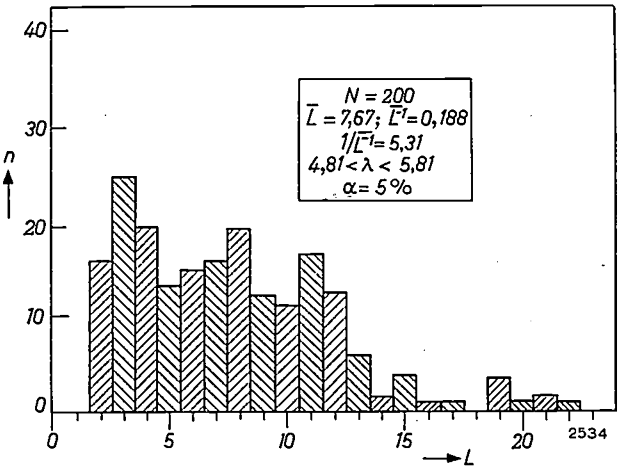


Fig. 2. Histogram of the lengths of 200 words, found by pricking the text blindly with a pin.

Table I. Comparison of the average word-lengths, $\overline{L}$, determined by three different methods. $\lambda$ is the true average length.

| Method | $\overline{L}$ letters/word | Confidence interval for confidence level of 95% | Width of confidence interval | Total number of counts |
|---|---|---|---|---|
| I. Sampling by random numbers | 5.34 | $4.88 < \lambda < 5.80$ | 0.92 | $200 \times 5.3 = 1060$ |
| II. Sampling by pinpricks | 5.31 | $4.81 < \lambda < 5.81$ | 1.00 | $200 \times 7.7 = 1540$ |
| III. Average number of letters per line divided by average number of words per line | 5.48 | $5.22 < \lambda < 5.74$ | 0.52 | $(7 \times 57.3) + (60 \times 10.5) = 1031$ |
| A | B | C | D | E |

many more counts were needed. Method II is thus decidedly less efficient than method I. In the third method the number of letters was counted in 7 lines, and the number of words in 60 lines, amounting altogether to 1031 counts, about the same as in method I. The resulting confidence interval, however, is only half as wide; method III therefore provides by far the most accurate result for the same number of counts. In this way, then, a statistical analysis enables us to decide on the method of investigation to be preferred.

The second question, concerning the accuracy to be aimed at, is countered by the statistician with the further question: For what purpose do you want to use the result? If it is simply a matter, as it is here, of collecting demonstration material for a lecture, no particular demands need be made on the accuracy, and the data given above are more than ample. But if we want to use the average word-length as a personal characteristic of the author, and to compare the discourses of various authors on that point, the position is entirely different.

From counts made (by the third method) on various published inaugural lectures I have not been able to discover any differences. That is not surprising, for the texts were all in Dutch, and differences, if any, would certainly be very slight. Much larger samples would be needed to find them.

The average word-length, then, is unsuitable as a characteristic of the literary efforts of different professors. In the average *sentence-length*, i.e. the average number of words per sentence, we may expect more variation. This characteristic may be found by a method similar to method III, that is by taking the product of the total number of printed lines and the average number of words per line, and dividing it by the total number of sentences. We can skip the details. An investigation of seven texts showed the overall average sentence-length to be about 25 words, with individual averages varying from 22 to 32 words. Here the differences are quite marked.

One might try to pursue this line of inquiry even further. I fear, however, that the reader will lose patience at this point, feeling that a study of the length of words and sentences can only be of academic interest. To bring out more plainly the practical significance of the procedures described, I shall therefore apply a transformation to our problem. I now change professors into machines. The words uttered by professors shall be products brought forth by these machines, and the length of the words shall be a dimension characterizing the quality of these products, for example a diameter.

After this drastic transformation we still find ourselves faced with the same problems. Although the dimensions of the products are not as variable as the lengths of words, they are not all exactly the same; some variability is always present in mass production. Further, the number of products in many cases is so large that we cannot feasibly test them all, and we must again be content with a sample.

There are, however, marked differences between the machine problem and the professor problem. Whereas the average word-length was not subject to any special requirements, a nominal value is usually specified for the average diameter; deviations from that value tell us whether a machine is properly adjusted. Again, the words in a piece of prose are neatly arrayed side by side, and long and short words succeed one another with a certain regularity. But the products turned out by a machine are jumbled together in a heap, and so we can no longer apply method III to find the average diameter. Moreover, whilst no or only minor disparities were found in the average length of the words produced by different professors, and no variation is to be expected in the course of time, the average diameters of the products from the several machines certainly will differ, and these values are also subject to variations over a short period as a result of wear. A single, elaborate investigation as represented in figs. 1 and 2 is out of place where machines are concerned; the appropriate method here is to take small samples at regular intervals, in order that prompt action can be taken if misalignment becomes too serious.

Nor do machines run on indefinitely. From time to time a breakdown occurs, or operations have to be stopped to make adjustments or feed in fresh material. We may thus imagine the lengths of the sentences to be transformed into time intervals during which the machines operate without interruption. Now, in any discourse short sentences generally improve the clarity and long sentences are better avoided, whereas long operating periods are just what we want from a machine; machines with a propensity for unduly short periods are consigned to the workshop for overhaul. Here again we have an evident difference.

Finally, we can go a step further and ask in how far discrepancies in the quality of the products, turned out by different machines, are attributable to the machines themselves, to the workers operating the machines, or to the batches of raw material processed. An answer to this and similar questions can best be supplied by an experiment, along the lines, for example, of a "Latin square", as represented in *fig. 3*. We choose, say, 4 machines, 4 operators and 4 batches of raw material. Each batch is divided into 4 portions and the experiment is arranged so that each batch is processed once on each of the 4 machines and once by each of the 4 operators, while each operator also works once on each machine. With an experimental design of this type, the influence of the machine, the operators and the raw material can be separately determined by an appropriate analysis of the data — a typical instance of the way in which the statistician interweaves several factors in a single experiment. With the Latin-square design the information required is obtained with a minimum number of observations.

|  | A | B | C | D |
|---|---|---|---|---|
| I | 2 | 4 | 1 | 3 |
| II | 3 | 1 | 2 | 4 |
| III | 4 | 2 | 3 | 1 |
| IV | 1 | 3 | 4 | 2 |

2535

Fig. 3. Statistical design for an experiment, a "Latin square", used for investigating the influence of machines, labour and raw materials on the quality of a product.
*A*, *B*, *C*, *D* machines.
*I*, *II*, *III*, *IV* operators.
*1*, *2*, *3*, *4* batches of material.

The reader will understand that experiments of this kind are not appropriate to the study of the lengths of words or sentences. One might ask, for example, in how far the average sentence-length, which — as we have seen — is characteristic of the individual professor, is also dependent on his subject matter. But an experiment in which a number of professors were each persuaded to write a number of inaugural addresses on different subjects, is out of tbe question.

The radical transformation I have applied thus demonstrates the point emphasized earlier: the statistical nature of the problems is not affected by such a transformation but the questions to which we seek an answer, and the conditions under which we must work, may be substantially altered. Each case, then, demands an individual approach to arrive at a satisfactory solution.

Finally, I should like to touch briefly on the question in what respect applied statistics differs from theoretical or mathematical statistics. There is reason for doing so in that the practitioners of these two branches of statistical science tend very much to keep their distance. A journal of such broad scope as the American "Annals of Mathematical Statistics", for example, prints many articles that are beyond the grasp of most applied statisticians, and give them the feeling that this is a special kind of mathematics for the sake of mathematics, but without any connection with the practical problems of daily life.

In our earlier count of 200 words we established a confidence interval for the average word-length $\lambda$ with a confidence level of 95%. The question turns now on what degree of accuracy may be attached to this quantity. The applied statistician is interested only in the order of magnitude. For him a confidence level of 95% means that the risk of reaching a false conclusion is of the order of 5%; but it is of no serious consequence if the real risk is not 5% but 2% or 10%.

The mathematician, on the other hand, would require the risk to be exactly 5%, or, if that is mathematically unattainable, he requires 5% to be an upper limit to the risk; on no account should the risk be greater than 5%.

The science of statistics owes a great deal to this striving after mathematical rigour. It has led, for instance, to the development of the elegant and often very simple "distribution-free" methods, to which Van Dantzig and his pupils at Amsterdam have contributed so much [3]).

But an exaggerated insistence on rigour has its practical disadvantages. It does not always result in simplifications; on the contrary, it produces a multiplicity of methods and formulae which may hamper rather than assist the effective application of statistical techniques.

What, then, it may be asked, is the real value of statistical analysis of observations? It does not, as I see it, lie in the exactness of the risks attached to conclusions. Statistical techniques provide us with internationally accepted standards for dealing with questions that arise in all research work. It is the international acceptance of uniform procedures that determines the real value of statistical methods. It leads to the economical formulation of the results and interpretation of scientific investigations. Undoubtedly international acceptance is more likely when the risks in a statistical procedure are clearly established. But simplicity and uniformity of methods and formulae are equally essential. If the pursuit of scrupulous accuracy produces a multitude of methods and complicated formulae, uniformity goes begging and it may therefore be preferable to relax the demands on exactness.

The difference in viewpoint between the theoretician and the practitioner has far-reaching effects. If we want to analyse the separate influences of operator, machines and raw materials on the quality of the products, using a Latin square as in fig. 3, the assumption underlying the experiment is that these three factors operate *independently*. We assume, for example, that the differences between the operators are always the same for any given combination of materials and machines. But if each operator is specially accustomed to his own machine, we are not entitled to interchange men and machines freely, and the Latin square is no longer the proper design for the experiment.

Every statistical experiment is thus based on a *model*, that is to say a mathematical formula which indicates in what way the factors that are varied influence the result. The strict theoretician wants this model to be completely established before the experiment is embarked upon, and he will not allow any modification of the model on the grounds of the experimental results; for any such subsequent alteration would mean that the risks inherent in the statistical conclusion could no longer be exactly evaluated and would certainly be increased. This the mathematician will not allow.

The implication, however, would be that statistical methods were only applicable in experiments whose outcome was a foregone conclusion and whose purpose was merely to establish numerically the parameters occurring in the model. In my view the experimenter can never accept such a standpoint. The object of many experiments is exploratory: it is often those experiments that yield unexpected results that are the most valuable. If the theoretical statistician washes his hands of such experiments as being beyond the scope of a statistical analysis, the experimenter for his part will have little interest in statistics.

Another way of trying to circumvent the choice of confidence limits is to interpret the problem "econometrically". It is then assumed that a decision has always to be taken at the end of an investigation; for example, a sampling inspection to determine the quality of a batch of products ends in a decision to accept the batch or to reject it. Considered statistically, there is always some risk of taking a wrong decision, that is of rejecting a good batch or of accepting a bad one. We can reduce these risks, however, by taking a larger sample. Now if we know the economic damage caused by wrong decisions, we can weigh the risk of damage against the costs of the inspection, and settle the size of the sample so as to minimize the average total costs.

But this approach leads to endless complications. The costs which a wrong decision may entail are difficult to determine and are subject to a high degree of uncertainty. They depend, for instance, very markedly on available stocks if the objects concerned are components intended for further production. Where stocks are plentiful, we need have no scruples about rejecting bad or doubtful lots, but if stocks are nearly exhausted, the rejection of a lot may soon force a manufacturer to stop production for lack of material. In a short time, then, the damage due to a wrong decision may vary over a very wide range.

When a supplier subjects a batch of finished products to sampling inspection, he does so partly to ensure that his good name will not suffer from the delivery of an inferior consignment. But the economic value of a good name and the harm done to it can be little more than a conjecture.

[3]) See e.g. the bibliography in Statistica neerlandica **13**, 432, 1959 (No. 4).
"Distribution-free" methods (also termed, less appropriately, "non-parametric" methods) are those which are not based on the assumption that the error distribution is necessarily normal or gaussian.

Furthermore, we must know what happens to rejected batches of products. They may, for example, be returned to the supplier, or subjected to a more thorough inspection, or sold at a reduced price, or written off as a total loss. It also makes a difference whether we consider the costs from the standpoint of the customer or of the producer, or of both together.

Here again, the conflict between theory and practice is clearly apparent. If 5 out of a sample of 100 products, i.e. 5%, are found defective, we can assert with a confidence of 95% that the lot from which the sample was taken will contain at least 1.9% and at the most 10.7% defective products. This is a purely statistical assertion, which is arrived at with the aid of an elementary table and leaves us free to adopt whatever course of conduct we may think reasonable. However, as soon as we attempt to define that reasonable course of conduct more exactly on a mathematical basis, we are up against not one but a multiplicity of situations, and we soon find ourselves in a maze of theories through which only an experienced theoretician can find his way. These theories may be successfully applied in specific cases, but they lack the general usefulness and wide compass of the confidence interval.

Between these rocks the applied statistician must learn to navigate. His starting point should be that the function of statistics in experimental science is that of a servant. The experimental scientist is guided by the stepping-stones of experience and intuition. That, in my opinion, is essentially characteristic of all research, and we have to accept the fact that experience and intuition are vague concepts that cannot be pinned down in mathematical formulae and numerical constants.

If the precise formulation of statistical theories is pursued too far, one is driven to express the vague concepts of experience and intuition in terms of exact numerical parameters. Doubtless the theories thus produced may be interesting, but I fear they will generally prove sterile in their application.

The applied statistician, then, is bound to accept the experience and intuition of the experimenter as his working basis, though at the same time preserving a critical attitude. For experience and prejudice are often intermingled and difficult to disentangle; and where prejudice is suspected, verification is required.

When a statistician is consulted on the solution of a practical problem, his first job is to find out, from a discussion with his principal, what may be regarded as well-founded experience and where preconceived opinions may have crept in. When it comes to interpreting the observational data, constant cooperation is called for between statistician and experimenter. If the latter is led by the observations to modify his views on the nature of the phenomena investigated, the statistician must adapt his model accordingly. But he should be aware of the dangers of such a procedure, and warn against it where necessary. Practised in this way, applied statistics constitutes a fascinating and valuable chapter in the development of experimental science.

## BIBLIOGRAPHY

Some publications by the author on the subject of applied statistics:

Beispiele zur Anwendung statistischer Untersuchungsmethoden in der Industrie, Mitt. Bl. math. Stat. **5**, 211-229, 1953.

De betekenis van de statistiek voor de ontwikkeling van de experimentele wetenschap, Sigma **1**, 55-58, 1955.

Naar efficiënte experimenten, Statistica neerl. **9**, 7-25, 1955.

Experimental design in industry, Biometrics **11**, 257-286, 1955.

Infusing statistics into industry, Bull. Inst. Intern. Stat. **35**, 433-443, 1957.

Statistiek en experiment, Statistica neerl. **12**, 119-130, 1958.

De recente ontwikkeling in proefopzetten met kwantitatieve factoren, Statistica neerl. **12**, 201-212, 1958.

**Summary.** In many fields of inquiry, especially those concerned with living beings, "exact" observations are not possible and it is necessary to investigate the effect of several factors at the same time. This has led to the design of experiments on a statistical basis, in which several factors may be varied simultaneously and which require an appropriatc statistical analysis for their interpretation (R. A. Fisher). In this his inaugural lecture as professor at the Eindhoven Technische Hogeschool, thc author advances the view that the introduction of statistical principles is bringing about a fundamental change in experimental science in general. Though the statistical principles are always the same: the methods adopted depend on the nature of the investigation. In agriculture, medicine, sociology, industry, etc., quite different techniques may be needed. To illustrate the statistician's approach, the author inquires into the average length of the words and sentences in his own discourse and in similar discourses by others. Three methods for determining the average word-length are discussed and compared. Next, the same principles are applied to problems of quality control of machine-made products. Though the problems are similar, the circumstances are different, particularly in regard to conditions (e.g. variation in time), aims and experimcntation possibilities. Finally, the author considers the marked differences between applied and theoretical statistics. The true value of statistical methods is that they provide universally accepted standards for the conduction and interpretation of investigations. It is of great value that with statistical procedures the inevitable risk of drawing a wrong conclusion is known. For the applied statistician, however, only the order of magnitude of this risk is of interest, while the mathematician requires that the risk shall be known exactly. Experimental science proceeds on the basis of experience and intuition, vague concepts that cannot be brought under the rules of precise mathematical formulae and constants. The applied statistician should take this fact as his point of departure and should not attempt to achieve a precision which does not correspond to the actual conditions of experimentation. Applied in this way statistics will prove to be a valuable aid to the experimenter.

# WAVEGUIDE EQUIPMENT FOR 2 mm MICROWAVES

## I. COMPONENTS

by C. W. van ES, M. GEVERS and F. C. de RONDE.

621.372.8

*Microwaves of millimetre wavelengths are an important tool in plasma research and in research on the solid state. Special equipment is therefore needed for the measurement of frequency, phase, power, absorption and reflection coefficient in the microwave region. The article below is part I of a survey of equipment developed at Philips for 2 mm waves. The equipment consists partly of scaled-down versions of equipment for longer waves, and partly of entirely new designs. Part II, to be published later, will describe various set-ups for microwave measurements at wavelengths of 2 mm.*

Microwave techniques are nowadays the subject of considerable interest, for general scientific reasons as well as in connection with the development of technical applications. The main branches of science in which microwave techniques are of importance are plasma physics, microwave gas spectroscopy and solid-state research.

In plasma physics, interest is focused on the electron density and temperature of plasmas. The electron density may be derived from the propagation characteristics of millimetre waves which are passed through the plasma. One way of determining the temperature is to compare the energy radiated by the plasma in a certain millimetre wave region with that of a radiation source of known intensity (noise generator) in the same waveband. Such measurements are of particular importance in experiments relating to nuclear fusion [1]).

Microwave gas spectroscopy is concerned in particular with the study of molecular rotational energy. In principle a microwave gas spectrometer consists of a microwave generator, a vacuum-tight waveguide filled with the gas under investigation, and a detector. The equipment measures the absorption occurring at specific frequencies [2]).

A third application is in the field of solid-state research. When an electromagnetic wave is directed upon a solid, such as germanium or silicon, in which a magnetostatic field is present perpendicular to the electric field, cyclotron resonance will occur at certain frequencies [3]), yielding information, e.g., on the charge carriers. Millimetre waves are also used in research on superconductivity [4]).

For investigations of this kind, suitable equipment is needed for generating the microwaves, for propagating them and performing measurements with them. As regards microwave generators, the Philips 4 and 2.5 mm reflex klystrons have proved very satisfactory [5]). The reflex klystron used for the present range of equipment is the 4 mm DX 151, the wavelength being reduced to 2 mm by means of a frequency doubler (point-contact diode).

For propagating the 2 mm waves, waveguides of rectangular cross-section are generally used ($0.83 \times 1.66$ mm or $0.0325'' \times 0.065''$).

The measurements involved usually relate to frequency, phase, power, absorption and reflection coefficient. Components for measuring these quantities in the centimetre wave region were described in this journal ten years ago [6]). Since then, comprehensive series of components have been developed for wavelengths of 3 cm, 8 mm and 4 mm. At the latter wavelength the utmost care is needed to keep the dimensions within the required tolerances. Even so, it has proved possible, with the same relative tolerances, to develop a range of components for 2 mm waves [7]), the more important of which will be reviewed in this article. Some are scaled-down versions of comparable components for longer waves; others differ mechanically from the familiar designs, and three are entirely new, namely: the *PIN modulator*, the *variable impedance* and the *rotary directional coupler*.

All components for 4 and 2 mm waves are fitted with claw flanges (to be discussed presently), by

[1]) M. A. Heald, Microwave measurements in controlled fusion research, Nat. Conv. Rec. Inst. Radio Engrs. **6**, No. 9, 14-18, 1958.

[2]) W. C. King and W. Gordy, One-to-two millimeter wave spectroscopy, Phys. Rev. **93**, 407-412, 1954.
M. Cowan and W. Gordy, Further extension of microwave spectroscopy in the submillimeter wave region, Phys. Rev. **104**, 551-552, 1956.

[3]) C. J. Rauch, J. J. Stickler, H. J. Zeiger and G. S. Heller, Millimeter cyclotron resonance in silicon, Phys. Rev. Letters **4**, 64-66, 1960 (No. 2).

[4]) M. A. Biondi and M. P. Garfunkel, Millimeter wave absorption in superconducting aluminum, Phys. Rev. **116**, 853-867, 1959 (No. 4).

[5]) B. B. van Iperen, Reflex klystrons for wavelengths of 4 and 2.5 mm, Philips tech. Rev. **21**, 221-228, 1959/60 (No. 8). Another method of generating millimetre waves is described by M. Yéou-Ta, Carcinotrons du type O fonctionnant sur une longueur d'onde de 2 mm, Onde électr. **39**, 789-794, 1959 (No. 391).

[6]) A. E. Pannenborg, A measuring arrangement for waveguides, Philips tech. Rev. **12**, 15-24, 1950/51.
A. E. Pannenborg, Some aspects of waveguide technique, Communic. News **11**, 65-75, 1950.

[7]) We are particularly indebted to the instrument-maker J. van Ostade for his contribution to this work.

which their respective waveguides are joined together. This construction makes for compactness and permits ready assembly and dissembly of microwave systems.

### Generation, detection and modulation of 2 mm waves

*The generator, the frequency multiplier and the detector*

As mentioned above, the 2 mm waves are generated in our equipment by doubling the frequency of a reflex klystron, type DX 151, designed for 4 mm waves. Mounted between the klystron and the frequency multiplier are an isolator *I* and a tuner *T* (*fig. 1*). The isolator prevents feedback from the load to the generator, whose frequency and power output would otherwise vary with the load. The principle of this isolator depends on gyromagnetic resonance [8]). The tuner, combined with a shorting plunger $P_1$, ensures that virtually the whole power output of the klystron reaches the frequency multiplier. The design of the tuners used will be described later.

The frequency multiplier consists of a point-contact diode situated at the junction between the 4 mm waveguide *1* and the 2 mm waveguide *2* (*fig. 2*). Projecting through these two waveguides is a tungsten catswhisker *3* which, together with a silicon crystal *4*, forms the point-contact diode. The lower end of the catswhisker is welded to an insulated pin *5*. The other end is etched to a fine point. The silicon wafer, which can be finely adjusted in the vertical direction by a differential screw, is brought into contact with the whisker. The part of the catswhisker running through the two waveguides acts as a coupling probe. The resistance of the diode (and also the capacitance) depends on the applied voltage [9]). As a result of this non-linearity the current induced in the coupling probe by the incident 4 mm waves contains a component with twice the frequency. This produces a 2 mm wave in

[8]) H. G. Beljers, Resonance isolators for millimetre waves, Philips tech. Rev. **22**, 11-15, 1960/61 (No. 1).

[9]) H. C. Torrey and C. A. Whitmer, Crystal rectifiers, M. I. T. Radiation Lab. Series, **15**, McGraw-Hill, New York 1948, in particular pages 97-107, Rectification at high frequencies. R. S. Ohl, P. P. Budenstein and C. A. Burrus, Improved diode for the harmonic generation of millimeter and submillimeter waves, Rev. sci. Instr. **30**, 765-774, 1959 (No. 9).

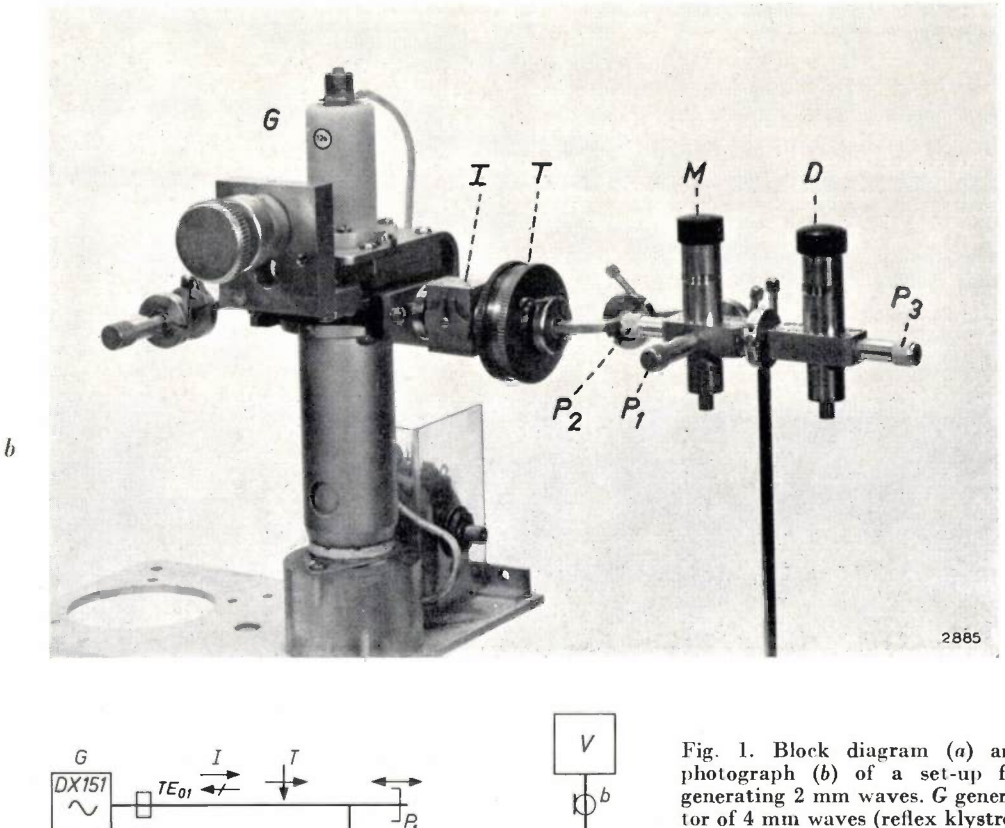


Fig. 1. Block diagram (*a*) and photograph (*b*) of a set-up for generating 2 mm waves. *G* generator of 4 mm waves (reflex klystron DX 151). *I* isolator. *T* tuner. $P_1$, $P_2$, $P_3$ adjustable waveguide plungers. *M* frequency multiplier. *D* detector for 2 mm waves, with millivoltmeter *V*. The sign *a* means "rectangular waveguide" and the sign *b* means "coaxial lign".

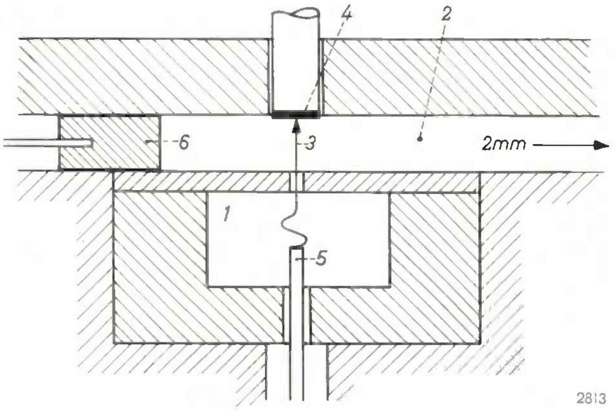


Fig. 2. Cross-section of frequency multiplier. *1* is the 4 mm waveguide, *2* the 2 mm waveguide. *3* catswhisker and *4* silicon crystal, together forming a point-contact diode. *5* insulated pin carrying the catswhisker. *6* shorting plunger in the 2 mm waveguide ($P_2$ in fig. 1).

the upper waveguide. *Fig. 3* is a photograph of the frequency multiplier.

To verify that the equipment is functioning properly up to the frequency multiplier, the latter can be used temporarily as a 4 mm-wave detector. For this purpose a DC voltmeter (e.g. type GM 6020 electronic millivoltmeter) is connected between the central pin *5* and earth — i.e. in parallel with the point-contact diode. If the meter reading is satisfactory, a crystal detector $D$ (fig. 1) is then connected to the 2 mm waveguide for the purpose of detecting the 2 mm waves. The detector is virtually identical with the frequency multiplier, except of course that it has no 4 mm waveguide.

### *The pivoting screw tuner*

The matching for the frequency multiplier and the 2 mm detector is effected by waveguide plungers ($P_1$, $P_2$ and $P_3$ in fig. 1) in conjunction with a tuner of special design. At longer wavelengths a tuner of the sliding screw type is ordinarily used. This consists of a metal stub *1* (*fig. 4a*), whose depth of penetration into the waveguide *2* can be adjusted by turning the screw, and which can also be moved along the waveguide, in a slot, over a distance of about $\frac{3}{4}\lambda_g$ (the wavelength in the waveguide, $\lambda_g$, is greater than the corresponding wavelength $\lambda$ in free space).

The relation between $\lambda_g$ and $\lambda$ is given by:

$$\frac{1}{\lambda^2} = \frac{1}{\lambda_g^{\,2}} + \frac{1}{\lambda_c^{\,2}},$$

where $\lambda_c$ is the cut-off wavelength. (See p. 17, formula (2) of the first article mentioned in reference [6].) For the mode used in the waveguide — the $TE_{01}$ mode — $\lambda_c$ is equal to twice the width $a$ of the rectangular waveguide. For $\lambda = 2.00$ mm and $\lambda_c = 2a = 3.32$ mm, the wavelength $\lambda_g$ is 2.51 mm.

The depth to which the stub penetrates into the waveguide regulates mainly the extent to which the stub reflects the wave; displacement of the stub along the guide varies the phase of the reflected wave. Reflections in the equipment can thus be cancelled by generating with the sliding screw tuner a reflection of equal magnitude but opposite phase. In this way a travelling wave is produced.

The 2 mm-wave screw tuner is based on a new design. The stub does not slide along the waveguide, but turns about an axis *3* (fig. 4*b*), so that it describes a pivoting motion, hence the term "pivoting screw tuner". In the millimetre wave region this design has special mechanical and electrical advantages:

1) The objection to the sliding screw tuner, that the stub may move slightly up and down as a result of surface irregularities, is entirely avoided here; there is less friction and adjustment is smoother.
2) The axial length of the tuner, from flange to flange, is limited to 12 mm, as a result of which

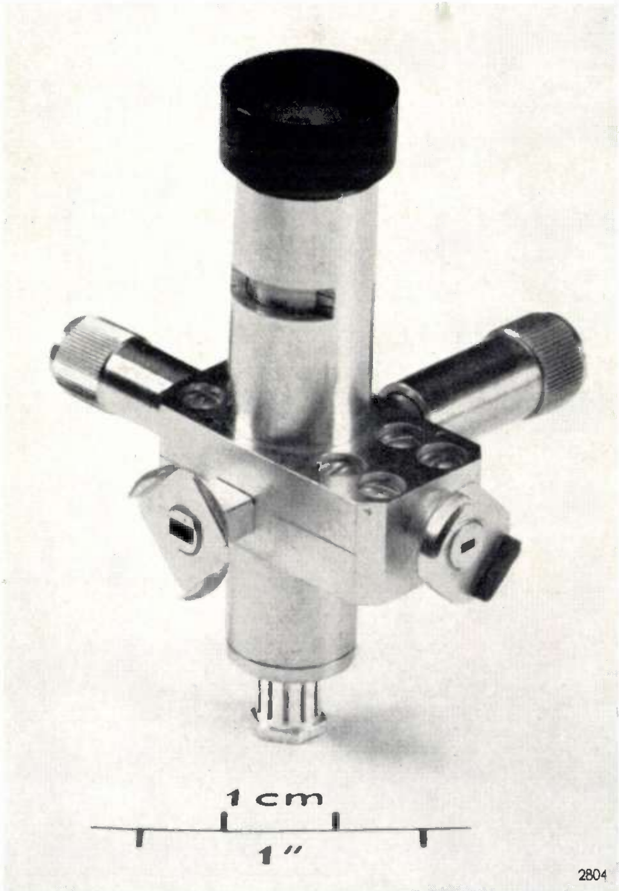
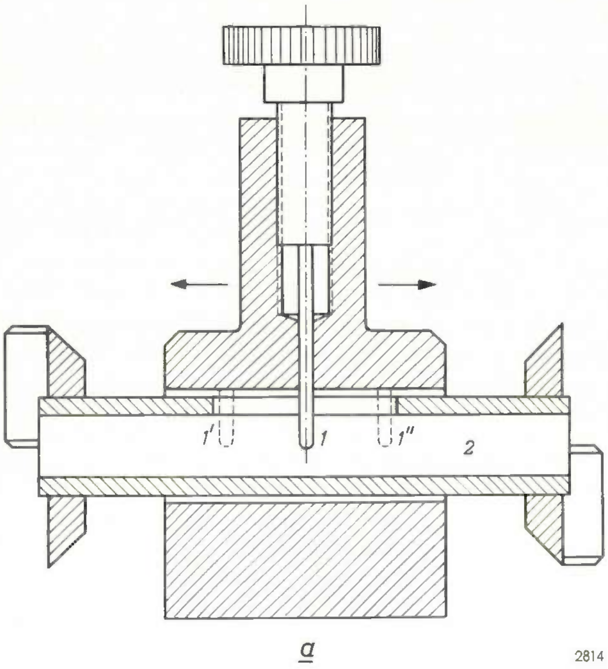


Fig. 3. The frequency multiplier. Foreground, left: connection for the 4 mm waveguide; right, connection for the 2 mm waveguide. The arms on the opposite side are fitted with waveguide plungers. The vertical arm contains the point-contact diode, and is surmounted by the adjusting knob. At the bottom is the coaxial connection for the millivoltmeter. The conversion loss is 15 to 20 dB.

the pivoting screw tuner causes particularly little attenuation (no more than 0.1 dB).

At first sight it might seem a drawback that, since the stub is pivoted about the axis *3*, its depth as well as its position varies in the waveguide, implying that the amplitude of the reflected wave will also vary slightly with the phase. This effect is so small, however, that it gives no trouble in practice.

A photograph of the pivoting screw tuner appears in *fig. 5*.



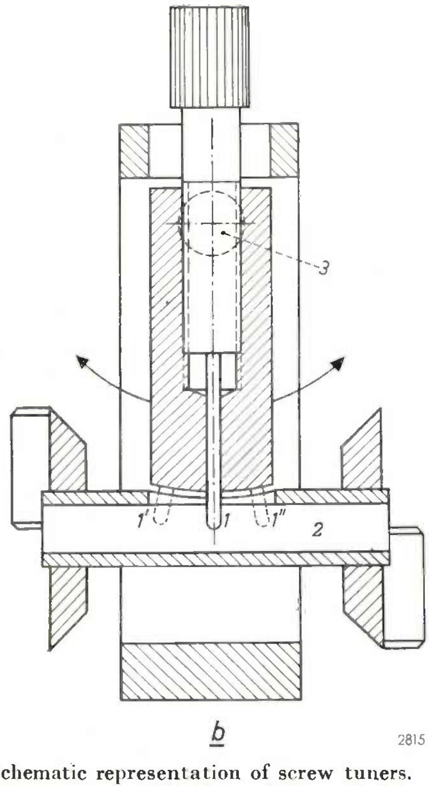


Fig. 4. Schematic representation of screw tuners.
*a*) Sliding screw tuner. The dept of the stub *1* in the waveguide *2* can be adjusted by turning a screw, and its position along the waveguide can be adjusted between *1'* and *1''*.
*b*) Pivoting screw tuner. Here too the stub *1* can be screwed in and out, but instead of sliding lengthwise it now pivots about spindle *3*. The extreme positions are again *1'* and *1''*.

### *Methods of modulation*

At an input power of, say, 100 mW the frequency multiplier delivers a 2 mm output of up to 2 mW. To detect the 2 mm wave at various points in the equipment, where the power may be much lower still, a highly sensitive method of measurement is therefore required.

For this purpose, in accordance with a widely used principle, the wave is modulated in amplitude and a selective indicator is employed behind the detector. The indicator consists of an amplifier tuned to the modulation frequency (e.g. 800 c/s), a low-frequency detector and a DC voltmeter. With a modulation depth of 100% the meter will then give a full scale deflection at a microwave power of the order of 1 μW. The sensitivity can be raised still further by the use of synchronous detection; in that case, to keep the noise level down, the modulation frequency should be fairly high (e.g. 8000 c/s).

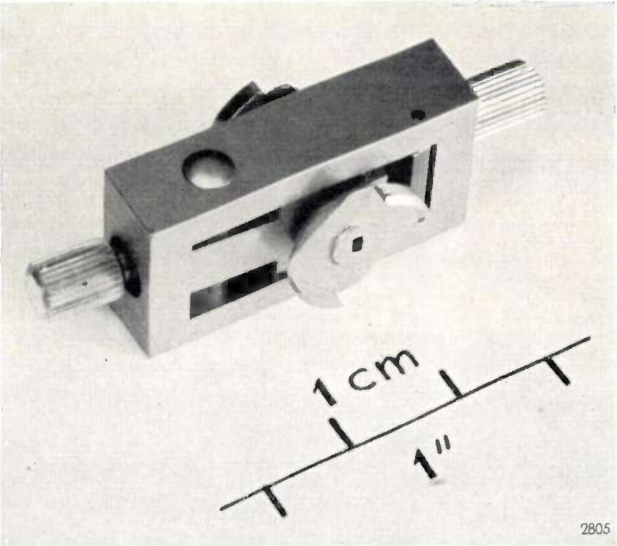


Fig. 5. Pivoting screw tuner for 2 mm waves. Left, the knob for screwing the stub in and out of the waveguide; right, the knob for altering its position (by pivoting about *3* in fig. *4b*). This component is capable of effecting a match in all cases encountered in practice. The attenuation is no more than 0.1 dB.

We shall now consider briefly some of the modulation methods used. The most common method is to periodically vary the repeller voltage $V_r$ of the reflex klystron between the value for optimum oscillation and a value at which the klystron does not oscillate. An objection to this method is that the oscillation *frequency* also depends on $V_r$ [10]); to avoid unwanted frequency modulation one would have to give $V_r$ a square waveform. In practice, however, it is very difficult to make the edges sufficiently steep and to keep $V_r$ constant enough in the oscillation interval. Some frequency modulation therefore remains, which can be particularly undesirable wherc bridge circuits are involved.

In our case, where a frequency multiplier is used, the modulation can be simply effected, without frequency modulation arising, by applying a low-frequency alternating voltage across the diode of the frequency multiplier, in series with any biasing voltage present. An alternating voltage of about 1 V is quite sufficient to modulate the 2 mm wave 100% in amplitude.

Without a frequency multiplier, pure amplitude modulation is possible by making use of a new device called the *PIN* modulator [11]). This is an electrically controlled attenuator which is inserted in the waveguide. It consists of a germanium wafer in which a *P* region and an *N* region are separated by an *I* region of pure germanium, possessing extremely low "intrinsic" conductivity. Only the *I* region is contained inside the waveguide. In the normal state it behaves like a low-loss dielectric. When, however, a voltage is applied between the *P* and *N* regions in the forward direction, holes and electrons are injected into the *I* region, which thereby becomes an absorbing medium. In this way, using a control current of 15 mA (control power approx. 10 mW), the losses in the *I* region can be made so high as to increase the attenuation by 25 dB. If an alternating control current is used, the wave is thus modulated in amplitude. A *PIN* modulator for 2 mm is illustrated in *fig. 6*.

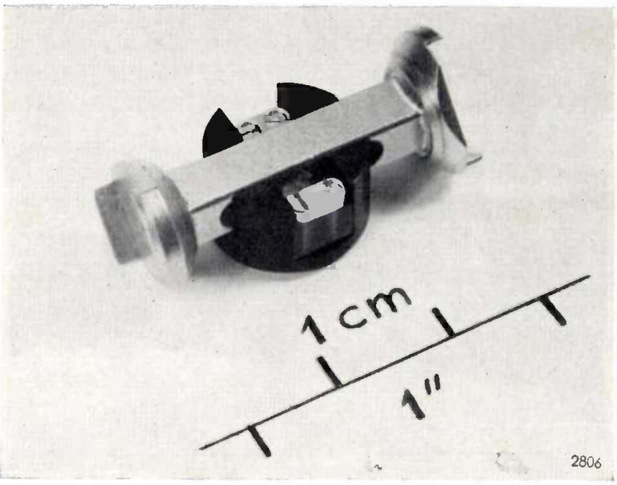


Fig. 6. *PIN* modulator for 2 mm waves. The attenuation is adjustable from about 5 dB to about 30 dB. The modulation frequency at which the modulation depth has dropped by 3 dB is approximately 50 kc/s.

### Waveguide and claw flanges

Except in the rotary attenuator, the variable impedance and the rotary directional coupler, where circular waveguides are used for reasons presently to be discussed, the waveguides in the various components are all rectangular.

The rectangular waveguides are constructed of a U-section, made by milling a groove in rectangular bar stock, to which a flat cover plate is screwed or soldered. In most cases the material used is brass; the waveguides are gold-plated inside and out in order to keep the surface properties constant. Silver is used only where minimum attenuation is required.

The internal dimensions of the 2 mm waveguides are 0.83×1.66 mm, with a tolerance of 0.01 mm. The outside dimensions are determined solely by the mechanical considerations applicable to the part in question. The internal cross-section must be uniform, otherwise a discontinuity can arise between coupled components, giving rise to reflections. But even when the cross-sections are uniform, reflections may still occur if the walls of the coupled sections of waveguide are not well-aligned (*fig. 7a, b* and *c*) [12]). Such misalignment points to shortcomings of the waveguide coupling (the flanges and the means by which they are fixed together). An essential requirement of a waveguide coupling is that the walls should be in true alignment with one another.

Other desirable properties of a waveguide coupling are:

1) The axial length should be small in order to mini-

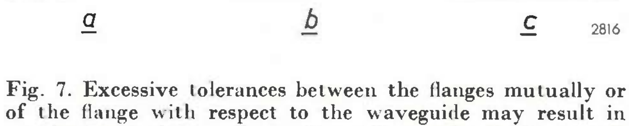


Fig. 7. Excessive tolerances between the flanges mutually or of the flange with respect to the waveguide may result in misalignment of the walls, e.g. the walls may be slightly turned in relation to one another (case *a*), or displaced at right angles to the electric field *E* (case *b*), or displaced in the direction of *E* (case *c*).

[10]) See article by Van Iperen in reference [5]), page 224 (fig. 6 and note [4])).

[11]) F. C. de Ronde, H. J. G. Meyer and O. W. Memelink, The *PIN* modulator, an electrically controlled attenuator for mm and sub-mm waves, Trans. Inst. Radio Engrs. on microwave theory and techniques MTT **8**, 325-327, 1960 (No. 3). See also Dutch patent application No. 229 531 of 11th July 1958.

[12]) U. von Kienlin and A. Kürzl, Reflexionen an Hohlleiter-Flanschverbindungen, Nachr.techn. Z. **11**, 561-564, 1958.

mize attenuation and frequency-dependence [13]). This applies both to the flange and to the space between flange and component. The construction should be such that no tools need be used between flange and component; for this reason the joint should preferably be secured from a direction perpendicular to the waveguide axis.

2) The losses should be low and also reproducible, i.e. the losses should not be dependent on the force applied to draw the two flanges together. This force should preferably be applied by means of no more than one screw. The construction illustrated in *fig. 8*, where the flanges are fixed together by four screws, is not only unpractical but fundamentally wrong, because it is not possible to check whether the screws have all been tightened equally.

3) The coupling should be "universal", i.e. it should be possible to reverse the components in the waveguide system and also to connect them in reversed positions about the axis of the waveguide.

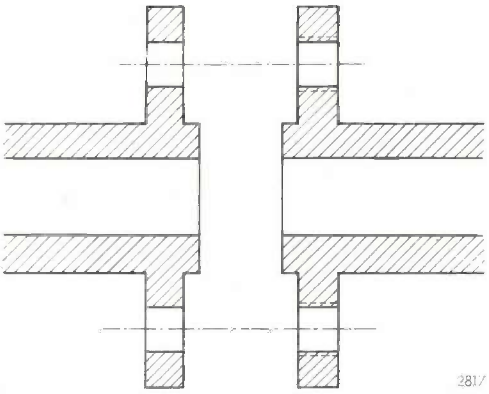


Fig. 8. Conventional waveguide joint. The flanges are joined by four bolts. The objections to this system are:
1) It is impossible to check whether the bolts have all been equally tightened.
2) The waveguide section between flange and component is necessarily relatively long, owing to the presence of the clamping bolts.

A flange that meets all these requirements is the *claw flange*, which is fitted to all our 2 mm components. *Fig. 9* shows two claw flanges and the clamping ring consisting of two hinged halves, by which the flanges are drawn tightly together by means of a single screw perpendicular to the waveguide axis. The whole coupling — the two flanges and the clamping ring — has a total length of twice the flange thickness, namely 5 mm. As appears from the various photographs in this article, the flanges are fitted close up against the components, making for a particularly compact assembly.

## Description of some 2 mm components

Some components of the microwave equipment will now be discussed in somewhat more detail, viz. three adjustable components, provided with dials: the rotary attenuator, the variable impedance and the rotary directional coupler.

### *The rotary attenuator*

It is sometimes necessary in microwave measurements to adjust the power level of the transmitted wave. There is also often a need to introduce a known attenuation. In both cases, use is made of a variable attenuator. One of the oldest types is the vane attenuator, the principle of which is illustrated in *fig. 10*. An insulating plate $V$ (the vane), coated with an absorbent resistive layer, is inserted in the waveguide, to an adjustable depth, parallel to the direction of propagation and to the electric field $E$. The absorption caused by the vane is greater the deeper the vane extends into the waveguide and the greater its axial length $l$ with respect to the wavelength. The drawback of the latter fact is that the attenuation is partly dependent on frequency. If $l$ is not large with respect to the wavelength, the attenuation will moreover be dependent on the standing-wave ratio in the waveguide.

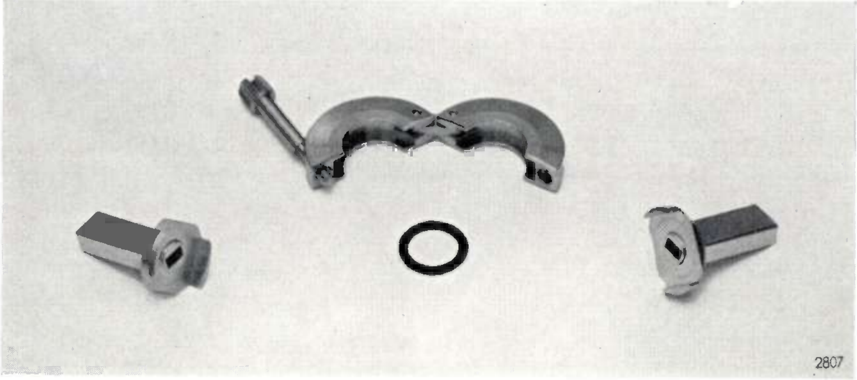


Fig. 9. Left and right, sections of waveguide with claw flanges; top centre, the hinged clamping ring which joins the flanges together; below, a rubber ring. (True size.) The clamping ring is tightened round the flanges with a single screw, perpendicular to the axis of the waveguide. The rubber ring ensures an airtight seal; the air in the whole waveguide system can then be put under pressure to prevent the entry of moisture.

The modulus of the reflection coefficient of the claw-flange coupling is smaller than 0.005.

[13] The frequency range in which a rectangular waveguide operating in the $TE_{01}$ mode can be used is from about 1.2 to 1.9 times the cut-off frequency $f_c = c/2a$ ($c$ = velocity of light, $a$ = width of waveguide).

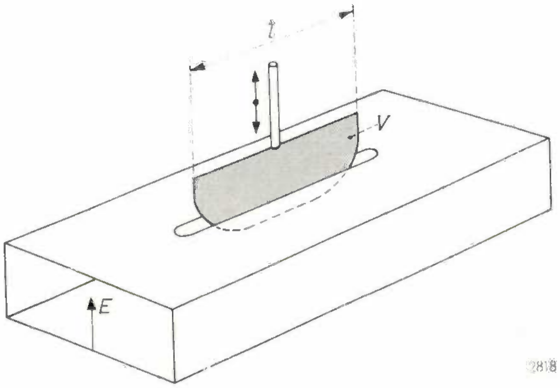


Fig. 10. Principle of the vane attenuator. An absorbent vane $V$, parallel to the electric field $E$, can be inserted to a variable depth in the waveguide through a narrow slot. The attenuation increases with the depth of insertion, but also depends on the ratio of the length $l$ of the vane to the wavelength, i.e. it is frequency-dependent.

These drawbacks are entirely absent in the *rotary attenuator*, a device invented in America during the war by Bowen [14]). A further advantage of this attenuator is that it is an absolute instrument, that is to say the scale giving the attenuation in dB obeys a mathematical law, and thus can be engraved at once to the instrument without previous calibration.

A schematic cross-section of the rotary attenuator is given in *fig. 11a*. The ends *1* and *2* are rectangular, and the middle section *5-7-6* is circular. The transition between the rectangular and the circular cross-sections takes place in the regions *3* and *4*. In the circular section the centre piece *7* is rotatable about the waveguide axis. In each of the (fixed) regions *5* and *6*, an absorbing vane (*8*, *9*) is mounted perpendicular to the electric field, i.e. parallel to the long sides of the rectangular waveguide section (fig. 11*b*). The central, rotatable part *7* of the waveguide also contains an absorbing vane, *10*. This can be rotated from a position parallel to *8* and *9* to a position perpendicular thereto; as we shall see, the extreme positions correspond to the minimum and maximum attenuation, respectively.

The attenuation as a function of the angle $\alpha$ between the plane of the vane *10* and that of the vanes *8* and *9* is found as follows. The electric field $E$ of a wave entering at *1* is perpendicular in the fixed section *5* to the vane *8*; the wave here therefore passes through unattenuated. In the rotary section *7* we resolve $E$ into a component $E \sin \alpha$ in the plane of vane *10* and component $E \cos \alpha$ perpendicular thereto (fig. 11*c*). The component $E \sin \alpha$ is entirely absorbed in the vane, leaving only the component $E \cos \alpha$. Finally, in the fixed section *6* the field is $E \cos \alpha$, whose component $E \cos \alpha \sin \alpha$, in the plane of *9*, is absorbed by this vane so that only component $E \cos^2\alpha$, perpendicular to *9*, remains (fig. 11*d*).

The emergent wave thus has an electrical field-

[14]) G. C. Southworth, Principles and applications of waveguide transmission, Van Nostrand, New York 1950, p. 374.
B. P. Hand, A precision waveguide attenuator which obeys a mathematical law, Hewlett-Packard Journal **6**, Jan. 1955.

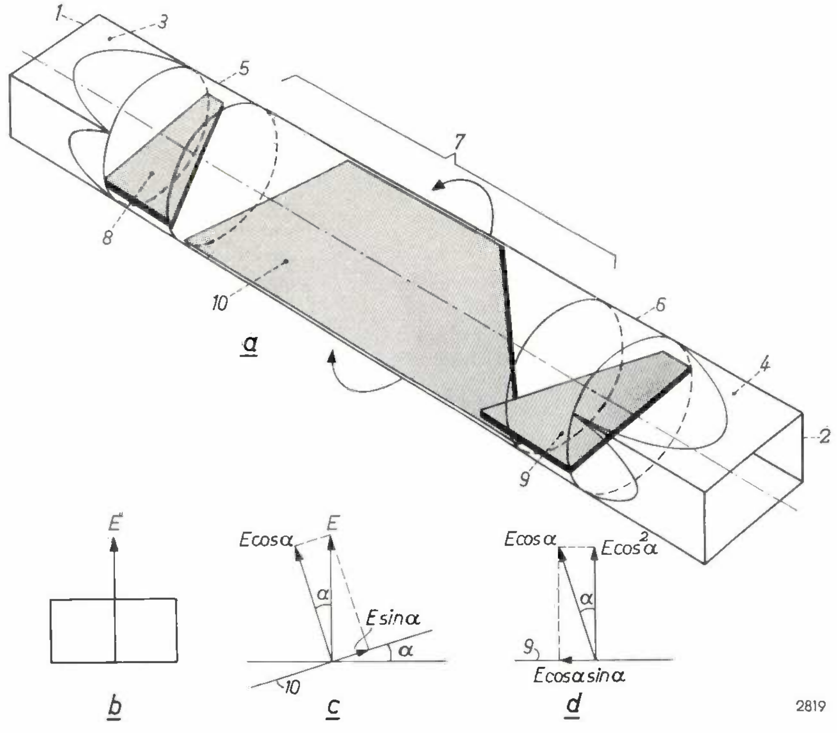


Fig. 11. *a*) Schematic diagram of the rotary attenuator. *1*, *2* rectangular ends for connection to rectangular waveguides. *3*, *4* transitions to circular waveguide. *5*, *6* fixed sections of circular waveguide, *7* rotary section. *8*, *9* and *10* vanes. *b*) Rectangular cross-section at *1* and *2* (long sides parallel to the vanes *8* and *9*), $E$ electric field.
*c*) The component $E \sin \alpha$ of the electric field of a wave entering at *1* is absorbed in vane *10*, leaving only the component $E \cos \alpha$ ($\alpha$ is the angle made by the plane of *10* with that of the fixed vanes *8* and *9*).
*d*) In section *6* the component $E \cos \alpha \sin \alpha$ of the remaining wave is absorbed by vane *9*, leaving only the component $E \cos^2\alpha$. The total attenuation is thus $\cos^2\alpha$, or $-20 \log \cos^2\alpha$ dB.

strength which is $\cos^2 \alpha$ times that of the incident wave. The attenuation is therefore solely a function of the angle $\alpha$ and amounts to $-20 \log \cos^2 \alpha$ dB. Without calibration, then, a dial can be fitted whose scale gives an absolute reading of the attenuation.

It is important that the vanes *8*, *9* and *10* should reflect as little energy as possible. For this purpose they are tapered at the ends to form trapezia and are extremely thin, consisting of metallized slivers of mica. The thickness of the metal coating is small compared to the skin depth, so that the vane acts as a resistive layer which strongly absorbs the wave. Due to the tapering of the ends and the thinness of the vanes, the reflection coefficient of the attenuator is less than 0.02.

The 2 mm rotary attenuator is illustrated in *fig. 12*.

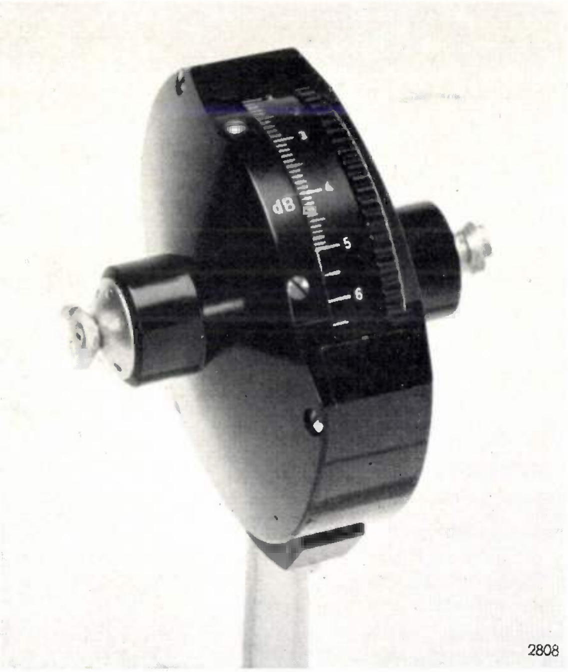


Fig. 12. Rotary attenuator for 2 mm waves. The attenuation, which is independent of the frequency, can be read on a scale covering a range from 0 dB to 50 dB. The insertion loss is 3.5 dB. The standing-wave ratio is less than 1.05.

### *The variable impedance*

At the end of a waveguide there is generally some reflection of energy [6]). The incident and the reflected waves give rise to a standing wave, as a result of which the electric field-strength in the waveguide shows a minimum ($E_{\min}$) at certain places and a maximum ($E_{\max}$) at others. The standing-wave ratio, $E_{\max}/E_{\min}$, is a measure of the mismatch (the extent to which the terminating impedance differs from the characteristic impedance), and in many cases this is an important indication of the quality of a microwave component.

In the centimetre range the standing-wave ratio is measured in waveguides in the same way as in lecher lines, that is with a standing-wave indicator. In the millimetre wave range, however, it is difficult to produce a standing-wave detector having the required accuracy. Instead of the standing-wave ratio we can measure the *reflection coefficient*, which is simply related to the standing-wave ratio (see page 18 of the first article mentioned in reference [6])). With the aid of a variable impedance, which gives a direct reading of the reflection coefficient, this measurement can be readily performed with a form of bridge circuit. The bridge is balanced by making the reflection coefficient of the variable impedance equal to that of the unknown impedance.

The device employed here as the bridge is a hybrid T, a schematic representation of which is shown in *fig. 13a*. A wave entering the arm *1* divides equally at the branching point into two components, which enter the symmetrical arms *2* and *3*. There is no direct transfer of energy from arm *1* to arm *4*. Arm *2* is terminated by the unknown impedance $Z_2$, and arm *3* by the variable (known) impedance $Z_3$. The waves reflected from $Z_2$ and $Z_3$

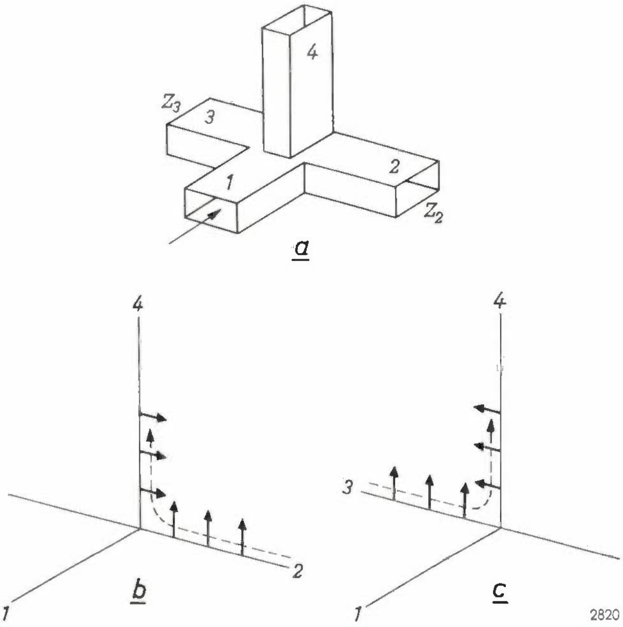


Fig. 13. *a*) Schematic representation of a hybrid Tee. A wave entering arm *1* divides into equal components which enter arms *2* and *3*. These waves are reflected from the impedances $Z_2$ and $Z_3$ terminating these arms, and return partly into arm *1* (not considered here) and partly into arm *4*. Here the waves cancel each other if $Z_3$ is equal in modulus and argument to $Z_2$.
*b*) The short arrows represent the electric field of the wave which is reflected from $Z_2$ and enters arm *4*; the dashed arrow denotes the direction of propagation.
*c*) The same for the wave reflected from $Z_3$. When $Z_3 = Z_2$, the two waves cancel out in arm *4*.

return to the branching point and enter arm *4*. When the modulus and the argument of $Z_3$ are made equal to those of $Z_2$, the electrical field-strength of the wave reflected from $Z_2$ into arms *2* and *4* will be as shown in fig. 13*b*, while that for the wave reflected from $Z_3$ is as shown in fig. 13*c*. In arm *4* the field-strengths are equal and opposite, and the waves thus cancel each other. An indicator connected to a detector mounted in arm *4* will therefore give no deflection when the bridge is balanced, but does give a deflection as long as $Z_3$ is not equal to $Z_2$.

*Fig. 14* shows a hybrid T for 2 mm waves. A particularly important requirement for hybrid T's is that the common planes of symmetry of the Tee *1-2-3* and of the Tee *4-2-3* *exactly* coincide.

The principle of the variable impedance is illustrated in *fig. 15* [15]). At the position of flange *1* the waveguide has a rectangular cross-section; via a transition *2*, this becomes a circular cross-section *3*. The latter contains two absorbing vanes, *4* and *5*, which again consist of thin metallized slivers of mica. Vane *4* is fixed, perpendicular to the electric field $E$ of the incident wave; vane *5* is mounted on a metal plunger *6*, which can be both rotated and displaced axially. If the planes of *4* and *5* subtend an angle $\alpha$, the component $E \sin \alpha$, parallel to vane *5*, is absorbed by this vane; the component $E \cos \alpha$, perpendicular to *5*, is reflected by the plunger *6*. Now $E \cos \alpha$ has a component $E \cos \alpha \sin \alpha$, which is absorbed in vane *4*, and a component $E \cos^2 \alpha$, which emerges from *1* as a reflected wave, so that the modulus of the reflection coefficient is equal to $\cos^2 \alpha$. This modulus is varied by turning the vane *5* by means of the plunger spindle. The argument of the reflection coefficient is varied by displacing the reflecting surface, i.e. by moving the plunger axially, in or out. A displacement $\delta l$ alters the phase $\varphi$ of the reflected wave by

$$\delta\varphi = 2 \times \frac{\delta l}{\lambda_{\text{gc}}} \times 2\pi \text{ radians},$$

where $\lambda_{\text{gc}}$ is the wavelength in the circular waveguide; the factor 2 is due to the fact that the wave

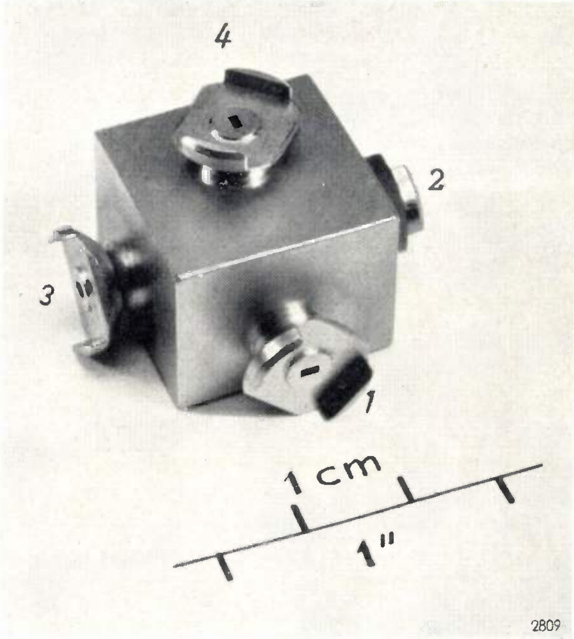


Fig. 14. Hybrid T for 2 mm waves. Numbering of the arms as in fig. 13. The decoupling between arms *1* and *4* is greater than 30 dB.

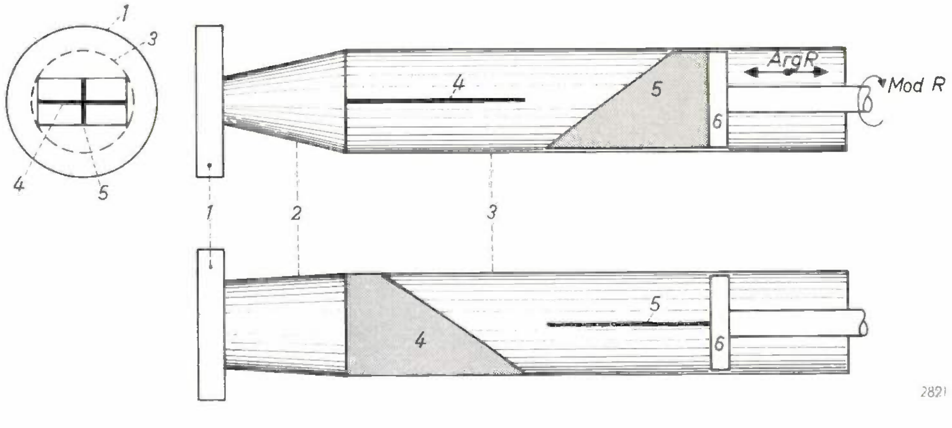


Fig. 15. Axial sections of the variable impedance along two mutually perpendicular planes, and end view. *1* flange with rectangular opening. *2* transition from rectangular to circular cross-section. *3* circular waveguide. *4* fixed vane. *5* vane capable of being rotated and axially displaced, fixed to plunger *6*. When *5* is rotated, the modulus of the complex reflection coefficient is changed; when *5* is displaced axially, the argument is changed.

[15]) See also F. C. de Ronde, A simple component for impedance measurements at cm and mm waves: the direct-reading variable impedance, Communic. Congrès internat. Circuits et antennes hyperfréquences, Paris 1957, Part I (Suppl. Onde électr. 38, No. 376 bis), pp. 294-295, 1958.

traverses the waveguide twice (incident and reflected waves).

As can be seen in *fig. 16*, the instrument gives a direct reading of both modulus and argument. The modulus scale, like the attenuation scale on the rotary attenuator, requires no previous calibration.

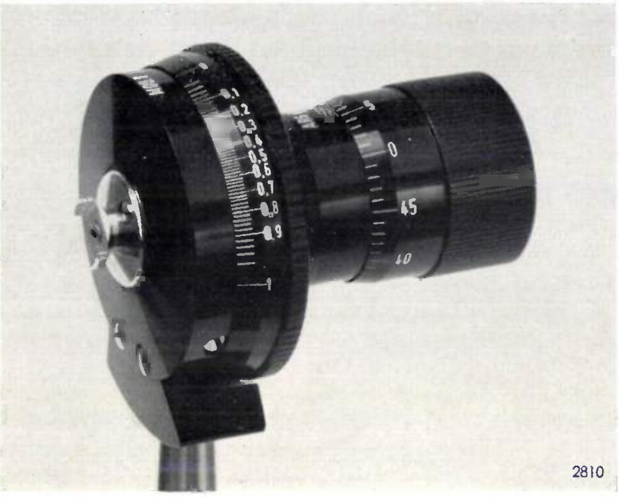


Fig. 16. Variable impedance as in fig. 15, for 2 mm waves. The modulus and argument of the reflection coefficient can be read from separate scales. If reflection losses are eliminated, the absolute error in the modulus reading is less than 0.05.

### *The rotary directional coupler*

A directional coupler is a device that makes it possible to sample either the forward or backward wave in a waveguide, and thus to demonstrate, e.g., the presence of reflection. The most familiar type contains coupling holes between the main and a subsidiary waveguide. Its operation is illustrated in *fig. 17*. A wave travelling to the right through the main waveguide $W_1$ in fig. 17*a* produces no response in the detector mounted in the subsidiary waveguide $W_2$; the detector does respond, however, to a wave travelling in the opposite direction (fig. 17*b*). In this way one can detect the presence of a reflected wave.

A drawback of such directional couplers is that they function only in the frequency range where the (fixed) distance between the coupling holes is roughly equal to $\frac{1}{4}\lambda_g$. Instead of two large holes, it is the usual practice for various reasons to employ a series of small coupling holes, whose spacing must be extremely accurate; this involves considerable difficulties in manufacture. A third drawback is that the coupling is not variable. None of these disadvantages apply to the new directional coupler which we shall now describe.

The *rotary directional coupler* [16]) (*fig. 18a*) consists of a circular waveguide $W_1$ (only part of which is shown in the figure, for simplicity) which terminates at both ends in a rectangular cross-section. The circular part, at end *I*, contains a metal strip *Pol*, the polarization strip, which can be regarded as lossless. At the left end (near *I*), this strip is exactly parallel with the broad faces of the waveguide, but it can be twisted into a helical surface by means of a rotatable section of the circular waveguide. The polarization strip terminates in a tapered absorption vane *V*, of the type already mentioned. This vane lies in exactly the same plane in which the polarization strip ends. Beyond the vane the circular waveguide contains a thin rod *A*, which acts as a coupling probe, and situated a quarter wavelength further on is a metal plate *R*, which acts as a reflector (see below). The probe and the reflector pass through the axis of the circular waveguide and are parallel to the broad sides of the rectangular ends. Transverse slots in the polarization strip and in the reflector prevent the occurrence of higher modes of oscillation.

The probe couples the circular waveguide with a rectangular subsidiary waveguide $W_2$. To direct

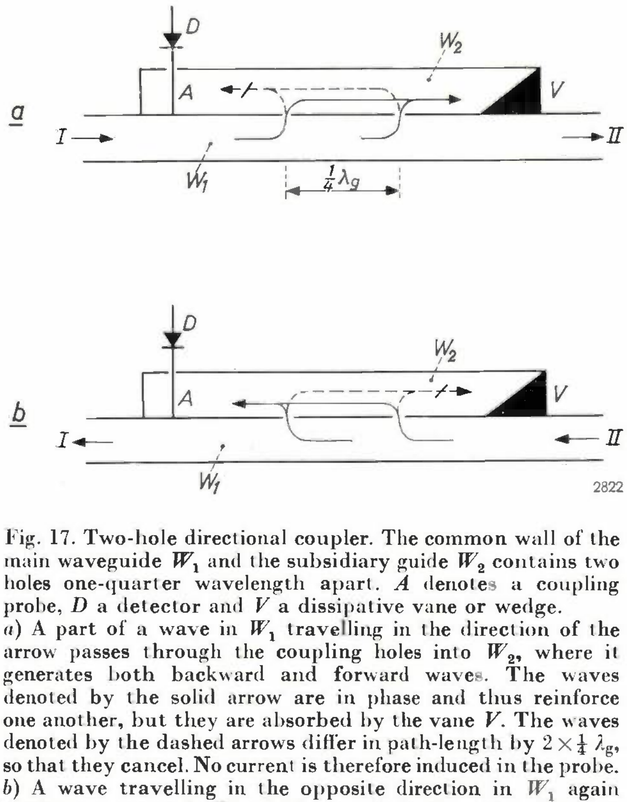


Fig. 17. Two-hole directional coupler. The common wall of the main waveguide $W_1$ and the subsidiary guide $W_2$ contains two holes one-quarter wavelength apart. *A* denotes a coupling probe, *D* a detector and *V* a dissipative vane or wedge.
*a*) A part of a wave in $W_1$ travelling in the direction of the arrow passes through the coupling holes into $W_2$, where it generates both backward and forward waves. The waves denoted by the solid arrow are in phase and thus reinforce one another, but they are absorbed by the vane *V*. The waves denoted by the dashed arrows differ in path-length by $2\times\frac{1}{4}\lambda_g$, so that they cancel. No current is therefore induced in the probe.
*b*) A wave travelling in the opposite direction in $W_1$ again produces a wave in $W_2$ which travels in the same direction as in $W_1$, and now induces a current in the probe.

[16]) Belgian patent application No. 464 685 of 16th December 1959.

all energy in this waveguide to one end — where there is usually a detector — the other end of $W_2$ is terminated by an adjustable plunger $P$.

The rotary directional coupler works as follows. We consider first a wave entering the coupler at $I$. The electric field $E$ is thus perpendicular to the beginning of the polarization strip (fig. 18*b*). The wave then passes along this strip, which is twisted through an angle $\alpha$. The direction of polarization

The reflector $R$ ensures that the component $E \sin \alpha$ is coupled as fully as possible to the waveguide $W_2$. For this purpose, $R$ is situated a quarter wavelength behind the probe, and consists of a vertical metal plate, its function being to reflect vertically polarized waves back to the probe and let horizontally polarized waves pass through. The joint effect of the reflector $R$ and the plunger $P$ is to couple the component $E \sin \alpha$ almost entirely to $W_2$.

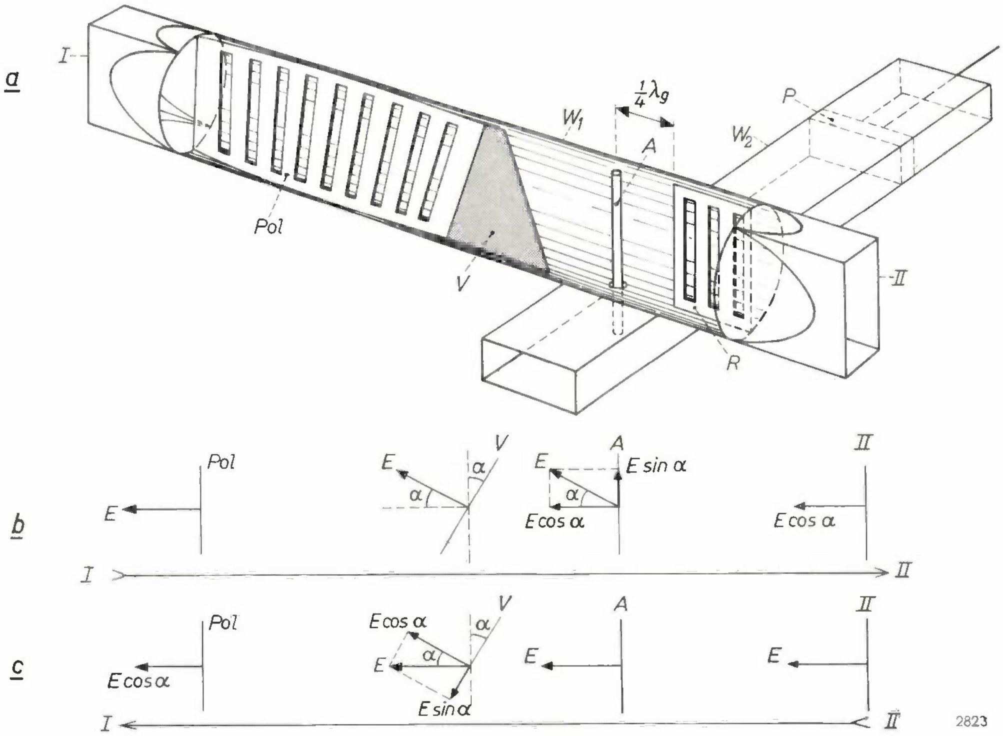


Fig. 18. *a*) Rotary directional coupler (schematic). $W_1$ circular waveguide terminating at both ends in rectangular openings *I* and *II*. *Pol* polarization strip which can be twisted from 0 to 45° by a central rotatable section of the waveguide wall. *V* vane. *A* coupling probe. *R* reflector. $W_2$ subsidiary waveguide with rectangular cross-section. *P* movable plunger. A wave passing from *I* to *II* is partly transferred to $W_2$, the coupling being continuously variable and independent of frequency. No energy is coupled to $W_2$ in the case of a wave passing from *II* to *I*.
*b*) Electric field of a wave travelling from *I* to *II*, at the beginning of the polarization strip *Pol*, at the absorption vane *V*, at the probe *A* and in the opening *II*. The angle of twist is $\alpha$.
*c*) The same for a wave from *II* to *I*; from right to left: the field at *II*, at *A*, at *V*, and at the end of *Pol*.

thereby undergoes a rotation $\alpha$, since the vector $E$ remains perpendicular to the lossless strip. At the end of the strip, then, the field is unattenuated and is at right angles to the absorption vane, so that no energy is absorbed by the latter. Beyond the strip the field possesses a vertical component $E \sin \alpha$ and a horizontal component $E \cos \alpha$. The component $E \sin \alpha$ is coupled, via the probe, with the waveguide $W_2$, whilst the component $E \cos \alpha$ passes on to the rectangular opening $II$.

A wave entering $II$ in the opposite direction, with $E$ again perpendicular to the broad sides of the rectangle (fig. 18*c*, right), passes the reflector, induces no current in the probe (so that energy in this direction is not transferred to $W_2$) and arrives at the absorption vane at a certain angle. The component $E \sin \alpha$ is absorbed by the vane, and the component $E \cos \alpha$ perpendicular to the vane reaches the twisted polarization strip, after which, still perpendicular to the strip and unattenuated, it is turned back through an angle $\alpha$ to the position perpendicular to the broad sides of the rectangular guide, in which position it emerges from the rectangular opening $I$.

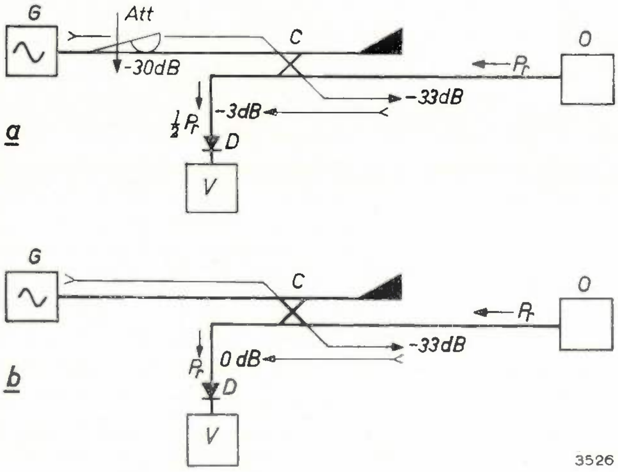


Fig. 19. *a*) Illustrating the principle of a gyromagnetic resonance measurement. $G$ microwave generator. *Att* variable attenuator. $C$ fixed directional coupler. $O$ specimen (solid in resonant cavity, in variable magnetostatic field). $D$ detector. $V$ millivoltmeter.

The directional coupler gives a coupling of, e.g., $-3$ dB in the path from $G$ to $O$. To avoid saturating $O$, an extra attenuation of 30 dB is introduced, bringing the total attenuation to 33 dB. Of the total reflected power $P_r$, only half enters the detector via $C$.

*b*) The same layout, now with a rotary coupler. The required attenuation of 33 dB in example (*a*) can now be produced without a variable attenuator by adjusting the coupling from $G$ to $O$ to $-33$ dB. The coupling from $O$ to $D$ is then strong enough for almost the entire reflected power $P_r$ to enter the detector.

The result is thus as follows.

*a*) When waves travel in the direction from $I$ to $II$, the component $E \sin\alpha$ is transferred to $W_2$, whilst the component $E \cos\alpha$ passes through.

*b*) When waves travel in the direction from $II$ to $I$, no energy is transferred to $W_2$, the component $E \sin\alpha$ is absorbed and the component $E \cos\alpha$ passes through.

The factor that indicates what fraction of the field of the waves travelling from $I$ to $II$ is transferred to the subsidiary waveguide is referred to as the coupling. The power coupling in this case is thus $\sin^2\alpha$ or, in decibels, $10 \log \sin^2\alpha$ dB. The coupling effected by the rotary directional coupler is therefore continuously variable and moreover solely a function of the angle $\alpha$, irrespective of the frequency. The value of the coupling can be read from a scale. In all other existing directional couplers the coupling is not variable and is to some extent frequency-dependent.

The usefulness of a continuously variable coupling may be illustrated by an example, in this case measurements of paramagnetic resonance in solids [17]. (Such measurements are ordinarily performed in the centimetre wave region, but this does not detract from the advantages of a variable coupler over a fixed type.)

The principle of the measurement is illustrated in *fig. 19a*. Microwave energy from a generator $G$ is conducted through a directional coupler $C$ to the sample under investigation $O$, which is contained in a resonant cavity, itself situated in a variable magnetostatic field. Through the same directional coupler a part of the reflected power $P_r$ is transferred to the detector $D$. When the magnetostatic field is varied, resonance is observed as a sharp dip in the reflected power. The sensitivity is optimum when the coupling is $-3$ dB. This value is obtained with a hybrid T (preferably a magic T, i.e. a hybrid T free from reflection). For this reason, and also because of its good directivity (this term is explained below), the magic T is often used as a directional coupler in this kind of measurement.

A coupling of $-3$ dB, however, may give rise to saturation

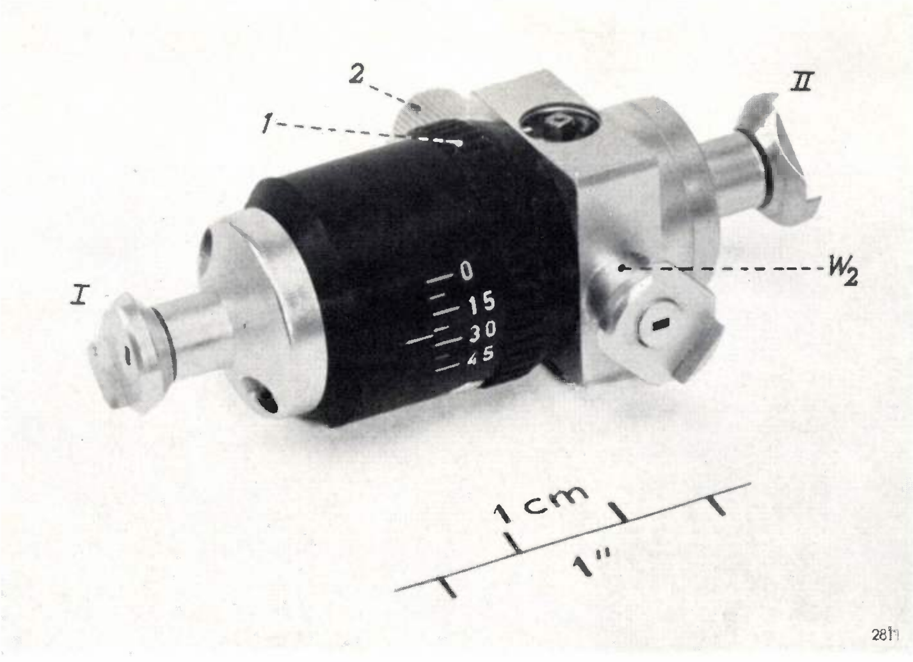


Fig. 20. Rotary directional coupler for 2 mm waves. $I$, $II$ and $W_2$ as in fig. 18*a*. *1* rotatable ring for twisting the polarization strip; the angle of twist $\alpha$ is read from the scale. *2* knob for adjusting the plunger in $W_2$. Directivity $>25$ dB, standing-wave ratio $<1.16$.

[17] See e.g. J. S. van Wieringen, Paramagnetic resonance, Philips tech. Rev. **19**, 301-313, 1957/58.

in the specimen $O$, that is to say the incident power is so high that the reflected power is virtually independent of it. To remain below the saturation region one must therefore reduce the incident power. The most obvious method is to replace the fixed directional coupler by one which gives a weaker coupling, but this is cumbersome and entails stepwise regulation. Another method is to insert a variable attenuator (*Att*, fig. 19*a*) between the generator and the directional coupler. Suppose that the power must be attenuated by 30 dB to remain below the saturation limit. The attenuation between the generator and the specimen is then 33 dB (fig. 19*a*). This can be achieved more simply by substituting for the attenuator and the fixed 3 dB directional coupler a variable rotary coupler adjusted to —33 dB (fig. 19*b*). This dispenses with the variable attenuator. A further advantage is that, with a weak coupling from $G$ to $O$, the coupling from $O$ to $D$ is strong enough to cause virtually all the reflected power $P_r$ to arrive in $D$. Where a fixed directional coupler of —3 dB is used (fig. 19*a*), half of $P_r$ is lost.

In fig. 19*b* the rotary coupler is connected as follows (cf. fig. 18*a*): the generator at the free end of the subsidiary waveguide $W_2$, the object at side *I* and the detector at side *II*.

For the rotary coupler it is found to be sufficient if $a$ is variable from 0 to 45°. Theoretically, the coupling is then variable from $-\infty$ dB to —3 dB. Owing to mechanical imperfections, the weakest coupling in practice is between —30 and —40 dB.

For the same reason $W_2$ still receives a small fraction of the energy of a wave passing through the rotary coupler from *II* to *I*. This imperfect directional effect — found in all such devices — is expressed quantitatively in the *directivity*. This is the ratio of the power coupled to the subsidiary waveguide as a result of a wave from *I* to *II*, to that of an equally strong wave from *II* to *I*. The directivity of the rotary coupler is thus better the tighter the coupling.

A photograph of the 2 mm rotary directional coupler can be seen in *fig. 20.*

Some measuring arrangements using the components described in the foregoing will be the subject of Part II of this article. This will also provide an opportunity to describe various other components not dealt with here.

**Summary.** For measurements in the 2 mm wavelength region a comprehensive range of waveguide components has been developed, the principal of which are reviewed in this article. The 2 mm waves are produced by using a silicon diode to double the frequency of the 4 mm waves generated by a reflex klystron, type DX 151. The detector used is also a silicon diode. The matching is effected by means of waveguide plungers and a pivoting screw tuner (a variant of the sliding screw tuner). Use is made either of synchronous detection or of a selective indicator, the 2 mm waves being modulated in amplitude, e.g. by superposing an audio-frequency voltage of about 1 V on the frequency multiplier. Without a frequency multiplier the *PIN* modulator may be used, which is an electrically controlled attenuator consisting of a germanium wafer having a *P*, an *I* and an *N* region. The waveguides are generally rectangular in cross-section (0.83 × 1.66 mm, or 0.0325″ × 0.065″) and gold-plated inside and out. They are joined together by claw flanges, which allow particularly compact assemblies. Three components are described in some detail: the rotary attenuator, the variable impedance and the rotary directional coupler. The properties of these components are entirely or largely independent of frequency. The rotary directional coupler gives a coupling which is continuously variable from —3 to —30 or —40 dB, and can be read from a dial. The variable impedance is used in conjunction with a hybrid T and gives a direct reading of the modulus and argument of the reflection coefficient.

# X-RAY DETERMINATION OF CRYSTAL STRUCTURES

by P. B. BRAUN and A. J. van BOMMEL.

548.735

*Since the fundamental work of Von Laue and Bragg (1912), X-ray diffraction analysis has grown into an extremely important technique for studying the solid state. It has led to an impressive series of structure determinations, including such spectacular ones as vitamin B 12, strychnine and penicillin.*

*The article below deals with the principles underlying X-ray diffraction, discussing the relation between the structure and the diffraction patterns, and how the one can be derived from the other. By way of illustration, examples of structures determined in Eindhoven are discussed, special mention being made of work on a substance from a class of magnetic materials known as ferroxplana.*

Crystals of differing structure also differ in the way they scatter X-rays. This fact underlies a method of identifying crystalline substances, by means of the characteristic diffraction patterns to which these substances give rise when irradiated with X-rays [1]). (Diffraction patterns have therefore been described as the finger prints of crystals.) Diffraction patterns can moreover be used for *determining the crystal structures of substances.* The detailed analysis of the crystalline and molecular structure of penicillin is one of the remarkable results achieved in this way. As appears in *fig. 1*, the structure of the penicillin molecule is very complicated, and its elucidation was therefore a landmark in the development of X-ray analysis. Other outstanding results in this field have been the determinations of the structure of strychnine and of vitamin B 12 [2]).

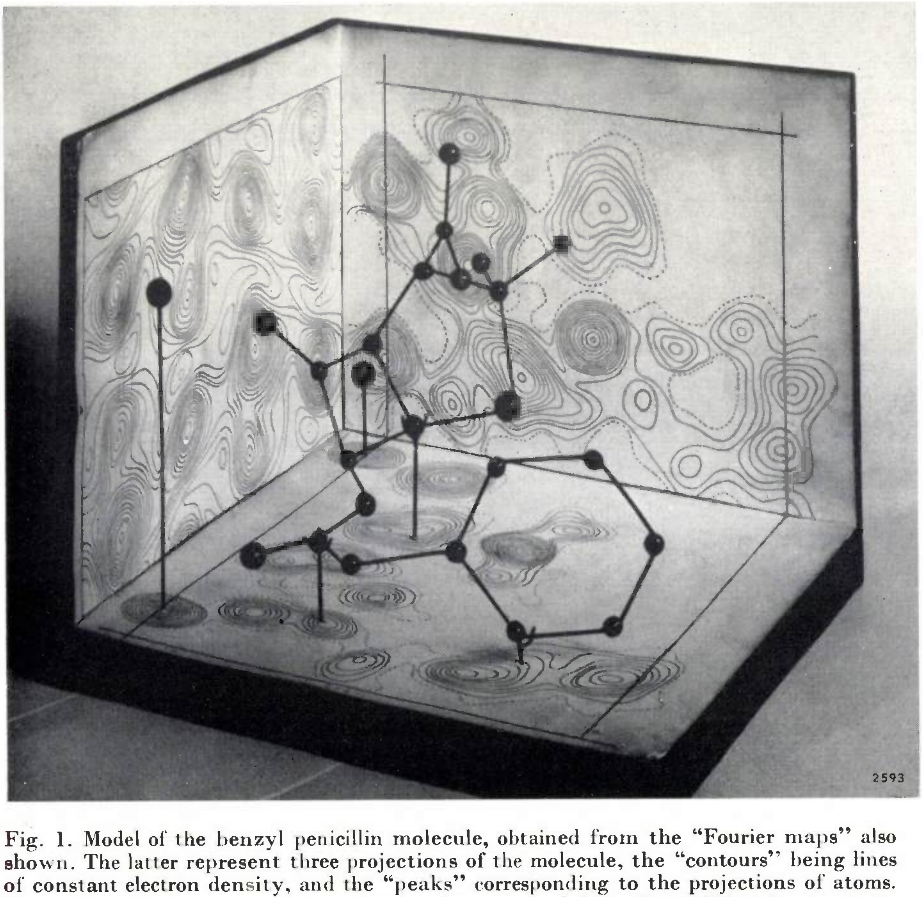


Fig. 1. Model of the benzyl penicillin molecule, obtained from the "Fourier maps" also shown. The latter represent three projections of the molecule, the "contours" being lines of constant electron density, and the "peaks" corresponding to the projections of atoms. These Fourier maps were constructed from data obtained from X-ray diffraction patterns. (From: E. Chain, Endeavour No. 28, Oct. 1948, p. 152.)

At the same time X-ray diffraction studies can lead to a better understanding of those physical and chemical properties that are related to the crystal structure of solids and molecules. Structure analysis is accordingly indis-

[1]) See e.g. W. G. Burgers, Philips tech. Rev. **5**, 157, 1940; W. Parrish and E. Cisney, Philips tech. Rev. **10**, 157, 1948/49.

[2]) D. Crowfoot, C. W. Bunn, B. W. Rogers-Low and A. Turner-Jones, The chemistry of penicillin, Princeton University Press, 1949. J. H. Robertson and C. A. Beevers, The crystal structure of strychnine hydrogen bromide, Acta cryst. **4**, 270-275, 1951. C. Bokhoven, J.C. Schoone and J. M. Bijvoet, The Fourier synthesis of the crystal structure of strychnine sulphate pentahydrate, Acta cryst. **4**, 275-280, 1951. D. Crowfoot Hodgkin et al., The structure of vitamin B 12, Proc. Roy. Soc. **A 242**, 228-263, 1957.

pensable in the study of ceramic magnetic materials, such as ferroxdure and ferroxplana [3]).

To understand the principles on which X-ray structure analysis is based, it is necessary first of all to know what information is contained in an X-ray diffraction pattern, and how the structure of a substance can be deduced from such a pattern. This will be the subject of the first half of this article [4]). The second half deals briefly with three examples taken from researches carried out in recent years at the Philips Research Laboratories in Eindhoven.

Electron and neutron diffraction analysis are analogous to X-ray diffraction analysis. These methods are all based on the same principles, but differ in technique. They also differ to some extent in the information they yield, so that they supplement one another very usefully. An article on neutron diffraction will appear in these pages in the near future; the various differences from X-ray diffraction will be further discussed there. The examples of structure determinations given in both articles will relate to the same substances, so that it will be readily possible to compare the two methods.

One final prefatory remark: X-rays are scattered solely by electrons. This means that the data derived from X-ray diffraction patterns can only relate to these particles. To determine a crystalline structure it is necessary to find the sites of the atoms, but this can be done by localizing the electron clouds. We can thus formulate the direct objective of the X-ray analysis of crystal structure as *the determination of the spatial distribution of the electron density in a crystal.*

### A useful way of describing the electron density in a crystal

It is characteristic of a crystal that it possesses three-dimensional periodicity. This means that it is possible to choose a volume element, by the regular stacking of which the whole crystal can be built up. The smallest volume element with which this can be done is called the *unit cell.*

The three-dimensional periodicity is, of course, shared by the *electron density.*

Now a periodic function having a period $d$ can always be resolved into a series of sine waves whose periods are $d$, $d/2$, $d/3$, etc. The process by which a function is resolved into these components is called *Fourier analysis*, whilst the building-up of a function from these components is referred to as *Fourier synthesis*.

In view of its periodicity, the electron density in a crystal can also be resolved into Fourier components. These constitute a collection of plane stationary waves having different directions. All the waves having any one direction comprise a set having wavelengths which are sub-multiples of the periodicity $d$ in that direction. The Fourier components are referred to here as *density waves* or *density components* (see *figs. 2* and *3*).

The electron density may thus be expressed as follows:

$$\varrho(x, y, z) = \sum_{h} \sum_{k=-\infty}^{+\infty} \sum_{l} A_{hkl}\, \mathrm{e}^{2\pi i\left(h\frac{x}{a}+k\frac{y}{b}+l\frac{z}{c}\right)}. \qquad (1)$$

[3]) See e.g. J. J. Went, G. W. Rathenau, E. W. Gorter and G. W. van Oosterhout, Philips tech. Rev. **13**, 194, 1951/52; G. H. Jonker, H. P. J. Wijn and P. B. Braun, Philips tech. Rev. **18**, 145, 1956/57.

[4]) Details of this subject will be found in such works as: J. M. Bijvoet, N. H. Kolkmeyer and C. H. MacGillavry, Röntgenanalyse van kristallen, Centen, Amsterdam 1948; H. Lipson and W. Cochran, The determination of crystal structures, Bell, London 1953; R. W. James, The crystalline state, Vol. II. The optical principles of the diffraction of X-rays, Bell, London 1948; J. Bouman, Theoretical principles of structural research by X-rays, Handbuch der Physik, Vol. **32**, 95-237, Springer, Berlin 1957. For the techniques used for recording diffraction patterns, see e.g. M. J. Buerger, X-ray crystallography, Wiley, New York 1942; H. P. Klug and L. E. Alexander, X-ray diffraction procedures, Wiley, New York 1954.

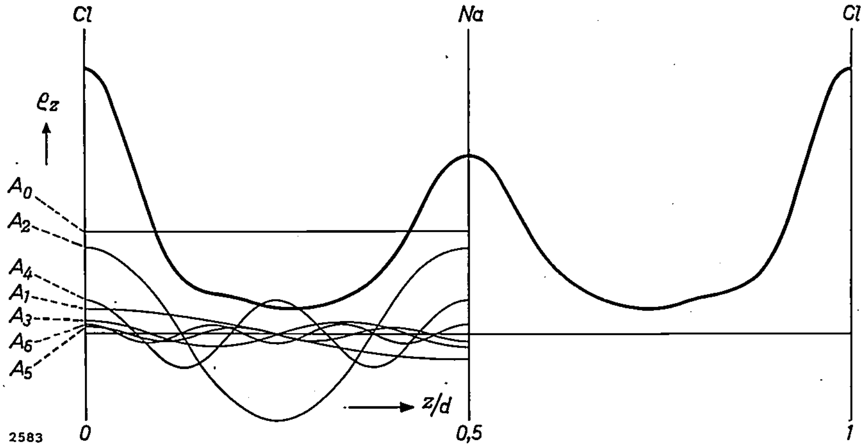


Fig. 2. The thick upper curve is obtained by adding together the lower series of cosine waves ($A_0$, $A_1$, $A_2$ etc.) whose wavelengths are all integral sub-multiples of $d$. The thick curve represents the electron density ($\varrho_z$) between the octahedral planes of rock-salt, while $A_0$, $A_1$, $A_2$, etc. are the amplitudes of the Fourier components of this curve, referred to here as "density waves". $z$ is the direction through the crystal (viz. normal to the octahedral planes), $d$ the lattice spacing in that direction.

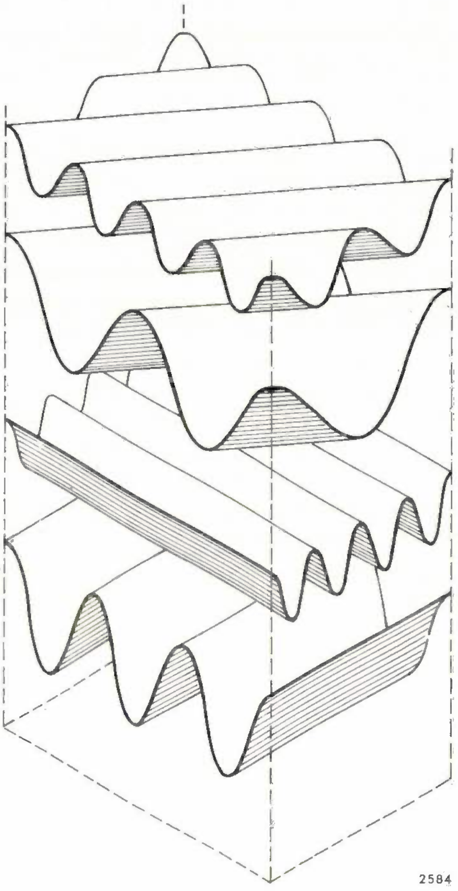


Fig. 3. Schematic representation of various density waves, in which the electron density is depicted by the geometrical amplitude. Two of the waves shown have the same direction, their wavelengths being in the ratio 1 : 2; the two other waves shown have different directions. To find the electron density distribution in a crystal, a number of such waves must be added together.

The term on the left is the electron density as a function of the coordinates $x$, $y$ and $z$. The terms on the right constitute the corresponding electron density waves. The wavelength and direction of each wave are expressed, with the aid of the integers $h$, $k$ and $l$, in terms of the longest repetition distances $a$, $b$ and $c$ of three waves whose directions do not lie in one plane [5]). $A_{hkl}$ is a complex quantity whose modulus and argument indicate the amplitude and phase, respectively, of the relevant density wave.

The volume of the parallelepiped whose edges are $a$, $b$ and $c$ will generally be taken as small as possible; in this way we thus arrive at the above-mentioned unit cell.

Sometimes to bring out the symmetry of the crystal, a unit cell is adopted that is an integral number of times larger than the smallest possible. In that case the quantity $A_{hkl}$ in formula (1) corresponding to given combinations of $h$, $k$ and $l$ is reduced to zero (a case of general extinctions).

We shall now show that the electron density waves can be deduced directly from diffraction patterns. We shall consider the simplest conceivable case, in which the electron density may be described in terms of a single density wave (this is never so in reality), and show what the form of the diffraction pattern is in that case. We shall then do the same for an electron distribution consisting of a number of density waves, as found in an actual crystal.

### Diffraction by electron density waves

The question in what directions a single density wave scatters is answered most easily if we consider the effect of successive "cross-sectional layers" of the density wave, as depicted in *fig. 4*. Each of these layers acts as a kind of mirror. The scattered radiation is thus to be sought in the direction corresponding to simple reflection from these layers.

In general, the total yield in reflected radiation from all layers is zero. A perceptible reflection occurs only under the condition that the angle of incidence $\Theta$ satisfies the relation:

$$2d \sin \Theta = \lambda, \qquad (2)$$

where $\lambda$ is the wavelength of the radiation and $d$ the wavelength of the density wave. In that case the amplitude of the reflected waves has a value proportional to the amplitude of the density wave.

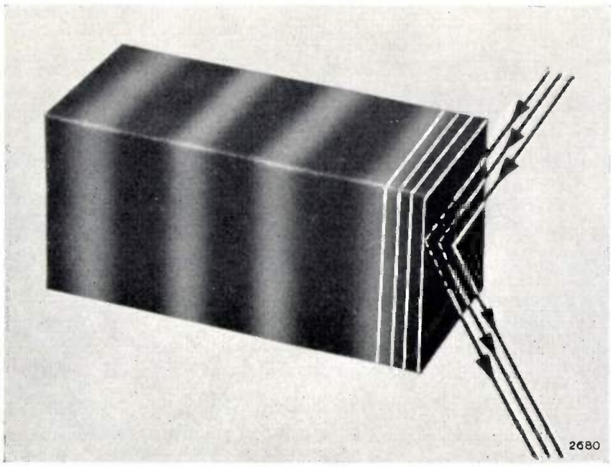


Fig. 4. Representation of a density wave with an X-ray beam incident upon it. The variation of the electron density in the wave is indicated by the shading. On the right, three "layers" of the density wave are represented, each of which acts as a reflector for X-rays. Rays reflected from the various layers interfere with one another and cause, in general, extinction, except at one specific angle of incidence, where a reflection of finite intensity occurs.

[5]) In the text books quoted the symbols $h$, $k$ and $l$ are given the meaning of indices of lattice planes. In a brief discussion of the principles of structural analysis, this is neither necessary nor desirable. No mention will therefore be made of lattice planes in this article, and the Bragg law will accordingly appear in a somewhat modified form.

These two relations, concerning the angle and the amplitude of reflections, are the foundations of the X-ray analysis of crystal structures. They show that a reflection provides information on no less than three properties of a density wave, viz. the amplitude, the wavelength and the direction. The intensity of the reflection is a measure of the amplitude of the density wave. Given the wavelength of the radiation, we can calculate the wavelength of the density wave from the angle of reflection. Thirdly the position of the crystal together with the direction and angle of the reflection determines the direction of the density wave.

The above can easily be demonstrated vectorially with the aid of *fig. 5*.

The difference in path length of the rays drawn in fig. 5 is equal to the component of the vector r along S minus the component of r along $S_0$, or:

$$\mathbf{r}\cdot(\mathbf{S}-\mathbf{S}_0).$$

The corresponding phase difference is:

$$\frac{2\pi\,\mathbf{r}\cdot(\mathbf{S}-\mathbf{S}_0)}{\lambda}.$$

The total amplitude $F$ of the radiation, caused by the scattering volume, is obtained by adding together the contributions from all volume elements d$V$, each having an electron density $\varrho(\mathbf{r})$, and taking the mutual phase differences into account:

$$F=\int_V \varrho(\mathbf{r})\,\mathrm{e}^{-2\pi i\frac{\mathbf{r}\cdot(\mathbf{S}-\mathbf{S}_0)}{\lambda}}\mathrm{d}V. \quad \dots\dots \quad (3)$$

Regarding the scattering body as an electron density wave, we can write for $\varrho(\mathbf{r})$:

$$\varrho(\mathbf{r})=A\,\mathrm{e}^{2\pi i\frac{\mathbf{r}\cdot\mathbf{z}}{d}}, \quad \dots\dots\dots \quad (4)$$

where z is the unit vector indicating the direction of the density wave, and $A$ and $d$ represent the amplitude and the wavelength, respectively. Using (4), the expression (3) becomes:

$$F=\int_V A\,\mathrm{e}^{2\pi i\,\mathbf{r}\left(\frac{\mathbf{z}}{d}-\frac{\mathbf{S}-\mathbf{S}_0}{\lambda}\right)}\mathrm{d}V. \quad \dots \quad (5)$$

Only when the exponent is zero will this integral have a value that is significantly different from zero.

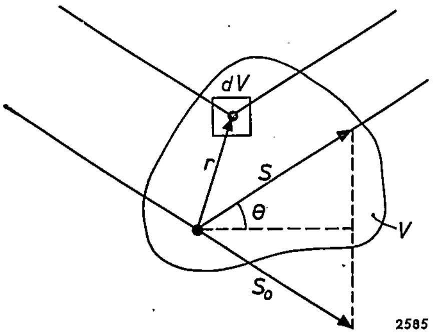


Fig. 5. Determination of the reflection condition for a density wave. $S_0$ and S are unit vectors denoting respectively the directions of incidence and scattering of an X-ray beam, two rays of which are shown. r is the vectorial distance between two scattering points, d$V$ is a volume element of the total volume $V$, and $\Theta$ is the semi-angle of reflection.

This gives us the condition for reflection:

$$\frac{\mathbf{z}}{d}=\frac{\mathbf{S}-\mathbf{S}_0}{\lambda}.$$

This means:

*a*) The two vectors ($\mathbf{S}-\mathbf{S}_0$) and z must have the same direction, i.e. the directions of incidence ($S_0$) and reflection (S) lie with z in one plane and make the same angle with z. In other words, the layers into which we have divided the density wave act as reflectors.

*b*) The magnitudcs of thc vectors must be equal. Since $|\mathbf{S}-\mathbf{S}_0| = 2\sin\Theta$ (see fig. 5), $1/d = 2\sin\Theta/\lambda$ or $2d\sin\Theta = \lambda$. These conditions being fulfilled, the integral gives:

$$F=VA, \quad \dots\dots\dots \quad (6)$$

i.e. the amplitude of the reflected ray is proportional to the amplitude of the electron density wave considered [6]).

We have seen that the electron density in a crystal can be described in terms of a group of density waves. Each such wave, independent of the other waves, is capable of "reflecting" in the manner discussed. A further point of importance is that each reflection can generally be intercepted separately from the other reflections. (Since no two waves are identical in direction as well as wavelength, they all "reflect" under different conditions.) This makes it possible to derive from each density wave the information present in its reflection.

To obtain this information it is convenient to use various different ways of irradiating the crystal(s). Weissenberg patterns are particularly valuable in structural analysis, but Laue and Debye-Scherrer patterns and rotation photographs are also often useful for this purpose (see *fig. 6*). The reflections may be recorded photographically or they may be measured with the aid of Geiger counters, proportional counters or scintillation counters [7]).

### The underlying principle of structure analysis

Summarizing the foregoing, the principles of the X-ray determination of crystal structures may be expressed as follows.

The electron density in a crystal can be described as the sum of a number of Fourier components. Each

[6]) In the derivation given here we have assumed that a ray, once reflected, leaves the crystal without again being reflected. The so-called "dynamic theory", which takes account of multiple reflections, produces slightly different results from the "kinematic theory" used here. The latter theory proves to be adequate for most crystals in practice, but corrections may sometimes be called for (extinction corrections). The situation is different where perfect crystals with no defects are concerned. In their case "dynamic" effects are very important. See e.g. L. P. Hunter, Anomalous transmission of X-rays by single crystal-germanium, Proc. Kon. Ned. Akad. Wet. **B 61**, 214-219, 1958.

[7]) See e.g. W. Parrish, E. A. Hamacher and K. Lowitzsch, The "Norelco" X-ray diffractometer, Philips tech. Rev. **16**, 123-133, 1954/55; P. H. Dowling, C. F. Hendee, T. R. Kohler and W. Parrish, Counters for X-ray analysis, Philips tech. Rev. **18**, 262-275, 1956/57.

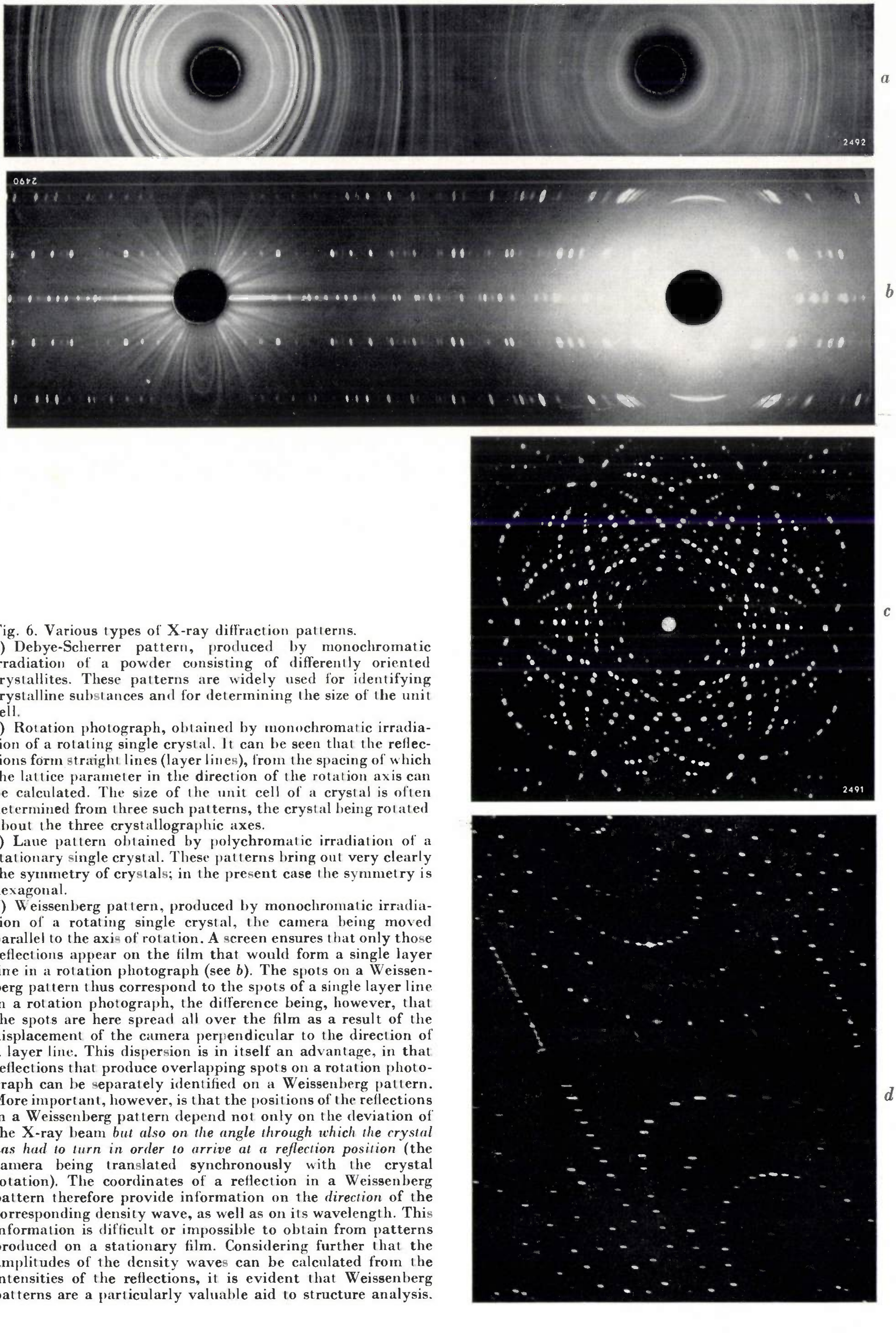


Fig. 6. Various types of X-ray diffraction patterns.
*a*) Debye-Scherrer pattern, produced by monochromatic irradiation of a powder consisting of differently oriented crystallites. These patterns are widely used for identifying crystalline substances and for determining the size of the unit cell.
*b*) Rotation photograph, obtained by monochromatic irradiation of a rotating single crystal. It can be seen that the reflections form straight lines (layer lines), from the spacing of which the lattice parameter in the direction of the rotation axis can be calculated. The size of the unit cell of a crystal is often determined from three such patterns, the crystal being rotated about the three crystallographic axes.
*c*) Laue pattern obtained by polychromatic irradiation of a stationary single crystal. These patterns bring out very clearly the symmetry of crystals; in the present case the symmetry is hexagonal.
*d*) Weissenberg pattern, produced by monochromatic irradiation of a rotating single crystal, the camera being moved parallel to the axis of rotation. A screen ensures that only those reflections appear on the film that would form a single layer line in a rotation photograph (see *b*). The spots on a Weissenberg pattern thus correspond to the spots of a single layer line in a rotation photograph, the difference being, however, that the spots are here spread all over the film as a result of the displacement of the camera perpendicular to the direction of a layer line. This dispersion is in itself an advantage, in that reflections that produce overlapping spots on a rotation photograph can be separately identified on a Weissenberg pattern. More important, however, is that the positions of the reflections in a Weissenberg pattern depend not only on the deviation of the X-ray beam *but also on the angle through which the crystal has had to turn in order to arrive at a reflection position* (the camera being translated synchronously with the crystal rotation). The coordinates of a reflection in a Weissenberg pattern therefore provide information on the *direction* of the corresponding density wave, as well as on its wavelength. This information is difficult or impossible to obtain from patterns produced on a stationary film. Considering further that the amplitudes of the density waves can be calculated from the intensities of the reflections, it is evident that Weissenberg patterns are a particularly valuable aid to structure analysis.

density component gives rise to an X-ray reflection which provides information on the amplitude, wavelength and direction of that particular component. From all the reflections from a crystal we have a (virtually) complete picture of all the electron density components, and hence in principle of the electron distribution in a crystal. The determination of this electron distribution — and with it the structure of the crystal — amounts to carrying out a Fourier synthesis, i.e. adding all components of the Fourier series to yield the required electron distribution function. It should be noted that the synthesis involved is three-dimensional, and is therefore extraordinarily complicated. A common practice is therefore to carry out one-dimensional and two-dimensional Fourier syntheses, which produce projections of the structure (projections on a line and on a plane, respectively). From two or three such projections it is possible to deduce the spatial configuration of the atoms. (This is well brought out in fig. 1.)

### The phase problem

However close the solution may now seem, there is still a lot to be done before we can determine the structure of a crystal from the summing of the density components. Although at first sight it might seem that a mere addition is called for, closer examination shows that one quantity is still missing. This brings us up against an obstacle briefly referred to as the "phase problem". Overcoming this obstacle often costs the crystallographer most of the time he spends on analysing the structure of a crystal.

To carry out the Fourier summation correctly we must know, apart from the amplitude, wavelength and direction of the constituent waves, the *phase* of the waves, which is given by the argument of $A_{hkl}$ in equation (1). This is demonstrated in *fig. 7.*

Unfortunately, the phase of the density components cannot be observed; unlike the amplitude, wavelength and direction, it cannot be deduced from the reflections. The intensity of a reflected ray is a measure of the amplitude of the corresponding density wave, but tells us nothing about its phase.

It is for this reason that the diffraction pattern of a crystal cannot be directly interpreted to give a pattern of the electron density in a crystal. Not until the phases have been determined, by the methods about to be discussed, is it possible to carry out the superposition of Fourier components and so produce the required picture of the electron clouds which show the arrangement of the atoms or molecules in a crystal.

### Phase determination

Just as knowledge of the phases of the reflections found by experiment would enable us to solve the structure of a crystal, it is possible conversely to calculate the phases from a known structure. This implies a possible method of solving the phase problem, for if we can produce a model representing a reasonable approximation of the true structure, the phases of the strongest reflections can often be determined with a fair degree of accuracy. These data are then used for carrying out a (provisional) Fourier synthesis. Although the picture obtained in this way is usually very rough, it often leads to a better approximation to the actual structure than the original model. With this improved approximation, better calculation of the phases can be made, and so on. In this procedure, phase calculation alternates with Fourier synthesis, thus producing step by step an increasingly more accurate picture of the electron density in the crystal.

The question now is how to arrive at an initial model to start the above procedure. The search for a suitable model leads along well-trodden paths, each attempt requiring from the investigator considerable crystallographic experience, deductive ability and intuition. For each attempt we have to

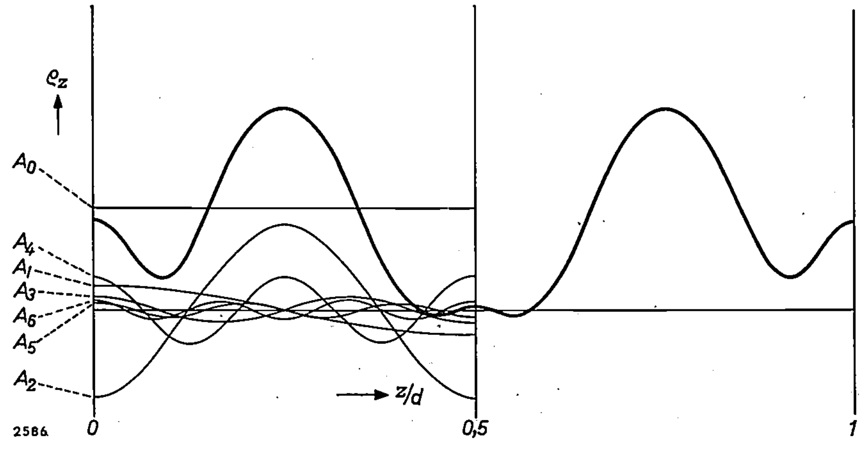


Fig. 7. Result of a Fourier synthesis, done with the same Fourier components as in fig. 2, but with the component $A_2$ having a different phase. The result is entirely different from the density distribution obtained in fig. 2.

ask the question: what is known of the crystal under investigation, apart from the X-ray data, which can help us to make a provisional guess at the structure?

In the first place the constituent atoms of a crystal are usually known from chemical analysis. The scattering contribution of each atom ("atomic structure factor") has been calculated for all elements and these are published in tables. We can also measure the density of crystals and calculate the number of atoms contained per unit volume. For our purposes it is particularly interesting to know the number of atoms contained in the unit cell, since the dimensions of this cell can be derived from the diffraction pattern. Having ascertained the number of atoms of each kind contained in this volume of known dimensions, it remains to find the spatial arrangement of the atoms.

Obviously, only those arrangements will be considered that agree with the properties of the crystal (chemical, optical, magnetic, morphological and particularly the symmetry properties). Taking these data into account, many possibilities can be eliminated, but it is hardly ever possible to reduce the remaining possible structures to one. Nevertheless, if only a few models are left after this elimination, it is now usually a practicable proposition to try each of them in turn. This is referred to simply as the "trial and error" method.

Knowledge of the *interatomic distances* can be of great help in the construction of a model. This knowledge is useful in the process of elimination described, and is sometimes decisive for the success of the analysis.

Information on the distances (magnitude and direction) between the atoms in a crystal (but not the sites of the atoms) can be obtained by Patterson's method. This is based on the dependence of the intensities of the reflections on the interatomic distances. The method consists in carrying out a so-called Patterson synthesis, which is a Fourier synthesis in which the components to be summed are simply the reflection intensities. This can be done *without* the phases of the reflections being known.

To show that something can indeed be learnt about interatomic distances from the reflection intensities, we shall first write formula (1) in a simpler notation:

$$\varrho(\mathrm{r}) = \sum_{\mathbf{H}} A_{\mathbf{H}}\, \mathrm{e}^{2\pi i\, \mathbf{H}\cdot \mathrm{r}}\,, \quad \ldots\ldots \quad (7)$$

where r is the vector from the origin to the point $x, y, z$ and $\mathbf{H}\cdot\mathrm{r}$ represents $h\,\frac{x}{a} + k\frac{y}{b} + l\,\frac{z}{c}$.

From this equation we shall now derive another in which the intensities occur. In this connection the following points should be noted: *a*) $A_{\mathbf{H}}$ is in general a complex quantity; *b*) the square of the modulus of this quantity, $|A_{\mathbf{H}}|^2$ or $A_{\mathbf{H}}A_{\mathbf{H}}^*$, where * denotes the complex conjugate of $A_{\mathbf{H}}$, is related, from eq. (6), to the intensity $I_{\mathbf{H}}$ ($\propto FF^*$) of the corresponding reflection by:

$$A_{\mathbf{H}}A_{\mathbf{H}}^* = \frac{I_{\mathbf{H}}}{V^2}. \quad \ldots\ldots\ldots \quad (8)$$

Of course, eq. (7) is also valid when $A_{\mathbf{H}}$ is replaced by its complex conjugate. Moreover, the equation also remains valid when the independent variable r is replaced by r–u, where u is any arbitrary vector. (By this device, as will be shown, we introduce the interatomic distance u.) We may thus write:

$$\varrho^*(\mathrm{r{-}u}) = \sum_{\mathbf{H}} A_{\mathbf{H}}^*\ \mathrm{e}^{-2\pi i\mathbf{H}\cdot(\mathrm{r}-\mathrm{u})}. \quad \ldots \quad (9)$$

Multiplication of (7) and (9), distinguishing the vectors **H** in (7) and (9) by **H** and **H′**, gives:

$$\varrho(\mathrm{r})\varrho^*(\mathrm{r{-}u}) = \sum_{\mathbf{H}}\sum_{\mathbf{H}'} A_{\mathbf{H}}A_{\mathbf{H}'}^*\ \mathrm{e}^{2\pi i\mathbf{H}'\cdot\mathrm{u}}\ \mathrm{e}^{2\pi i(\mathbf{H}-\mathbf{H}')\cdot\mathrm{r}}. \quad \ldots\ldots \quad (10)$$

We shall henceforth omit the * sign in $\varrho^*$(r–u), since this quantity is real and thus identical to its complex conjugate. Integrating both sides over the whole volume of the unit cell gives:

$$\int_V \varrho(\mathrm{r})\varrho(\mathrm{r{-}u})\ \mathrm{d}V(\mathrm{r}) = $$
$$= \sum_{\mathbf{H}}\sum_{\mathbf{H}'} A_{\mathbf{H}}A_{\mathbf{H}'}^*\,\mathrm{e}^{2\pi i\mathbf{H}'\cdot\mathrm{u}} \int_V \mathrm{e}^{2\pi i(\mathbf{H}-\mathbf{H}')\cdot\mathrm{r}}\ \mathrm{d}V(\mathrm{r}). \quad (11)$$

The integral on the right is always zero, except when $\mathbf{H} = \mathbf{H}'$, in which case the integral is equal to $V$. Using (8) we may thus write for (11):

$$\int_V \varrho(\mathrm{r})\varrho(\mathrm{r{-}u})\,\mathrm{d}V(\mathrm{r}) = \frac{1}{V}\sum_{\mathbf{H}} I_{\mathbf{H}}\ \mathrm{e}^{2\pi i\mathbf{H}\cdot\mathrm{u}}. \quad \ldots \quad (12)$$

The right-hand side represents a Fourier summation, performed with intensities instead of amplitudes; compare (12) with (7). The significance of this Patterson function, as it is called, appears from the left-hand side: it is a volume integral of the product of two electron densities at places lying at vector distances of u apart. This implies that the integral in question will assume high values when u is a distance between peaks in electron density, i.e. when u is an interatomic distance really present in the crystal. By performing a Fourier synthesis, using the intensities of the reflections as Fourier coefficients, we thus obtain a function whose maxima provide indications of the atomic distances.

The clues to interatomic distances obtained by the Patterson method are sometimes so clear that they reveal the whole structure; sometimes they throw light only on certain details of the structure. Much depends on the ingenuity of the researcher's interpretation of the Patterson synthesis. However this may be, information is always present, which can help in the construction of the model.

We should mention finally that other methods exist which lead directly, i.e. not via models, to the phase constants required. One such is the "method of inequalities", which is based on the fact that the electron density in a crystal can nowhere be negative, and another is the "statistical method", where use is made of knowledge concerning the constituent atoms of the crystal. Particulars of these methods will be found in the treatise by Bouman, mentioned in footnote [4]).

These, then, are the broad lines of structure determination. We shall now say something about the "strategy".

We have seen that each reflection corresponds to a density componcnt, providing information on its amplitude, wavelength and direction, and that to determine the phases a provisional model is devised. Now every incorrect model is shown to be wrong only after a timeconsuming procedure. The less probable models should therefore be avoided as far as possible. The time can better be used to look far a more reasonable model; again, the search for such a "good" model should not take too long. An important aspect of structure analysis is thus the art of finding, with the available data, the shortest way to this "good" model. Since the data involved are different in every case, one should always be prepared for other possibilities and difficulties, and be ready to try one way after another. The question often arises later whether one should persist with a choice once made, when the work based on the model adopted does not lead directly to results. In this way, experimenting and doubting, the crystallographer may be busy for months before he sees the first signs of success, indicating that he may at last start on the actual determination of the structure. This element of uncertainty imparts to structure determinations the interest and excitement of adventure.

## Examples

Having sketched the broad outlines of the X-ray analysis of crystal structure, we shall now give a more detailed description of three cases, the first two being quite simple and making only partial use of the above procedures, and the third relating to a complicated structure, which necessitated a broader approach.

### *Analysis of the compound $Al_{0.89}Mn_{1.11}$*

It was known as early as 1900 that various alloys of manganese are ferromagnetic, even though manganese itself is not ferromagnetic (Heusler alloys [8])). The compound $Al_{0.89}Mn_{1.11}$ having turned out to be a material suitable for permanent magnets, a study was made of the relation between the magnetic saturation and the structure of this substance.

The analysis of this structure went very smoothly. The Debye-Scherrer diagram showed the (tetragonal) unit cell to be very small, which indicated that there were not many different possibilities as to the atomic positions. It further appeared from the density and the chemical formula that the unit cell could in fact contain no more than two atoms. Finally, the symmetry found left no other choice but to place the atoms at the corners and at the centre.

It is very rare that the atomic sites can be determined with such accuracy at such an early stage of an analysis. Usually, it may be recalled, only the approximate positions of the atoms would be known at this stage, and on the basis of this rough model one would try by Fourier syntheses to obtain a more accurate picture of the structure. In this case.that was not necessary because the symmetry had already pointed the way to the precise positions for the atoms.

This would have concluded the structural analysis had it not been that the non-stoichiometric formula $Al_{0.89}Mn_{1.11}$ showed that the aluminium and manganese atoms could not possibly occur in a completely regular arrangement in the crystal. If, for example, all aluminium atoms were situated at the corners and all manganese atoms in the centres, the formula would be AlMn exactly. This suggested that the atoms were not so strongly bound to their own position in the cell, in other words that aluminium atoms might well occupy some of the sites of manganese atoms, and vice versa (the latter case prevailing).

In such cases we want to know the ratios in which the various kinds of atoms occur at the various sites. This information, too, can be deduced from the intensities of the reflections. If in this case we calculate the reflection intensities on the basis of a unit cell that is everywhere identical — containing for example one manganese atom at site A (corner) and one aluminium atom at site B (centre) — the results obtained will differ from the observed values. The effect on the intensities of the random atomic distribution described is as if each lattice site contained not one single particular atom but fractions of different atoms. These fractions on one site must, of course, together add up to one; and if we add the fractions of the same atoms in their different positions, the sum must obviously be in agreement with the chemical formula. In our case we can therefore write:

[8]) F. Heusler, Über Manganbronze und über die Synthese magnetisierbarer Legierungen aus unmagnetischen Metallen, Z. angew. Chem. 17, 260-264, 1904.

| | Site A | Site B |
|---|---|---|
| Aluminium | $r$ | $0.89 - r$ |
| Manganese | $1 - r$ | $0.11 + r$, |

where $r$ is the fraction of the number of A sites occupied by aluminium atoms.

We now calculate the reflection intensities for various values of $r$, and find the value of $r$ that best agrees with the observed intensities. In our case this value was 0.03.

Finally, there was the crucial question whether this structure indeed corresponded to the magnetic properties of the substance. It was found, for example, that the magnetic saturation decreased when the substance was subjected to mechanical deformation. An X-ray investigation of this phenomenon revealed that $r$ increases under these conditions to 0.13.

This meant that the drop in magnetic saturation must be accompanied by an increase in the number of manganese atoms on the "wrong" B sites (only the manganese atoms are considered, since they alone possess a magnetic moment). The conclusion was drawn that the magnetic moment of the manganese atoms on the B sites must be opposite in orientation to that of the manganese atoms on the A sites.

We shall not go further into this subject, but it may be mentioned that neutron diffraction analysis has confirmed this conclusion, making it possible to relate the magnetic saturation quantitatively with the above picture of the structure.

### *Analysis of the compound $Th_2Al$*

In the course of research on thorium-aluminium compounds, which are important for their gettering properties, the structure of $Th_2Al$ was analysed [9]. This is an example of a fairly simple "trial and error" analysis, calling for no Fourier syntheses.

In this case the unit cell (again with tetragonal symmetry) contained four aluminium and eight thorium atoms. On grounds of crystal symmetry, the aluminium atoms could immediately be assigned to special positions in the cell, as was done for all atoms in the first example. It was not immediately clear where the eight thorium atoms should be located. The task of determining the 24 coordinates of these atoms was fortunately simplified by the fact that the symmetry severely limited the number of possibilities, inasmuch as a choice of one coordinate established the other 23.

The procedure thereafter was as follows. Various values of this coordinate were chosen, the reflection intensities were calculated with the models produced and the results were compared with the observed intensities. In this way the structure was determined by trial and error.

Here, too, the structure found threw light on to the other properties of the substance. The structure contains fairly large tetrahedral interstices (see *fig. 8a* and *b*). These interstices may be supposed to

[9] See also P. B. Braun and J. H. N. van Vught, Acta cryst. **8**, 246, 1955.

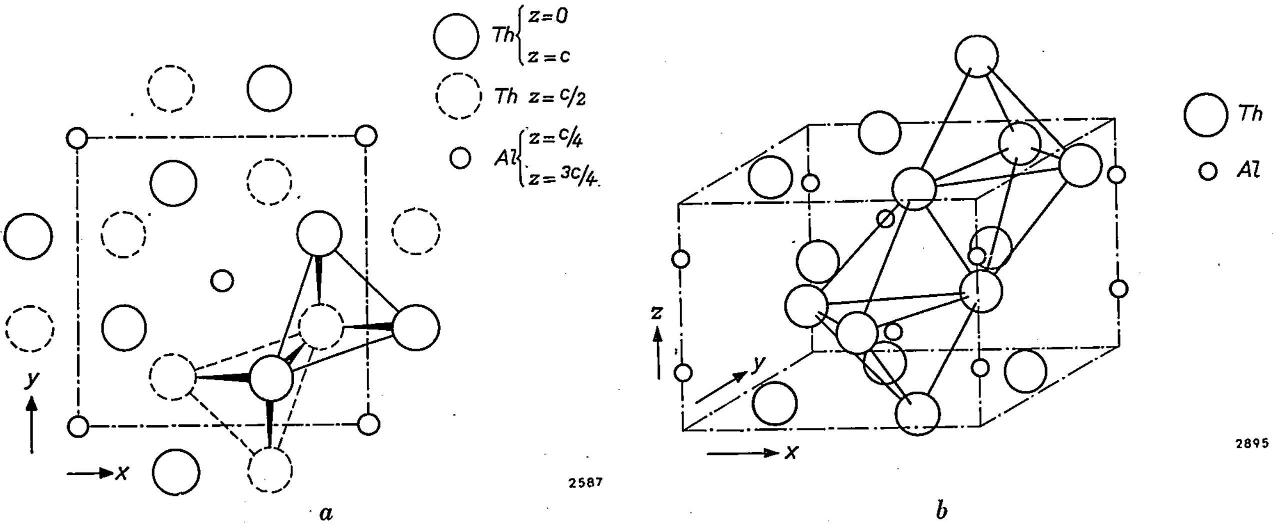


Fig. 8. The unit crystal cell of $Th_2Al$, a substance with gettering properties. The thorium atoms form tetrahedrons, a number of which are shown. Absorbed gas atoms or ions take up positions in the interstices of these tetrahedrons. *a*) Projected positions of the atoms looking along the $c$ axis (5.86 Å). The levels of the various atoms ($z$) are given in fractions of the $c$ axis. The $x$ and $y$ directions are those of the $a$ and $b$ axes, respectively (both 7.62 Å). *b*) Perspective view of unit cell, showing atomic positions. The $x$, $y$ and $z$ directions are again the $a$, $b$ and $c$ axes.

have some connection with the gettering properties of the substance. This supposition was later confirmed by nuclear magnetic resonance measurements [10]) and by neutron diffraction analysis, from which it was possible to demonstrate the presence of gas atoms or ions in these interstices.

### *Analysis of the compound* Y-$Ba_2Zn_2Fe_{12}{}^{III}O_{22}$

As our last example we shall take the analysis of a highly complex structure. Y-$Ba_2Zn_2Fe_{12}{}^{III}O_{22}$ is a substance from the class of ceramic magnetic materials which have been given the collective name "ferroxplana" [11]). These materials, whose properties resemble those of the familiar ferroxcube materials, can be used up to even higher frequencies than the latter (>100 Mc/s).

The fact that the unit cell in this case contains 114 atoms, viz. 6 barium, 36 iron, 6 zinc and 66 oxygen atoms, makes it evident that this analysis was a great deal more difficult than the two described above. The following outline of the analysis will give a good general picture of how such a complicated structure can be tackled.

One could not start on such a structure if it were not for the knowledge gained of simpler related structures. In this case, for example, we assumed that parts of the structure must resemble the structure of spinel [12]).

With the aid of Debye-Scherrer patterns, Laue patterns and rotation photographs (the first obtained with an X-ray diffractometer) it was found that the crystal is hexagonal and that the unit cell has one very long edge (*c* axis, 43.6 Å) and two short edges (*a* and *b* axes, 5.9 Å).

Our first object was to find the position of the heaviest atoms, barium, iron and zinc. These contribute most to the scattering and therefore dominate the intensity pattern of the reflections. For this reason the position of such atoms must be accurately known. Only then is there any sense in looking for the sites of the lighter atoms. At this stage, however, no distinction could be made between iron and zinc atoms, the difference in their scattering power being relatively small.

It was assumed, from analogies and other indications, that the structure concerned would be one of layers perpendicular to the hexagonal axis, consisting mainly of closely packed oxygen and barium atoms, most of the other atoms being in the interstices. First of all we tried to establish in which layers the barium atoms occurred. To do this it was sufficient to investigate a one-dimensional projection of the unit cell perpendicular to the hexagonal axis. The projection of the structure on to a plane was determined at a later stage, and finally the three-dimensional structure.

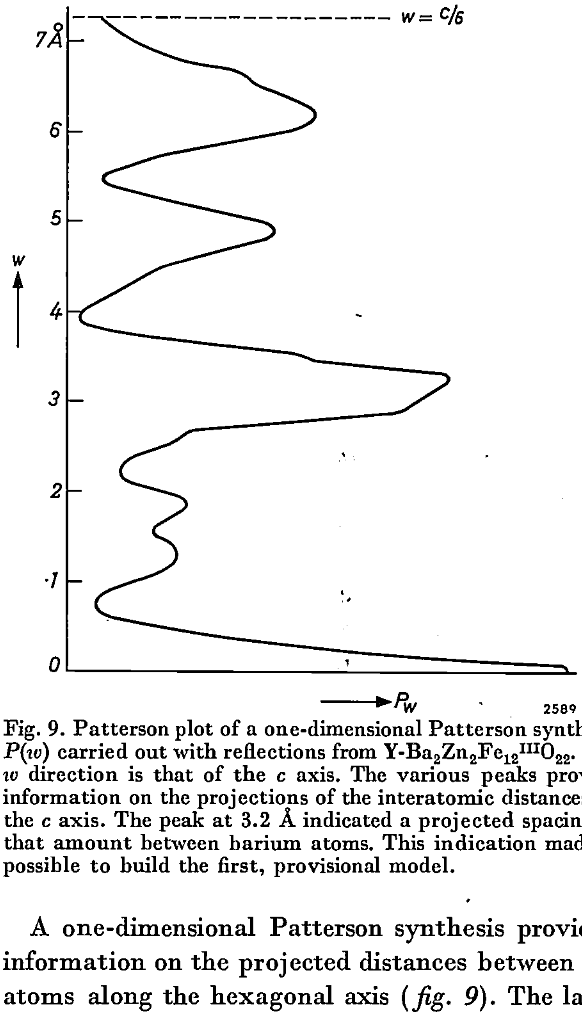


Fig. 9. Patterson plot of a one-dimensional Patterson synthesis $P(w)$ carried out with reflections from Y-$Ba_2Zn_2Fe_{12}{}^{III}O_{22}$. The *w* direction is that of the *c* axis. The various peaks provide information on the projections of the interatomic distances on the *c* axis. The peak at 3.2 Å indicated a projected spacing of that amount between barium atoms. This indication made it possible to build the first, provisional model.

A one-dimensional Patterson synthesis provided information on the projected distances between the atoms along the hexagonal axis (*fig. 9*). The large peak at 3.2 Å was taken as an indication of the projected spacing between the heavy barium atoms, || to the *c* axis. Knowledge of this spacing (however inaccurate due to the width of the peak) decided the choice between a number of models already drawn up on the basis of analogies. With the model thus obtained we set out to calculate the phases of the reflections.

This model represented, of course, a very rough structure, devoid of detail: all the light atoms and even part of the iron or zinc atoms had been disregarded. What is more, it was still only a projection. Using this incomplete model we determined the phases of 15 reflections and then carried out a Fourier synthesis. The resultant picture contained,

[10]) See D. J. Kroon, Nuclear magnetic resonance, Philips tech. Rev. **21**, 286-299, 1959/60 (No. 10), in particular page 297.

[11]) See P. B. Braun, The crystal structures of a new group of ferromagnetic compounds, Philips Res. Repts. **12**, 491-548, 1957, and G. H. Jonker, H. P. J. Wijn and P. B. Braun, Philips tech. Rev. **18**, 145-154, 1956/57.

[12]) Cf. E. J. W. Verwey, P. W. Haaijman and E. L. Heilmann, Philips tech. Rev. **9**, 185, 1947/48.

as we had hoped, numerous indications regarding the possible situations of the other atoms. This enabled us to venture on to new models, which were successively supplemented with more of the iron atoms still unused. When devising these models we made repeated checks with the Patterson synthesis (as we did on several occasions later) to ensure that the models chosen were not in conflict with the Patterson plot.

The model containing all the iron atoms was a good step forward, but still not sufficient to allow us to distinguish between the iron and the zinc atoms and to find sites for the light oxygen atoms. We had the impression that this was because one of the iron atoms was not yet on its proper site. By analogy with the structures of other similar materials, we had allocated it a position in the same plane as the barium atoms. Various other positions were tried, at first with little success. Finally it was found that the answer was to place this iron atom just outside the plane containing the barium atoms. After this it was a fairly straightforward procedure to complete

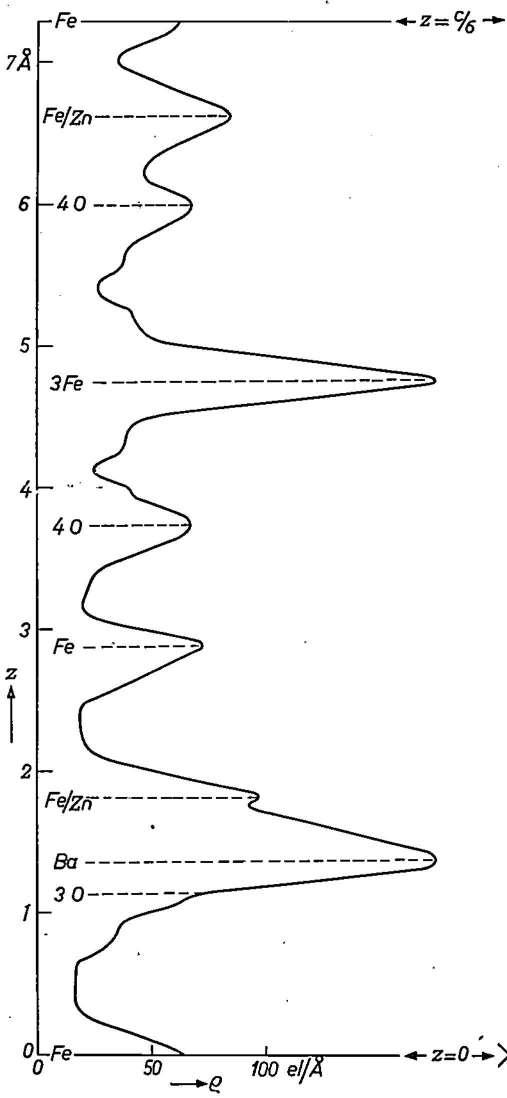


Fig. 10

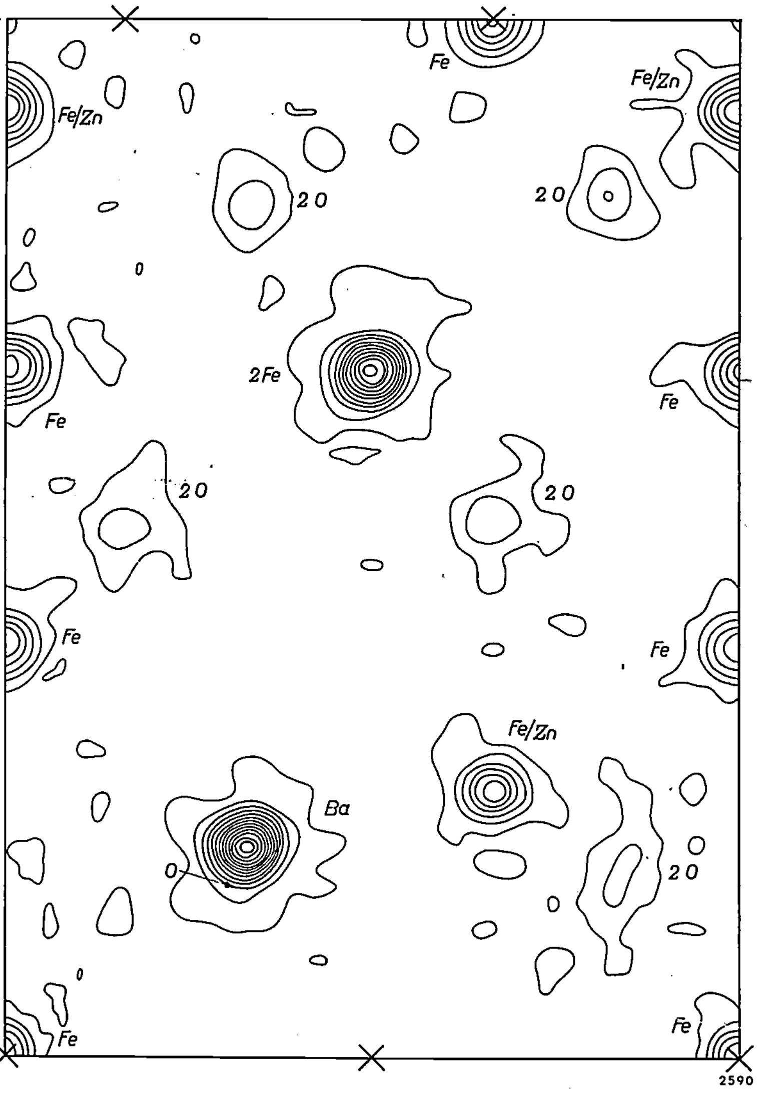


Fig. 11

Fig. 10. Result of a one-dimensional Fourier synthesis, obtained from the reflections from Y-$Ba_2Zn_2Fe_{12}^{III}O_{22}$. The curve represents the variation of electron density (in electrons per ångström) in one sixth of the unit cell projected on to the $c$ axis ($z$ direction). The positions of the various projected atoms are denoted by their chemical symbols. This concluded the first stage of the analysis, in which only the projection on the $c$ axis was considered. In order to reach this result, however, it was necessary to make many trials to determine the location of the iron (or zinc) atom, the peak of which is just perceptible beside the large peak for the barium atom. The next stage in the investigation is illustrated by fig. 11.

Fig. 11. Map of a two-dimensional Fourier synthesis, obtained from the reflections from Y-$Ba_2Zn_2Fe_{12}^{III}O_{22}$. The content of one sixth of the unit cell is projected on to a plane. The contours shown represent lines of equal electron density. The "peaks" indicate the presence of atoms, denoted by their chemical symbols. The small peaks in the map are due to unavoidable experimental errors. The crosses are centres of symmetry. The $z$ direction represents the $c$ axis. The relation of the map to the projection of fig. 10 is seen directly simply by projecting the atoms on to the $c$ axis. (It is only necessary to determine the one-sixth of the unit cell shown, because the rest of the unit cell can be obtained by inversion operations with respect to the centres of symmetry.)

the structure determination and extend it into two and three dimensions, thereby localizing the zinc and oxygen atoms. It is interesting to note that the zinc atoms and some of the iron atoms were found to compete for the same sites (exactly as in the case of the manganese and aluminium atoms in $Al_{0.89}Mn_{1.11}$). In this way, then, the final structure was arrived at step by step. *Figures 10, 11* and *12* mark the three most important steps of the determination.

The result of a one-dimensional Fourier synthesis (projection on the $c$ axis), carried out after all the required phases had been calculated, is shown in fig. 10. Note the small peak beside the large peak of the barium atom: this is the iron atom that caused so much trouble. A two-dimensional Fourier synthesis — obtained analogously with the aid of two-dimensional models and the phases calculated on the basis of these models — gives the electron density of the structure projected on to a plane (fig. 11). (Projection of the peaks in this figure on to the $c$ axis again produces the one-dimensional projection of fig. 10.) Finally, fig. 12 illustrates the three-dimensional structure found. Looking along the $a$ axis we can recognize, between $z = 0$ and $z = c/6$, the projection given in fig. 11.

Attention may be called to some of the more important features of the structure thus determined. The iron atoms are located in octahedral interstices (atom $C$ in fig. 12) or in tetrahedral interstices (atom $B$ in fig. 12). Some of the tetrahedral

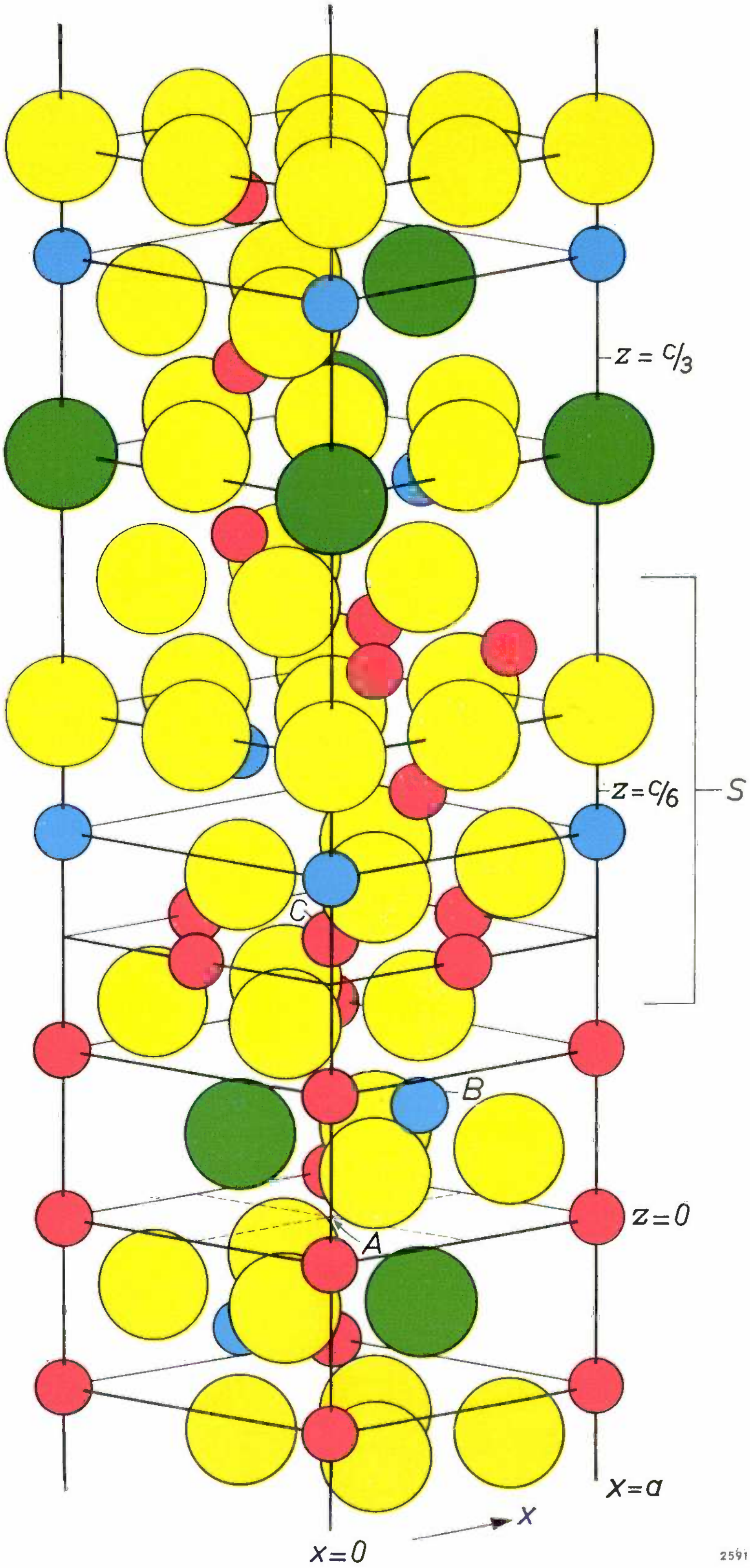


Fig. 12

Fig. 12. Three-dimensional model of part of the unit cell of $Y\text{-}Ba_2Zn_2Fe_{12}^{III}O_{22}$. The coloured spheres indicate the positions of the various atoms: green = barium atoms, red = iron atoms in octahedral interstices, blue = iron or zinc atoms in tetrahedral interstices, yellow = oxygen atoms. The long $c$ axis runs in the vertical ($z$) direction. A projection of the cell between $z = 0$ and $z = c/6$ looking along the $a$ axis ($x$ direction) produces the map of fig. 11. In this three-dimensional representation it is possible to distinguish layers consisting of close-packed oxygen and barium atoms. The iron or zinc atoms are contained in certain of the oxygen interstices. Other important features of the structure are the presence of centres of symmetry (shown only at $A$, but others shown by × in fig. 11) and of three 3-fold axes parallel with the $c$ axis (not indicated). $S$ denotes a block possessing the same structure as spinel. The fact that the interstices may be surrounded by either four or six oxygen atoms (tetrahedral or octahedral interstices) is clearly to be seen at $B$ and $C$, respectively. $B$ is also the iron (or zinc) atom which, as mentioned in the text, caused such trouble.

interstices are occupied by zinc atoms instead of iron atoms. Further, at either side of centres of symmetry (such as $A$ in fig. 12) barium atoms are found which are relatively close together and are located in adjacent layers. (This had already been established, it may be recalled, from the result of the one-dimensional Patterson synthesis in fig. 9.) Owing to the proximity of these "large" atoms, the iron atom (or zinc atom) $B$ is "pushed aside" slightly, so that this atom lies outside the plane, perpendicular to the $c$ axis, through the layer of barium atoms (this, too, had emerged at an earlier stage, see fig.10). [13]

The knowledge gained of the structure has been used for estimating the magnetic interactions between the various iron atoms, as described in this journal some years ago [11]. It was found that all iron atoms on tetrahedral sites have parallel oriented magnetic moments, while those of the iron atoms on octahedral sites are for the most part anti-parallel. Such a distribution of the magnetic orientations explains the low magnetic saturation of this substance.

Another investigation concerned the property which this substance shares in common with other materials of the "ferroxplana" group, namely the existence of a preferred plane of magnetization. In this plane the magnetization is fairly free to rotate, enabling these substances to be used as a "soft" magnetic material in rapidly alternating magnetic fields. The magnetic moments are tightly bound to the plane perpendicular to the $c$ axis. Various calculations have been made to explain this effect, but they cannot be dealt with here [14]. It will suffice to remark that, as illustrated by the above "case histories", structure determinations can evidently lead to a better understanding of the properties of the substance investigated.

[13] A full description is not possible in the compass of this article. No mention is made, for example, of the method of least squares, of difference Fourier synthesis, etc. A full description will be found in the article in Philips Research Reports quoted under [11].

[14] See § 39 in "Ferrites" by J. Smit and H. P. J. Wijn, Philips Technical Library, 1959.

**Summary.** The principles and some applications of the X-ray determination of crystal structures are discussed. The electron distribution in a crystal can be described as the sum of a number of Fourier components, here termed "density waves". The diffraction by a crystal can be regarded as due to the scattering caused by these density waves. When a density wave is irradiated by X-rays, a perceptible reflection occurs only at a particular angle of incidence. The resultant reflections make it possible to determine the amplitude, wavelength and direction of these density waves. It then remains to determine the phase of the waves in order to calculate the electron distribution in a crystal by means of Fourier synthesis. The phases are found by successive approximation, on the basis of provisional models of the structure. Various means of arriving at such models are discussed, including the Patterson projection, which provides information on interatomic distances. Finally, various structure analyses carried out in the Philips Research Laboratories are described. Analysis of the structure of the compound $Al_{0.89}Mn_{1.11}$ made it possible to infer that the manganese atoms are present in two non-equivalent sites in the unit cell and that their magnetic moments at these sites are oppositely oriented. The structure of $Th_2Al$ was found to contain fairly large interstices, which explained the gettering properties of this substance. The structure of $Y$-$Ba_2Zn_2Fe_{12}{}^{III}O_{22}$, a ceramic magnetic material (ferroxplana), proved to be exceptionally complicated (114 atoms in the unit cell). The successful analysis of its structure contributed to a deeper understanding of the magnetic properties of this substance.

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

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, The Netherlands, where a limited number of reprints are available for distribution.

**2767:** M. Koedam: Sputtering of copper single crystals bombarded with monoenergetic ions of low energy (50-350 eV) (Physica **25**, 742-744, 1959, No. 8).

Experiments on the sputtering of copper atoms ejected from Cu single-crystal surfaces of various orientations bombarded with rare-gas ions at normal incidence show an anisotropy in the directional distribution. The ejection patterns in the case of a bombardment of Cu (111) and (110) surfaces are given.

The number of copper atoms ejected per incident ion in the normal (110) direction is given as a function of the ion energy for a copper (110) surface bombarded with rare-gas ions at perpendicular incidence. The number varies from 0.01 at 75 eV to about 0.04 at 350 eV, depending on the bombarding ion and its energy. For comparison the sputtering yield of silver is given as a function of the ion energy. The silver surface was bombarded with normally incident $He^+$, $Ne^+$, $Ar^+$ and $Kr^+$ ions, with energies ranging from 50 to 250 eV. The yield varies from 0.1 at 50 eV to about 1.5 at 250 eV ion energy.

**2768:** A. Venema: The measurement of pump speed (Le Vide **14**, 113-120, 1959, No. 81; in French and in English).

Discussion of the various factors, in particular the definition and measurement of pressure, involved in the measurement of pump speeds at low pressures. A number of conclusions are reached concerning the location and nature of the pressure gauge to be used in order to obtain meaningful results.

**2769:** K. H. Hanewald: Analysis of fat soluble vitamins, IV. Discussion of the chemical routine determination of vitamin D (Rec. Trav. chim. Pays-Bas **78**, 604-621, 1959, No. 8).

Details of the already published procedure for the chemical routine determination of vitamin D have been investigated, especially the colorimetric determination with antimony trichloride and the elimination of tachysterol by addition of maleic anhydride.

**2770:** P. Westerhof and J. A. Keverling Buisman: Investigations on sterols, XII. The conversion of dihydrovitamin $D_2$-I and $D_2$-II into dihydrotachysterol$_2$ (Rec. Trav. chim. Pays-Bas **78**, 659-662, 1959, No. 8).

The results are presented of investigation on preparing dihydrotachysterol$_2$ from dihydrovitamins $D_2$-I and $D_2$-II.

**2771:** E. J. W. Verwey: Synthetische keramiek (Chem. Weekblad **55**, 553-556, 1959, No. 41). (Synthetic ceramics; in Dutch.)

In the last decennia, especially in the last 10 years, there has been a rapid development of a new branch of ceramics, called here "synthetic ceramics", characterized by the use of synthetic starting materials and by the circumstance that the final composition corresponds to compound materials synthetized purposely in view of specific physical properties. These materials are mainly applied in high-frequency electrical technique. They comprise insulators, semiconductors and magnetic materials. A survey is given of this branch of ceramics. A number of materials developed in the Philips laboratories are discussed in some detail.

**2772:** J. te Winkel: Drift transistor (Electronic and Radio Engr. **36**, 280-288, 1959, No. 8).

Equivalent circuits for a drift transistor are developed starting from a set of parameters derived from the physical principles underlying the device. It is shown what approximations are possible if limited frequency ranges or large values of the drift field are considered. Further simplifications are obtained from the introduction of suitably chosen frequency parameters. The resulting equivalent circuits appear to be simply related to those commonly used for a normal alloy transistor. The form is the same and the values of the circuit elements can be found by means of a number of multiplying factors that depend on the drift field only. These are given in graphical form.

**2773:** C. Jouwersma: Die Diffusion von Wasser in Kunststoffe (Chem. Ing. Technik **31**, 652-658, 1959, No. 10). (The diffusion of water in plastics; in German.)

The application of classical diffusion theory to the water uptake of plastics is described. In the cases investigated experimentally, it was found that Fick's law was obeyed. In such cases, the variables — viz. form and size of specimen, nature of surrounding medium, time and temperature, and solubility and diffusion coefficients of the material — can be separated. This enables one to make exact predictions concerning the water uptake of specimens of any form and size provided the characteristic constants of the plastic material in question are known.

**2774:** P. Penning and G. de Wind: Plastic creep of germanium single crystals in bending (Physica **25,** 765-774, 1959, No. 9).

The vertical displacement of the centre of a bar supported on two knife-edges and loaded in the centre, caused by plastic flow, is measured as a function of time. First the creep rate increases gradually and then becomes constant. The parameters describing the behaviour in these two regions have been determined as a function of the applied stress, for crystals of low and high dislocation densities and for crystals doped with oxygen.

**2775:** H. G. van Bueren: Theory of creep of germanium crystals (Physica **25,** 775-791, 1959, No. 9).

To explain the shape of the creep curves of germanium single crystals loaded in tension and in bending, a simple kinetic model is proposed, in which the dislocations are generated by (surface) sources and move with a uniform velocity over their glide planes. In this model a quantitative interpretation of the parameters of the creep curve in terms of the velocity of the dislocations, the incubation time of the sources and the density of the sources is possible. From the observations at "high" stress levels the velocity of dislocations in the germanium lattice can be determined; from those at "low" stress levels the rate of generation of the dislocations from the sources. The observed stress and temperature dependence of the creep process leads to similar dependences of incubation time and velocity. These dependences are used to form a quantitative theory of dislocation production and motion in the germanium lattice. This theory is shown to reflect semi-quantitatively various observed peculiarities of the creep phenomenon. The influence of other dislocations and of oxygen as an impurity on the elementary creep process can now also be qualitatively understood.

**2776:** J. S. C. Wessels: Dinitrophenol as a catalyst of photosynthetic phosphorylation (Biochim. biophys. Acta **36,** 264-265, 1959, No. 1).

The insensitivity of photosynthetic phosphorylation to dinitrophenol shows that the mechanisms of photosynthetic and respiratory generation of ATP may be quite different. The present note reports that, surprisingly enough, DNP is able to catalyse the generation of ATP by illuminated chloroplasts. In this respect it is more active than FMN but less active than vitamin $K_3$.

**2777:** H. Koelmans, J. J. Engelsman and P. S. Admiraal: Low-temperature phase transitions in $\beta$-$Ca_3(PO_4)_2$ and related compounds (Phys. Chem. Solids **11,** 172-173, 1959, No. 1/2).

On measuring the temperature dependence of the luminescence of $\beta$-$Ca_3(PO_4)_2$ activated with divalent Sn, strong hysteresis effects were found. These effects are interpreted as being due to phase transitions occurring at about 20 °C and at about — 40 °C. The presence of these two phase transitions has been confirmed calorimetrically. A number of modified Sr-orthophosphates, which have a structure closely related to that of $\beta$-$Ca_3(PO_4)_2$, show similar behaviour.

**2778:** J. Davidse: Transmission of colour television signals (T. Ned. Radiogenootschap **24,** 255-272, 1959, No. 5).

The paper discusses the transmission of colour-television signals according to the NTSC system. The choice of the chrominance signals, of their bandwidths and of the subcarrier frequency is discussed. The consequences of the method of gamma correction and of deviations from the constant-luminance principle are considered. The significance of the statistics of the chrominance signal is pointed out. This article has since been published in I.R.E. Trans. on Broadcasting **BC-6,** 3, Sept. 1960.

**2779:** A. G. van Doorn: Studio equipment for colour television (T. Ned. Radiogenootschap **24,** 237-253, 1959, No. 5).

This paper gives a broad survey of the equipment constructed in the Philips Research Laboratories for the generation of colour-television signals and the testing of the different pick-up systems. It describes three colour cameras, one using image orthicons as pick-up tubes, while in the other two experimental vidicons are used. Further the principle of the flying-spot scanner is described, as well as the colour-film camera. The special problems encountered in designing simultaneous pick-up systems and concerning colour-image registration, signal uniformity and gamma correction are discussed in more detail. In conclusion more is said about the different pick-up tubes and their use in colour-television cameras, their sensitivity, picture quality and overall performance.

VOLUME 22, 1960/61, No. 5 pp. 141-180 Published 6th March 1961

Philips Technical Review

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

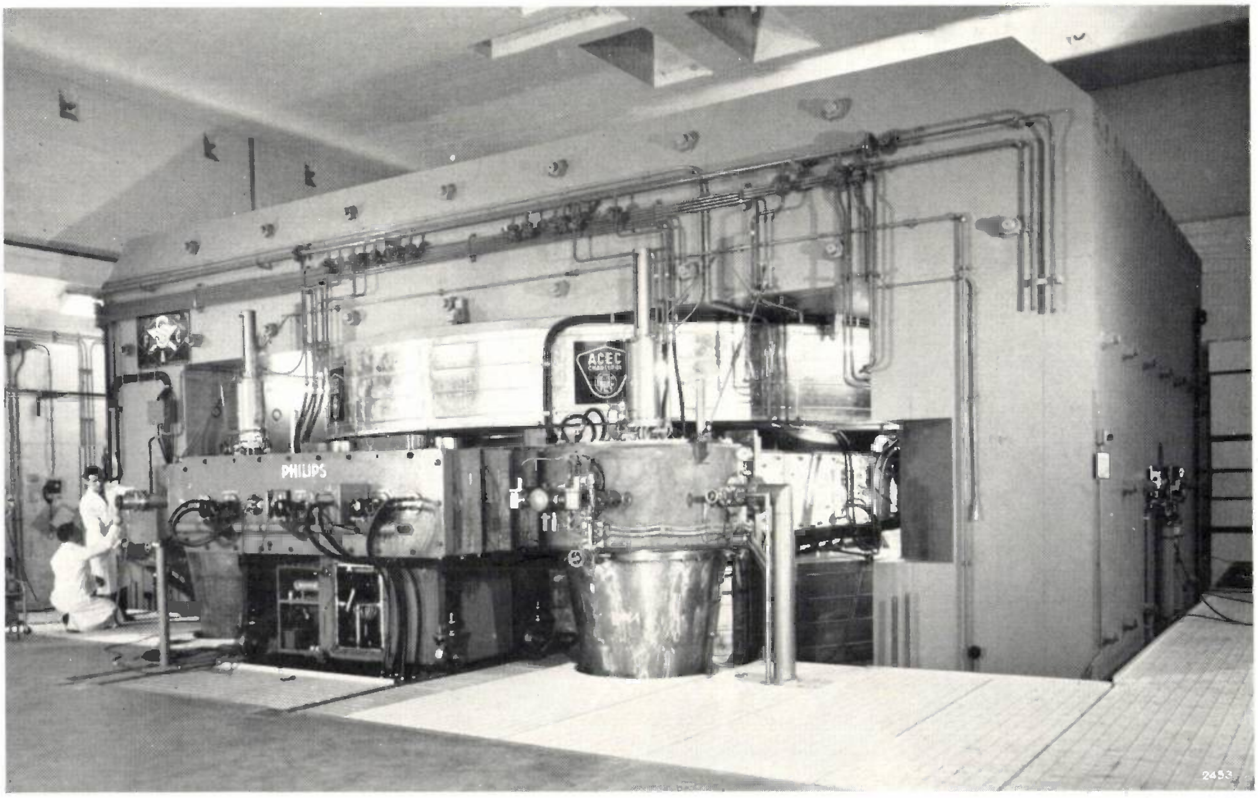


Photo CERN

# THE CERN 600 MeV SYNCHROCYCLOTRON AT GENEVA

I. OBJECT AND DESIGN.
II. THE RADIO-FREQUENCY SYSTEM.
III. THE TUNING-FORK MODULATOR.

*On 1st August 1957 the synchrocyclotron of the Organisation Européenne pour la Recherche Nucléaire (CERN) in Geneva entered into operation for the first time at maximum particle energy. The machine delivers protons with an energy of 600 million electron-volts, which makes it the third largest of its kind in the world (the largest is in Berkeley, California, and delivers proton energies of 740 MeV). Several European firms have contributed to the construction of the CERN cyclotron: the 2500-ton magnet was supplied by Schneider (Le Creusot, France); the energizing coils for the magnet by ACEC (Belgium); the vacuum chamber, in which the particles are accelerated by the radio-frequency field, was made by Avesta (Sweden); the large vacuum pumps for this chamber came from Leybold (Germany); etc. The entire radio-frequency system, with its modulator, was developed and manufactured by Philips Eindhoven in collaboration with the CERN.*

*It will be known that the principle of the cyclotron in its original form, i.e. with an accelerating voltage of constant frequency, can be used only up to relatively low values of particle energy (some tens of MeV), since the particles spiralling in the magnetic field fall out of phase with the voltage as a result of the relativistic increase in their mass. Higher energies can be achieved by frequency-modulating the RF voltage: if the frequency, during a part of the modulation period, is decreased in correspondence with the increase in mass, the particles can be kept in phase with the voltage up to the edge of the magnetic field. In this way, owing the to "phase focusing", not only the particles that have*

*originated at one given instant are accelerated but a whole group of particles that have originated both before and after this instant (principle of the synchrocyclotron) *).*

*In most synchrocyclotrons built hitherto the accelerating voltage is frequency-modulated with the aid of a rotating capacitor. A different system was adopted for the CERN machine, a vibrating capacitor being used in the form of a giant tuning fork. A description of this new modulator system is given here. The description is prefaced by an article dealing with the general object and principal design data of the synchrocyclotron, and by a brief account of the complete radio-frequency system.*

*The entire installation was built under the direction of Professor W. Gentner, the plans having been drawn up mainly under the direction of Professor C. J. Bakker, afterwards Director General of the CERN, whose death in April 1960 as the result of an air crash was such a tragic loss. This series of articles was planned with the kind help of the late Professor Bakker. We are indebted to Professor Gentner for the introductory article.*

# I. OBJECT AND DESIGN

by W. GENTNER **). 621.384.611.2

## History

Serious discussions concerning the foundation of a European laboratory for studying the physics of energetic particles were begun in 1951. At the first meeting of the provisional Conseil Européen pour la Recherche Nucléaire (CERN), agreement was reached on the general objectives, which were subsequently laid down in the Charter of the Organisation Européenne pour la Recherche Nucléaire, established in 1954. The laboratory was to be equipped with two large accelerators, on a scale not then existing in Europe. The first machine envisaged was a large synchrocyclotron, whose construction was to be based on installations of the same type and comparable energy as already proved elsewhere. The second machine was to be a pioneering project, both in regard to its particle energy (above 10 000 MeV) and its construction [1]).

## Principles of the design

We shall here describe the first, smaller accelerator — later commonly referred to as the SC machine. The main object was to provide the CERN laboratory in the shortest possible time with a powerful source of pi and mu mesons, so as to enable the laboratory to play a full part in meson research. For this purpose the energy of the accelerator had to be substantially higher than the threshold value for meson production, which is about 180 MeV when protons are used, for it was already known that the production of pi mesons does not initially rise very steeply as a function of proton energy (*fig. 1*). Moreover, it was important that the mesons produced should have the widest possible energy spectrum with a view, for example, to experiments on the scattering of mesons in matter as a function of their velocity. An upper limit was set to the choice of energy by the steeply rising costs involved, and also by purely technical considerations. Like the classic

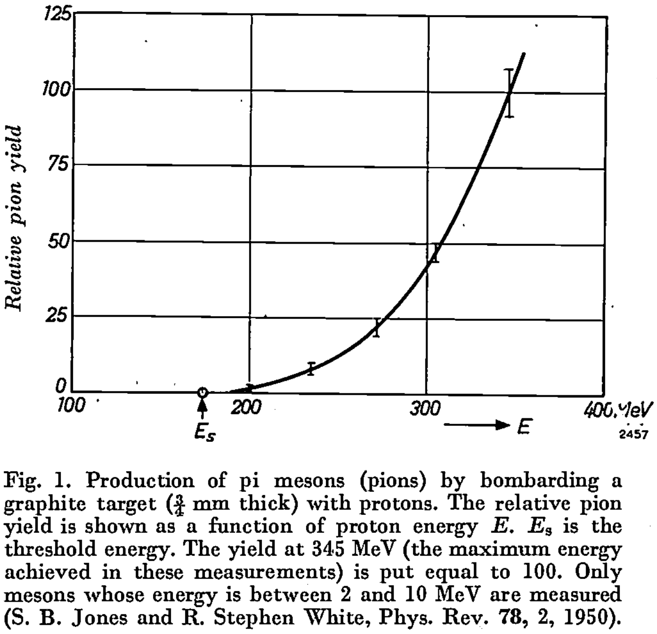


Fig. 1. Production of pi mesons (pions) by bombarding a graphite target (¾ mm thick) with protons. The relative pion yield is shown as a function of proton energy $E$. $E_s$ is the threshold energy. The yield at 345 MeV (the maximum energy achieved in these measurements) is put equal to 100. Only mesons whose energy is between 2 and 10 MeV are measured (S. B. Jones and R. Stephen White, Phys. Rev. 78, 2, 1950).

*) For further details of this principle, see W. de Groot, Cyclotron and synchrocyclotron, Philips tech. Rev. 12, 65-72, 1950/51. Fundamental problems involved in the construction of a synchrocyclotron are dealt with in the following series of articles: F. A. Heyn, The synchrocyclotron at Amsterdam, I. General description of the installation, II. The oscillator and the modulator, III. The electromagnet, IV. (with J. J. Burgerjon) Details of construction and ancillary equipment, Philips tech. Rev. 12, 241-247, 247-256, 349-364, 1950/51, and 14, 263-279, 1952/53.

**) Director of Max Planck Institut für Kernphysik, Heidelberg. Director of the Synchrocyclotron department of the CERN, Geneva, from 1955 to 1959.

[1]) This accelerator has been completed and was officially put into operation on 5th February 1960. The proton energy achieved is 28 500 million electronvolts (28.5 GeV).

cyclotron, the synchrocyclotron requires a constant magnetic field in which the circular accelerating chamber is placed. The size of the pole pieces of the magnet was to be such as to allow them to be made on the largest available lathes in Europe. Furthermore, the pole pieces and also the energizing coils were to be transported by road. All these physical, technical and financial considerations led finally to a pole diameter of about 5 metres (cf. *fig. 2*). With a field of suitable form, it is possible with this size of magnet of obtain a maximum beam radius of 2.27 m, which, in the case of protons and with an induction at the centre of 1.9 Wb/m$^2$ (19 000 gauss), corresponds to an energy of approximately 600 MeV. This energy was regarded as sufficient for pi and mu meson research, particularly if high beam currents can be achieved, which was a further feature of the design. Further, the intention was not only to produce mesons on targets inside the cyclotron, but also to extract a large fraction of the internal beam of particles, in order to generate mesons on external targets.

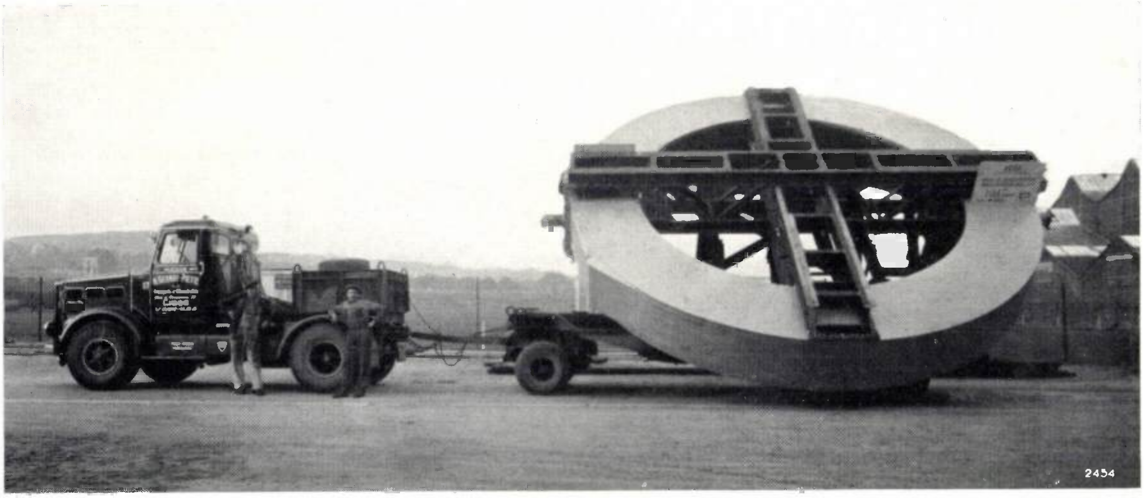


Photo CERN

Fig. 2. One of the two energizing coils (made by ACEC, Belgium) on the way to Geneva by road. The internal diameter (i.e. the pole diameter of the magnet) is roughly 5 metres.

The design of the installation was carried out by a team of engineers and physicists, largely drawn from nuclear research laboratories in the twelve member countries [2]). Numerous specialists and designers in the field of particle accelerators acted as advisers in the early years to the permanent team, which has since been steadily expanded. Finally, tenders were invited from European firms for the manufacture of the cyclotron components and ancillary equipment, orders being placed with those firms whose tenders were most satisfactory. The entire project was thus a striking example of the European cooperation which CERN set out to achieve.

## General layout

The layout of the synchrocyclotron and the experiment rooms and beam channels can be seen in *fig. 3*. The cyclotron is situated centrally in the building, and is surrounded by a radiation shield of barium-concrete walls 6 metres thick (the barium content increases the absorption). Diametrically opposite each other, left and right in the figure, are the two rooms for experiments: left, in $P$, the protons are available, and right, in $N$, the neutrons and mesons produced inside the cyclotron. The protective partitions between the rooms and the cyclotron are built up from movable blocks, and can be made to sink into the floor, thus greatly facilitating the setting-up of new experimental arrangements. When irradiation is in progress in one of the rooms, independent experiments can be prepared in the other one. Room $N$, with its wall $B_2$, is built as close as possible to the cyclotron, since the mesons are so short-lived that they have already considerably decayed on their way to the experiment bay.

[2]) The member countries of CERN are Belgium, Denmark, Federal Germany, France, Great Britain, Greece, Italy, Jugoslavia, The Netherlands, Norway, Sweden and Switzerland. Austria has also been a member since 1959.

About 30 metres from the source, high earth walls provide extra protection for the other parts of the SC department. Any excessive radiation in particular directions is absorbed by concrete blocks mounted at suitable locations within the area (*fig. 4*).

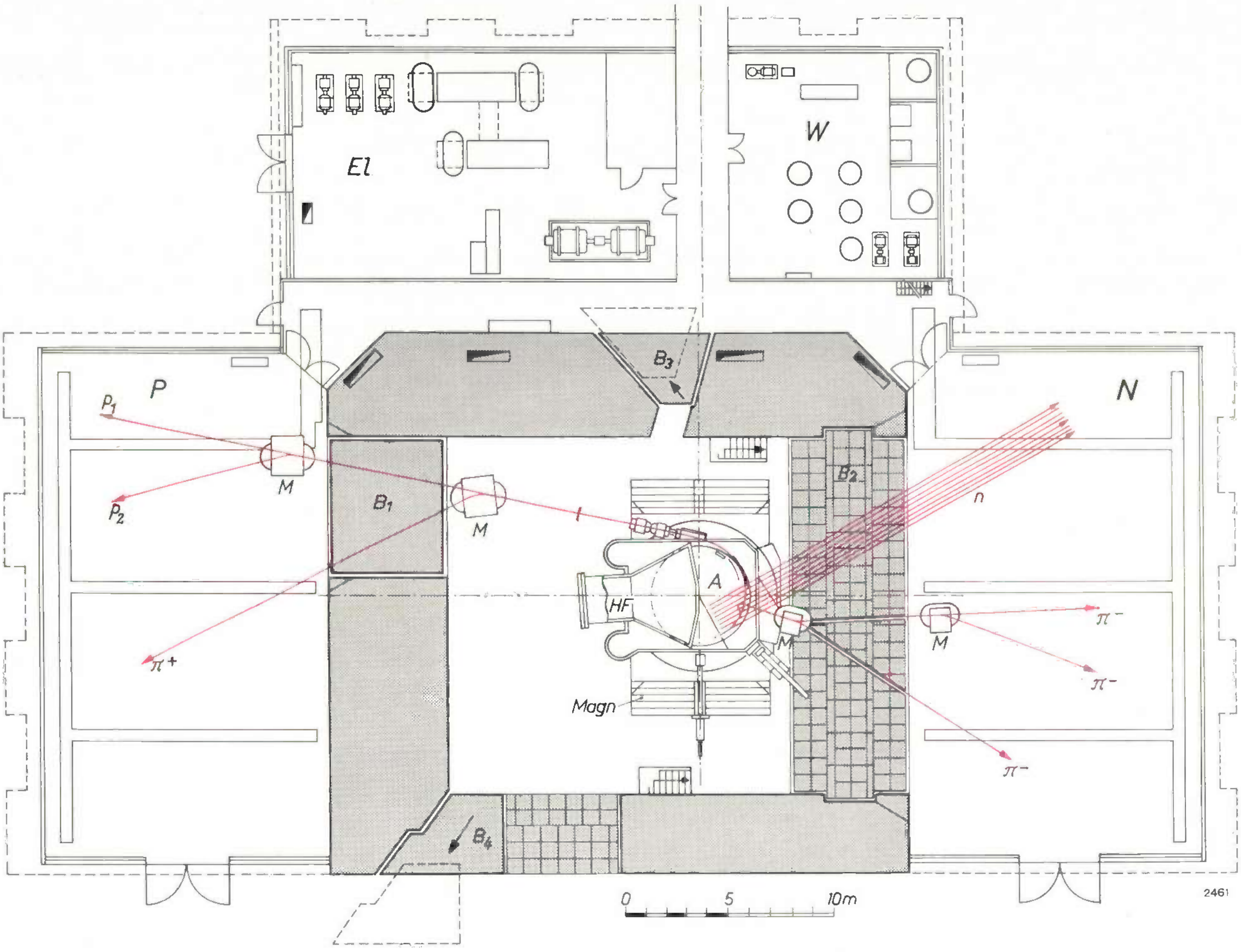


Fig. 3. Layout of the CERN synchrocyclotron, showing the concrete protective walls (shaded), the beam channels and experiment rooms (*P* for protons, *N* for neutrons). All particle beams are drawn in red: $p_1$ proton beam, $p_2$ extracted proton beam, *n* series of neutron beams generated at different radii of the proton path in the accelerating chamber *A*, $\pi^+$ and $\pi^-$ beams of the positive and negative pions, respectively, extracted by suitable means and focused by magnets *M*. *Magn* cyclotron magnet. *HF* radio-frequency system. $B_1$ and $B_2$ protective partitions built up from separate concrete blocks and capable of being sunk into the floor and raised hydraulically. $B_3$ and $B_4$ concrete blocks acting as doors, movable by means of electric motors. *El* power house containing the energizing-current generators and other ancillary electrical apparatus. *W* pump house supplying the cooling water and containing the cooling installation.

A passage 60 m long leads from the cyclotron and the experiment rooms to the control room (*fig. 5*) and the "counter rooms", which contain the equipment for counting the particles in the experiments (most experiments boil down to counting operations). It is here that the cyclotron and the experiments are controlled, since it is not as a rule permissible for anyone to be in the cyclotron building when irradiations are in progress in both experiment rooms.

The fact that the cyclotron is set up at ground level — and not, as many similar installations are, underground — has the great advantage that the outside walls of the experiment rooms proper may be thin. As a result, there is very little back scatter of the radiation used in the experiments. This in no way detracts from the effectiveness of the external protection, and has proved to be of great practical value. No other comparable installation has such a low background of scattered radiation in the experiment rooms. Indeed, in many experiments this fact is of decisive importance. The thin walls offer the further

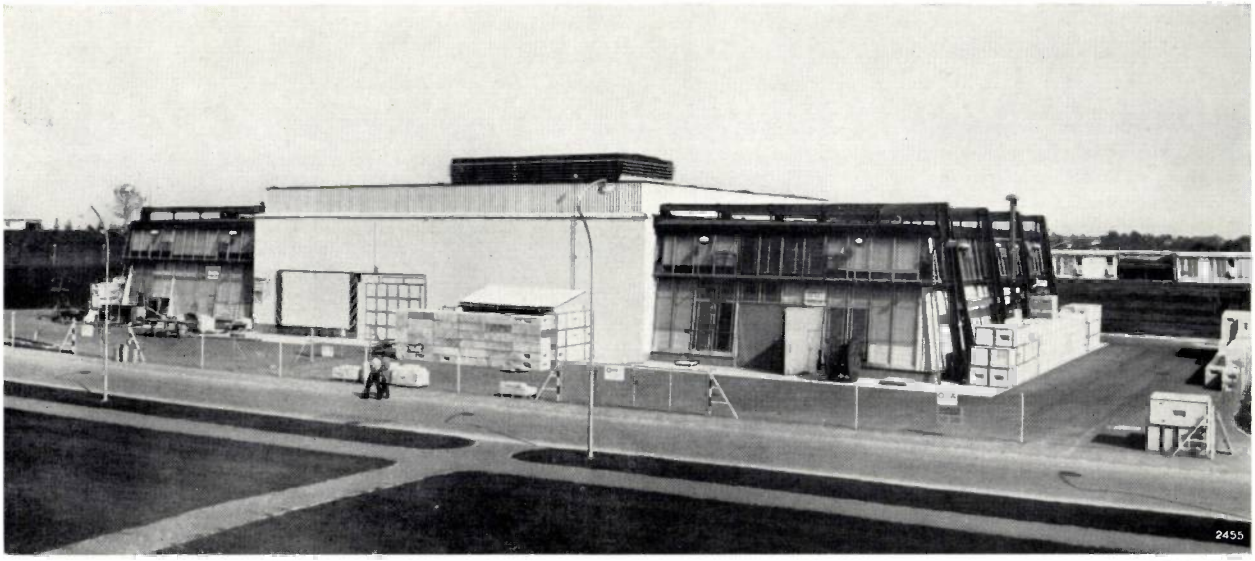


Photo CERN

Fig. 4. The building which houses the synchrocyclotron. The experiment rooms, seen right and left, in which remotely controlled instruments are subjected to bombardment by high-energy particles when the cyclotron is in operation, are enclosed by thin walls with the object of reducing back-scatter. Special directions in which radiation might penetrate outside the rooms are shielded off at some distance by concrete blocks.

advantage of making it easily possible to extend the beam channel outside if the experiments should make this necessary.

## Main components

The construction of the magnet was first studied on a 1 : 10 scale model. (This model is now being used in a small cyclotron for further investigation of the acceleration process.) The magnet consists of 54 blocks, most of them weighing 46 tons each. The total weight of 2500 tons calls for a very reliable foundation. The distance between the

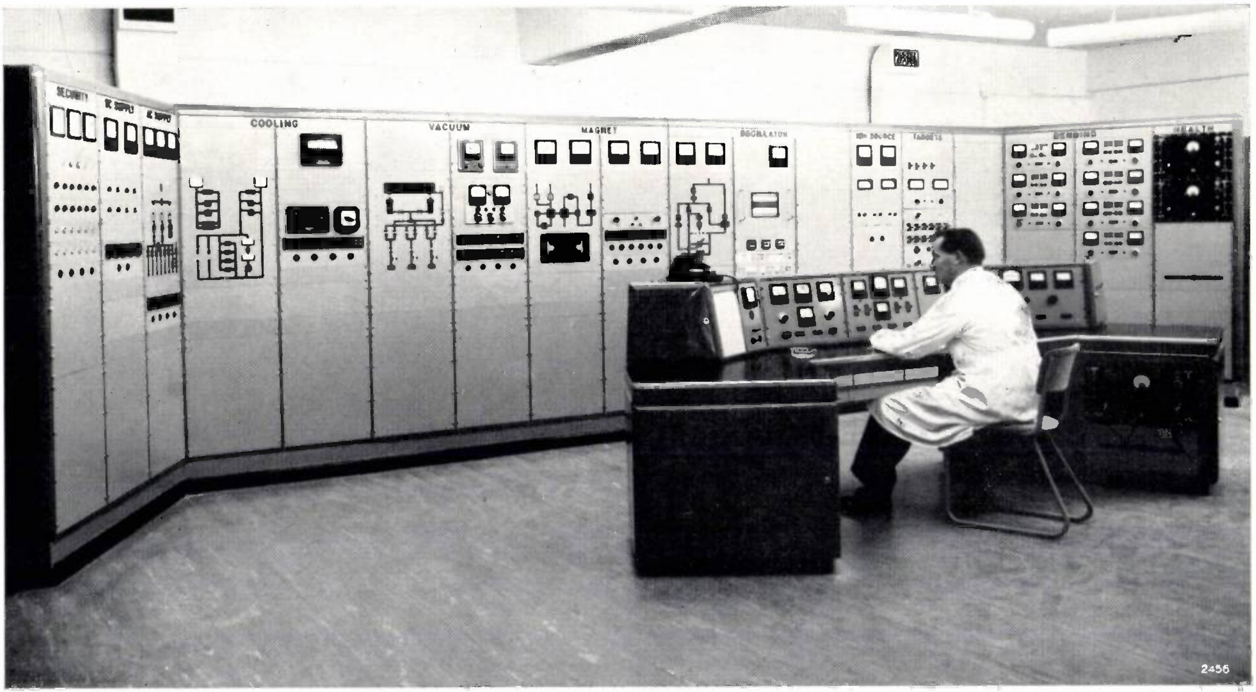


Photo CERN

Fig. 5. The master control room of the CERN synchrocyclotron.

poles varies between 45 and 35 cm. The two energizing coils are wound with aluminium tubing of rectangular cross-section, through which cooling water flows. The coils are ordinarily fed with a current of 1750 A, giving the magnetic induction at the centre of 1.9 Wb/m$^2$ mentioned above. *Fig. 6* shows the magnetic induction in the median plane between the pole pieces as a function of the radius.

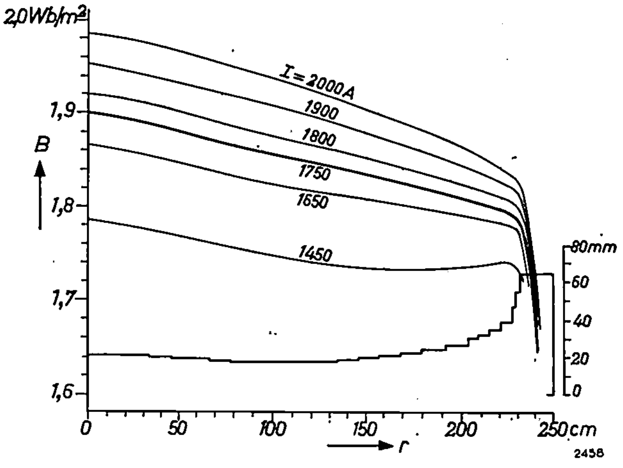


Fig. 6. Induction $B$ in the plane midway between the poles of the magnet, as a function of the radius $r$, for different currents $I$ in the energizing coils. The field shape in a synchrocyclotron is adjusted by means of ring-shaped shims of soft iron, increasing stepwise in thickness towards the edge, so as to keep the orbits of the particles stable up to the edge. The critical point where stabilization fails (i.e. where the gradient $n = (\mathrm{d}B/\mathrm{d}r)\,(r/B)$ reaches the value 0.2) lies here at the radius $r = 2.27$ metres.

The profile of the pole pieces produced by the shims is shown below the curves; see scale on right.

At a current of 1750 A the power converted into heat in the coils is approximately 750 kW. The cooling water which removes this heat is itself cooled, and is recirculated. In the title photo the upper part of the magnet can be seen, with the upper energizing coil and the accelerating chamber.

The vacuum in the accelerating chamber, which must be of the order of $10^{-6}$ mm Hg, is maintained by two oil-diffusion pumps, each with a pumping rate of 12 m$^3$/sec at $10^{-4}$ mm Hg. The whole chamber, including the connections to the pumps, is made of stainless-steel plates welded to a frame assembly. No trouble has ever been experienced with the vacuum.

The proton source originally envisaged was a hot-cathode arc ion source. This was later replaced by a Penning cold-cathode ion source, which is much simpler and can also be pulsed without difficulty.

The radio-frequency system, on whose reliability the operation of the whole cyclotron depends, is described at some length in the following articles. The principle of the tuning fork, chosen as the variable capacitance of the modulator, providing a frequency variation between 29 and 16.5 Mc/s, finally proved its value after a good deal of difficult development work. The protons are accelerated in the RF field in the established way with only one complete Dee, which leaves one half of the acceleration chamber entirely free for the positioning of targets. The general construction of the dee and further particulars of the accelerating chamber are shown in *fig. 7*.

The machine is provided with the following target systems. For the production of neutrons of different energy, eight "flip targets" (see below, fig. 7) can be "flipped" into the beam at varying distances from the centre. A movable "probe target" can also be positioned at a varying distance from the centre; the use to which it is put will be discussed presently. Further, there is a universal target or "Fermi trolley" available, which can be moved around the outermost proton orbit and is used for the production of mesons. This system permits easy adjustment of the meson source in relation to the deflection system for the meson channels.

The proton beam is deflected in a magnetic channel in which an adjustable perturbation of the magnetic field produces suitable oscillations of the beam. This beam-extraction system, developed by Le Couteur, had already proved its effectiveness in the Liverpool synchrocyclotron [3]). Upon extraction the proton beam immediately passes through two quadrupole lenses, each of which has a focal length of about 3 m. These are followed, near the radiation shield, by a double-focusing deflection magnet, which accurately directs the beam through the channel to the experiment room.

The principal operating data of the machine are collected in *Table I*.

### Yield of fast protons and mesons

The current of the internal proton beam can be measured with a thermocouple fixed to the probe target. As the radius increases, however, and with it the particle energy, the beam passes through the thermocouple more than once. At the same time,

[3]) K. J. Le Couteur, The extraction of the beam from the Liverpool synchrocyclotron, I. Theoretical, Proc. Roy. Soc. A **232**, 236-241, 1955. — The required magnetic channel for the CERN machine was designed with the assistance of N.F. Verster of the Philips Laboratory at Eindhoven, making use of a laboratory computer.
*Editorial note:* This design work, which has also been applied successfully to the synchrocyclotron built by Philips at Orsay near Paris, will in due course be the subject of a special article in this journal.

**Table I.** Principal data of CERN synchrocyclotron.

| | |
|---|---|
| Maximum proton energy | 600 MeV |
| Maximum radius $R$ of proton path ($n = 0.2$) | 2.27 m |
| Magnetic induction at $R = 0$, midway between the poles | 1.88 Wb/m$^2$ |
| Magnetic induction at $R = 2.27$ m | 1.79 Wb/m$^2$ |
| Ampere turns, normal | $1.2 \times 10^6$ |
| Ampere turns, maximum permissible | $1.35 \times 10^6$ |
| Energizing power, normal | 750 kW |
| Weight of magnet | 2500 tons |
| Frequency sweep of accelerating voltage | 29-16.5 Mc/s |
| Modulation frequency | 55 c/s |

the energy loss of the particles per unit length of path is reduced. Consequently, at radii above 1.5 metres, the proportionality factor that gives the relation between the thermo-e.m.f. and the beam current is dependent on the radius. Where an absolute measurement is required, this method can therefore only be used with certain corrections, which are not so simple to estimate. For relative measurements, however, the method is most convenient. It has been shown, for example, that there is no appreciable

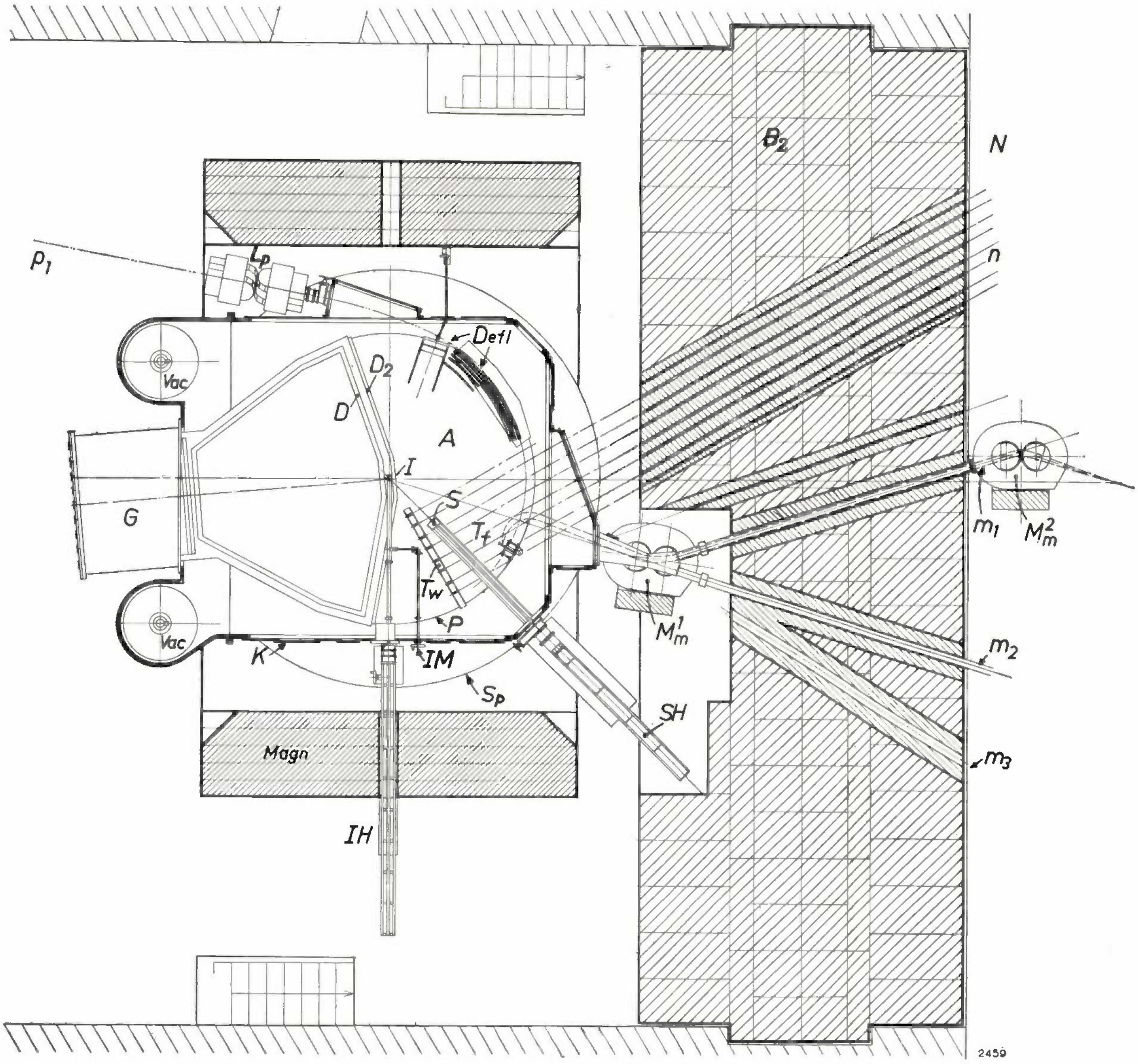


Fig. 7. The accelerating space $A$ in the vacuum chamber $K$, with channels through the protective wall $B_2$ leading to the neutron experiment room $N$. Only one of the dee electrodes ($D$) is connected to the radio-frequency system: the second is earthed and is therefore a simple slotted strip ($D_2$), leaving the whole of the right half of the chamber free for target assemblies. $I$ ion source attached to arm $IH$ and with adjusting mechanism $IM$. $T_w$ eight "flip targets", which can be set upright for producing neutrons of different energies (beams $n$). $S$ probe fitted to arm $SH$, one purpose of which is to measure the intensity of the proton beam at various distances from the centre. *Defl* magnetic channel with devices for extracting the protons from the accelerating chamber. $L_p$ lens for strong focusing of the extracted proton beam $p_1$. $T_f$ Fermi trolley, i.e. universal target, which can be moved around the outermost proton orbit, for the production of mesons; in the position illustrated, negative pions are extracted by the deflection magnets $M_m{}^1$ and $M_m{}^2$. The meson beams are $m_1$, $m_2$, $m_3$. *Magn* yoke of cyclotron magnet. $P$ magnet pole. *Sp* energizing coil. *Vac* pumps. $G$ radio-frequency source and tuning-fork modulator.

loss of particles during the acceleration. The method was also used, with a differential thermocouple, to get the spiral paths of the protons to lie very accurately in the median plane between the pole pieces. For this purpose the energizing currents of the two large magnet coils had to be made slightly different by means of an additional current generator: the magnetic median plane was found to be originally 1.2 cm lower than the geometrical median plane.

The absolute intensity of the external beam was measured with a graphite target, in which $^{11}$C is produced by the process $^{12}$C(p,pn)$^{11}$C. Since $^{11}$C emits only positrons, the number of these reactions per unit time can easily be determined by comparing the emitted radiation with the radiation from a standard gamma source. Moreover, the effective cross-section — which gives the ratio between the number of reactions and the number of bombarding protons — is well known for this reaction, and is virtually energy-independent for protons in the energy range involved. In this way the internal proton beam current at a radius of 2 m was found to be 0.3 μA.

Extensive magnetic measurements in the deflection channel were needed in order to arrive at the highest possible intensity in the deflected beam. With the lenses referred to above, the beam in the experiment room can be focused to a diameter of 3 cm at a point 5 m beyond the radiation shield. The total external proton beam current is $1 \times 10^{11}$ protons per second, i.e. about 6% of the internal current. These protons possess an energy of 600 MeV.

The experiment room at the other side receives the positive and negative pi mesons (pions) from the Fermi trolley. The energy of these pions is between 70 and 320 MeV. The negative pions describe a path as illustrated in fig. 7. The maximum intensity is reached at 150 MeV and amounts to $4 \times 10^5$ pions per sec. Since the path of the positive pions is curved in the same direction as that of the protons, only those positive pions can be extracted into an external channel that emerge from the target backwards. The Fermi trolley is moved into the appropriate position for this purpose (see *fig. 8*). Owing to this less favourable direction of emission (the majority of the pions in this energy range leave the target in the forward direction) the intensity of the positive pion beam is much lower. Other things being equal, it amounts to roughly $10^4$ pions per sec at 70 MeV. It is also possible to produce pions in an external target, and to extract them from the proton beam by means of a deflection magnet (see the $\pi^+$ beam in bay *P* in fig. 3).

**Results to date and future plans**

After the cyclotron was first put into operation on 1st August 1957 at its maximum energy, experiments could soon begin. There was then very little time available to make the modifications which had been shown to be desirable in the first month of operation. After having worked for more than two years, the cyclotron was shut down for a while to allow these modifications to be made. Improvements were introduced in the RF system, and changes are being made that will enable the cyclotron to operate on short pulses, whilst at the time raising the intensity of the proton beam. The protective wall with the beam channels is being rebuilt, and quadrupole lenses are to be incorporated in the wall in an attempt to raise the intensity of the pion currents. Another series of quadrupole lenses is being set up to produce a powerful beam of mu mesons(muons) for scatter experiments.

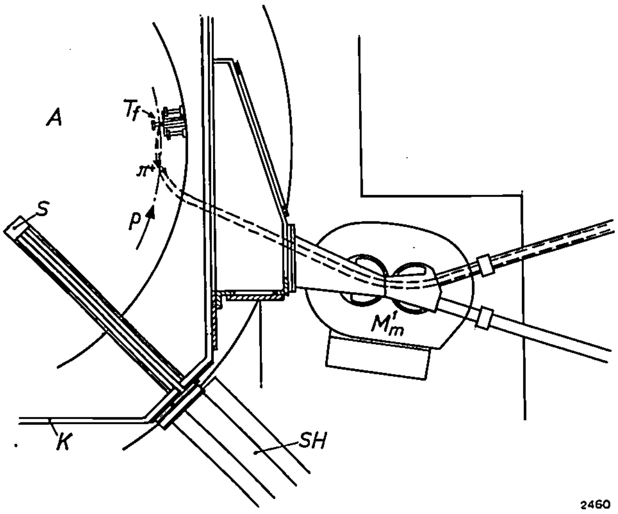


Fig. 8. Position of universal target $T_f$ (Fermi trolley) for producing positive pions with the proton beam *p*. Meaning of letters as in fig. 7.

In the first three years the machine was in operation 24 hours a day over long periods. During this time, many interesting experiments were successfully completed. One group of investigators demonstrated that, in addition to the normal decay of the pion into a muon and a neutrino, it is also possible to observe the much rarer decay of the pion into an electron and a neutrino [4]). Other groups have worked on the K-capture of muons in $^{12}$C, with the object of studying the interaction of muons and electrons with the atomic nucleus. Experiments have been

[4]) F. Fazzini, G. Fidecaro, A. W. Merrison, H. Paul and A. V. Tollestrup, The electron decay of the pion, Phys. Rev. Letters 1, 247, 1958.

carried out to test the hypothesis that meson processes are independent of their charge. A chemical team has been engaged on the study of various nuclear reactions involving the ejection of numerous fragments (spallation). In addition, guest teams from the member countries have started on their own research programmes, and these are being given special prominence.

**Summary.** The synchrocyclotron of the CERN laboratory at Geneva, which has been in operation since August 1957, produces protons of 600 MeV. This article discusses briefly the principal considerations governing the choice of this energy and the general design. Of particular interest is the layout of the building, which made it possible to minimize the general background of radiation in the experiment rooms. After mentioning the salient features of the construction of the synchrocyclotron, the author touches on the targets and the beams of particles produced. The internal proton beam current is approx. 0.3 μA. An efficient deflection system makes it possible to extract about 6% of this current. Further, intensive beams of pions and muons are available (per second $4\times10^5$ negative pions of 150 MeV or $10^4$ positive pions of 70 MeV). The article concludes with a reference to the research programmes in progress.

# II. THE RADIO-FREQUENCY SYSTEM

by K. H. SCHMITTER *) and S. KORTLEVEN **).

621.384.611.2

In the following account of the radio-frequency system of the CERN synchrocyclotron we shall describe the main technical features of the system and also touch briefly on the theoretical considerations underlying its design. Some parameters on which the design is based have already been mentioned in the previous article (particle energy, magnetic induction, dimensions of the magnet poles, etc.); various other conditions were imposed by the design of the modulator, which is described in article III.

## Principles underlying the design

The orbital frequency $\omega$ of the particles in a cyclotron decreases during the acceleration process, for two reasons:

1) because of the relativistic increase of the mass $m$ of the particles,

2) because the magnetic induction $B$ falls off radially. This radial diminution of the field in the (rotationally symmetric) cyclotron is necessary in order to stabilize the orbits of the particles. Let $m_0$ be the rest mass of the particles, $e$ their charge, $r$ the radius of the orbit and $c$ the velocity of light, then the familiar equation applicable to the classic cyclotron,

$$\omega_0 = \frac{eB}{m_0}, \quad \dots\dots\dots \quad (1)$$

should be replaced by the more general equation:

$$\omega(r) = \frac{e\,B(r)}{m_0\sqrt{1+\left(\frac{r\,e\,B(r)}{m_0c}\right)^2}}. \quad \dots \quad (2)$$

For a particle at $r$ which possesses the kinetic energy $E_\mathrm{k}$ corresponding to that position, the dee voltage must have the frequency given by equation (2) if the orbiting particle is to remain exactly in phase with this voltage. Equation (2) thus gives the ideal frequency variation in an acceleration cycle, and that variation should be achieved in the synchrocyclotron by modulating the frequency of the dee voltage. *Fig. 1* shows the induction curve $B(r)$ — see article I, fig. 6 — and the proton frequency curve $\omega(r)$, together with the curve representing the kinetic energy of the protons $E_\mathrm{k}(r)$. As can be seen, the frequency $\omega/2\pi$ for protons must initially be about 29 Mc/s for a magnetic induction at the centre of 1.9 Wb/m$^2$, and at the energy of 600 MeV, which corresponds to the boundary radius of 2.27 m of the stable orbit (see I), this frequency must have dropped to 16.5 Mc/s.

Broadly speaking, the above shows that the frequency sweep must be greater the higher the

*) Max Planck Institut für Physik und Astrophysik, Munich; formerly attached to CERN, Geneva.

**) Industrial Equipment Division, Philips, Eindhoven.

carried out to test the hypothesis that meson processes are independent of their charge. A chemical team has been engaged on the study of various nuclear reactions involving the ejection of numerous fragments (spallation). In addition, guest teams from the member countries have started on their own research programmes, and these are being given special prominence.

**Summary.** The synchrocyclotron of the CERN laboratory at Geneva, which has been in operation since August 1957, produces protons of 600 MeV. This article discusses briefly the principal considerations governing the choice of this energy and the general design. Of particular interest is the layout of the building, which made it possible to minimize the general background of radiation in the experiment rooms. After mentioning the salient features of the construction of the synchrocyclotron, the author touches on the targets and the beams of particles produced. The internal proton beam current is approx. 0.3 μA. An efficient deflection system makes it possible to extract about 6% of this current. Further, intensive beams of pions and muons are available (per second $4\times10^5$ negative pions of 150 MeV or $10^4$ positive pions of 70 MeV). The article concludes with a reference to the research programmes in progress.

# II. THE RADIO-FREQUENCY SYSTEM

by K. H. SCHMITTER *) and S. KORTLEVEN **).

621.384.611.2

In the following account of the radio-frequency system of the CERN synchrocyclotron we shall describe the main technical features of the system and also touch briefly on the theoretical considerations underlying its design. Some parameters on which the design is based have already been mentioned in the previous article (particle energy, magnetic induction, dimensions of the magnet poles, etc.); various other conditions were imposed by the design of the modulator, which is described in article III.

## Principles underlying the design

The orbital frequency $\omega$ of the particles in a cyclotron decreases during the acceleration process, for two reasons:

1) because of the relativistic increase of the mass $m$ of the particles,

2) because the magnetic induction $B$ falls off radially. This radial diminution of the field in the (rotationally symmetric) cyclotron is necessary in order to stabilize the orbits of the particles. Let $m_0$ be the rest mass of the particles, $e$ their charge, $r$ the radius of the orbit and $c$ the velocity of light, then the familiar equation applicable to the classic cyclotron,

$$\omega_0 = \frac{eB}{m_0}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

should be replaced by the more general equation:

$$\omega(r) = \frac{e\,B(r)}{m_0\sqrt{1+\left(\frac{r\,e\,B(r)}{m_0c}\right)^2}}. \quad . \quad . \quad (2)$$

For a particle at $r$ which possesses the kinetic energy $E_k$ corresponding to that position, the dee voltage must have the frequency given by equation (2) if the orbiting particle is to remain exactly in phase with this voltage. Equation (2) thus gives the ideal frequency variation in an acceleration cycle, and that variation should be achieved in the synchrocyclotron by modulating the frequency of the dee voltage. *Fig. 1* shows the induction curve $B(r)$ — see article I, fig. 6 — and the proton frequency curve $\omega(r)$, together with the curve representing the kinetic energy of the protons $E_k(r)$. As can be seen, the frequency $\omega/2\pi$ for protons must initially be about 29 Mc/s for a magnetic induction at the centre of 1.9 Wb/m², and at the energy of 600 MeV, which corresponds to the boundary radius of 2.27 m of the stable orbit (see I), this frequency must have dropped to 16.5 Mc/s.

Broadly speaking, the above shows that the frequency sweep must be greater the higher the

*) Max Planck Institut für Physik und Astrophysik, Munich; formerly attached to CERN, Geneva.

**) Industrial Equipment Division, Philips, Eindhoven.

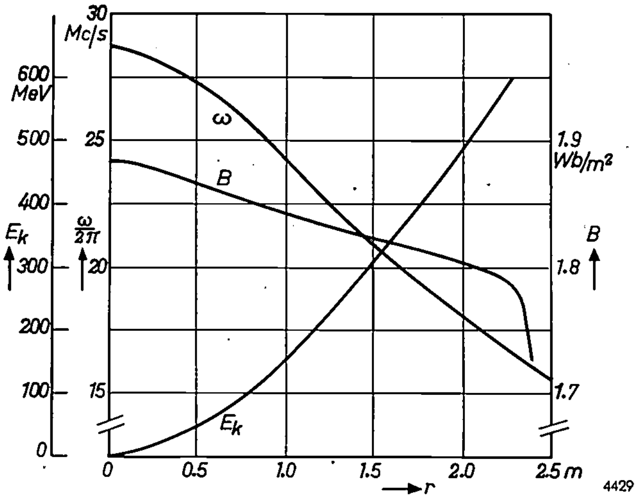


Fig. 1. Curve showing the radial variation $B(r)$ of the magnetic induction in the CERN cyclotron, the corresponding frequency variation $\omega(r)$ for the acceleration of protons, and the radial increase in the kinetic energy $E_k(r)$ of the protons.

energies aimed at [1]. That is one of the reasons why the building of a large synchrocyclotron is no easy task from the point of view of high-frequency engineering.

The above equation does not unambiguously establish the manner in which the frequency of the dee voltage should vary as a function of time between the specified limiting values. A further condition remaining to be satisfied is that the frequency variation of the particles per orbit ($\dot{\omega}/\omega$) must exactly correspond to the energy gain of the particles in each complete orbit. For any given frequency curve one can thus derive an ideal programme for the amplitude $\hat{U}$ of the dee voltage, since the latter governs the energy gain per orbit. If we want to let the radius of the orbit increase linearly with time — in which case the curve $\omega(r)$ in fig. 1 would also represent the variation with *time* (the frequency programme) — then $\hat{U}$ = const. would be the corresponding ideal amplitude programme.

In reality the choice of amplitude and frequency programme is not entirely free, in particular because — as we shall see — the dee voltage necessarily varies with frequency. Fortunately it is not necessary that frequency and amplitude should be exactly in step with one another: if this *were* so it would mean that a particle, in each revolution, would pass the dee gap in the same phase $\varphi$ with respect to the dee voltage, namely at $\varphi = -30°$. (Such particles are then 60° in phase behind those particles which pass the dee gap at the exact moment of maximum dee voltage, for which, by definition, $\varphi = -90°$.) However, because of the phase focusing, which characterizes the operation of the synchrocyclotron, a certain variation in phase is permissible; the only consequence of a deviation of $\varphi$ from the chosen value is that the phase range of the particles carried along by the focusing will change slightly, thereby also changing the beam current of the cyclotron. The value $\varphi = -30°$ is generally regarded as the most favourable. It is true that at $\varphi = -90°$ the acceleration benefits from the maximum instantaneous value of the dee voltage, but the effect of the phase focusing is then zero since in that situation there is no available reserve of acceleration for the particles travelling too slowly. At $\varphi = -30°$ the focusing embraces all particles in a phase range from —150° to +40° [2]; only 10% of the available particles are lost and the instantaneous value of the accelerating dee voltage still amounts to half the maximum.

The practical consequence of this is that the actual amplitude of the dee voltage should always, to be on the safe side, exceed the "ideal" amplitude. True, the accompanying increase in the beam current by no means comes up to the higher demands made on the equipment to produce a higher dee voltage, but there is at least the certainty that the dee voltage will never fall *below* the ideal value, even if a dip should occur in the amplitude curve, as may happen at certain frequencies: in the latter case the beam current would fall rapidly, since many particles not accurately in phase would be lost when traversing the frequency region in which the dip appeared.

The amplitude of the dee voltage and the amplitude programme approximately establish the admissible repetition frequency of the acceleration process (modulation frequency). Given an average phase $\varphi = -30°$ and a dee voltage $\hat{U}$ of say 4 kV at 29 Mc/s and 8 kV at 16.5 Mc/s, each proton has an average energy gain of about 200 MeV per millisecond; it should thus reach the final energy of 600 MeV in about $\frac{1}{300}$th second. Since the form of the frequency curve is suitable only during roughly a third part of the modulation period, it follows that the maximum permissible value of the modulation frequency would be about 100 c/s. Unfortunately, owing to the limited mechanical strength of the material used for the mechanical modulator (see article III), it is not possible to achieve such a high

[1] For comparison it may be mentioned that a frequency sweep of only 4% was needed in the synchrocyclotron built by Philips for the Kernfysisch Instituut (IKO) at Amsterdam, and which supplies deuterons of 28 MeV; see Philips tech. Rev. **12**, 244, 1950/51.

[2] D. Bohm and L. L. Foldy, Phys. Rev. **72**, 649, 1947. See also W. de Groot, Philips tech. Rev. **12**, 65, 1950/51.

modulation frequency [3]). Originally, therefore, 50 c/s was decided upon, the idea being to use the mains frequency. When it appeared, however, that the mains frequency was not sufficiently constant, it was decided to use a special generator, and the modulation frequency was fixed at 55 c/s in order to avoid beat-frequency interference from the mains.

The next step in the design of the RF system was to estimate the power required. As in most particle accelerators, the power needed for the actual acceleration is of secondary importance. The average particle current envisaged was 0.5 μA, which meant a power consumption of 300 W at a final energy of 600 MeV. Taking into account the particles lost during the acceleration process, the average RF power required for the acceleration was thus estimated at roughly 1 kW. The joule losses in the dee system, however, and other losses whose nature is not entirely clear but which generally occur in cyclotrons, are very much larger than the power needed for the actual acceleration, so that the oscillator power decided on was 10 to 20 kW [4]).

## The design of the dee system

### *The dee system as a resonant transmission line*

In a synchrocyclotron the accelerating electrodes normally consist of only one dee and an earthed strip (see I). With its supply line the dee constitutes the inner conductor of a coaxial line, which must be excited into resonance (the inner and outer conductors of the coaxial line in this case have a rectangular cross-section, and the inner conductor is hollow). Other forms of resonators are not suitable because of radiation losses. The resonator is excited in its fundamental mode. Its "electrical length" is varied by varying the boundary conditions: the modulator does this at the repetition frequency (55 c/s). The variation in the present apparatus is brought about by the variable capacitance $C_M$ of the tuning fork.

In the simplest case, illustrated in *fig. 2*, the fundamental mode for the limiting value $C_M = 0$ (line open at both ends) corresponds to a half-wavelength along the transmission line; for the other limiting value, $C_M = \infty$ (line short-circuited at one end), it corresponds to a quarter-wavelength along the line. The corresponding resonant frequencies would be in the ratio 2 : 1. Since the practicable capacitance variation lies between much narrower limits than between 0 and ∞ — between 256 and 2580 pF with the envisaged tuning-fork modulator — it was obvious that the required frequency sweep from 29 to 16.5 Mc/s could *not* be achieved with this arrangement.

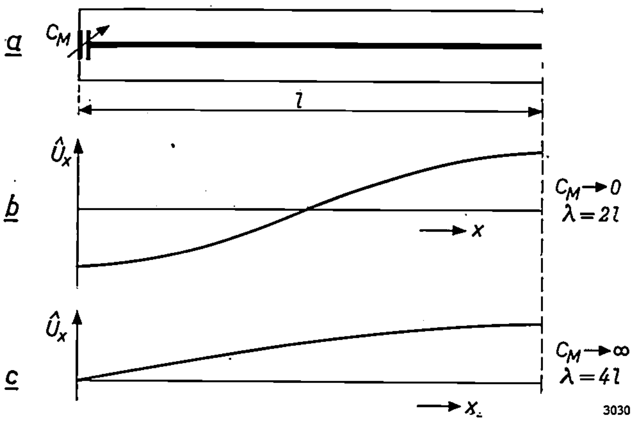


Fig. 2. *a*) Simplest form of a coaxial transmission line coupled to modulating capacitor $C_M$.
*b*) and *c*) At the limiting values $C_M = 0$ and $C_M = \infty$, a standing wave of $\frac{1}{2}\lambda$ and $\frac{1}{4}\lambda$, respectively, is set up in the resonator (the voltage amplitude $\hat{U}_x$ is plotted against distance $x$).

A greater frequency variation can be obtained if the modulation capacitor is made into a series resonant circuit by the addition of an inductance (*fig. 3a*) [5]). The inductance used is in the form of a shorted section of coaxial line (a "stub", fig. 3*b*). Assuming that the line has everywhere the same characteristic impedance, we then obtain the voltage distributions shown in fig. 3*c*-*g* for the limiting values $C_M = 0$ and $C_M = \infty$ and for various values in between. The whole transmission line will then have a $\frac{3}{4}\lambda$ standing wave in the one limiting case and a $\frac{1}{4}\lambda$ wave in the other, so that the extreme frequencies in the ideal case are in the ratio of 3 : 1. This brings us in sight of the desired frequency sweep (29 : 16.5), but it is evident from the cases (*d*) and (*f*) in the figure that only a frequency ratio of 5 : 3 (= 29 : 17.5) is reached at a capacitance ratio of $C_{M\max} : C_{M\min} = \pi^2 = 9.85$. We were thus still some way from our objective, and the last steps were the most difficult.

Before dealing with them, however, we shall dwell for a moment on the case of fig. 3*f*, which corresponds to the largest obtainable capacitance

[3]) If a rotating capacitor had been used instead of the tuning fork, the properties of the material would have imposed a similar limit on the frequency.

[4]) In the Amsterdam synchrocyclotron the "efficiency" (power on the target divided by transmitter power) is about 15%, which may be regarded as exceptionally high.

[5]) A parallel resonant circuit is also feasible, but the results are not so favourable. The intricate considerations involved are described by K. H. Schmitter, CERN report 59-33 of 22th September 1959, in which other questions concerning the resonator system are also discussed. See also M. Morpurgo, CERN report SC 136, 1955.

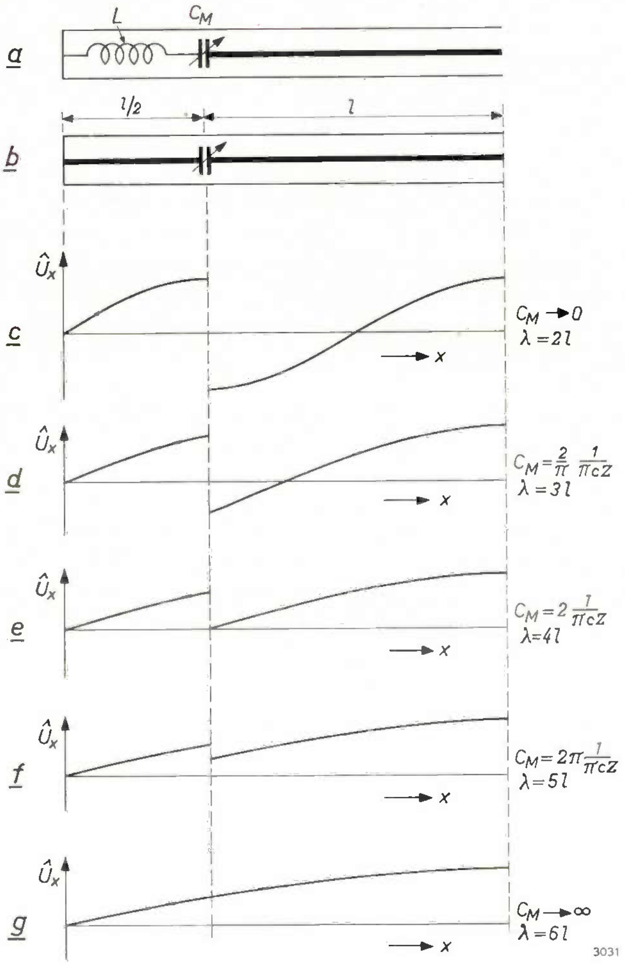


Fig. 3. *a*) The modulating capacitor is turned into a series resonant circuit by the addition of an inductance *L*.
*b*) This inductance is in reality a short-circuited section of line (stub).
*c*) to *g*) Voltage distribution on the resonator thus formed, at different values of capacitance $C_M$. Cases (*d*) and (*f*) roughly correspond to the extreme values attainable with the tuning-fork capacitor.

$C_{M\max}$, that is to the lowest frequency. We see that in this case the resonator from the lips of the dee to the modulator must have an electrical length $l = \lambda_{\max}/5$. With $f_{\min}$ = 16.5 Mc/s this gives $l$ = 3.75 m. Since the length of the dee is roughly equal to the radius of the magnet pole, i.e. 2.5 m, the dee had to be provided with an extension — a stem or neck. For constructional reasons this was in fact most welcome, 1) because it created more room for the modulator (the larger a cyclotron, the more likely that the whole RF system has to be squeezed in the rather inadequate space between the coils of the large magnet), and 2) because the modulating capacitor is then situated farther away from the stray field of the magnet (see III).

The way in which the frequency sweep was finally raised to the required value may be explained by considering the influence of the characteristic impedance. To begin with, at a given variation $\Delta C_M$ of the modulator capacitance the frequency sweep will be greater the smaller is the capacitance in parallel with the modulator. It is therefore important to keep the capacitance of the dee as small as possible, that is the dee should not be larger than is absolutely necessary (half the area of the magnet poles). The characteristic impedance of the dee, which for a uniform line would be $\zeta = \sqrt{L_I/C_I}$ ($L_I$ and $C_I$ being respectively the inductance and capacitance per unit length), is then as large as possible. Further considerations must take into account the fact that the dee can certainly not be treated as a uniform line — this is evident on purely geometrical grounds, see the model in *fig. 4*. This at once raises the question whether the situation might not perhaps be more favourable if, also for the dee stem and stub, we were to drop the assumption hitherto made that the system is a uniform line, possessing the same characteristic impedance everywhere along it. It turns out that it is indeed better to make the characteristic impedance of the dee stem greater than that of the dee itself, and similar considerations apply to the stub, as explained in the first report cited in reference [5]). (The situation as regards the stub may be roughly understood if we remember that a large characteristic impedance implies an approach to the ideal case of a pure inductance, as in fig. 3.) Admittedly, it is not possible to take full advantage of the resultant possibilities of increasing the frequency sweep, for one reason because the latter is also limited, at a given $\Delta C_M$, by the current and voltage load of the modu-

Fig. 4. Model of resonator on 1 : 10 scale. The photograph shows the outer conductor of the "coaxial" system, which, however, is removed on the left-hand side in order to show the modulating capacitor and the stub. Inside the outer conductor on the right is the dee. By means of the adjusting screw on the bracket on the extreme left the capacitance of the simulated tuning fork (meshed comb capacitor) can be changed.

lating capacitor. Nevertheless, after empirically arriving at a suitable compromise, it proved possible along these lines to achieve the required frequency sweep.

The fact that the current and voltage load affects the frequency sweep is due to the direct relation between the frequency variation at a given $\Delta C_M$ and the variation of the wattless power in the capacitor. This relation remains, no matter how the capacitor circuit may be varied; this can easily be verified on a number of simple examples.

One further remark before we enter into details on the design of the resonator. We have considered the dee system as a one-dimensional system, that is to say we have assumed that the voltage varies in the same phase and with the same amplitude at all places in the mouth of the dee. A glance at fig. 4 will make it obvious that vibration modes must also exist for which this is not the case, i.e. transverse vibrations in the dee. Such parasitic oscillations are of course unwanted. Since transversal parasitic oscillations having an antinode on the axis of symmetry are particularly likely to occur, countermeasures have been taken in some cyclotrons by cutting a slot in the dee along the axis of symmetry. Longitudinal modes are not affected by this slot. In the present case, such a measure was found to be not necessary in view of the high degree of symmetry of the system. The only dangerous parasitic oscillations were those due to the supply lines to the oscillator (these were longitudinal modes of oscillation differing from those in fig. 3). After we had adopted the principle of the flywheel oscillator, to be discussed below, even parasitic modes of this kind occasioned no further difficulties, whereas in the construction of other cyclotrons it has often taken months of hard work to eliminate the effects of parasites.

*Design of the resonator*

From the desired frequency programme and the variation of the tuning-fork capacitance we can now, along the lines discussed above, calculate the dimensions of a resonator as in fig. 3, provided we know the characteristic impedance of the dee. As we have seen, the dee is by no means a uniform line. Satisfactory approximation is obtained by substituting for the dee five consecutive sections of line of equal length, each in itself being a uniform line. The first section is loaded with the 200 pF capacitance of the dee mouth (*fig. 5*). We can now calculate the characteristic impedance of each section of line, and hence the reactance of the whole system as a function of frequency (*fig. 6*). Comparison of the reactance curve with that calculated when the dee is replaced by a uniform line (dashed curve in fig. 6) demonstrates the effect of the improved approximation.

Fig. 5. Approximation to the dee by a line whose characteristic impedance $\zeta$ varies in steps, being constant in each section. The first section of line is loaded by the capacitance of the dee mouth. The steps drawn in the dee are merely a symbolic representation of the variation of $\zeta$.

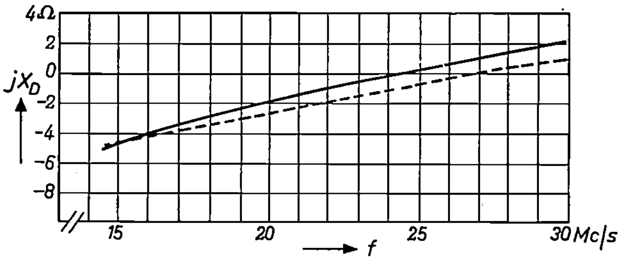


Fig. 6. Calculated reactance of the dee in fig. 5 as a function of frequency. The dashed curve represents the reactance calculated when the dee is treated as a uniform line.

The further design of the resonator was carried out, as is usual with such complicated systems, on partly empirical lines, that is the calculation outlined above was continuously supplemented and improved by a series of concurrent tests on models. One model was made on a scale of 1 : 10 (i.e. for frequencies ten times as high), and one full-scale model was made. The small model appears in fig. 4, and the full-scale model in *fig. 7*. Measurements were also made on a model of 1 : 1 scale in the longitudinal direction and 1 : 7 scale in the transverse direction. The full-scale model was also used for developing and proving the RF generator. This model — made of copper sheets on a wooden frame — was constructed in such a way as to enable the characteristic impedances of stem and stub to be varied independently of one another by changing the dimensions of the outer conductor (the "liner"). It was also possible to vary the length of the stub. In this model the tuning-fork modulator was replaced by a stationary condenser of similar shape, the capacitance of which could be varied from 250 to 2700 pF by adjusting an air gap.

The principal dimensions of the final version of the

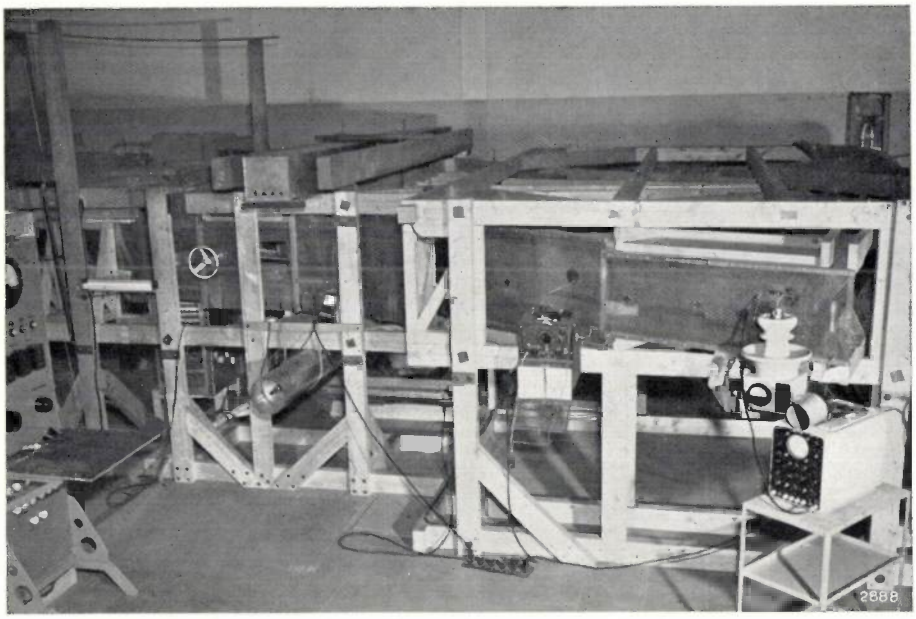


Fig. 7. Full-scale model of the resonator (wooden frame clad with sheet copper), built in Eindhoven, together with instruments for testing the RF generator.

resonator are given in *fig. 8*. The variation of the resonant frequency measured on this version as a function of the capacitance of the tuning fork is shown in *fig. 9*. For this measurement the connection terminal of the generator was placed as near as possible to the shorted end of the stub. This must be done because the capacitative reactance of the generator lowers the frequencies obtained but especially the highest frequency, the more so the farther the connection point is removed from the shorted end. This is easiest to understand if one remembers that the effect of the inductance, which was specially introduced to increase the frequency sweep (fig. 3*a*), is partly destroyed by the capacitance of the generator (primarily the anode capacitance of the transmitting valve). On the other hand, the connection terminal must not be too near the shorted end, otherwise the impedance in the anode circuit will be too small, making it impossible to satisfy the oscillation condition in the whole range of the frequency variation. The necessary minimum distance was determined empirically on the full-scale model. The dashed curve in fig. 9 represents the frequency variation calculated, taking into account the known generator reactance. *Fig. 10* shows how the frequency varies as a function of time when the tuning fork is made to vibrate sinusoidally (see III).

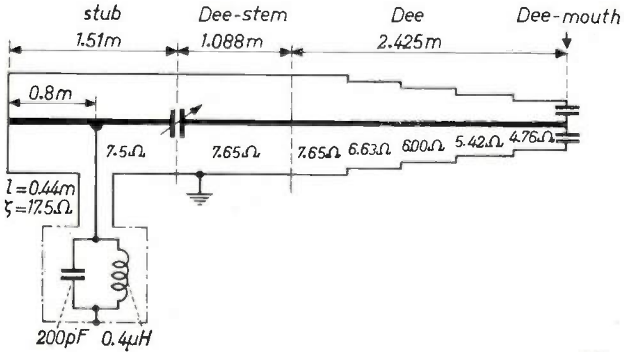


Fig. 8. Principal dimensions and characteristic impedances of the resonator in its final form.

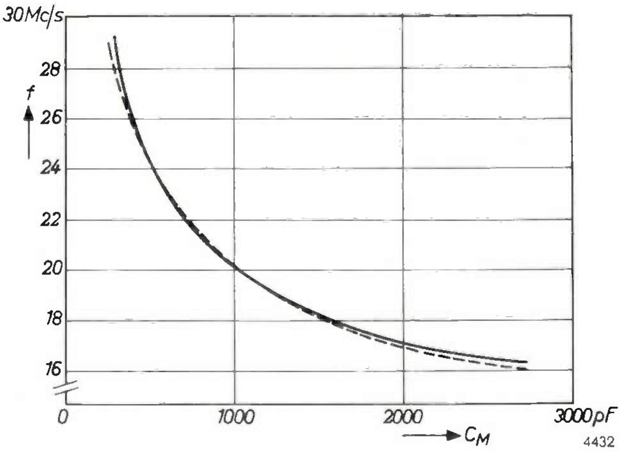


Fig. 9. Measured variation of the natural frequency $f$ of the resonator as a function of tuning-fork capacitance $C_M$. The dashed curve represents the calculated variation.

*Fig. 11* illustrates the voltage and current distribution over the system at the two extreme frequencies, on the assumption that the amplitude of the

RF supply voltage is constant at 5.4 kV (which corresponds to a direct anode voltage of 5 kV on the transmitting valve). It can be seen that current amplitudes of 3000 A occur in the stub, and that the voltage at the dee mouth rises from about 5 kV at 29 Mc/s to about 19 kV at 16.5 Mc/s. If we calculate the dee-voltage amplitude for the frequencies in between these values, we obtain the curve shown in *fig. 12*. This curve should be compared with the ideal amplitude programme corresponding to the frequency curve in fig. 10 if we wish to keep the phase constant at —30° (see page 150). As already explained, the deviations occurring enable the acceleration to act on the maximum number of particles, and also provide the reserve needed in the event of dips appearing in the amplitude curve of the dee voltage.

In reality, it is difficult to supply the system with a generator voltage of constant amplitude, owing to the effect of load variations on the excitation of the generator. In practice, therefore, the amplitude variation of the dee voltage differs from the calculated curve, as shown by the dot-dash curve in fig. 12. The curve bends over at the lowest frequencies because the generator in this region is faced with a circuit of relatively large capacitance and low inductance; such a circuit is difficult to set in oscillation, as will be seen by comparing it with the mechanical analogue of a stiff spring with a very small mass. At the higher frequencies the generator is presented with a more normal oscillatory circuit. The relatively large initial voltage which this gives rise to is very useful for drawing large numbers of ions out of the space-charge region near the axis of the cyclotron and making them take part in the acceleration process.

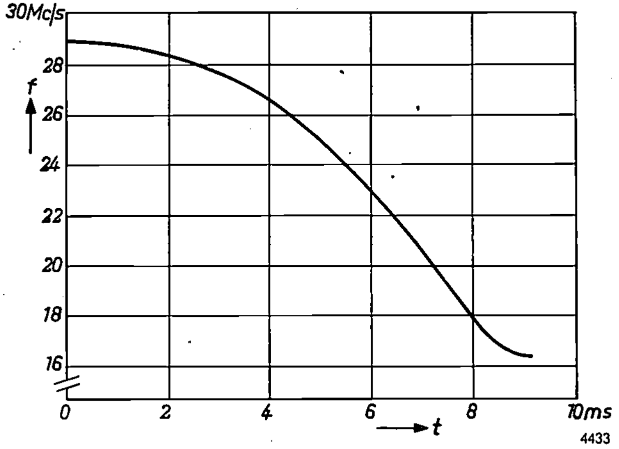


Fig. 10. Curve showing the natural frequency *f* of the resonator coupled to the oscillator as a function of time, for sinusoidal vibration of the tuning fork.

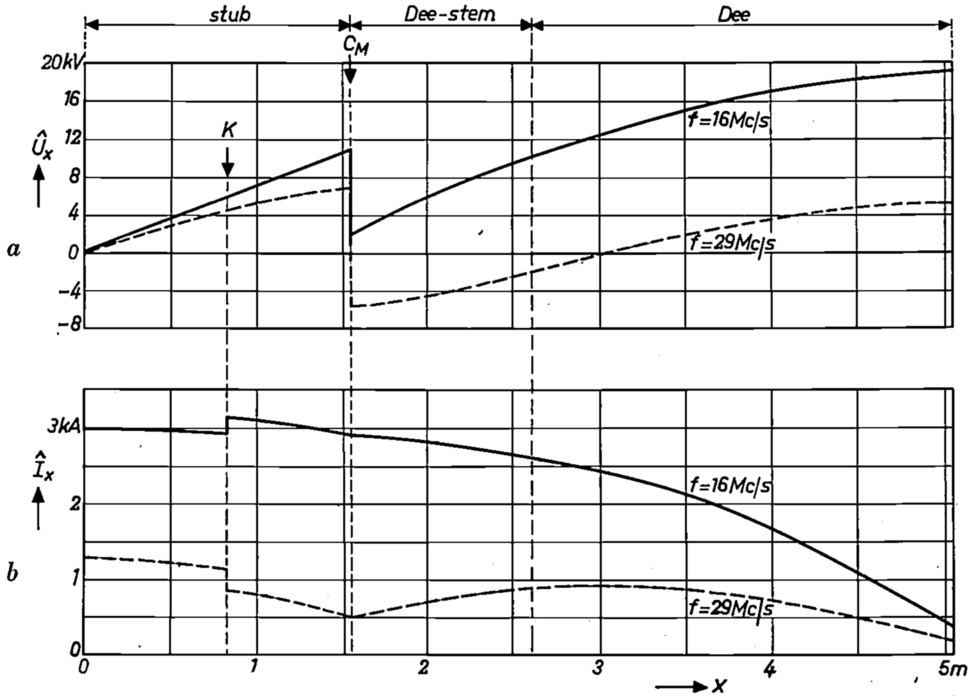


Fig. 11. *a*) Distribution of the voltage amplitude $\hat{U}$, *b*) of the current amplitude $\hat{I}$ over the resonator at the frequencies *f* = 16 Mc/s and *f* = 29 Mc/s. The abscissa *x* is the distance from the end of the stub. The RF voltage from the oscillator is applied at point *K*, it being assumed for both frequencies that the amplitude of the RF voltage on the oscillator and of the connecting line is 5.4 kV.

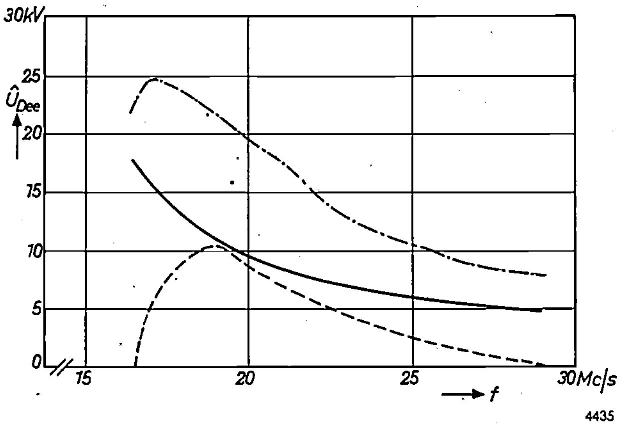


Fig. 12. Solid curve: amplitude of the voltage $\hat{U}_{\text{dee}}$ at the dee mouth, calculated as a function of frequency $f$. Dashed curve: the ideal amplitude programme corresponding to the frequency variation shown in fig. 10. Dot-dash curve: actual variation of voltage amplitude.

## Construction of the resonator

The whole inner conductor of the resonator, consisting of dee, stem, tuning fork and stub, is under high vacuum; the stub is accommodated in an extension housing connected to the evacuated accelerating chamber of the cyclotron. *Fig. 13* shows the resonator seen from above, and *fig. 14* represents a longitudinal section. The stem $H$ and the dee itself, $D$, form a single structural assembly. It is supported on two spherical insulators $Is_1$, at two places near the ends of the dee mouth, and at the stem it is suspended at two places by insulators $Is_2$. The spherical insulators rest on the lower pole plate $P_2$ of the magnet, which is the floor of the accelerating chamber. The insulators can roll slightly on this plate. With this mounting of dee and stem, deformations due to thermal expansion or to magnetic forces cause no significant mechanical strains in the long axis. (When the enormous magnet is switched on, the upper and lower pole plates approach each other by several millimetres!) Transverse strains are taken up by the insulators $Is_2$, which are specially designed for that purpose. The reinforcement ribs carried by the upper and lower plates of the dee (fig. 13) are bent upwards slightly, to compensate for the bending that occurs with the assembly supported by only two spherical insulators situated far apart. *Fig. 15* shows a photograph of the dee mounted between the pole plates; note the relatively narrow dee mouth.

The spherical insulators are surrounded by metal shields to protect them from fast particles. The dee mouth, the sides of the dee and of the stem, and the contact faces of the spherical insulators are all water-cooled. For this purpose the dee system is fitted with two parallel cooling circuits, in each of which the cooling water is supplied and returned through two hoses of hard polyvinyl chloride about 10 metres long, which insulate the system from earth. The hoses are rolled up and stored in the compartment $Q$. They pass through the vacuum-tight insulators $Is_2$ to the dee system.

Perpendicularly below the insulators $Is_2$ are two identical insulators $Is_3$; these have no supporting function but act only as insulating ports for the lines to the tuning-fork feelers $X$ (see article III, fig. 24).

The stub $St$, with the tuning fork $T$ attached to it, is suspended from two insulators $Is_4$, which are mounted on top of the stub housing $G$ and are of the same construction as the $Is_2$ insulators. The possibility of passing lines through these insulators was not used in this case, but their construction has the advantage of allowing considerable freedom of movement in the suspension, which facilitates the initial adjustment of the tuning fork. The centre of gravity of the entire stub lies almost perpendicularly under the line joining the suspension points, so that virtually no bending moments act on the insulators. As an additional support the shorted end of the stub is secured to the floor of the housing via a simple insulator $Is_5$; the latter is necessary to keep this end insulated from earth for DC voltage (see below). The stub and the foot of the tuning fork (see fig. 22 in article III) are water-cooled, the supply lines passing through the rear wall of the stub housing. When the cyclotron is in operation, the entire stub housing

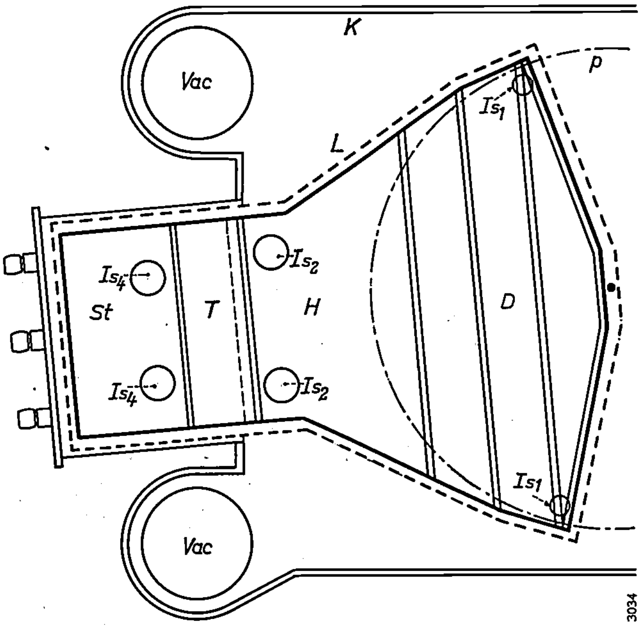


Fig. 13. Construction of the resonator and its positioning in the cyclotron. $D$ dee with reinforcement ribs, supported at the two points $Is_1$. $H$ dee stem, suspended at the two points $Is_2$. $T$ tuning fork. $St$ stub, suspended at the two points $Is_4$. $L$ outer conductor of resonator. $K$ accelerating chamber. $Vac$ pumps. $p$ outermost proton orbit.

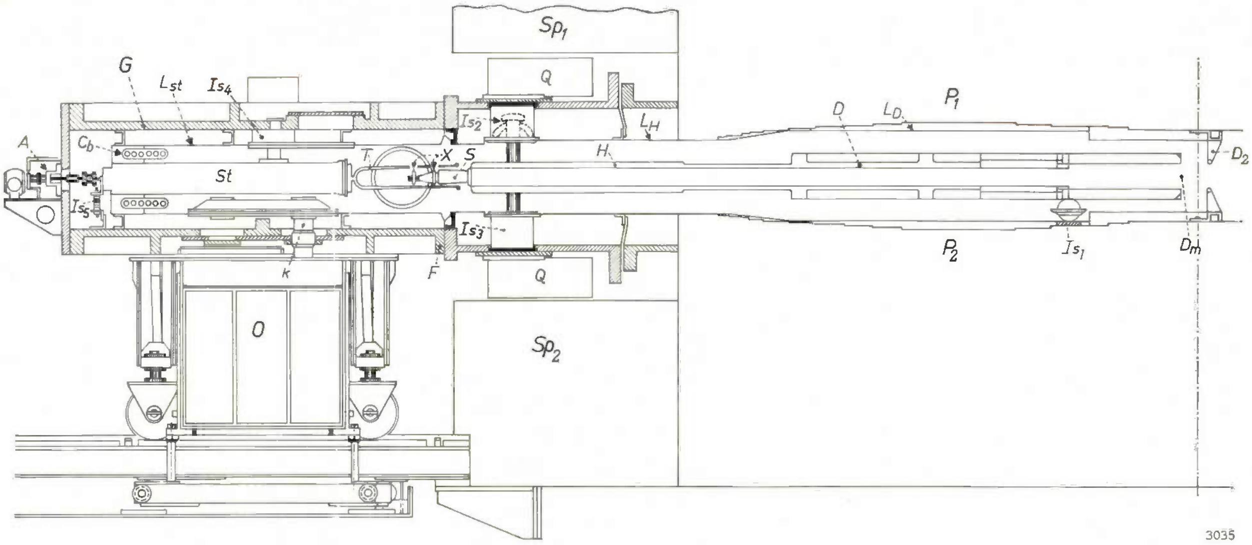


Fig. 14. Longitudinal section through the RF system. The dee $D$ and dee stem $H$ form an integral assembly. $D_m$ dee mouth. $D_2$ dummy dee. $P_1$ and $P_2$ pole pieces of cyclotron magnet with shims, likewise ceiling and floor of accelerating chamber. $Is_1$ spherical insulators. $Is_2$ and $Is_3$ vacuum-tight lead-in insulators for the dee stem. $S$ stator and $T$ tuning fork of the modulator, with feelers $X$. $St$ stub, suspended in the housing $G$ from the insulators $Is_4$ and secured by the insulator $Is_5$. $C_b$ shorting capacitor. $A$ servo mechanisms for adjusting the mounting of the tuning fork. $F$ flange, connecting the stub housing via an extension to the accelerating chamber. $L_D$-$L_H$-$L_{St}$ outer conductor of coaxial resonator system. $k$ oscillator connection. $O$ oscillator housing. $Q$ compartment containing the cooling-water hoses. $Sp_1$ and $Sp_2$ coils of cyclotron magnet.

is supported only at the flange $F$ by which it is connected to the accelerating chamber. This prevents floor vibrations, due to the operation of pumps, etc., from being transmitted to the stub housing and the tuning fork, and the housing is able to follow changes in the position of the main mass of the cyclotron (see above) without strains being set up. Deformations of the housing itself do not affect the accurate

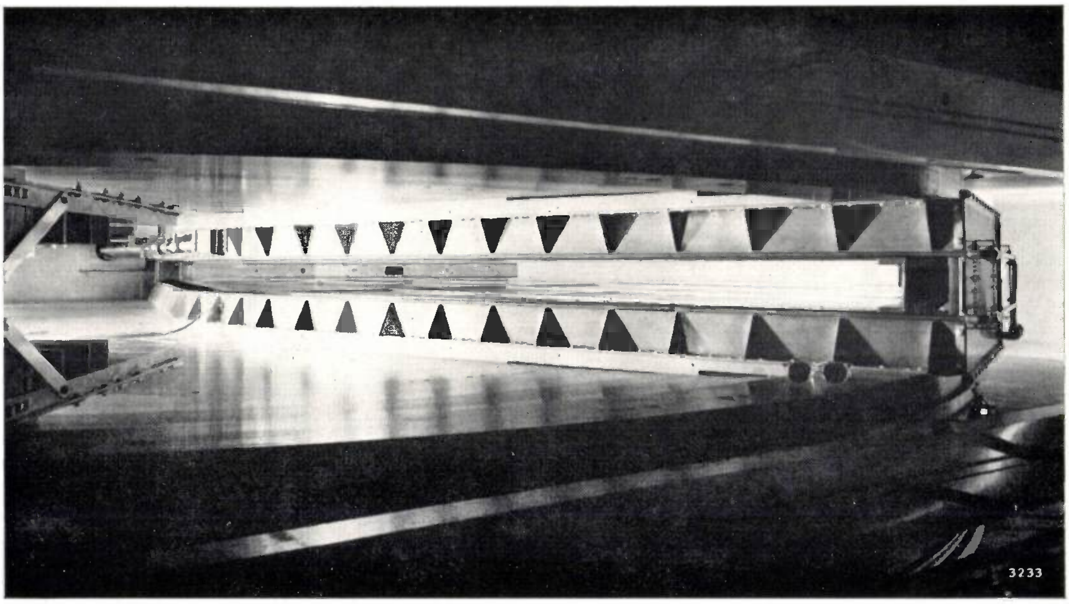


Photo CERN

Fig. 15. View of the dee, mounted between the pole plates. Note the relatively narrow mouth of the dee. To reduce the high demands on the stability of the orbits of the particles in the vertical direction, it would be a help to make the dee mouth wider. This has not been done, however, largely because of the small dee capacitance needed, which requires a large distance between the dee plates and the earthed outer conductor of the dee system. On the left, at the rear, the tube to which the ion source is attached can be seen projecting into the accelerating chamber.

maintenance of the mutual positioning of the tuning fork $T$ (on the stub) and the associated stator $S$ (on the stem), their position being constantly corrected by the feelers $X$ and the suspension mechanism driven by servo motors $A$ (see III, p. 176). To facilitate assembly and disassembly, a trolley is fixed underneath the stub housing, which can be lowered on to rails; after lowering this trolley and disconnecting the flange, the stub housing can be wheeled away.

Finally, a few comments on the assembly of copper plates which constitutes the outer conductor of the resonator (the liner). The part that surrounds the dee, $L_{\mathrm{D}}$, and the dummy dee, $D_2$, are electrically joined to the pole-piece plates $P_1$ and $P_2$ and directly secured to them. The part $L_{\mathrm{H}}$ around the dee stem is at about the same distance from it as at the transition from the dee to the stem. The side walls of these sections of the outer conductor are formed by perforated copper plates. The outer conductor $L_{\mathrm{St}}$ of the stub is carried by steel girders, which are bolted to the inside of the stub housing.

The stub at the shorted end cannot be connected directly but only capacitatively to its outer conductor, the reason being that direct voltages of about 1000 V must be applied to the dee and dee stem and also to the stub in order to avoid a discharge. The capacitance of the shorting capacitor ($C_{\mathrm{b}}$ in fig. 14) should be high and its inductance low. This apparently simple circuit element caused quite a lot of trouble in the construction (see also the end of this article). Use was made for this purpose of commercial ceramic disk capacitors of 1000 pF. The capacitor was built up from 20 groups of 126 disks each (*fig. 16*). The total capacitance is 2.52 μF. The heat generated by the losses in the capacitor is removed by cooling water. During operation the maximum permissible direct voltage is 1500 V.

Fig. 14 also shows the tap connection $k$ on the stub for the radio-frequency supply. By removing a few screws on the outside of the stub housing, this connection can be moved slightly in the longitudinal direction of the resonator for making initial adjustments. The RF generator with its transmitting valve is contained in a housing $O$, which is bolted to the bottom of the stub housing. For replacing the valve, it can be let down and wheeled away.

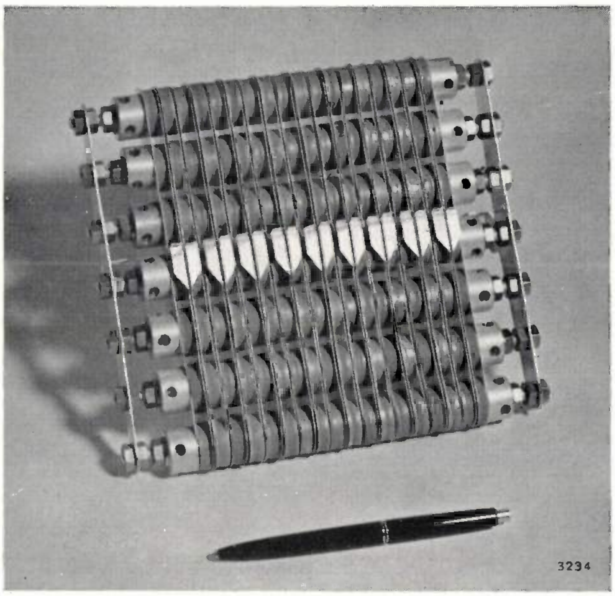


Photo CERN

Fig. 16. One of the twenty groups of ceramic capacitors from which the shorting capacitor $C_{\mathrm{b}}$ is built up.

## The RF generator

The RF voltage for exciting the resonator is supplied by a triode oscillator, fitted with a water-cooled triode TBW 12/100[6]. Even taking all possible losses into account (see p. 151), the power reserve of this 100 kW valve is very appreciable. As we shall see, this triode is particularly suitable for use in a grounded-grid arrangement.

In view of the wide range in which the resonator frequency is varied, the circuitry and design of the oscillator is somewhat unconventional. The oscillator is required to operate stably in this whole range (and preferably only in this range, in order to avoid the possible excitation of parasitic oscillations). Tests on the full-scale model had shown that this was not possible in our case with the conventional feedback circuit generally used in synchrocyclotrons; see *fig. 17*.

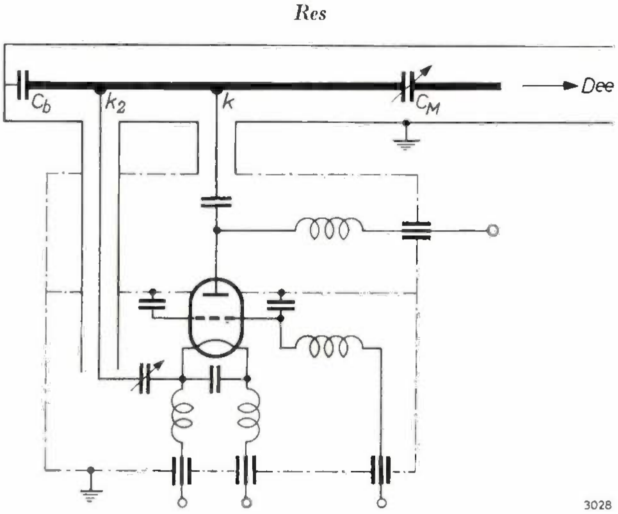


Fig. 17. Oscillator with conventional feedback circuit, as ordinarily used in synchrocyclotrons. The anode is connected at $k$ to the resonator *Res*. Since the feedback voltage in this case is taken from a second tap $k_2$ on the resonator, a loop is produced which can easily give rise to unwanted modes of oscillation. (In its further details the circuit corresponds to that in fig. 18.)

[6] This valve was described in Philips tech. Rev. 14, 226, 1952/53.

The primary reason is that the oscillator is liable to produce oscillations in the loop formed by the relatively long transmission lines to the two connection terminals on the resonator. Furthermore, this circuit makes it difficult to keep the feedback voltage in a suitable phase over the whole frequency range: the phase difference between anode and cathode voltage, which should be zero in the ideal case, becomes so large at the extreme frequencies that the oscillation condition for this circuit can no longer be satisfied.

A circuit that proved highly satisfactory is the one represented in *fig. 18*. The oscillator here is excited by a feedback voltage taken from an inductance $L$ in the anode circuit, but it is the resonator, which is connected at only one point, that determines the frequency. This virtue is attributable to the low losses (high wattless power) of the resonator (the $Q$ is particularly high, $> 2000$). The resonator functions, as it were, like a flywheel: hence the term *flywheel circuit*. The fact that the resonator need be connected to only one point is also an advantage from the point of view of construction, since it calls for only one high-vacuum lead-in for RF current. Moreover, the feedback factor can now be adjusted by shifting the connection tap $a$ in the oscillator itself, i.e. outside the vacuum tank, and the variation of the feedback factor with frequency can be controlled by the addition of circuit elements in the easily accessible feedback system. Since the grounded-grid arrangement, when properly designed, ensures the efficient internal decoupling of anode and cathode, the feedback is governed solely by the circuit elements outside the valve.

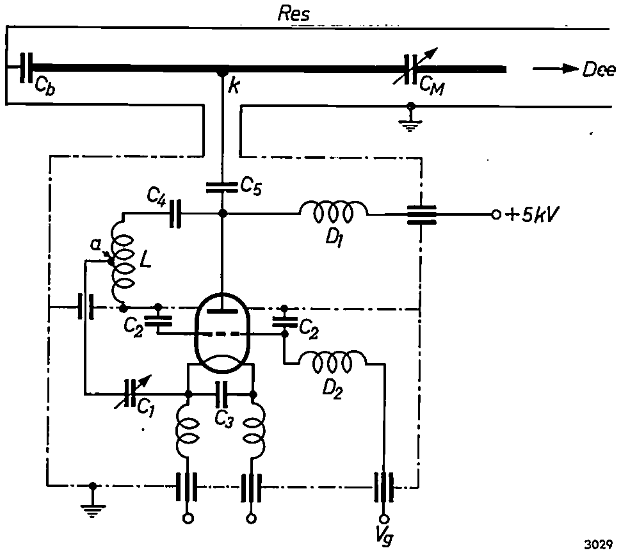


Fig. 18. Oscillator with flywheel circuit. The feedback voltage is tapped from a point $a$ on the inductance $L$; the resonator *Res*, which is connected at only one point $k$, itself determines the frequency owing to its high $Q$, i.e. its high wattless power. The inductances of the tube and supply lines are compensated by means of the remotely controlled variable capacitor $C_1$ in order to ensure that the alternating voltage on anode and cathode are exactly in phase. The capacitor is adjusted such that this is achieved at the centre frequency of the resonator frequency swing. $C_2$ capacitor for RF earthing of the grid, $C_3$ for shunting the filament. $C_4$ and $C_5$ anode-voltage blocking capacitors. $D_1$ and $D_2$ chokes.

This arrangement proved to be so effective, particularly in the avoidance of parasitic oscillations, that it was subsequently adopted by Philips for the 150 MeV cyclotron in Paris.

If the anode circuit, which in fig. 18 consists of an inductance $L$ shunted across the resonator, is to act as a resonant circuit, the resonator proper must possess capacitative reactance in the relevant frequency range. The frequency of the oscillator is therefore always slightly above the natural frequency of the free resonator. Because of the high quality of the resonator, however, the difference is negligibly small.

The self-inductances of the valve and its connections tend to give rise to differences between the alternating voltages applied to anode and cathode, which theoretically should be in phase. These inductances are compensated by a capacitor $C_1$, which consists of two variable vacuum capacitors connected in parallel. Of course, the compensation can only be perfect at one particular frcquency. When the cyclotron is in operation, this frequency is held in the middle of the modulation range. For this purpose, one of the vacuum capacitors can be varied from the control room, 60 metres away from the cyclotron (see I). The phase deviation at both ends of the modulation range is relatively small, owing to the relatively low $Q$ of the cathode portion of the feedback circuit (which, as we have seen, does not determine the frequency).

We shall now briefly consider some other details of the design. The TBW 12/100 triode is built in such a way as to allow an effective separation between the cathode and anode of the oscillator, as required in a grounded-grid arrangement. To this end the oscillator housing is divided into two parts by a horizontal partition at the level of the grid connection. Extremely effective RF earthing of the grid is achieved by fitting the partition with 96 ceramic capacitors of 1000 pF each, arranged in a circle around the valve ($C_2$ in fig. 18). Care was taken to avoid resonances from these capacitors in the range between 16 and 29 Mc/s. The same type of capacitor is used for shunting the filament ($C_3$). The coils used for the inductance $L$ in the anode circuit and the choke $D_1$ in the anode-voltage supply line are of tubing through which an air current is passed for cooling. A ceramic pot capacitor $C_5$ is included in the anode connection on the resonator for DC blocking. The line for the connection is kept as short as practical, and special arrangements are made to keep the inductance of the capacitor connections very small, so that the RF oscillator voltage is transmitted to the resonator with as little attenuation as possible.

Apart from the usual electrical screening with sheet copper, the generator housing is surrounded by a sheet-steel housing to shield the valve from the stray field of the cyclotron magnet (which would adversely affect the mutual conductance). For the same reason the anode can is enclosed in a 10 millimetre thick iron tube. For insulation from earth the cooling water is again passed through a double hose of polyvinyl chloride about 10 m long.

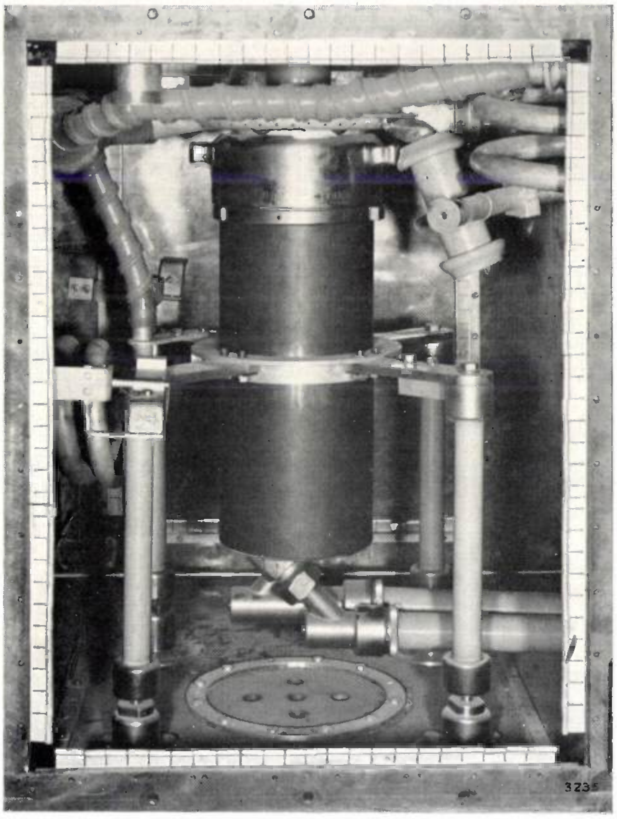


Photo CERN

Fig. 19. View inside the generator housing. Centre, the anode can with cooling jacket; top right, the feedback coil, consisting of three turns; left, one of the chokes.

*Fig. 19* gives a view of the interior of the generator housing, showing some of the components mentioned. The normal operating data of the oscillator are presented in *Table I*.

### Ancillary equipment

The radio-frequency system also comprises power supplies, a cooling installation for the cooling water, a device for pulsing the cyclotron, and instruments for controlling and monitoring the system. All this ancillary equipment, which is more or less conventional in design, is housed separately from the cyclotron in the power house which contains the converter set supplying the current for the large magnet (*fig. 20*; see *El* in fig. 3 in I). The RF system can largely be operated and controlled from here instead of from the master control room (fig. 5 in I). The operating controls are interlocked in such a way as to safeguard the system from damage due to manipulating the controls in the wrong order. The main switching operations can be checked on panels both in the master control room and in the converter room *El*. These operations include switching on and off the filamant current for the HT rectifier and for the oscillator valve; switching on and off the anode voltage; raising and lowering this voltage, and also the direct voltage for dee and stub; switching on and off the tuning-fork modulator and raising or lowering its amplitude. The equipment is provided with elaborate safety devices: all major faults are individually signalled, and if dangerous faults arise the relevant part of the installation is automatically switched off.

Table I. Oscillator data for normal pulsed operation.

| | |
|---|---|
| Anode voltage | 5-6 kV |
| Anode current (average value) | 3-4 A |
| Grid current (average value) | 1-1.5 A |
| Duty cycle (fraction of modulation period) | 60-70% |
| Frequency sweep | 16.5-29 Mc/s |
| Anode efficiency | 59-62% |

The object of the above-mentioned pulsed operation of the cyclotron is to obviate needless thermal loading, in particular of the modulator: the oscillator need only operate during that portion of the modulation period in which particles can be accelerated. The oscillator is controlled by pulses applied to its grid. The requisite switching pulses are supplied by the tuning fork itself (III, fig. 25). The fairly elaborate equipment required for producing the pulses and for effecting synchronization with the tuning fork or with external pulses (e.g. for experiments with a bubble chamber) will not be discussed here.

During the operation of the cyclotron the RF system has caused no particular difficulties. At certain frequencies, dips sometimes appeared in the amplitude curve of the dee voltage, and to compensate for these the anode voltage had to be chosen somewhat higher than normal (6 kV). It was soon discovered that these dips were attributable to reso-

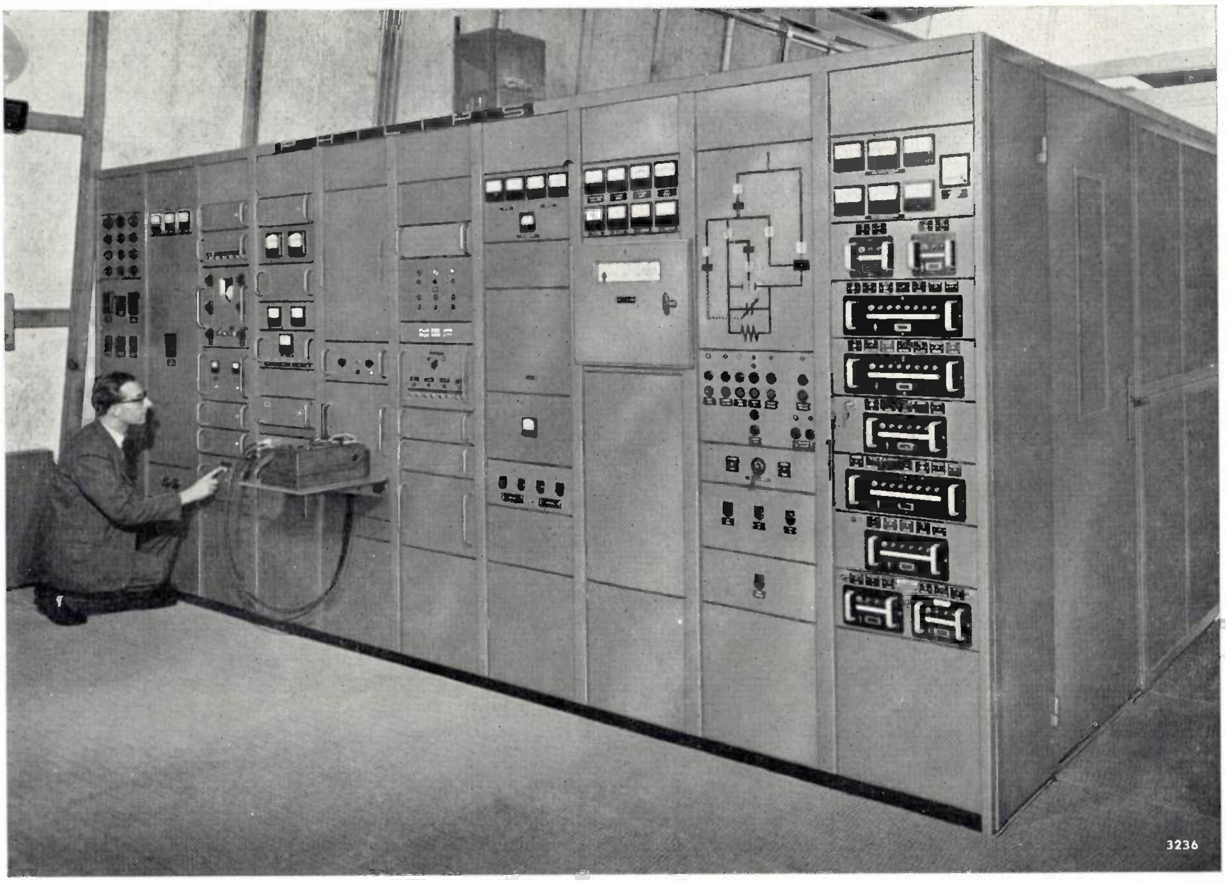


Photo CERN

Fig. 20. Part of the control and monitoring equipment for the RF system.

nance frequencies from the stub shorting capacitor ($C_b$ in fig. 14). Since the properties of the ceramic dielectric of this capacitor were somewhat temperature-dependent, the occurrence and location of the dips in the curve depended markedly on the time during which the cyclotron had been in operation and on the cooling of the capacitor. Not much could therefore be done about these dips. While the synchrocyclotron was shut down for a while, however, the original shorting capacitor was replaced by another ceramic capacitor specially designed for this purpose, and consisting of only a few large plates.

**Summary.** The RF system of the CERN synchrocyclotron, developed and built by Philips Eindhoven, in cooperation with CERN engineers, contains two main components which are briefly described. These are the resonator and the RF generator, with elaborate ancillary equipment. The required proton energy of 600 MeV calls for a very large frequency sweep, from 29 to 16.5 Mc/s. The tuning-fork capacitor used for this frequency modulation provides a capacitance variation between 256 and 2580 pF. By combining this capacitor with an inductance ("stub") to form a series resonant circuit, a dee system could be designed whose resonant frequency varies between the above-mentioned limits. Some underlying theoretical considerations are discussed, and details given of the design work on the dee system, for which three different models were used. The voltage and current distributions on the resonator at the highest and lowest frequencies are given on the assumption of an RF supply voltage of constant amplitude. The voltage amplitude at the dee mouth is shown to be considerably greater during the whole modulation period than accords with the "ideal" amplitude programme (for a phase $\varphi = -30^\circ$ of the accelerated particles). The construction of the resonator system is then described, mention being made of the measures taken to avoid mechanical strains in the resonator resulting from deformations in the cyclotron. The circuit chosen for the oscillator is a "flywheel" circuit, not previously used in cyclotrons (using a water-cooled 100 kW triode TBW 12/100 oscillating valve). The essential feature of this circuit is that the control voltage for the valve is taken from a feedback circuit in the oscillator, although the resonator — which is connected at only one point — still determines the frequency, owing to its high $Q$. This circuit largely overcomes the troubles experienced from parasitic modes of oscillation in the RF system, which have been such a source of difficulties in other cyclotrons. The construction of the oscillator is briefly discussed, and passing mention is made of the ancillary apparatus, including the equipment for pulsing the cyclotron.

# III. THE TUNING-FORK MODULATOR

by B. BOLLÉE *) and F. KRIENEN **).

621.376.32:621.384.611.2

The function of the modulator in the RF system of the CERN cyclotron is to cause the frequency of this system to swing periodically (55 times per second) from about 29 to 16.5 Mc/s (wavelength 10.3 to 18.2 metres). The frequency modulation in a synchrocyclotron is commonly effected by means of a rotating capacitor. The RF voltage which thereby appears on the bearings and on the vacuum-tight lead-in to the rotating shaft can, however, give rise to considerable difficulties. Another method is based on the use of a varying inductance, consisting of a coil with ferrite core whose permeability is varied by passing a periodically varying premagnetizing current through an auxiliary winding. In the present case it was to be foreseen that this method would flounder on the magnetic losses in the ferrite. These considerations led to the trial in the CERN machine of a new method of frequency variation hitherto little used in synchrocyclotrons, namely the use of a vibrating capacitor [1]). The development of this system was entrusted to the Philips Laboratories at Eindhoven, in 1952, that is before the official foundation of the CERN.

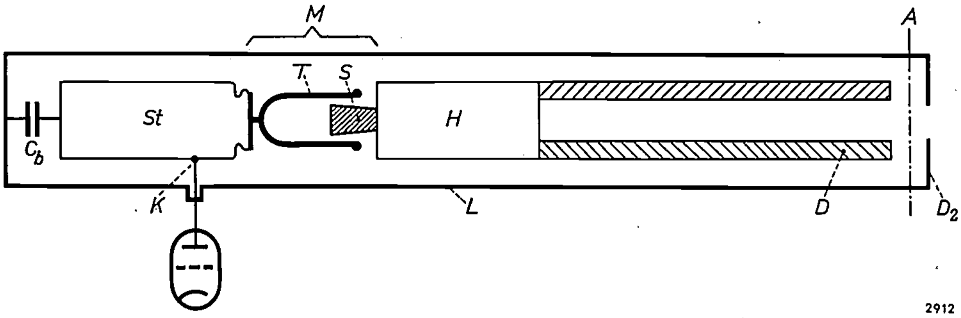


Fig. 1. The radio-frequency system of the CERN synchrocyclotron may be regarded as a coaxial transmission line. *A* centre line of the cyclotron (position of ion source). *D* dee. $D_2$ dummy dee (earthed). *H* stem of dee. *M* modulator, with tuning fork *T* and stator *S*. *St* "stub" (end section of transmission line). *L* outer conductor of transmission line. *K* coupling to anode of oscillating valve. $C_b$ capacitor giving RF connection of stub to outer conductor but enabling the stub to be brought to a DC potential with respect to earth (see II, p. 158).

First we shall give here a broad description of the system that was finally evolved after lengthy experiments. Numerous difficulties had to be overcome before arriving at the present satisfactory results. We shall then go on to discuss in more detail the structural design, the drive mechanism and the cooling of the vibrating system, together with various subsidiary problems.

## Outline of the vibrating capacitor system

The varying capacitance of the modulator is achieved by means of a vibrator in the form of a tuning fork, the prongs of which encompass a stator (see *fig. 1*).

Between the prongs of the tuning fork and the stator a voltage of 10 to 20 kV is applied. In view of the risk of flash-over, it was desirable to make the distance between the prongs and the stator not less than 1.5 mm. This meant that the prongs of the tuning fork had to be made very wide in order to achieve a sufficiently high maximum capacitance. This was also desirable for other reasons: the density of the RF current over the prongs had to be small enough to allow adequate removal of the heat thereby generated. The prongs of the tuning fork, or the *blades*, as we can now better call them, are therefore 2 metres wide, which was roughly the maximum width allowed by the available space between the two vacuum pumps of the accelerating chamber (see part I, fig. 7). The stator, situated symmetrically between these blades with an overlap of 10 cm, is a 2-metre long aluminium girder bolted to the stem of the dee. Limitations were also imposed on the other

*) Philips Research Laboratories, Eindhoven.

**) CERN, Geneva.

[1]) The 740 MeV synchrocyclotron at Berkeley in California, the world's largest, also uses a vibrating capacitor, though of different design (with vibrating tongues). See R. L. Thornton, CERN Sympos. 1956$^{I}$, p. 413; B. H. Smith, K. R. Mackenzie, J. Reidel *et al.*, Wescon Conv. Rec. I.R.E. **1**, 60, 1957.

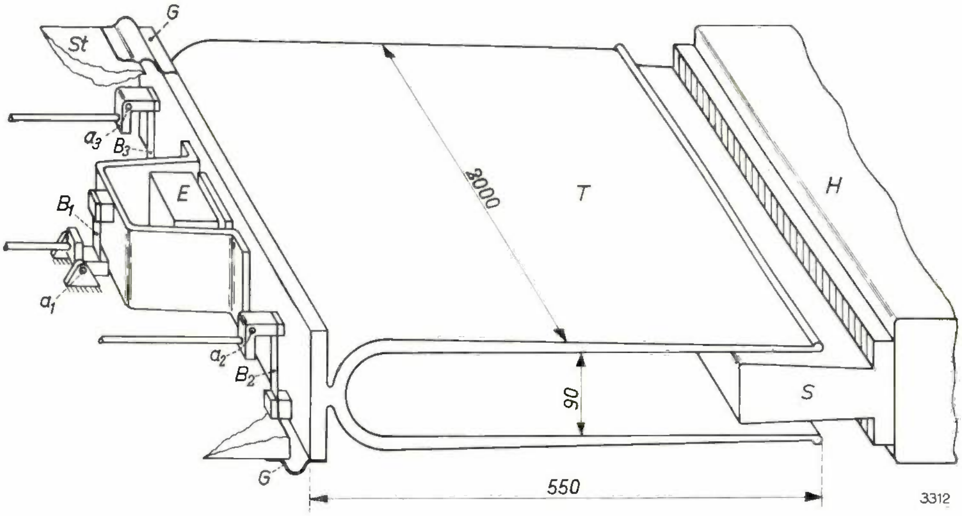


Fig. 2. Perspective sketch of tuning fork $T$ and stator $S$. $H$ dee stem. $E$ drive mechanism. $B_1$, $B_2$, $B_3$ three steel strips on which the tuning fork is suspended. $a_1$, $a_2$, $a_3$ spindles on bearings in the stub frame. $G$ leaf springs of beryllium bronze, fitted over the whole width of the tuning fork to effect electrical connection with the stub $St$.

dimensions of the tuning fork. The height was limited by the height of the vacuum chamber to about 20 cm; the length was limited to about 60 cm, having regard to the specified total length of the RF system and the optimum position for the coupling to the anode of the oscillating valve (see Part II, p. 154).

The tuning fork, which is attached by a short stem to a foot, was made from a solid block of aluminium to within an accuracy of $^1/_{10}$ mm. It weighs about 75 kg, and its natural frequency is 55 c/s. Further details of the shape and dimensions of the tuning fork are given in *fig. 2*, and *fig. 3* shows a photograph of the finished construction.

Fig. 2 illustrates how the tuning fork is set in vibration by an electromechanical converter, fixed to the foot. The operation depends on the converse of the well-known phenomenon that a vibrating tuning fork held to a sounding board transfers energy to the latter (and thereby becomes more audible). If, instead of the tuning fork, the sounding board is vibrated at the appropriate frequency, energy is transferred in the opposite direction and the prongs of the tuning fork are set in vibration. In the present case the foot — which acts as the sounding board — must be given a horizontal amplitude of about 0.15 mm if the blades are to vibrate at the required amplitude of 12.5 mm at each of the "lips". The tuning fork is mounted on steel suspension strips ($B_1$, $B_2$, $B_3$ in fig. 2), so that the horizontal movement is virtually unimpeded. The electrical connection to the stub (the end section of the resonator system, *St* in fig. 1) is effected by bent leaf springs ($G$) of beryllium bronze. No vibration energy of consequence is transmitted via the strips and leaf springs to the stub and its supporting insulators. The large amplitude of the tuning-fork blades is illustrated by the photograph in *fig. 4*.

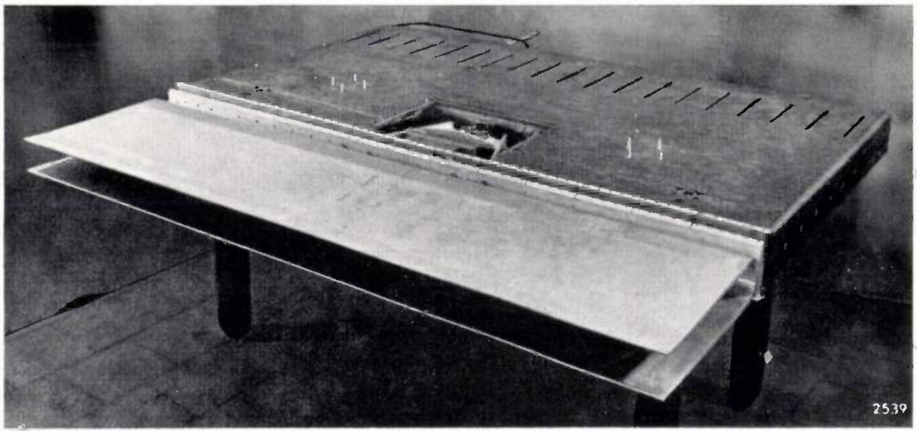


Fig. 3. The tuning fork mounted to the stub.

We shall now consider the properties of the tuning fork as a capacitor. In the "closed" position the inner face of the blades is parallel to the surface of the stator, the spacing $\delta$ between them being 1.5 mm. The capacitance is then 2580 pF. In the "opened" position, measured at the lips, $\delta$ is 26.5 mm and the

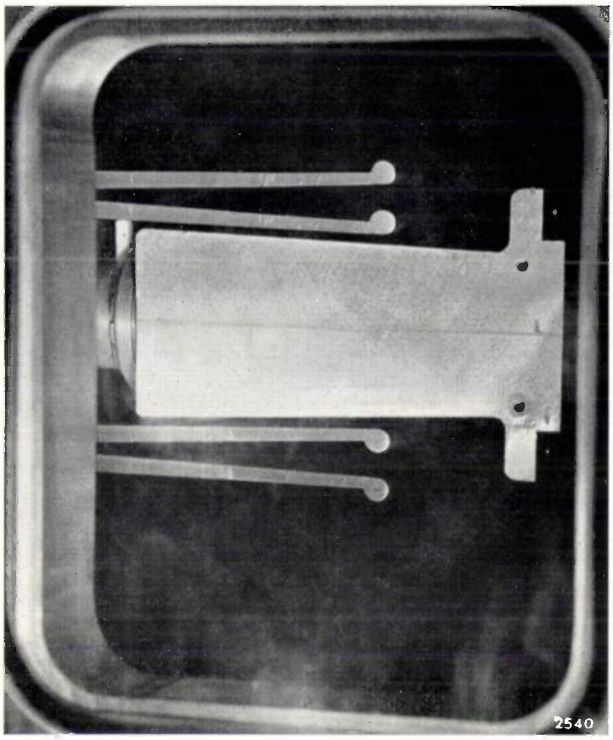


Fig. 4. Stroboscopic photograph, at 110 flashes per second, of the stator of the modulator and the overlapping ends of the blades of the vibrating tuning fork. Each blade is thus shown in its two extreme positions ("closed" and "open" state of tuning fork). Note the considerable amplitude of the blades.

capacitance 256 pF. These figures reveal that the relative variation of the capacitance (approx. 1 : 10) is considerably smaller than the relative variation of $\delta$ (approx. 1 : 18). This is due to the non-parallel motion and to the varying spread of the electrical field. This is evident from the pattern of the electrical lines of force and equipotential surfaces, from which the capacitance can be determined [2]); see *fig. 5a* and *b*.

If the synchrocyclotron is to function properly, the electrical oscillation frequency $f$ of the RF system must fall successively from the maximum to the minimum value in accordance with a certain time function (see part II). The mechanical vibration of the tuning fork, i.e. the variation of the distance $\delta$ between the blades and the stator as a function of time, is sinusoidal. The variation of the capacitance $C$ as a function of $\delta$ — between the two extreme values mentioned — is represented in *fig. 6a*. The variation of $f$ as a function of $C$ (fig. 6*b*) was found from calculations and from measurements on models of the RF system, where the vibrating capacitor was replaced by an adjustable plate capacitor. From

[2]) E. Weber, Electromagnetic fields, theory and applications, I. Mapping of fields, Wiley, New York 1950.

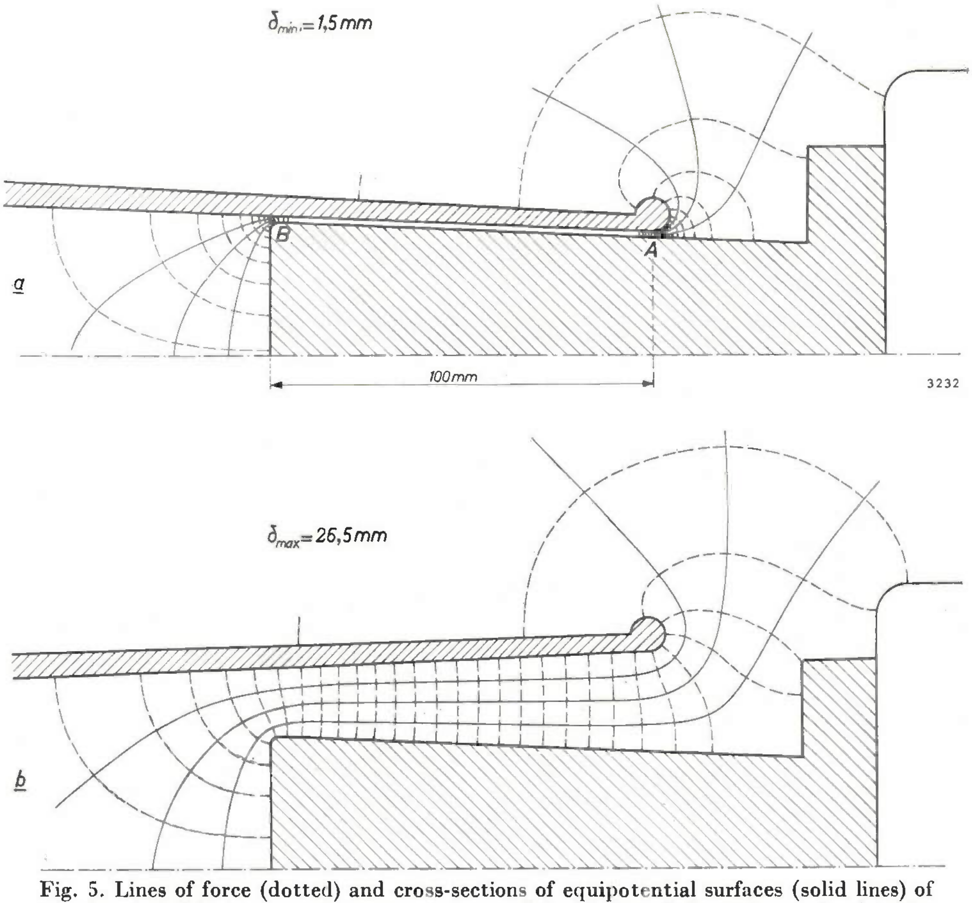


Fig. 5. Lines of force (dotted) and cross-sections of equipotential surfaces (solid lines) of the electrical field between the stator and one of the blades of the tuning fork, *a*) in the closed state, *b*) in the open state. In (*b*) the electrode faces are not parallel.

these three relations we can construct the resultant curve of $f$ as a function of time. This is shown in fig. 6c.

Structures of such large dimensions as this cyclotron are always subject during operation to perceptible deformations due to variations of temperature and pressure, to magnetic forces, to elastic after-effects in vacuum seals, and so on. These deformations affect the relative position of stator and tuning fork. The amplitude of the tuning fork is also subject to a variety of changing influences, like the electrostatic forces between tuning fork and stator and the damping caused by eddy currents, which are induced in the moving blades by the stray field of the cyclotron magnet. Variations in these factors are unavoidable, since nuclear experiments with the machine require that it may be possible to vary not only the magnetic field but also the magnitude of the RF voltage and the length of time it is switched on. Since it is essential to the proper operation of the cyclotron that the relative position of stator and tuning fork, and also the amplitude of the latter, should not vary significantly during operation (e.g. less than 0.1 mm), in order to maintain the limits of the frequency variation, it was necessary to install elaborate control equipment. This equipment, which will be briefly described below, serves the additional purpose of protecting the tuning fork from damage: in the event of a fault, it switches off the drive and in some circumstances the RF oscillator also.

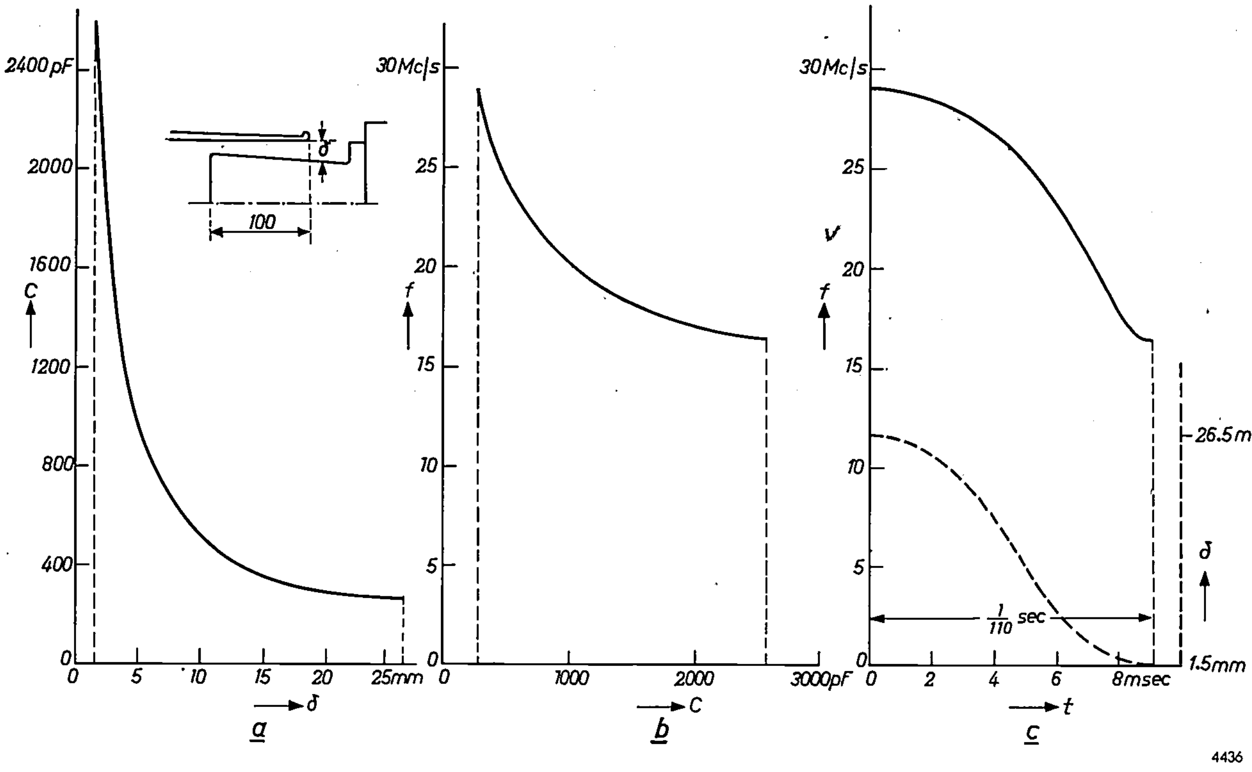


Fig. 6. *a*) Capacitance $C$ of vibrating capacitor, as a function of the distance $\delta$ between the blades and the stator.
*b*) Oscillation frequency $f$ of the RF system as a function of $C$.
*c*) Variation of oscillation frequency $f$ with time $t$, during one modulation period. This relation is derived from (*a*) and (*b*), assuming $\delta$ varies sinusoidally with time (dashed curve).

When the cyclotron is operating, the RF current on the outer surface of each blade, and also on the inside near the stator, is of the order of 1000 amperes. The removal of the heat generated by this current created various problems, which we shall also consider at some length.

### Mechanical design of the tuning fork

To illustrate the procedure adopted in designing the tuning fork, we shall take as our starting data a frequency of 55 c/s, the maximum and minimum capacitance required and the maximum available space of 200×60×20 cm. The tuning fork in its present form was developed in successive stages, partly empirically and partly on the basis of theoretical considerations stemming from the simple case of a vibrating bar of constant cross-section clamped at one end. It was necessary to take into account the material stresses and fatigue lifetime, the driving system, the weight and the method of support, the heating, and finally the problems of fabrication.

It will be useful first of all to confine our considerations to the form of the longitudinal section (the profile) of the tuning fork. The effect of the large width of the plates will be dealt with at a later stage.

*The vibrating cantilever bar*

When a uniform bar of free length $l$ and thickness $h$, clamped rigidly at one end (*fig.* 7), vibrates in its fundamental mode with a maximum amplitude $p$, the maximum bending stress $\sigma$ occurs at the place where the bar is clamped, and is given by:

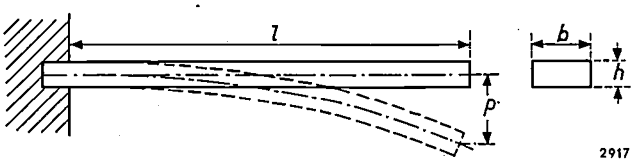


Fig. 7. Uniform bar clamped at one end. The width $b$ of the cross-section is irrelevant to considerations concerning the amplitude $p$, the frequency $f_m$ and the bending stress $\sigma$.

$$\sigma = 1.76\,\frac{E\,h\,p}{l^2}\,, \quad \ldots\ldots \quad (1)$$

where $E$ is the modulus of elasticity. The fundamental vibration (for which the bar deflects as in fig. 7) has the frequency

$$f_m = 0.162\,\frac{c\,h}{l^2}\,, \quad \ldots\ldots \quad (2)$$

where $c$ is the speed of sound in the bar material. From (1) and (2) we at once find the important relation:

$$\sigma = K_v\,\frac{E}{c}\,f_m p\,, \quad \ldots\ldots \quad (3)$$

in which the length and the cross-section of the bar have thus been eliminated. The numerical factor $K_v$ is 10.9. We write the equation in this form because it is found to apply equally to vibrating bars of other longitudinal profile though with different values for the "form factor" $K_v$.

Before drawing conclusions from formula (3) we shall deal with the choice of material. From the points of view of fatigue strength and manufacturing possibilities, the choice lay between an aluminium alloy and a (non-magnetic) steel. The speed of sound $c$ in both materials is roughly 5000 m/sec; thus according to equation (2) the same dimensions would apply to both cases. Aluminium was preferable because of its better electrical conduction, its lower mechanical damping and better workability, and the choice finally fell on an aluminium alloy, "Permandur", which possesses very high fatigue strength whilst retaining a reasonable electrical conductivity.

With reference to the bar, eq. (3) expresses the general fact that the natural frequency $f_m$ and the amplitude $p$ of a vibrating body cannot be raised independently, a limit being set by the permissible loading of the material. This is a problem encountered, for example, in mechanical sound-recording, where the problem is to record the highest audible frequencies with a reasonable amplitude [3]). It is perhaps surprising that, with our tuning fork, this problem arises at the relatively low frequency of 55 c/s. The reason is the very large amplitude required. The value of $p = 1.25$ cm was specified from an estimate of the capacitance variation obtainable with a vibrator as in fig. 7. If we substitute this value for $p$ in eq. (3), taking $f_m$ and $c$ at the values mentioned and with $E_{Al} = 0.7 \times 10^6$ kg/cm$^2$, we find a bending stress of $\sigma = 1050$ kg/cm$^2$. This is much greater than the value of 600 kg/cm$^2$, which is regarded as admissible in the case of aluminium required to have a fatigue life of about $10^9$ vibrations.

The fact that $p = 1.25$ cm at a frequency $f_m = 55$ c/s is a very large amplitude may perhaps best be appreciated by comparing the situation with that under static loading. For the bar to have the desired natural frequency it could be made, say, 1.5 cm thick and 46.8 cm long (see eq. 2). If the "bar" is given a width of 2 metres, like our tuning fork, the force that would have to act uniformly over its free edge to produce a deflection of 1.25 cm would be 4.4 tons.

With a uniform bar, then, it was not possible to meet the requirements.

As long ago as 1879 Kirchhoff showed that, given the same natural frequency and amplitude, the bending stress in bars of wedge-shaped profile (*fig. 8*) is substantially lower [4]). Eq. (3) is applicable here with the much smaller form factor $K_v = 2.78$. This would make it possible to achieve the desired capacitance variation with a bending stress of only 260 kg/cm$^2$. A "mathematical" wedge cannot be

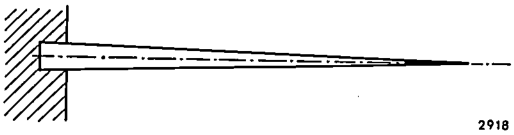


Fig. 8. Bar with wedge-shaped profile, clamped at one end.

[3]) See e.g. A. T. van Urk, The sound recorder of the Philips-Miller system, Philips tech. Rev. 1, 135-141, 1936, formula (3).

[4]) G. Kirchhoff, Über die Transversalschwingungen eines Stabes von veränderlichem Querschnitt, Ann. Wiedemann 10, 501-512, 1880.

used because a sharp edge would greatly increase the risk of electrical breakdown between vibrator and stator, quite apart from the manufacturing difficulties it would involve and its mechanical vulnerability. A compromise can be found, however, between the uniform and wedge shapes by giving the profile the form of a truncated wedge, the form factor for which will be between 10.9 and 2.78. The process of calculating the natural frequency and form factor of a bar of this shape is fairly complicated. The shape finally computed [5]) is shown in *fig. 9*, the form factor for which is $K_V = 4.87$. The maximum bending stress is 460 kg/cm², which is well below the permissible value of 600 kg/cm².

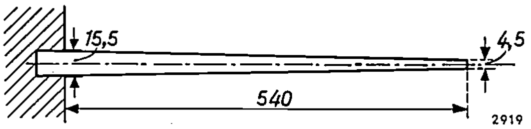


Fig. 9. Cantilever bar of truncated wedge profile chosen for the blades of the tuning fork.

For the interested reader we shall give a brief outline of the above-mentioned calculation. For a bar of density $\varrho$ and of linearly-varying thickness as in fig. 9, whose cross-section at unit distance from the vertex of the mathematical wedge has an area $A_0$ and a moment of inertia $I_0$, the deflection $y$ as a function of the distance $u$ to this vertex is given by the differential equation [6]):

$$4\pi^2 f_m{}^2 y = \frac{EI_0}{\varrho A_0}\frac{1}{u}\frac{\mathrm{d}^2}{\mathrm{d}u^2}\left(u^3\frac{\mathrm{d}^2y}{\mathrm{d}u^2}\right). \quad . \quad . \quad . \quad . \quad . \quad (4)$$

Changing to the new variable

$$v = 2\pi f_m u \sqrt{\frac{\varrho A_0}{EI_0}}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

we can express (4) in the simpler form:

$$vy = \frac{\mathrm{d}^2}{\mathrm{d}v^2}\left(v^3\frac{\mathrm{d}^2y}{\mathrm{d}v^2}\right). \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

It is easy to verify that all solutions of the differential equation

$$v\frac{\mathrm{d}^2y}{\mathrm{d}v^2} + 2\frac{\mathrm{d}y}{\mathrm{d}v} = y \quad . \quad . \quad . \quad . \quad . \quad . \quad (7a)$$

satisfy (6) and the same holds for all solutions of:

$$v\frac{\mathrm{d}^2y}{\mathrm{d}v^2} + 2\frac{\mathrm{d}y}{\mathrm{d}v} = -y. \quad . \quad . \quad . \quad . \quad . \quad (7b)$$

The general solution of (6) can therefore be written as a linear combination of the general solutions of (7a) and (7b), which are special forms of the Bessel differential equation. These solutions are:

$$y = \tfrac{1}{2}v^{-\frac{1}{2}}\,[AJ_1(2v^{\frac{1}{2}}) + BN_1(2v^{\frac{1}{2}})]$$

and

$$y = -\tfrac{1}{2}v^{-\frac{1}{2}}\,[jCJ_1(2jv^{\frac{1}{2}}) + DH_1^{(1)}(2jv^{\frac{1}{2}})],$$

where $J_1$, $N_1$ and $H_1^{(1)}$ are respectively first-order Bessel, Neumann and Hankel functions. With the new variable $w = 2v^{\frac{1}{2}}$ the deflection is finally given by

$$wy = AJ_1(w) + BN_1(w) - jCJ_1(jw) - DH_1^{(1)}(jw)\,. \quad . \quad (8)$$

The four integration constants $A$, $B$, $C$, $D$ have to be found from the boundary conditions. These are:

1) The angle of inclination $\mathrm{d}y/\mathrm{d}u$ at the clamped end = 0.
2) The deflection $y$ at the clamped end = 0.
3) The bending moment, i.e. $\mathrm{d}^2y/\mathrm{d}u^2$, at the free end = 0.
4) The transverse force, i.e. $\mathrm{d}^3y/\mathrm{d}u^3$, at the free end = 0.

Substitution of these boundary conditions in (8) yields four homogeneous equations in the four constants. This system of equations has a solution only if its determinant, consisting of 16 Bessel functions, is zero. From the equation thus obtained one can determine, though rather laboriously, the eigen values $w$ and hence, via eq. (5), the natural frequencies $f_m$. Of these we are only interested in the lowest frequency (the fundamental).

### *The bent bar*

Even with the profile thus derived, the cantilever (wedge) bar — or rather, in view of its width, the cantilever (wedge) plate — is still not immediately a practical proposition. Noticeable deformation is unavoidable at the edges of the clamping blocks, with all the disadvantages this entails. Furthermore, powerful longitudinal forces as well as transversal forces act on the clamped end during vibration, and whilst the transversal forces might be balanced in our case, namely by using two plates vibrating in antiphase and clamped in the same block, longitudinal balancing is not readily possible. The latter forces are thus transmitted via the clamped end to the stub and to the accelerating chamber in the cyclotron.

It can be seen that these difficulties are circumvented by building the vibrator in the form of a tuning fork. This may be regarded as consisting of two bars, each having a profile as calculated above (fig. 9) but bent to form a half U. At the point where the two halves are "joined" (in reality, of course, the whole U-bar is made from a solid block of material) the same boundary conditions are realized as in the case of clamping. The difference, however, is that no forces whatsoever need be transmitted to the surroundings, since the tuning fork can be mounted on weak suspension springs. For the fundamental mode the tuning fork then vibrates as sketched in *fig. 10*.

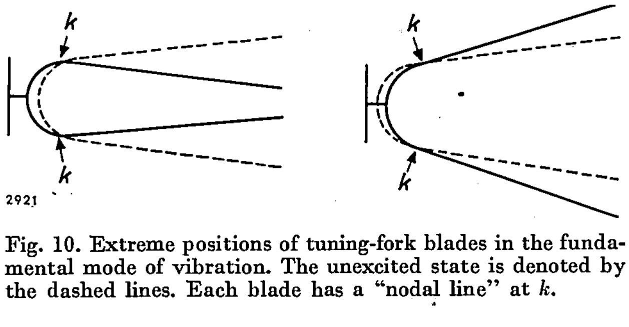


Fig. 10. Extreme positions of tuning-fork blades in the fundamental mode of vibration. The unexcited state is denoted by the dashed lines. Each blade has a "nodal line" at $k$.

[5]) See F. Krienen, Modulator CERN synchrocyclotron, CERN report 58-8, 23rd April 1958.

[6]) See e.g. J. W. Strutt (Lord Rayleigh), The theory of sound, Macmillan, London 1926 (2nd ed.), Part I, Chapter 8; P. M. Morse, Vibration and sound, McGraw-Hill, New York 1948.

The mode of vibration and frequency of the bar bent into a half U cannot be calculated exactly. Experiments show that the frequency of the bar is not much altered by the bending [7]. Investigation of the mode of vibration of a model (plate width 30 cm), using stroboscopic illumination, showed that there was not much change either in the deflections with respect to the centre plane of the bar profile. The important question remained, however, whether the bar would not be subject to considerably larger mechanical stresses $\sigma$ at a given frequency $f_m$ and amplitude $p$ than the straight bar. Since the modes of vibration were the same, we drew the inference that the maximum kinetic energy was the same in both cases, and hence also the maximum potential energy (work expended on bending). But to infer from this that the stresses in the bent bar were the same as in the straight bar, we needed to be certain that the neutral plane (plane of zero deformation) remained in the centre plane of the bar profile during the vibrations of the bent bar, as it does in the straight bar. This was by no means certain; on the contrary, there was reason to believe that the neutral plane shifts, since it is known that this is the case for the deformation of thick rings, and since the major part of the bending energy of the U-bar is accumulated in the bent portion where it most resembles such a ring.

The simplest way out of this difficulty was to measure the stresses directly with the aid of strain gauges affixed to the inside and outside of the model. The largest alternating bending stress was found to be 370 kg/cm$^2$ on the outside, and 510 kg/cm$^2$ on the inside. These values were thought to allow a sufficient margin of safety, having regard to the "permissible" stress of 600 kg/cm$^2$. To make doubly sure, a 1 : 4 scale model was made in which, at a quarter the amplitude and a four times higher frequency (220 c/s), the same bending stresses arise. This model was subjected to a life test of $10^9$ vibrations. The higher frequency reduces the length of the test to 52 days and nights, which the model withstood very satisfactorily.

In fact, it was these results that led to the decision to fix the modulation frequency at 55 c/s. It had been the aim to have as high a modulation frequency as possible, since this favours a high average proton current. Actually, the upper limit set to the repetition frequency by the requirement that the particles must have sufficient time to reach their final energy is considerably higher than 55 c/s (viz. above 100 c/s), as explained in Part II.

The actual dimensions of the tuning fork (fig. 2), which it was not yet necessary to include in these considerations concerning frequency, amplitude, mode of vibration and stresses, were established by the absolute magnitude of the capacitance to be achieved.

After these preparatory investigations, two tuning forks (one as a stand-by) were made with the required dimensions from two rolled blocks of aluminium. These blocks were ultrasonically tested for the presence of flaws. After machining, the tuning forks were finished by hand to an accuracy of 0.1 mm. In view of the marked effect which the state of the surface has on the fatigue strength, the bent part was polished on the inside and outside.

For the construction of the stator it was important to ascertain whether, in the closed state, the amplitude of the lips was sufficiently constant over the whole width of 2 metres. This again was checked stroboscopically. The amplitude, measured within 0.1 mm, was found to be constant within the permissible margin of 0.2 mm. The stator could thus be made in the form of a straight uniformly profiled bar. The profile had to be adapted to thc shape of the tuning-fork blades in their closed position. The part of the blades involved ($A$-$B$ in fig. 5) was found to remain straight within 0.1 mm.

## Drive mechanism

From fig. 10 it may be inferred that the tuning fork will vibrate as required if the foot is caused to vibrate periodically in the direction of the axis of symmetry at a frequency equal to the fundamental frequency of the tuning fork. First thoughts perhaps suggest transferring the necessary alternating force to the foot by means of a rigidly mounted system, driven magnetostrictively, electromagnetically or electrodynamically (like a loudspeaker). Model tests soon made it clear, however, that the vibrations transmitted to the surroundings are prohibitive in the case of a rigidly mounted drive system. We may illustrate this by making an estimate of the necessary alternating forces — which, of course, are exerted as reaction forces on the support of such a rigid drive system. In the two blades of the tuning fork a vibration energy of about 60 joules is accumulated. From very rough model tests it was estimated that about $\frac{1}{2}$% of this energy is dissipated per vibration. The power to be supplied by the drive at a frequency $f_m = 55$ c/s was accordingly estimated at $P = 15$ W (see also below). Assuming that the

[7] This has also been demonstrated theoretically for the prismatic bar (fig. 7); see C. H. Keulegan, On the vibration of U-bars, Bur. Stand. J. Res. 6, 553-592, 1931.

driving power $F$ is in phase with the velocity $\dot{x}$ at which the foot is displaced, we can write:

$$P = \tfrac{1}{2} \times 2\pi f_{\rm m} \hat{x}\ \hat{F}.$$

At the assumed amplitude of the lips (1.25 cm), $\hat{x} \approx 0.15$ mm, which yields $\hat{F} \approx 600$ newtons. In practice it is not generally possible to satisfy the above phase condition in full, so that the amplitude of the alternating drive force would be something like 1200 newtons (~ 120 kg).

Numerous other requirements and limitations made the drive system a difficult problem. The available space for the mechanism was restricted, the mechanism had to be readily replaceable in the event of a fault occurring, the heat generated (copper and iron losses) could effectively only be removed by conduction, because of the location in vacuo, for the same reason the use of sliding surfaces demanding lubrication was ruled out, and finally the drive had to operate at a position where the stray field of the cyclotron magnet was still fairly considerable.

The solution finally arrived at was an "internal" drive, utilizing an auxiliary mass which avoids a rigid mounting. The system is illustrated in *fig. 11*, and in the photographs of *fig. 12a* and *b*. An electromagnet with a laminated core is screwed to the foot of the tuning fork. The armature of the magnet, which acts as the auxiliary mass (further weighted by brass blocks to bring the total weight to 17 kg), is fixed to the foot by two very stiff springs of chrome-nickel steel; the resonant frequency of this system (auxiliary mass and springs) is virtually identical with the fundamental frequency of the tuning fork, 55 c/s. When the electromagnet is now energized with an alternating current of 27.5 c/s (half the resonant frequency, because the magnetic force is not reversed when the current changes direction), the tuning fork and the auxiliary mass are set in vibration in opposite phase at a frequency of 55 c/s. The amplitude of the relative displacement is about 1 mm. Because the system vibrates at its resonant frequency, the amplitude of the force $F$ which the springs transmit to the foot of the tuning fork is about six times greater than the amplitude of the force $F_{\rm m}$ with which the magnet attracts the armature. The value $\hat{F} \approx 1200$ newtons can thus be reached with a reasonable energizing current.

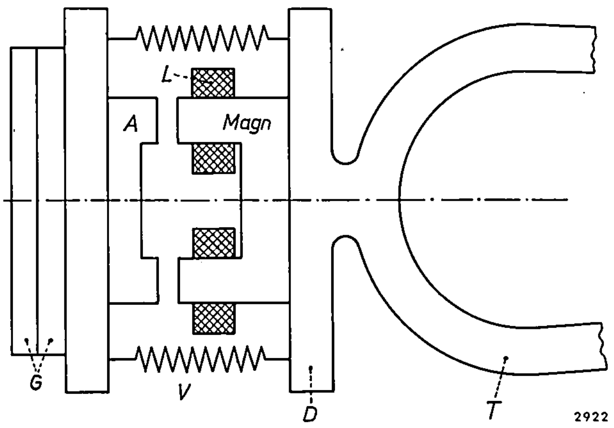


Fig. 11. Schematic representation of drive mechanism. $T$ tuning fork with foot $D$, to which is mounted the electromagnet *Magn* with coils $L$. $A$ armature, also connected to foot by springs $V$. The armature is weighted by brass blocks $G$. In reality, these are mounted partly around the magnet, owing to restricted space and also with a view to minimizing the bending moment acting on the springs.

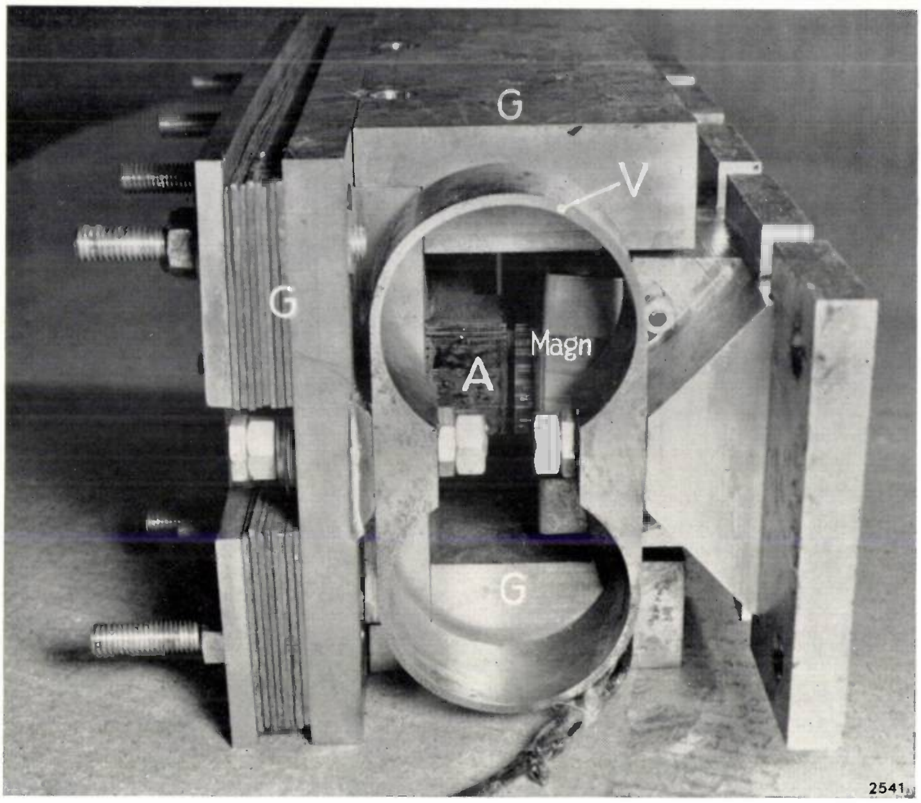
For the coil windings, good use was made of a new Philips technique for winding enamelled wires, in which no paper insulation is needed and which gives a particularly high filling factor (85%) and good heat dissipation (1 W/m$^2$ per °C/m) *) — properties which were important here, where the magnet had to be accommodated in a limited space in high vacuum.

Mention should be made of the special form of the springs, shown clearly in fig. 12*a* and *b*. The fabrication of the springs from a single piece of material presented some difficulties, but the result justified the trouble taken. The springs have a highly constant stiffness of 1000 N/mm each, provide rigid clamping, and moreover their stiffness is relatively much greater in the transversal direction, properties insufficiently attainable with the more conventional helical spring and leaf-spring assemblies. A secondary advantage is their easy mountability. The bending stresses occurring are fairly high, up to 3000 kg/cm$^2$, which calls for a highly polished finish to avoid premature fatigue failure due to surface discontinuities.

We shall briefly describe the method of working out the required mass of the auxiliary mass, the spring stiffness and the magnetic force. For this purpose the system consisting of tuning fork and drive mechanism was replaced by the model represented in *fig. 13*. $M''$ is the auxiliary mass, $S''$ the stiffness of the chrome-nickel steel springs. The friction $B''$ represents the hysteresis and eddy-current losses of the electromagnet and the deformation losses of the springs. The continuously distributed mass of the tuning fork is replaced by two concentrated masses $M$ and $M'$, corresponding to the parts lying respectively between, and outside (on either side), the nodal lines of the fundamental mode of vibration (cf. fig. 10). Further, $S'$ represents the lumped stiffness of the tuning fork, $B$ the lumped mechanical damping of the mounting (fig. 2), which

*) *Editorial note:* We hope to be able in the near future to publish an article on this winding technique, which has already proved to be of value in many applications.

can be disregarded in the first instance, and $B'$ allows for the internal damping of the tuning-fork material, for eddy-current losses in the stray field of the cyclotron magnet, and, in some experiments done during the development, for viscous damping by the atmosphere (air or hydrogen gas). The quantities $M$, $M'$, $S'$ and $B'$ were calculated from the fundamental frequency $f_m$, the total potential energy $W$ of the vibrating tuning fork, and the measured power $P$ absorbed at resonance. (It is found that $M$ may be put equal to the actual mass of the whole tuning fork, together with all parts rigidly mounted to it.) We now have the following three differential equations for the motion of the three masses:



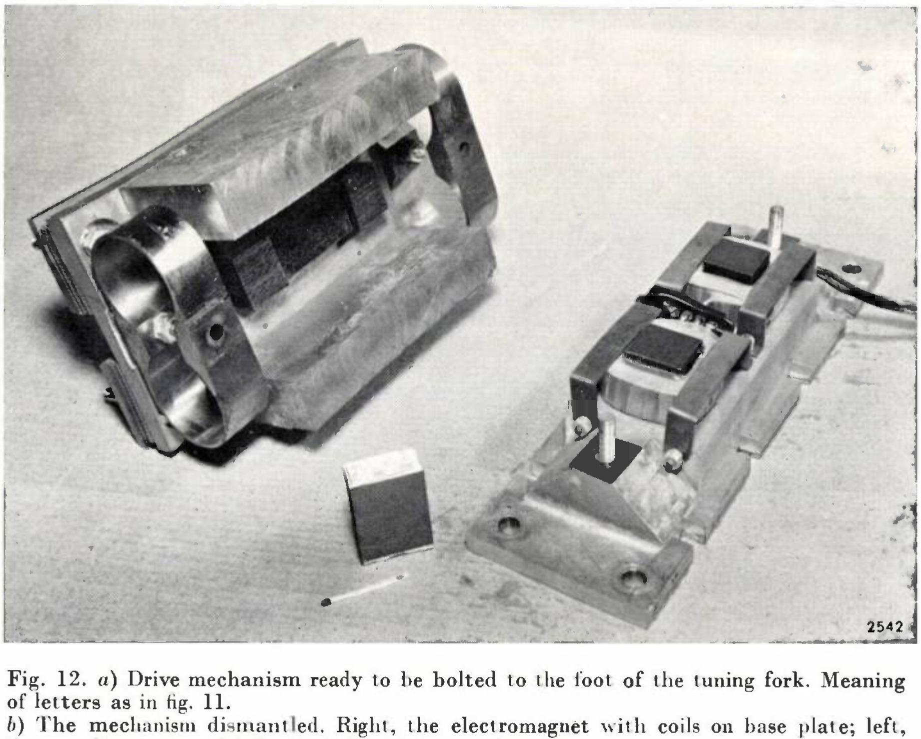


Fig. 12. *a*) Drive mechanism ready to be bolted to the foot of the tuning fork. Meaning of letters as in fig. 11.
*b*) The mechanism dismantled. Right, the electromagnet with coils on base plate; left, the auxiliary mass with laminated armature and springs.
The rectangular recesses in the edges of the base plate are for the cooling pipes, which remove the power dissipated in the electromagnet (about 30 W, see page 179).

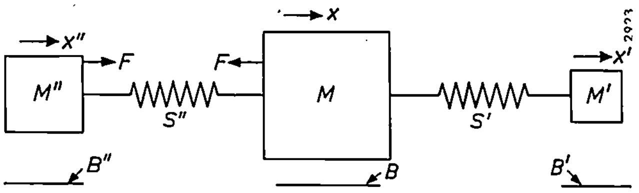


Fig. 13. Equivalent circuit of tuning fork and drive mechanism, for deriving the equations of motion and for studying the behaviour of the whole system.

$$\left.\begin{aligned} M\ddot{x} + B\dot{x} + S'(x-x') + S''(x-x'') &= F_{\mathrm{m}}, \\ M'\ddot{x}' + B'\dot{x}' + S'(x'-x) &= 0\,, \\ M''\ddot{x}'' + B''\dot{x}'' + S''(x''-x) &= -F_{\mathrm{m}}\,. \end{aligned}\right\} \quad . \; . \; (9)$$

With the aid of these equations, the behaviour of the vibrating system was examined in detail. In particular, the force $F$ acting on the foot of the tuning fork, and given by:

$$F = F_{\mathrm{m}} + S''(x''-x), \quad . \; . \; . \; . \; . \quad (10)$$

was calculated for varying conditions. We shall mention here only one rather unexpected result of these calculations, namely that stable operation is best ensured if the natural frequency of the system formed by the two masses $M$ and $M''$ and the spring $S''$, is about 0.5 c/s lower than the resonant frequency of the whole system. This was also confirmed experimentally.

To avoid the excitation of unwanted modes of vibration, and the needless dissipation of energy in the coils (which would complicate the removal of heat in the vacuum), the magnetic force on the armature should possess no higher harmonics of the fundamental frequency, i.e. it should be purely sinusoidal. The form of the energizing current needed for this purpose is by no means sinusoidal, since the magnetic force at a given current is still highly dependent on the length of the air gap of the magnet, which varies during the vibration. This is illustrated in *fig. 14*.

The energizing current with the required waveform is obtained by supplying the sinusoidal voltage from an *RC* oscillator to a push-pull amplifier in class C operation. The natural frequency of the *RC* oscillator is roughly 27.5 c/s. As the tuning fork vibrates, one of the blades reflects the light from a stationary electric bulb, thereby transmitting periodic flashes to a photocell. The voltage pulses thereby produced are used to synchronize the *RC* oscillator. As a result of this feedback, which is automatically kept in the correct phase, the tuning fork vibrates precisely at its fundamental frequency.

The photocell signal is also used for synchronizing the periods during which the RF oscillator for the cyclotron is switched on: so that no energy shall be wasted, the RF oscillator is operated only in that half of each modulation period during which the particles can be usefully accelerated ($f$ decreasing). Moreover, the signal can be used to synchronize the ion source — if the latter is pulsed — and any other devices that may be involved in experiments with the particle beams.

### Parasitic modes of vibration

The mode of vibration in which we are interested — the fundamental mode of the tuning fork — is purely two-dimensional. For this reason, in the foregoing, we have mainly had to concern ourselves with the *profile* of the tuning fork. However, its third dimension (its width of 2 metres) may give rise to three-dimensional modes of vibration, such as occur in membranes and plates. By exciting the tuning fork at its various resonant frequencies, and

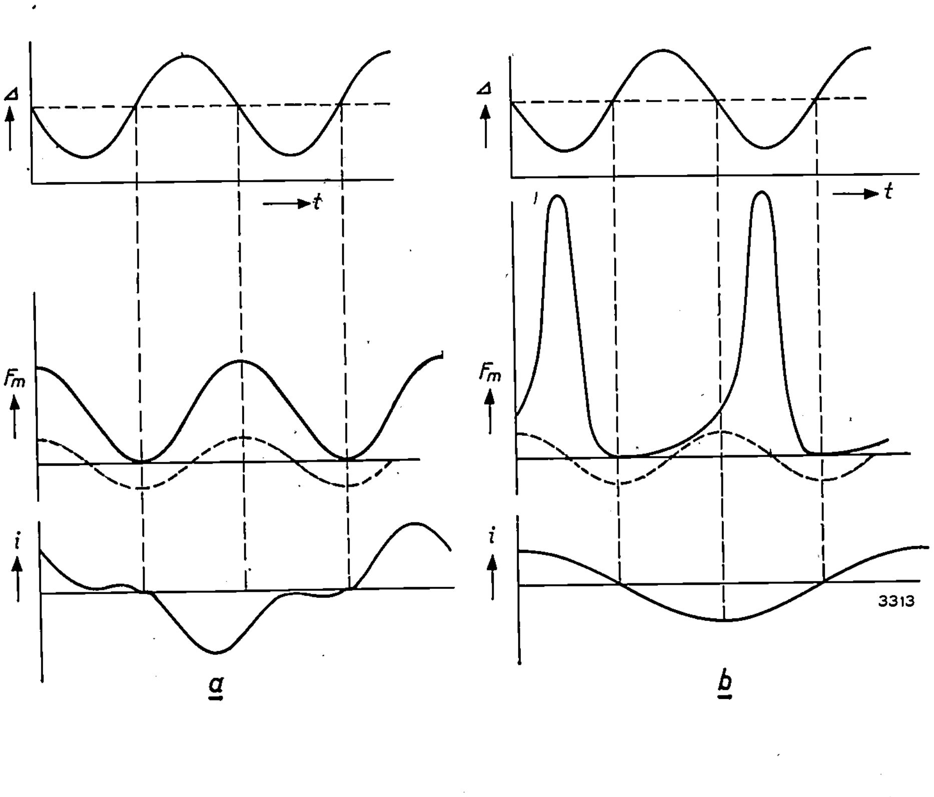


Fig. 14. Electrical supply for drive mechanism. From top to bottom are shown, as a function of time $t$: the relative displacement of the armature (length $\Delta$ of air gap), for which a sinusoidal variation is assumed; the magnetic attractive force $F_{\mathrm{m}}$; the current $i$ for energizing the electromagnet. The dashed curve represents the velocity of the armature.

Curves ($a$) relate to a current waveform designed to produce a sinusoidal variation of the magnetic force. Curves ($b$) represent the case of a sinusoidal current; the magnetic force here is badly affected by higher harmonics of the fundamental frequency.

sprinkling sand on to the plates, Chladni's figures can be produced of the different modes of vibration (see *fig. 15*) [8]). If the natural frequency of a particular mode of vibration is equal or almost equal to a multiple of the fundamental frequency, this vibration may be directly excited by the driving force or via an internal coupling with the fundamental mode of vibration (e.g. due to the Poisson contraction). It is then superposed on the fundamental vibration, thereby upsetting the operation of the modulator.

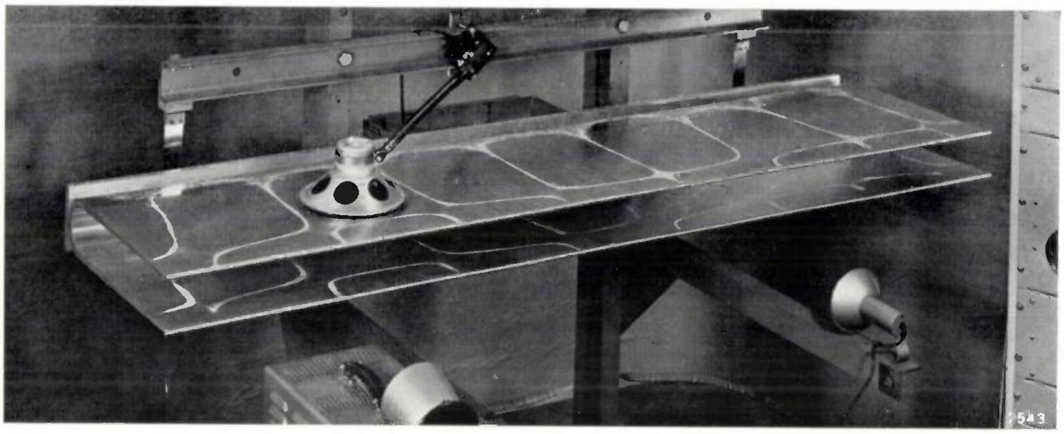


Fig. 15. Chladni's figure obtained when the tuning fork is made to vibrate in a particular undesired mode by excitation with the appropriate resonant frequency [8]).

The chance of such a dangerous coincidence of frequencies is small, but we did in fact happen to stumble on it. To illustrate the counter-measures adopted, we shall examine the possible modes of vibration at somewhat greater length.

We denote the modes of vibration by two index figures, which indicate for a single blade of the tuning fork the number of nodal points at the sides (but not counting the inevitable fundamental node near the foot, see fig. 10) and at the front edges, respectively. We also add to these figures an *s* or an *a*, depending on whether the two blades vibrate symmetrically or anti-symmetrically (see *fig. 16*); both cases are possible, since anti-symmetric vibration does not shift the centre of gravity of the system as a whole. In this notation, *s00* is the fundamental mode of vibration, which is the one we want. *Fig. 17* illustrates various modes of vibration together with their resonant frequencies, measured on a 1 : 4 scale model of the tuning fork. (We are concerned only with vibration modes of the first kind, i.e. forms without node at the sides, since modes with side nodes are found to have much higher frequencies which are not troublesome.) It can be seen in fig. 17 that the frequencies of the modes *a04* and *s04* are fairly close to twice the frequency of *s00*. With the actual tuning fork, mounted in the vacuum tank, the frequency of the *s04* mode was even more troublesome than in the model; it was this mode that gave rise to the trouble referred to above. This occurred particularly when the tuning fork was warming up (see next section). The effect of heating causes a slight shift of all resonant frequencies — not so much because of the expansion of the material but because of the drop in the modulus of elasticity. At a particular tem-

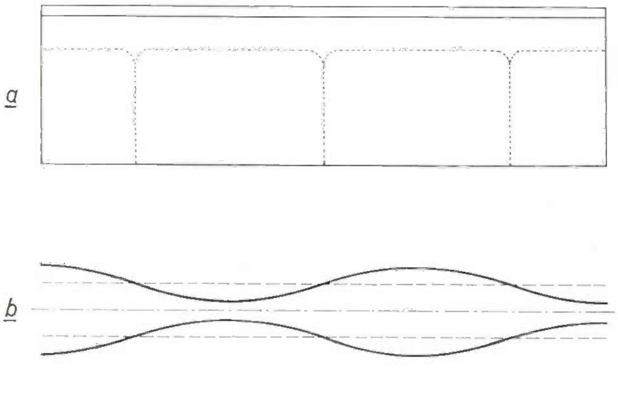


Fig. 16. *a*) Chladni's figure of a vibration of the tuning fork, with three nodes at the front edge and none at the sides of the blades (except the fundamental node as in fig. 10). The two blades can vibrate here in two modes: symmetrically (*b*), called the *s03* mode; or anti-symmetrically (*c*), mode *a03*.

[8]) A short account of these was published some time ago in this journal: B. Bollée, Chladni's figures on the vibrating capacitor of a synchrocyclotron, Philips tech. Rev. **19**, 84-85, 1957/58.

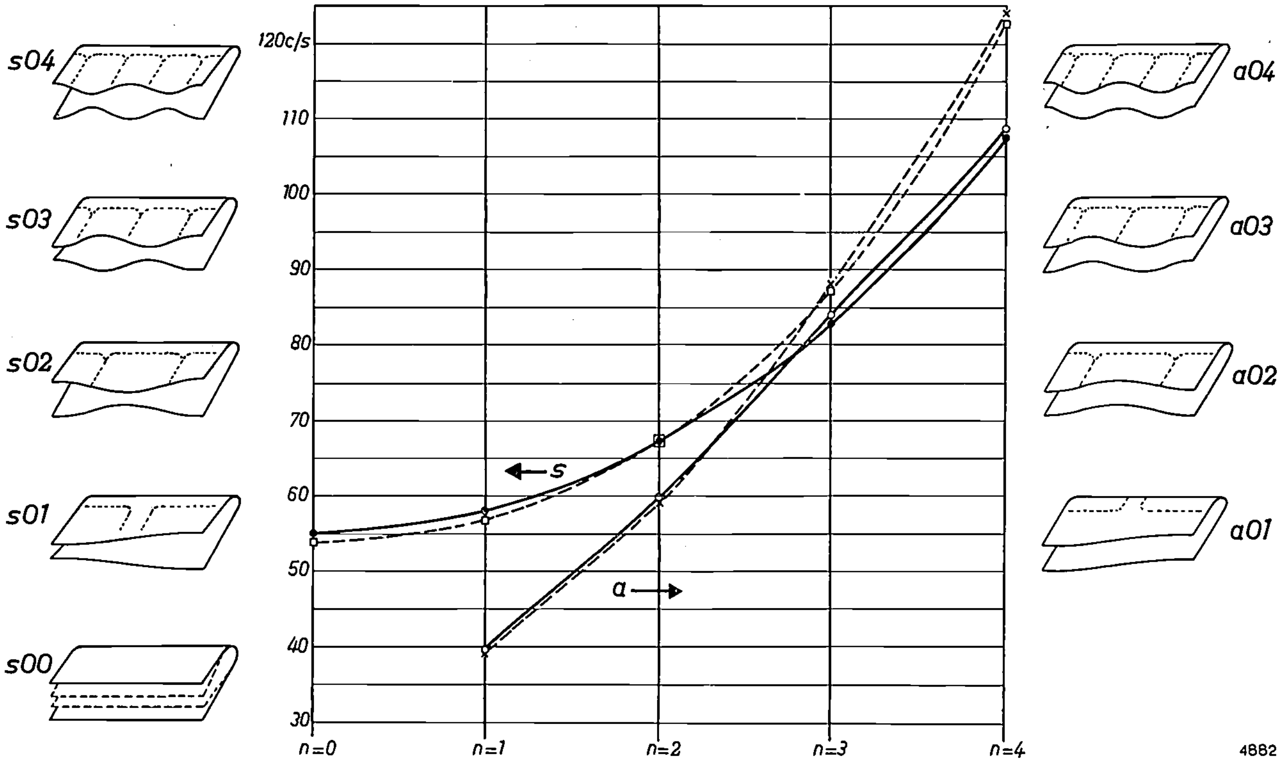


Fig. 17. Resonant frequencies of the modes of vibration "of the first kind" (i.e. without nodes at the sides of the blades, first digit = 0), for $n = 0$ to 4. The black dots pertain to symmetrical modes of vibration, the open circles to anti-symmetrical modes. These modes, *s00* to *s04* and *a01* to *a04*, are illustrated at the sides of the graph. The solid curves drawn through the points have no physical significance in the sections between the whole values of $n$; they were useful, however, in enabling us, after finding the first three resonant frequencies, to find the others: the resonances are so sharp (width of resonance curves of the order of 0.01 c/s) that they are easily missed.

The dashed curves show the resonant frequencies of the tuning fork when stiffening ribs are fitted, as in fig. 18. The small squares pertain to symmetrical modes, the crosses to anti-symmetrical modes of vibration.

perature distribution, the *s04* vibration came to resonance.

Apart from making a new tuning fork with modified dimensions, which was obviously not an attractive proposition, there were various possible ways of dealing with this interference, namely: artificial damping, "decoupling" by cutting slots into the blades, and finally shifting the dangerous resonant frequency by reducing the width of the blades or by mounting a stiffening rib on each of them.

After careful consideration of the drawbacks and risks of these measures and the time they would probably cost, the choice fell on the stiffening ribs. *Fig. 18* shows how the rib is mounted over the whole width of a blade. It was necessary to secure the ribs with particular care, since they undergo accelerations of 120 $g$ during vibration. Of course, the ribs not only change the ratio of the fundamental frequency to the frequency of *s04*, but also to the frequency of all other modes of vibration. The chance that this will give rise to other resonances can, however, be minimized, for it can be shown that, at the temperature distributions found in practice, no troublesome resonances will occur provided that all resonant frequencies, assuming a *homogeneous* temperature distribution, differ by at least one per cent from any neighbouring whole

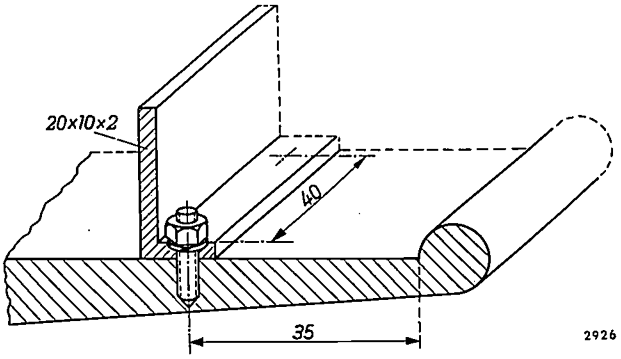


Fig. 18. Location and form of the stiffening rib mounted to each of the two blades to prevent the excitation of undesired modes of vibration. (The ribs are clearly visible in fig. 24.)

multiple of the fundamental frequency. This follows from an estimate concerning the expected effect of the locally varying decrease in the modulus of elasticity, and concerning the (very small) width of the resonance curves for the various modes of vibration. With the aid of data derived from a model, the approximate displacements of all resonant frequencies, brought about by the stiffening ribs, was calculated for various positions and dimensions of the ribs. In this way the dimensions and location were determined which satisfy the above 1% condition (fig. 18). After the ribs had been fitted, no further undesired resonances were encountered. As a consequence of the increased weight the fundamental frequency itself dropped to 54 c/s, and it was necessary to adjust the drive mechanism accordingly.

It would be going too far to describe here the method of calculation adopted. It should be mentioned, however, that the solution of the problem, although it apparently concerns a minor detail, had a decisive bearing on the success of the whole modulator project.

### Cooling the tuning fork

Part of the tuning fork (*AB* in fig. 5) functions as a variable capacitance in the RF system and thus carries an RF current (instantaneous frequency *f*). For various reasons it was necessary to give very careful consideration to the removal of the heat thereby generated. We shall go into these reasons and discuss the solution adopted.

The heat likely to be generated can be fairly accurately calculated. Since the length of the tuning fork is small in relation to the wavelength, the current flowing in the tuning fork may be regarded to a good approximation as quasi-stationary. This current is given by:

$$\hat{I} = \frac{\hat{U}_D}{\zeta} \sin \frac{2\pi f l}{c}, \quad . \quad . \quad . \quad . \quad (11)$$

where $l$ is the distance between the tuning fork and the dee mouth, $\zeta$ is the characteristic impedance of dee and dee stem, $U_D$ the voltage at the dee mouth and $c$ the velocity of light. $\hat{U}_D$ varies with the frequency $f$, from about 8 kV at 29 Mc/s to about 25 kV at 16.5 Mc/s. Given $l = 3.5$ m and the value $\zeta = 6$ ohms, as assumed at a preliminary stage of the design, it follows that $\hat{I}$ varies between 3100 and 1550 amperes. If $s$ is the length and $b$ the breadth of the tuning fork, $d$ the penetration depth of the current and $\varrho_{Al}$ the resistivity of the aluminium alloy, then the resistance is given by:

$$R = \frac{s\varrho_{Al}}{2bd} . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

We can fill in $s = 60$ cm, $b = 2$ m, $\varrho_{Al} = 4.926 \times 10^{-8}$ Ωm at room temperature, but the penetration depth $d$ also varies with the frequency, from 22 microns at 29 Mc/s to 29 microns at 16.5 Mc/s. The average power is found by integrating $\frac{1}{2}I^2R$ over one period of modulation, using the known time-variation of the frequency (fig. 6*c*) and taking into account that the RF oscillator is operated only for about half of every modulation period (see above). Graphic integration led to the value $P = 240$ W.

Apart from this electrical dissipation, some heat is also generated by the mechanical losses in the vibrating tuning fork, but these losses are roughly estimated, as mentioned earlier, at only about 15 W.

The removal of a dissipated power of some 250 W from the large surface of the tuning fork would be a simple matter if convection were possible; in a vacuum, however, it becomes a problem. Relying on the large surface area, one might in the first place think of removing the heat by radiation. A simple calculation shows that this is not a feasible solution. The total radiating surface may be put at $2 \times 200 \times 60$ cm$^2$ (the heat spreads easily enough throughout the tuning fork). Between the copper outer conductor of the RF system and the aluminium tuning fork the temperature difference should be such as to allow a heat transfer of about 0.01 W/cm$^2$. Having regard to the absorption coefficients of copper and aluminium (both of which are about 0.2), we arrive at a temperature rise of approximately 120 °C in the aluminium. Such a temperature rise is out of the question. Not only would it reduce the natural frequency of the tuning fork by about 1 c/s, thereby somewhat changing the tuning with the drive (see the frequency condition mentioned on page 171, at the end of the small print), but what is more, the material of the tuning fork would not be able to withstand it: at temperatures above about 100 °C, slow changes in the metallographic structure of the aluminium alloy can occur, as a result of which the material loses its favourable properties.

An attempt to improve the radiation cooling by artificially blackening the tuning fork and the copper outer conductor (thus increasing the absorption coefficient) came up against practical difficulties. Black anodizing of the tuning fork is ruled out because of its known disastrous effect on the endurance limit. To spray on a black coating is rather risky because the coating might flake in the long run or give off gases in vacuo, and also because it might make things even worse by increasing the generated heat (owing to dielectric losses).

It was therefore decided to remove the heat not by radiation but by water-cooling. This meant that cooling pipes had to be fitted to the foot of the tuning fork. Now, however, we were faced with another danger. The non-uniform cooling gives rise in the tuning fork to a temperature gradient which, by causing non-uniform expansion and consequent mechanical stresses, might easily warp the blades. This would have the most serious consequences on the capacitance variations and on the operation of the positioning control (see the next section). An inquiry was therefore made into the temperature distribution and deformation to be expected in each of the blades.

The temperature distribution can be calculated to a good approximation by treating the profile of the tuning fork as a *linear* truncated wedge, as in its mechanical design (see fig. 9), having a length equal to the path along which the heat-flow takes place to the cooling pipes. The calculated temperature curve is shown in *fig. 19*. A correction has been applied here to allow for the effective constriction at the stem of the tuning fork, and allowance is also made for the heat generation on the inside of the blades near the stator, which was not taken into account in the power calculation given above. *Fig. 20* shows the heat streamlines and the contours of equal temperature. (Following the rules of field theory [2]), a sketch like this can be drawn simply by inspection, and with some practice and patience the result is often just as accurate as by calculation, and takes very much less time than measurements on an electrolytic tank, for example.) The outcome shows that a temperature difference of less than 13 °C was to be expected between foot and lips. As regards the deformations caused by this temperature difference, it was necessary to experiment on a full-size tuning fork. For this experiment the tuning fork (not vibrating) was set up in an air atmosphere, the foot being kept at room temperature by the cooling pipes, whilst the whole surface was artificially heated by strips of electrically conductive paper which were stuck to the outer surface of the blades with an insulating underlayer of thin nylon fabric

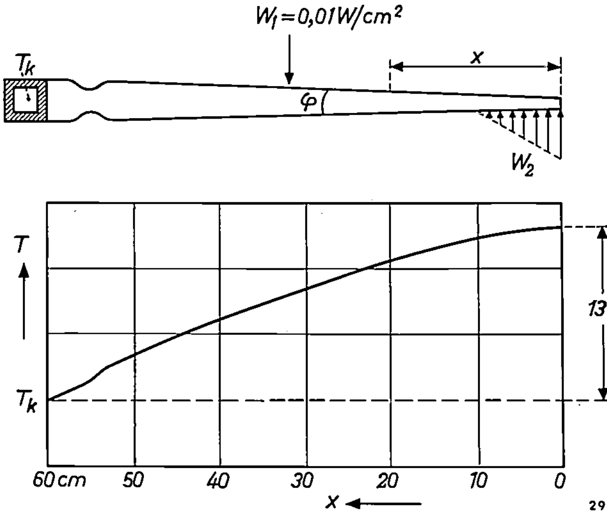


Fig. 19. Variation of temperature $T$ over the profile of a tuning-fork blade, calculated by treating the profile as a straight truncated wedge with a constriction at the stem of the tuning fork. From the outer surface a power $W_1$ (= 0.01 W/cm$^2$) goes inwards, and from the inner surface a power $W_2$ (at the position of the stator). The foot of the tuning fork is kept at the temperature $T_k$ by a cooling-water pipe.

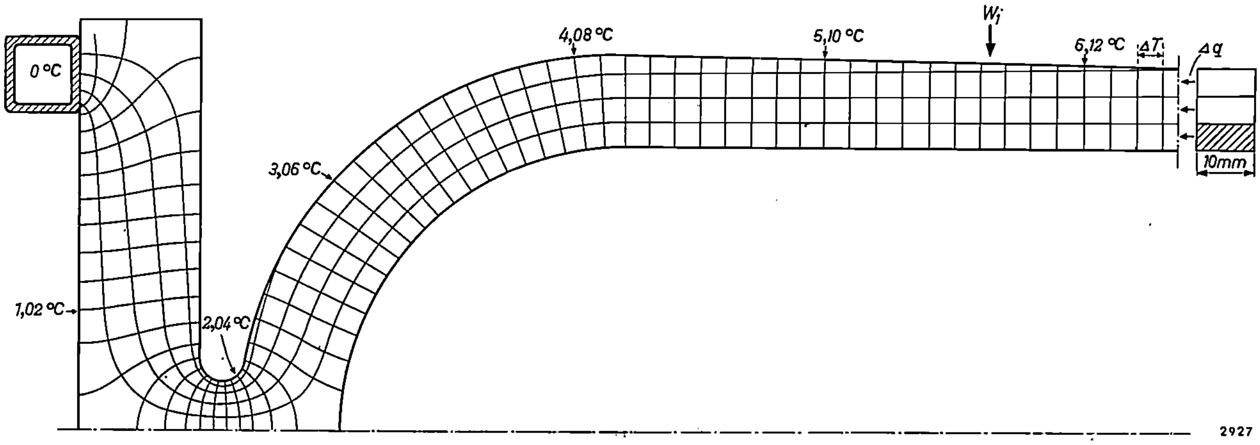


Fig. 20. Streamlines of heat-flow and equal-temperature contours in the profile of a blade of the tuning fork (only the curved section is reproduced, the temperature variation in the straight section being almost linear). In each stream tube (shown hatched in cross-section) a power $\Delta q$ flows of 0.14 W. Since only the temperature differences are important here, the temperature in the cooling-water pipe (top left) is assumed to be 0 °C.

(*fig. 21*). By supplying each of the strips with a suitable current, the theoretically derived temperature gradient was fairly accurately simulated. It was found that the lips of the blades were less straight and parallel than when the temperature of the tuning fork was uniformly raised. The deviations were not readily reproducible, presumably because of various residual mechanical stresses in the rolled and machined plates, but they were so small (no more than about 0.2 mm) that they do not have any adverse consequences.

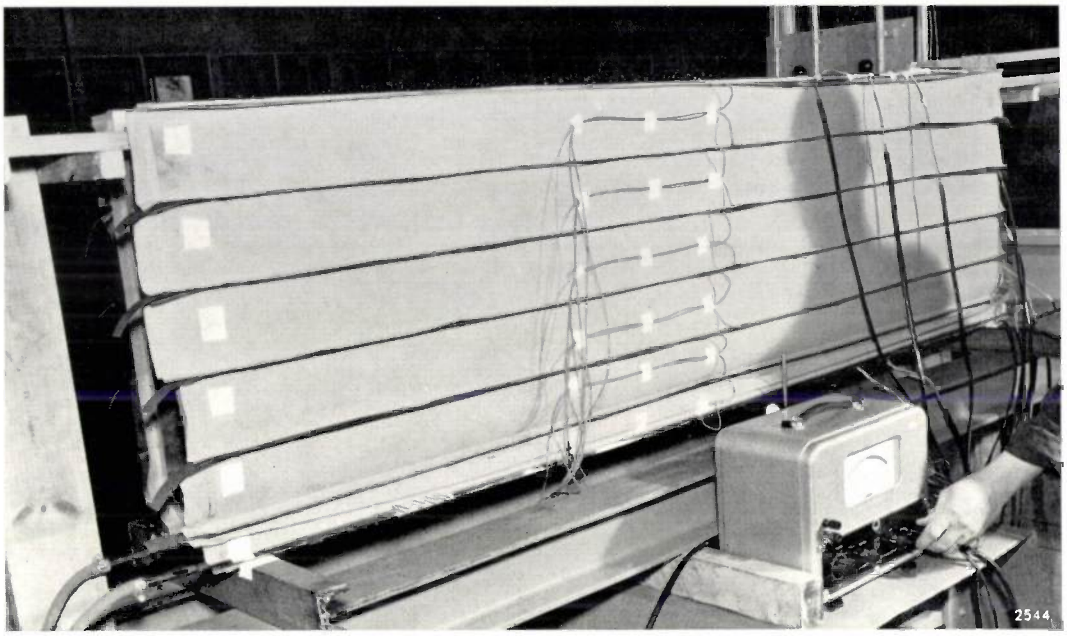


Fig. 21. Experimental arrangement for measuring the expected deformations of the tuning fork as a result of heating during actual operation, where only one end (the foot) is cooled with flowing water. The calculated temperature distribution is simulated by heating the blades with strips of electrically conducting paper, each fed with a specific current.

*Fig. 22* shows the tuning fork mounted to the stub; note the flexible connections for the cooling pipes along the foot of the tuning fork.

It should be pointed out that in the foregoing we have considered the temperature distribution in the stationary state. The time constant of the process of heating and cooling the tuning fork is calculated to be about 1700 seconds. This means that each point of the tuning fork will have reached $\frac{2}{3}$ of its final value about half an hour after switching on the cyclotron.

### Automatic positioning control

In the introduction we referred to the deformations that occur in the whole RF system when the cyclotron is in operation, and to the fact that the correct position of the tuning fork in relation to the stator must be maintained in spite of these deformations. There are two possible changes in their relative position: changes in the overlapping of tuning fork and stator, which amounts nominally to exactly 10 cm over the whole width; and changes in the "gap", by which we mean the distance within which the vibrating blades can approach the stator at any point. In the ideal case the position and the amplitude of vibration are so regulated that the gap at all points is 1.5 mm.

The relative horizontal displacements caused by deformations are of the order of 1 mm. They have hardly any effect on the gap but only on the overlapping, which causes no noticeable change in the capacitance *variation*. As regards the relative position of tuning fork and stator in the horizontal direction, it was therefore thought to be sufficient to make an adjustment once and for all when installing the modulator in the system at Geneva. To facilitate this adjustment, the three suspension points on the tuning fork are mounted to carriages, which can be slid over a distance of 5 mm (fig. 22).

The relative displacements that cause changes in the *gap* can be analysed into three movements (see *fig. 23*): *a*) vertical displacement of the tuning fork; *b*) turning of the tuning fork about an axis parallel to the lips; *c*) turning of the tuning fork about the centre line of the whole RF system. We may add to these *d*) incorrect vibration amplitude of the tuning fork. To correct the changes in the gap resulting from these deviations it was necessary to introduce, as mentioned in the introduction, an automatic positioning control system. This operates with three servomotors, which cause vertical displacement of the three suspension points (by pivoting about the spindles $a_1$, $a_2$ and $a_3$, respectively, in fig. 2). By suitably actuating the servomotors, separately or simultaneously, the three movements *a-c* can be effected and compensated. The three servomotors can be seen on the front face of the stub tank in the title photo of Part I. A fourth servomotor, whose shaft is coupled with a potentiometer which controls the energizing current for the tuning-fork drive, is used to keep the amplitude of the tuning fork at the correct value (*d*).

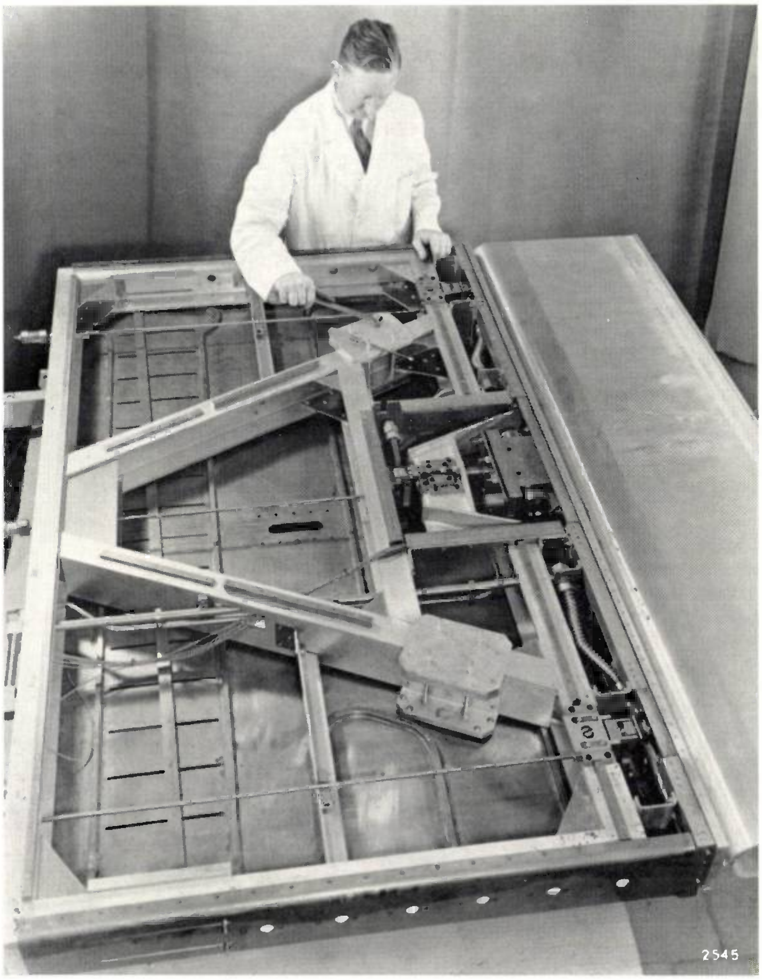


Fig. 22. The tuning fork (right) mounted to the stub, the top plate of the latter having been removed. Note the flexible connections for the cooling-water pipes, fitted to the foot of the tuning fork. The drive mechanism can also be seen, with the three rods which connect the points of suspension to the servomotors for the position control (see fig. 25 and fig. 2).

The four servomotors are controlled by a system which operates roughly as follows. The situation as regards the gap between tuning fork and stator is explored by feelers attached to the stator. There are altogether five such feelers, made of "Teflon" (*fig. 24*), each of which, if touched by the tuning fork, breaks an electrical contact. Three feelers are mounted directly on the end of the stator (two of these are visible in fig. 24, the third being about 2 metres away at the other end of the stator). From time to time, these three feelers, upon a "command" from the control system, can be slid further out of their metal sheaths in two steps of 0.1 mm by a small electromagnet. If one of these feelers is touched in its fully withdrawn position, the control system immediately switches off the oscillator, for in that case the gap is less than 1.4 mm, which was considered both mechanically and electrically dangerous. If the tuning fork touches one of the three feelers only after the first step, then the gap at that position is between 1.4 and 1.5 mm; if a feeler is touched only after the second step, the gap is between 1.5 and 1.6 mm; if a feeler is not touched at all, the gap is greater than 1.6 mm. The twenty-seven possible signal combinations from these three feelers correspond to various kinds of deviations from the desired situation, namely of type *a* or *c* (or *d*) or combinations thereof. By means of a "translator" in the control system, each of the signals sets in momentary operation that combination of servomotors which leads to the cor-

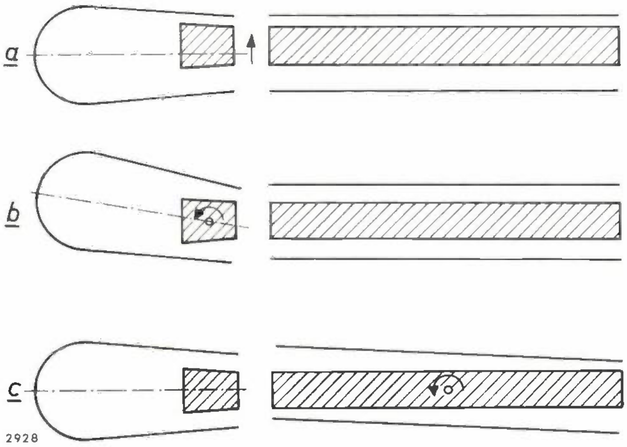


Fig. 23. The various misalignments in the position of the tuning fork relative to the stator. The arrows indicate the necessary corrections.
*a*) Vertical displacement.
*b*) Turning about an axis parallel to the lips.
*c*) Turning about the centre line of the RF system.

rection of the relevant deviation. The two Teflon feelers not yet discussed are mounted on an extension piece (partly visible in fig. 24) on the stator so that they check on the tuning fork at a point 100 mm nearer its foot than the first three. If one of these feelers is touched, the "translator" — depending on the information received from the first three feelers — actuates the servomotors in such a way that deviations of the type *b* or *d* are corrected.

*Fig. 25* shows a highly simplified block diagram of the control system. A photograph of the control racks can be seen in *fig. 26*. The block diagram also indicates the frequency control, earlier mentioned, for the drive mechanism of the tuning fork, which is also responsible for synchronizing all the relevant parts of the cyclotron [9]).

### Proving the modulator

It was not possible to say whether the modulator in all its complexity would come up to expectations until it had first been proved in the cyclotron itself, at Geneva, and then only after some months of normal operation. Various modifications were found to be necessary. In fact, some modifications had turned out to be desirable after the installation had passed through its preliminary trials at Eindhoven. In the latter tests the conditions of operation in the cyclotron could only be simulated to a limited extent, as far as the positioning in vacuo and the heating were concerned. The stub tank at Geneva was to be in communication with the accelerating chamber; at Eindhoven, after the tuning fork was mounted to the stub (fig. 22), the tank was provisionally sealed with a special cover plate to which the stator with built-in feelers was attached, and which was fitted with windows for observing the movements inside (*fig. 27*). The tuning fork was heated by twelve radiators each 2 metres long (the use of heating strips, as employed for investigating the deformation, was obviously not possible with the vibrating tuning fork because of the damping they would cause). The foot of the tuning fork was kept at room temperature by water-cooling. During the vibration of the tuning fork it was found necessary to make the back of the stub heavier with a block of lead weighing roughly 60 kg, the reason being that a resonance in the region of 50 c/s caused fairly considerable vibration of the stub together with its supporting insulators. The amplitude of the lips, measured with a stroboscope and cathetometer, showed local differences up to 0.3 mm. It was possible to reduce this deviation from straightness to about 0.15 mm by attaching a number of weights of 10 grammes to the ribs.

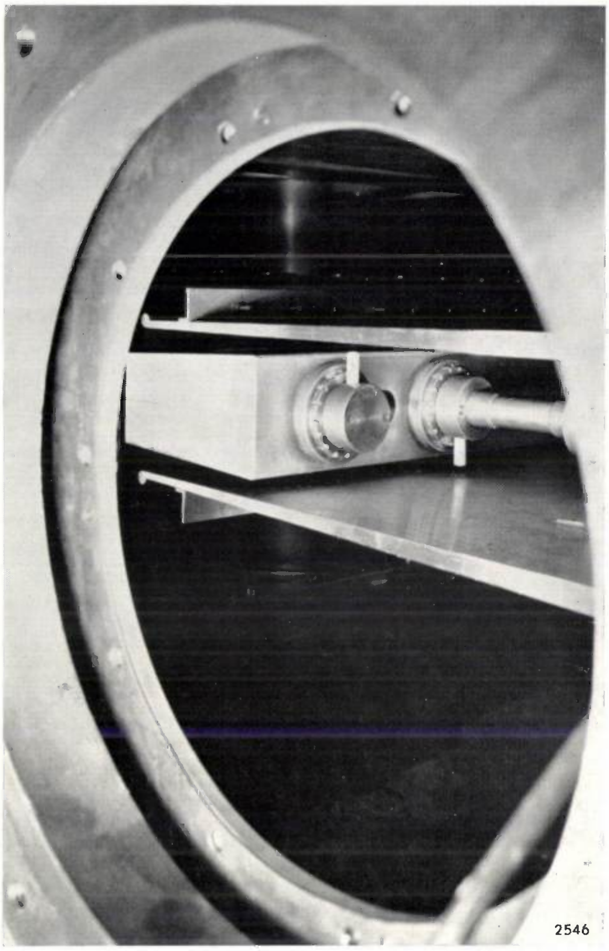


Photo CERN

Fig. 24. End of tuning fork and stator, mounted in the cyclotron and seen through a window in the accelerating chamber. The photograph shows two of the five "Teflon" feelers which are fitted to the stator for sensing the position and amplitude of the tuning fork.

In Geneva, however, other complications arose, in particular owing to the stray field of the large electromagnet of the cyclotron. The displacements of tuning fork and stator when the magnet was switched on were fortunately not too serious. Various components, like the feeler elements and filters in the control system which are within the stray field, had to be provided with better screening. The most disagreeable surprise, however, during the trial operation of the complete synchrocyclotron — in which only the ion source was not functioning, since of course no particle beams could be produced whilst people were still working inside the screened-

[9]) The tuning-fork modulator in its definitive form, as indeed the whole RF system for this cyclotron, was built under the direction of L. van Mechelen by a department of Philips Industrial Equipment Division at Eindhoven. Special mention is made of A. de Groot, now at the CERN in Geneva, who played an important part in the development of the control system for the tuning fork and of the entire electronic equipment for the RF system.

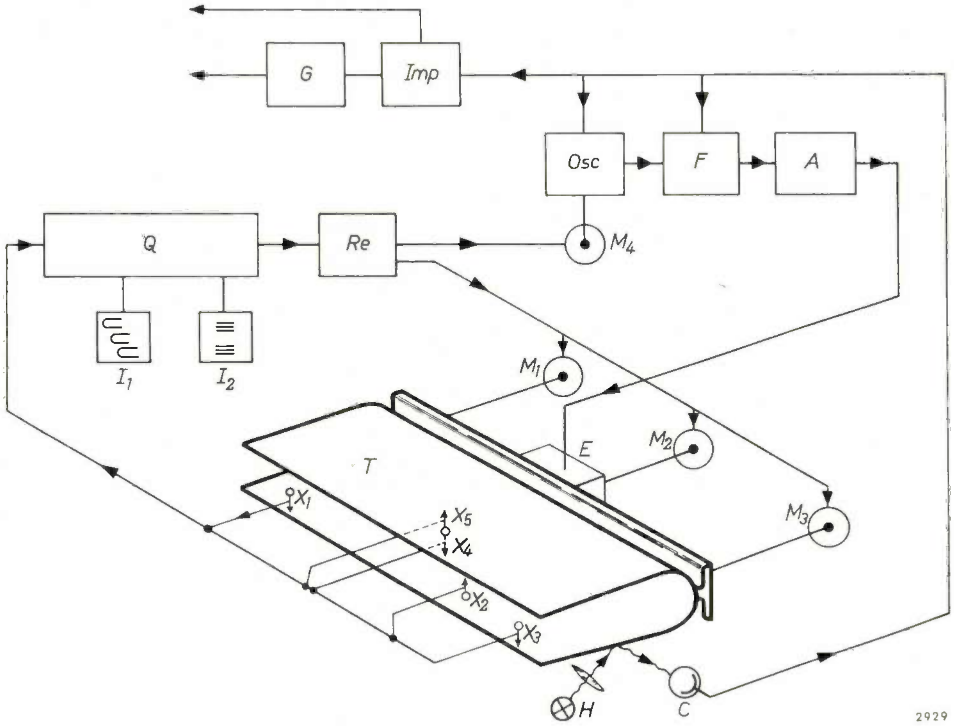


Fig. 25. Diagram of the system for automatically controlling the position and amplitude of the tuning fork. $X_1$-$X_5$ feelers fitted to the stator for sensing the situation of the tuning fork. *Q* "translator", which sets in operation the appropriate combination of the four servomotors $M_1$-$M_4$ via a series of relays *Re*, thereby operating the linkage for correcting the situation. $I_1$ and $I_2$ visual indicators. The three motors $M_1$-$M_3$ act on the suspension of the tuning fork *T*; $M_4$ acts on a potentiometer which controls the amplitude of the drive mechanism *E*. *Osc* RC oscillator, *F* phase-regulating circuit, and *A* amplifier for the drive system. *H* electric bulb and *C* photocell, which produce the feedback signal for frequency control in the drive system. This signal also serves for synchronizing the pulse generator *Imp*, which controls the RF generator for the cyclotron and (via *G*) the ion source and irradiation set-up.

off area — was occasioned by the energy losses in the tuning fork. Originally it was estimated that about 15 W would be needed to cover the losses due to damping, with perhaps the same wattage for the copper and iron losses of the electromagnet. During the trials at Eindhoven, with no electric or magnetic field, it was found that the damping consumed only 3 W, whilst the losses of the electromagnet were a mere 4 W. It therefore looked as if the drive mechanism, designed for about 30 W, was amply equipped for its task. In the proving run at Geneva, however, the drive mechanism consumed a total of 45 W! The reason turned out to be additional damping due to eddy currents produced in the blades of the tuning fork by a fairly strong horizontal component of the magnetic field at that position. A horizontal component of that strength had not been foreseen.

To reduce this component of the magnetic field an auxiliary winding was fitted around the stub tank, consisting of ten turns and carrying a current of 400 A. This reduced the eddy-current losses to about 6 W and the total power consumed by the electromagnet to about 30 W.

The RF voltage between stator and tuning fork, varying between 8 and 25 kV (see above), gave rise to no flashover. With a view to capturing stray ions, which can initiate discharges, stator and tuning fork were given DC potentials of —2200 and —800 V, respectively, with respect to earth.

The drift of minimum and maximum frequency as the tuning fork gets hot was found to be very slight indeed. The installation can also work for hours on end without the position control having to come into operation. The amplitude control comes

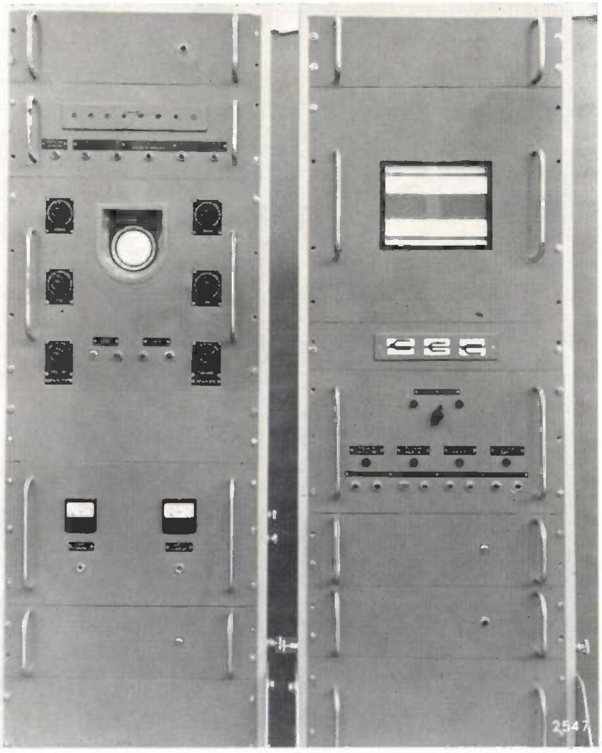


Fig. 26. Racks containing control equipment for the tuning fork. The panel on the right contains the instruments for visually indicating certain deviations in the position of the tuning fork: above, for deviation *c*; in centre, for deviation *a*; see fig. 23.

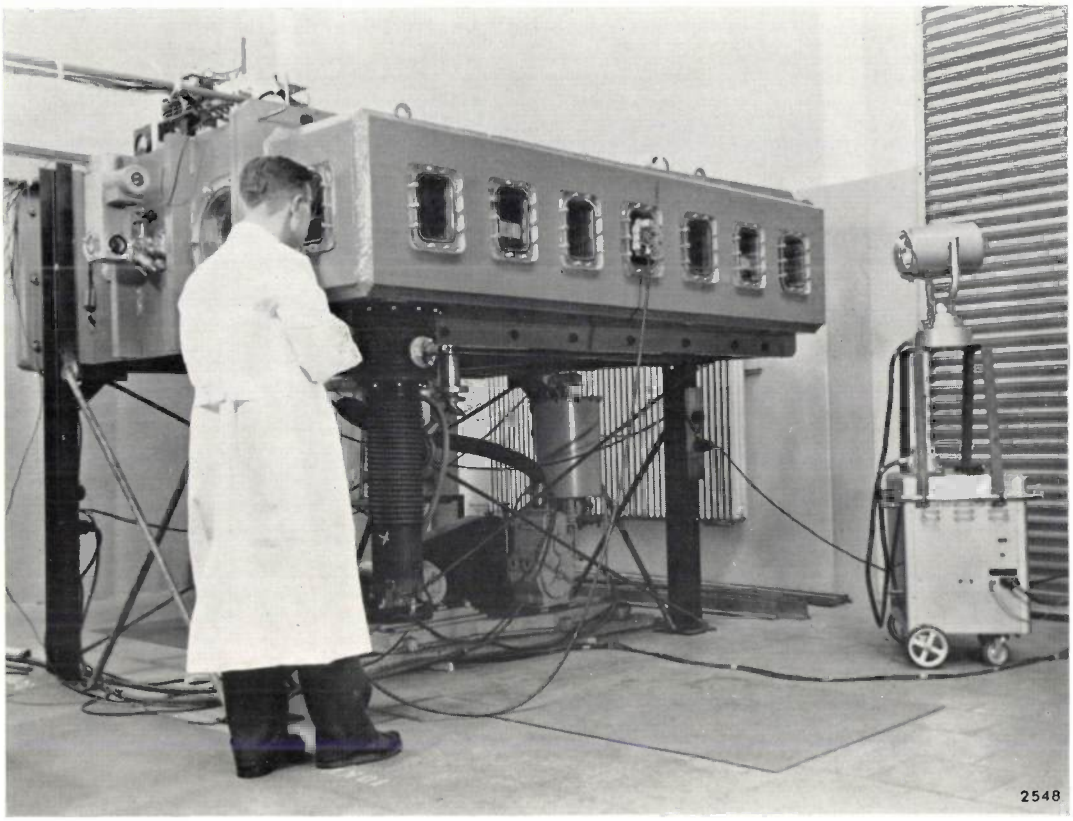


Fig. 27. Proving the modulator at Eindhoven. The stub tank, with the tuning fork mounted inside it, is provisionally sealed with a vacuum-tight cover plate provided with observation windows at the side where the dee stem will later be connected. Right: stroboscope for observing the vibration of the tuning fork.

on more frequently, but why this is so is not entirely clear.

The modulator has now been in operation some three years. Since it was not originally known whether the tuning fork would be satisfactory, the CERN had also built for all eventualities a modulator using a rotating capacitor. No call has been made on this, however. The first tuning fork has far exceeded its estimated life; the second remains as a stand-by, so that a third has now been constructed.

**Summary.** The frequency of the accelerating voltage of the CERN synchrocyclotron swings periodically from 29 to 16.5 Mc/s. This frequency modulation is effected by a capacitor made to vibrate at 55 c/s, which was developed by the Philips Eindhoven laboratories in cooperation with the CERN. The capacitor consists of an aluminium "tuning fork" roughly 60 cm long and 2 metres wide, and a stator of the same width attached to the dee stem. The very large amplitude of vibration required of the tuning-fork blades (1.25 cm) gives rise, at a resonant frequency of 55 c/s, to bending stresses that approach the endurance limit. In this respect an adequate safety margin was achieved by giving each blade a profile in the shape of a truncated wedge (minimum fatigue life $10^9$ vibrations). The tuning fork is set in vibration by an electromagnetic drive system with an auxiliary mass, whereby a rigid mounting is avoided. As a result, virtually no vibrations are transmitted to surrounding parts. The drive system is fed by an *RC* oscillator operating at 27.5 c/s and an amplifier which delivers an output current of such form that the driving force (magnetic attraction) is almost purely sinusoidal. The *RC* oscillator is synchronized by a feedback signal which the vibrating tuning fork itself produces by periodically reflecting a beam of light on to a photocell. The drive electromagnet consumes about 30 W. During preliminary trials in the synchrocyclotron, the drive power required was found to be much greater, due to eddy-current losses in the tuning-fork blades caused by an unexpectedly large horizontal component of the stray field from the cyclotron magnet. This component was largely compensated by means of a DC coil consisting of ten turns and carrying a current of 400 A. The tuning-fork blades can vibrate not only in the desired fundamental mode, whereby the lips remain straight and parallel to the stator, but also in unwanted higher modes of vibration (observable by means of Chladni figures). It was found that, owing to the RF heating of the tuning fork, the resonant frequency of one of these higher modes coincided with a multiple of 55 c/s, resulting in the excitation of this mode. The coincidence was eliminated by fitting a stiffening rib of specific dimensions to each blade (care being taken that no other coincidences then arose). The heat generated by the current in the tuning fork (approx. 250 W) is removed by cooling the foot with water; cooling by radiation — the only alternative, since the tuning fork is contained in a vacuum — would lead to excessive temperatures. Non-uniform cooling involves the danger of warping. This was experimentally investigated by simulating with the aid of heater strips the calculated temperature distribution expected, and then measuring the deformations. The "lips" of the tuning fork were found to remain straight within about 0.2 mm. The relative position of the tuning fork and stator and the amplitude of the lips must be accurately maintained. For this purpose, the stator is fitted with five feeler devices. These supply signals which, via a "translator" and relays, actuate suitable combinations of four servomotors to automatically adjust the amplitude and/or appropriately displace the three points of suspension.

The first of the tuning forks constructed has operated satisfactorily for about three years, and has amply exceeded its estimated life.

VOLUME 22, 1960/61, No. 6 pp. 181-216 Published 14th March 1961

Philips Technical Review

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

# WAVEGUIDE EQUIPMENT FOR 2 mm MICROWAVES

## II. MEASURING SET-UPS

621.372.8:621.317.3

by C. W. van ES, M. GEVERS and F. C. de RONDE.

*In Part I of this article, a review was given of the equipment developed by Philips for 2 mm microwaves. Part II below contains a description of some measuring set-ups in which the use of the components discussed are considered in more detail; in addition, some components which were not considered in Part I are also described.*

*The first two set-ups discussed are for the measurement of the losses and the impedances of microwave components. The third set-up is a microwave gas spectrometer by means of which the absorption lines in gases can be experimentally determined.*

When designing components for a new frequency range, the microwave technician requires good measuring equipment in order to investigate the properties of the developed components. In this context, his preference will be for measuring instruments which do not need to be calibrated, i.e. instruments which are "absolute". The rotary attenuator is one such instrument. This latter can be used for the measurement of attenuations such as the dissipative loss in microwave components, which will later be discussed. The rotary attenuator itself was treated in Part I of this article [1]; likewise, the variable impedance with which reflection coefficients can be measured. This instrument is also absolute as far as the modulus of the reflection coefficient is concerned. It is used in the second measuring set-up discussed below for measuring the reflection coefficient of an unknown impedance in a bridge circuit (with a hybrid T as the bridge element). Here, the variable impedance serves as a reference impedance.

The third measuring set-up to be discussed is for the investigation of absorption spectra in gases. An example is carbonyl-sulphide gas (COS), which exhibits absorption at about 146 Gc/s. If the gas is irradiated at the correct frequency, the energy supplied allows the COS molecules to move to a higher rotational level. This transition is accompanied by absorption. The frequency at which absorption occurs can be determined with the set-up to be described.

### Measurement of the dissipative loss in microwave components

When a microwave component is inserted in a waveguide set-up which has been matched to both the load and generator sides, the losses will in general increase. This increase, which is called "insertion loss", is caused partly by reflections and partly by dissipation in the four-terminal network constituted by the inserted component. Reflection and dissipation are both characteristic properties of a four-terminal network. The dissipative loss can be measured in the manner now to be described; the reflective loss is derived from an impedance measurement using a second set-up which will be dealt with presently.

It is almost always desired to keep the dissipative loss as small as possible, and for that reason an effort has been made to keep the length of our components to a minimum. The need for short components becomes more evident as the frequency increases, since the loss per unit length increases as the $\frac{3}{2}$ power of the frequency. As mentioned in Part I, the claw flange with its associated lock ring is one of the devices used to achieve short constructions

[1] C. W. van Es, M. Gevers and F. C. de Ronde, Waveguide equipment for 2 mm microwaves, I. Components, Philips tech. Rev. 22, 113-125, 1960/61 (No. 4).

with correspondingly low dissipative losses. Moreover, at high frequencies the surface roughness has a considerable influence: the higher the frequency, and therefore the smaller the depth of penetration, the greater the resistance which irregularities and impurities will offer to the current.

The dissipative loss can be determined by two power measurements, one *with* the component being investigated in the set-up and one *without* it. The power can be measured by a water calorimeter; this is an absolute instrument but is cumbersome to use. In its place, use can be made of a thermistor, i.e. a thermal detector, which has been calibrated by a water calorimeter or (somewhat less accurately) by DC [2]). Also, a crystal detector (*fig. 1*) with a DC meter can be used as a power indicator. The crystal is a more or less square-law detector, so that the deflection of the meter is approximately proportional to the microwave power. The dissipation loss can be found from the ratio of the readings given by the two power measurements. Use of the rotary attenuator has the advantage that the two measurements can be made at the same crystal power, the measurements then being completely independent of the detection characteristic. The difference in position of the rotary attenuator when the component is present from that when it is removed from the set-up thus gives the dissipative loss directly.

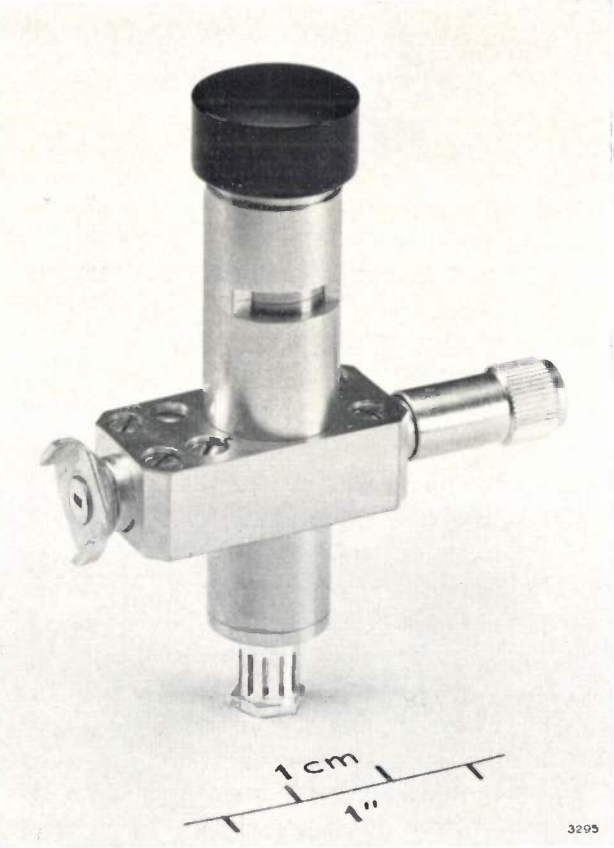


Fig. 1. Crystal detector for 2 mm waves. Left, claw flange, right, shorting plunger. The differential screw which is used to bring the silicon crystal into contact with the catswhisker is situated beneath the black cap; the displacement can be monitored through the window. Below: the coaxial connection for the millivoltmeter. The sensitivity is better than 10 mV per mW.

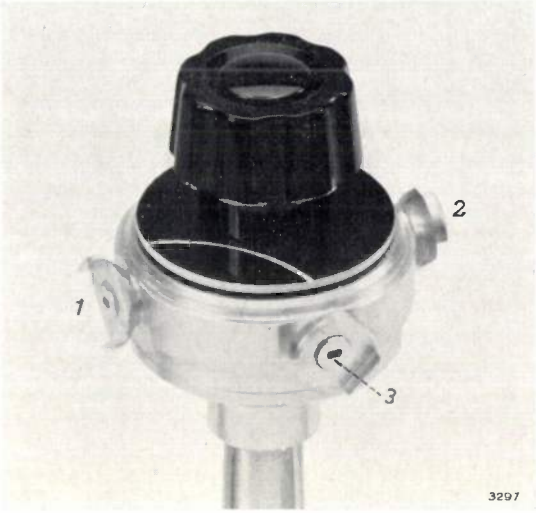


Fig. 2. Waveguide switch for 2 mm wavelength. A rotatable piece of waveguide, in the form of a quadrant, connects waveguide *3* to waveguide *1* in one position of the switch, and waveguide *3* to waveguide *2* in the other position. The standing-wave ratio is less than 1.02.

In these measurements, it is desirable to be able to switch rapidly from one condition to another since, in the interim period, the supplied power may vary for all sorts of reasons: the output power from the klystron, the conversion of the frequency multiplier and the sensitivity of the detector are subject to irregular fluctuations. Rapid switching-over is made possible by means of a waveguide switch (*fig. 2*). This contains a rotatable part consisting of a curved waveguide of quadrant form so that, according to requirements, a connection can be made between the waveguides *3* and *1* or *3* and *2*. It will be clear that particularly high demands are placed upon the mechanical construction and finish, in order to obtain a reproducible connection with a very low reflection coefficient (e.g. $|R|$ smaller than 0.01) between the rotatable and fixed parts of the waveguide.

The circuit with crystal detector and waveguide switch is reproduced in *fig. 3*; a photo of the whole set-up is given in *fig. 4* and a photo of the 2 mm part in *fig. 5*. Two branches are connected to the waveguide switch $S$. One consists of the component to be measured, in this case a piece of waveguide of

[2]) See Philips tech. Rev. **21**, 228 (Note [7])), 1959/60 (No. 8).

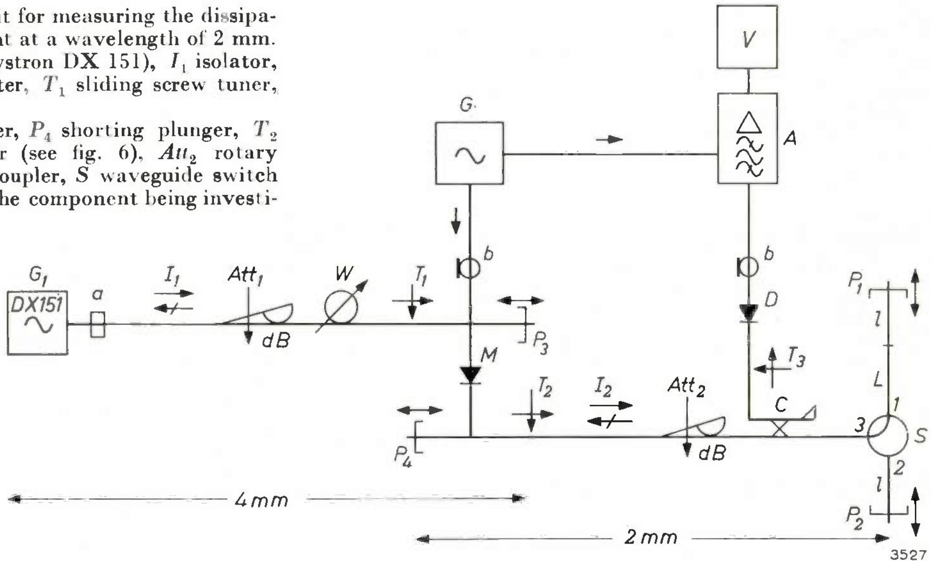


Fig. 3. Block diagram of the circuit for measuring the dissipative loss in a microwave component at a wavelength of 2 mm. *4 mm part*: $G_1$ generator (reflex klystron DX 151), $I_1$ isolator, $Att_1$ vane attenuator, $W$ wavemeter, $T_1$ sliding screw tuner, $P_3$ shorting plunger.
*2 mm part*: $M$ frequency multiplier, $P_4$ shorting plunger, $T_2$ pivoting screw tuner, $I_2$ isolator (see fig. 6), $Att_2$ rotary attenuator, $C$ rotary directional coupler, $S$ waveguide switch (see fig. 2). Branch *1* consists of the component being investigated (length $L$) combined with a line of length $l$ adjustable with plunger $P_1$. Branch *2* consists of a similar length $l$ adjustable with plunger $P_2$. $T_3$ pivoting screw tuner. $D$ detector (see fig. 1).
*Other equipment*: $G$ generator (8 kc/s) for synchronous detection, $A$ selective amplifier, $V$ millivoltmeter.
The sign *a* means "rectangular waveguide" and the sign *b* means "coaxial line".

length $L$ terminated at a distance $l$ from the flange by a shorting plunger $P_1$. The other branch is terminated directly by a plunger $P_2$, this latter being adjusted to the same distance $l$. By means of the waveguide switch, either the branch $L + l$ or the branch $l$ can thus be connected to the rest of the circuit. The circuit contains, in addition to a rotary attenuator, a rotary directional coupler $C$ — a component described in Part I. The rotary directional coupler is used to measure the reflected wave which comes back from the branch switched in by $S$.

The measurement of the dissipative loss is carried out as follows. With the waveguide switch in position *1*, adjustments are made until the meter $V$ shows full-scale deflection, and the rotary attenuator is set to zero. This done, $S$ is switched to position *2*, i.e. to the opposite branch from that carrying the component to be m:asured. The attenuation will

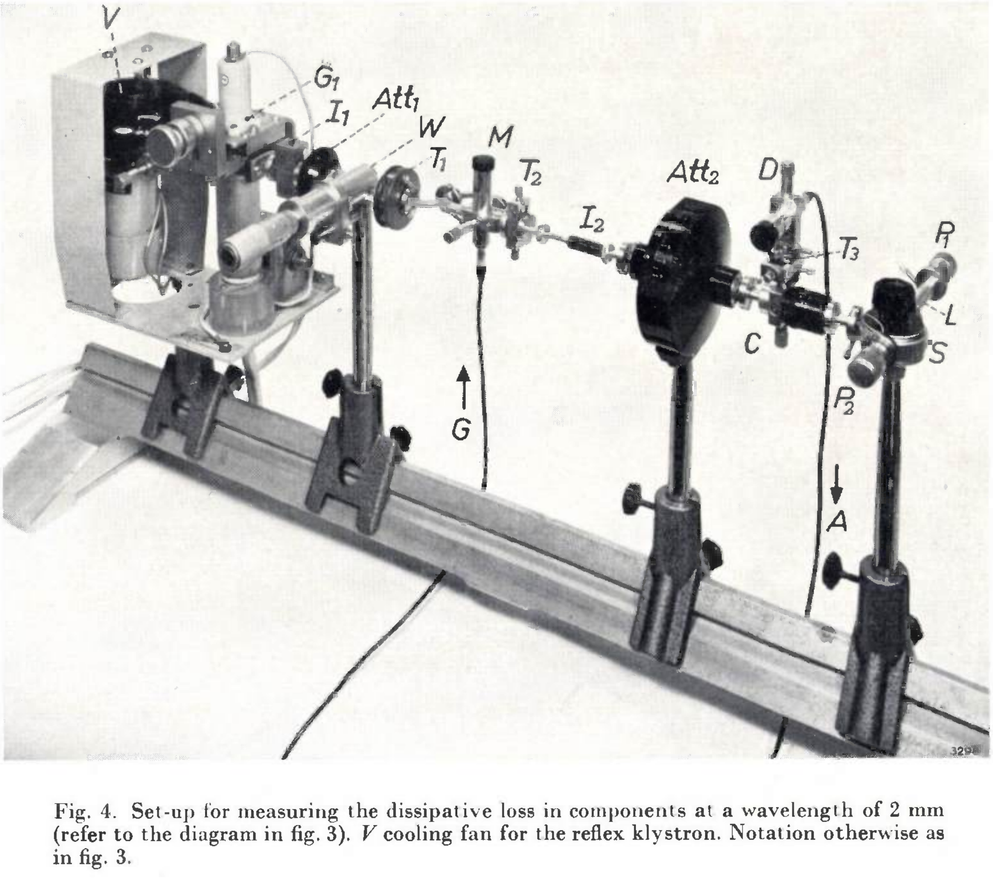


Fig. 4. Set-up for measuring the dissipative loss in components at a wavelength of 2 mm (refer to the diagram in fig. 3). $V$ cooling fan for the reflex klystron. Notation otherwise as in fig. 3.

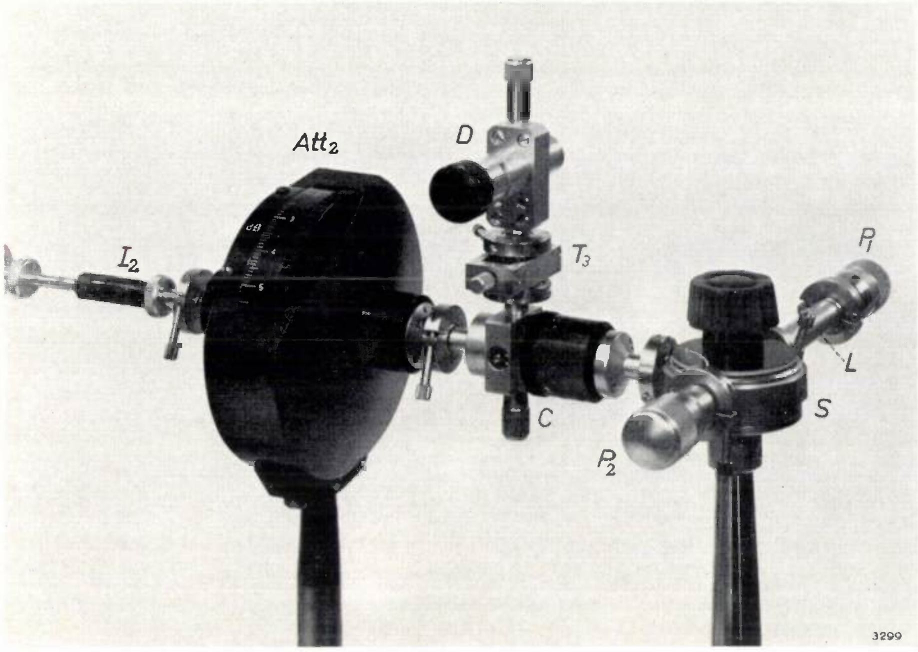


Fig. 5. The 2 mm part of the set-up reproduced in fig. 4. Notation as in fig. 3.

then be less, hence the reflected wave stronger and the meter will tend to deflect further. Now, using the rotary attenuator, the signal is attenuated sufficiently to restore the meter to its original deflection. We then read off from the rotary attenuator the indicated attenuation. Half this attenuation gives the dissipation in the component (half, since the wave traverses the length $L$ of the component twice).

In order to make the measurement as accurate as possible, a few precautions must be taken, one of which is to include a directional isolator.

Looking towards the generator, the transmission line will not generally be reflection-free. For this reason, on displacing the plunger $P_1$ or $P_2$, the deflection of the meter will vary. In order to keep this effect as small as possible, an isolator is connected between the attenuator and the multiplier ($I_2$ in fig. 3). This device, whose operation depends on the Faraday effect [3]), is depicted in *fig. 6*. However, even with these precautions some reflection will remain so that, in order to adjust the meter to the maximum or minimum deflection, the plungers $P_1$ and $P_2$ must both be adjustable.

At the beginning of the measurement, and before the component of length $L$ has been inserted in one of the branches, it must be ascertained that the two branches (length $l$) are identical, i.e. that, for both positions of the waveguide switch, the meter deflections are the same. When the component $L$ is subsequently fitted into the branch $l$, a small phase change will generally appear (unless the length $L$ is precisely a whole multiple of $\frac{1}{2}\lambda_g$, where $\lambda_g$ is the wavelength in the guide). This phase change can be corrected by adjusting the plunger $P_1$. In its turn, this adjustment causes some change in the dissipative loss, but only to a negligibly small degree.

Since the use of a directional coupler means that *reflections* from the component are also measured, it must first be ascertained that these are indeed negligibly small. If this is not the case, the reflection must first be compensated.

The measures taken make a very accurate measurement possible. The circuit described is particularly suited to the measurement of losses in sections of waveguide and residual losses in components, e.g. the losses in a vane attenuator (see Part I) in the zero position. With this method, losses of the order of 0.1 dB can easily be measured.

The dissipative loss in the components discussed in Part I was measured with the circuit of fig. 3. This does not, of course, mean that the method is only suitable for measurements on microwave components; it can be used equally well for measuring attenuations caused by any sort of physical

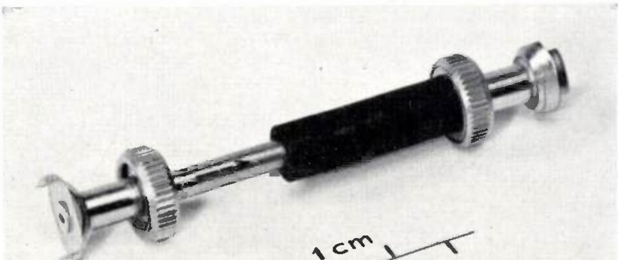


Fig. 6. Isolator for 2 mm wavelength, based on the Faraday effect [3]). The black rings are ferroxdure magnets. Attenuation in the forward direction (in the direction of the arrow): 2 dB; in the reverse direction: 15 dB.

[3]) H. G. Beljers, The application of ferroxcube in unidirectional waveguides and its bearing on the principle of reciprocity, Philips tech. Rev. 18, 158-166, 1956/57.

phenomena. It is, for example, very useful for measuring the absorption in superconductors [4]).

### Impedance measurements using the bridge method

In general, an impedance reflects some of the applied microwave energy when the impedance differs from the characteristic impedance of the line. A measure of this "mismatch" is either the standing-wave ratio or the reflection coefficient [5]). In order to determine the impedance by means of the standing-wave ratio, a standing-wave detector is necessary. However, it is extremely difficult to make a standing-wave detector for millimetre waves with reasonable accuracy. A more direct measurement of the impedance is possible by inserting it in a bridge circuit and comparing it with a standard impedance. The variable impedance discussed in Part I can serve this latter purpose. This instrument is absolute and the value of the reflection coefficient of the unknown impedance can be read off directly.

The circuit is reproduced in *fig. 7*. It consists of a 4 mm and a 2 mm part, the latter containing a hybrid T (see Part I), *HT*, as the bridge element. The input arm *1* of the T is connected to the frequency multiplier *M* via a pivoting screw tuner $T_1$. On the arm *4*, via another pivoting screw tuner $T_3$, a crystal detector *D* is connected, and on the arms *2* and *3* the variable impedance $Z_V$ and the unknown impedance $Z_3$, respectively. The latter consists of a pivoting screw tuner $T_2$ in conjunction with a matched load. Using this assembly, any required impedance can be made up.

The energy coming from the multiplier divides in the hybrid T into two equal parts, which travel along the arms *2* and *3*. $Z_3$ will cause a certain amount of reflection in arm *3*; the reflected wave returns to the branching point of the hybrid T and distributes itself over the arms *1* and *4*. If the reflection from $Z_V$ is now made equal in phase and amplitude to the reflection from $Z_3$, then a reflected wave of equal magnitude will likewise be split into two equal parts at the branching point and travel along the arms *1* and *4*. As explained in Part I, the geometric configuration is such that the waves issuing from the arms *2* and *3* cancel each other in arm *4*; therefore, the detector *D* receives no signal and the meter does not deflect. This is thus an indication that $Z_V = Z_3$, and the modulus of the reflection coefficient can be directly read off on the variable impedance, whilst its argument can be rapidly determined (see Part I, page 121).

A detail photo of the 2 mm part of this set-up is shown in *fig. 8*.

The sensitivity of this method is very high. Another property of the bridge circuit is that the sensitivity is considerably greater if the measurement is made after the bridge has been slightly unbalanced. This property is useful in the measurement of very weak reflections, such as those which occur at flanges, when values of $|R|$ less than 0.01 can still be fairly reliably measured. For this purpose, one arm of the hybrid T is terminated by a matched load and

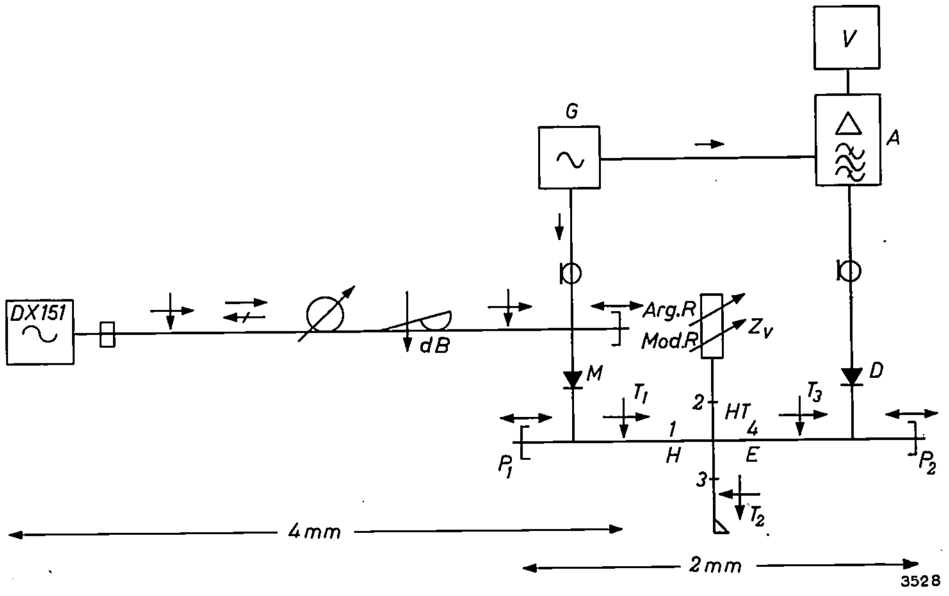


Fig. 7. Block diagram of the circuit for measuring the reflection coefficient of impedances at a wavelength of 2 mm.
*4 mm part*, from left to right: generator, sliding screw tuner, isolator, wavemeter, vane attenuator, sliding screw tuner, shorting plunger.
*2 mm part*: *M* frequency multiplier, $P_1$ shorting plunger, $T_1$ pivoting screw tuner, *HT* hybrid T, $Z_V$ variable impedance, $T_2$ pivoting screw tuner (with matched load), $T_3$ pivoting screw tuner, $P_2$ shorting plunger, *D* detector.
*Other equipment*: *G* generator (8 kc/s), *A* selective amplifier (tuned to the frequency of *G*). *V* millivoltmeter.

[4]) M. A. Biondi and M. P. Garfunkel, Millimeter wave absorption in superconducting aluminum, I and II, Phys. Rev. **116**, 853-867, 1959 (No. 4).

[5]) For the relationships between the quantities impedance, reflection coefficient and standing-wave ratio, see e.g. A. E. Pannenborg, A measuring arrangement for waveguides, Philips tech. Rev. **12**, 15-24, 1950/51.

a small reflection is introduced in the arm containing the variable impedance, so that the bridge is slightly out of balance. By altering the $|R|$ of the variable impedance by a certain amount with respect to the adjusted value (readable upon the scale "mod $R$", see fig. 16 of Part I), a specific variation in the deflection of the meter is obtained. The introduction of a flange coupling between the hybrid T and the matched load likewise causes a change in the deflection. The reflection coefficient can be found from the two changes in deflection. The fact that there are still other reflections of the same order of magnitude as those which are being measured, actually makes the measurement somewhat more complicated than is here described. It is necessary, for example, to compensate as well as possible for the reflections at other flange connections in the arms and for reflections at the transition, and the reflection at the matched load can no longer be neglected. When reasonably accurate measurements are required of very small reflection coefficients such as occur at claw-flange couplings, all this must be taken into account.

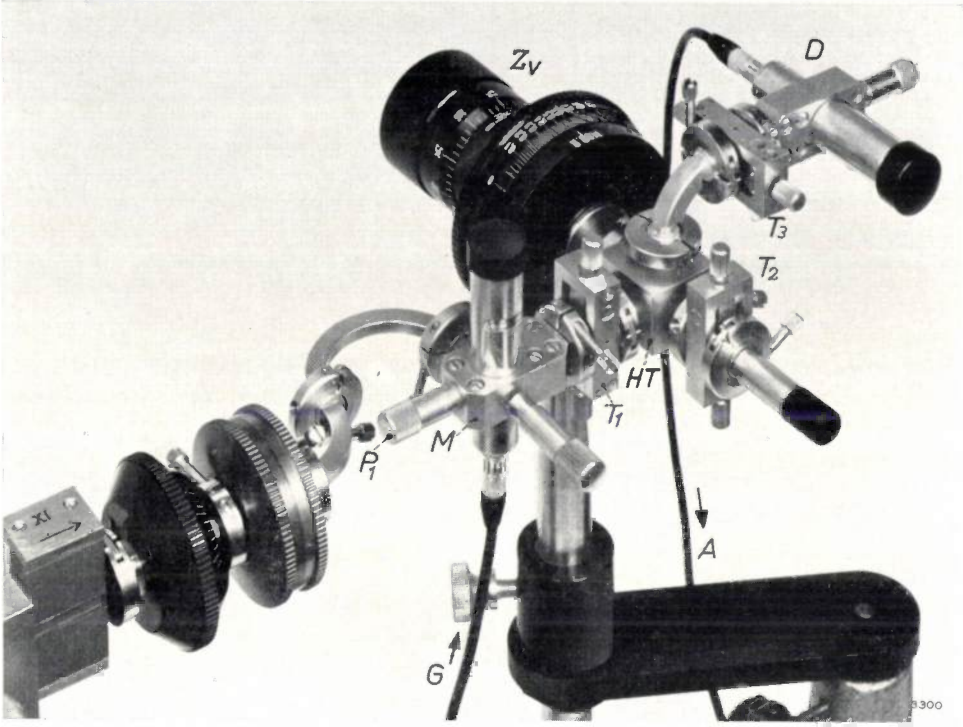


Fig. 8. The 2 mm part of the set-up shown in the diagram of fig. 7. Notation as in the latter. At the extreme left, a few 4 mm components: isolator, vane attenuator and sliding screw tuner.

The accuracy with which the reflection coefficient of an impedance can be determined is dependent upon the quality of two components: *a*) the hybrid T and *b*) the variable impedance. As has already been shown in Part I, as far as the hybrid T is concerned the accuracy is determined solely by the mechanical construction and the finish. In the variable impedance, still other factors play a part. The detrimental influence of all these factors upon the accuracy are particularly noticeable in the two extreme positions, mod $R = 1$ and mod $R = 0$. In the latter position, there is always some residual reflection, namely at the transition from the rectangular waveguide to the circular waveguide. If extremely accurate measurement is required, this residual reflection must be neutralized by means of a tuner.

In the maximum reflection position, $|R|$ is not quite 1 because of the losses in the waveguide. Here, again, is an advantage of the bridge circuit, in that these losses can be compensated by inserting between the impedance $Z_3$ and the hybrid T a section of waveguide which causes identical losses. For this purpose, the variable impedance is set to the position mod $R = 1$ and, in the other arm of the hybrid T, $Z_3$ is replaced by a short-circuited piece of waveguide of such length that, by adjustment of the phase, equilibrium can be established. Having balanced the bridge in this way for mod $R = 1$, we can then determine the dissipative loss $\alpha$ in components having low reflection. For this purpose the component is connected between the short-circuit and the added piece of waveguide. If balance is subsequently re-established, then:

$$a = -10 \log |R| \text{ dB}.$$

Thus, the dissipative loss in microwave components can be determined by the method just described as well as by the circuit of fig. 3. The measurement of large losses is, in both cases, limited by reflections.

## Measurement of an absorption line of COS at 2 mm

The binding forces existing between particles which together form a "system" (such as between the atoms in a molecule, between the nucleus and the electrons in an atom, between the protons and the neutrons in an atomic nucleus) give rise to a series of energy states which the system can occupy. On absorbing electromagnetic energy of certain frequencies, such a system can change from a given energy state to one of higher energy. A transition from the $n^{\text{th}}$ to the $(n + 1)^{\text{th}}$ energy level is made possible by absorption of radiation whose frequency $\nu$ is proportional to the energy difference $E_{n+1}$—$E_n$ between the two levels:

$$\nu = \frac{E_{n+1} - E_n}{h}.$$

Here, $h$ is Planck's constant (= $6.6 \times 10^{-34}$ joule seconds). The energy levels $E_n$ are determined by the nature of the forces and the kinds of particles. Thus, absorption in the gamma ray region is caused by changes of state within the atomic nucleus; absorption in the optical region is predominantly concerned with the binding forces between nuclei and electrons, and absorption in the infra-red and microwave regions are related to the force which the atoms of the molecule exert upon one another.

Thus, microwave spectroscopy can provide information concerning the chemical bonds in molecules (vibrational and rotational states). Here, infra-red spectroscopy is also important, but spectroscopy in the microwave region has the advantages that the radiation is purely monochromatic and that the sensitivity, the accuracy of determination of the absorption frequencies and the resolving power are greater. The latter is such that fine and hyperfine structures can be observed, enabling conclusions to be drawn concerning the behaviour of the nuclei of individual atoms. For further particulars the reader is referred to the literature [6]).

The third set-up, now to be discussed, is a microwave gas spectrometer used for determining the rotational spectra of gas molecules. The diagram of the circuit is reproduced in *fig. 9*. The main components in the 2 mm part — shown separately in *fig. 10* — are a gas cell *GC* and a 2 mm crystal detector $D_1$. The gas cell consists of a piece of 8 mm waveguide (this being used since it has a greater volume and smaller wall-losses per unit length) with at each end a transition to 2 mm guide. The

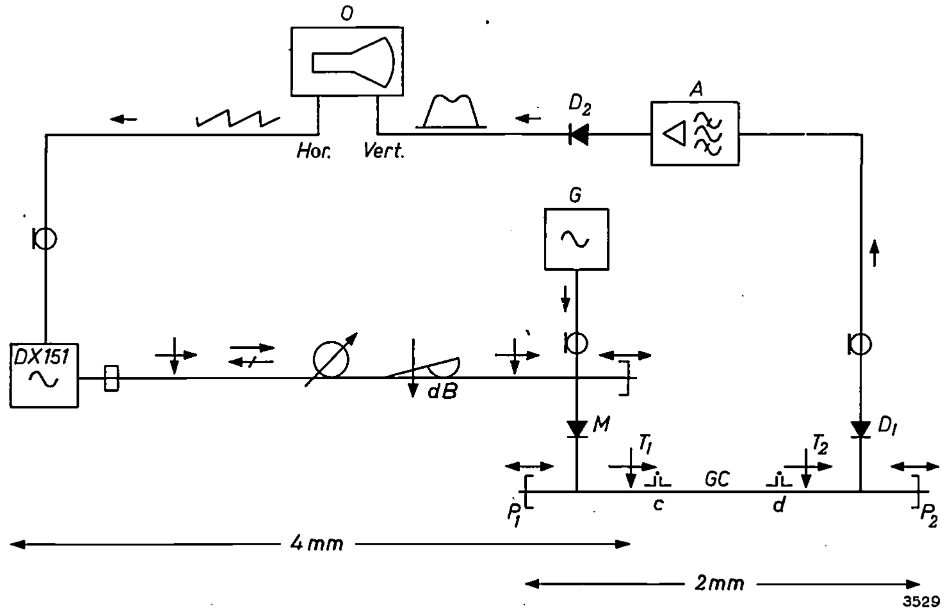


Fig. 9. Block diagram of the circuit for measuring the absorption line of carbonyl sulphide at about 146 Gc/s.
*4 mm part*, from left to right: generator, sliding screw tuner, isolator, wavemeter, vane attenuator, sliding screw tuner, shorting plunger.
*2 mm part*: *M* frequency multiplier, $P_1$ shorting plunger, $T_1$ pivoting screw tuner, *GC* gas cell (with mica windows *c* and *d*), $T_2$ pivoting screw tuner, $P_2$ shorting plunger, $D_1$ detector.
*Other equipment*: *G* generator (45 kc/s) which amplitude-modulates the 2 mm wave, *A* selective amplifier tuned to the frequency of *G*, $D_2$ detector, *O* oscilloscope. The sawtooth time-base voltage of *O* frequency-modulates the 4 mm generator. The form of the detected microwave is indicated at the input of *A*; the form of the signal after low-frequency detection by $D_2$ is shown at the input of *O*.

[6]) See, for example: W. Gordy, W. V. Smith and R. F. Trambarulo, Microwave spectroscopy, Wiley, New York 1957.
D. J. E. Ingram, Spectroscopy at radio and microwave frequencies, Butterworth's Scientific Publications, London 1955.
C. A. Burrus, Stark effect from 1.1 to 2.6 millimeters wavelength: $PH_3$, $PD_3$, DI, and CO, J. chem. Phys. 28, 427-429, 1958.
M. Cowan and W. Gordy, Precision measurements of millimeter and submillimeter wave spectra: DCl, DBr, and DI, Phys. Rev. 111, 209-211, 1958.

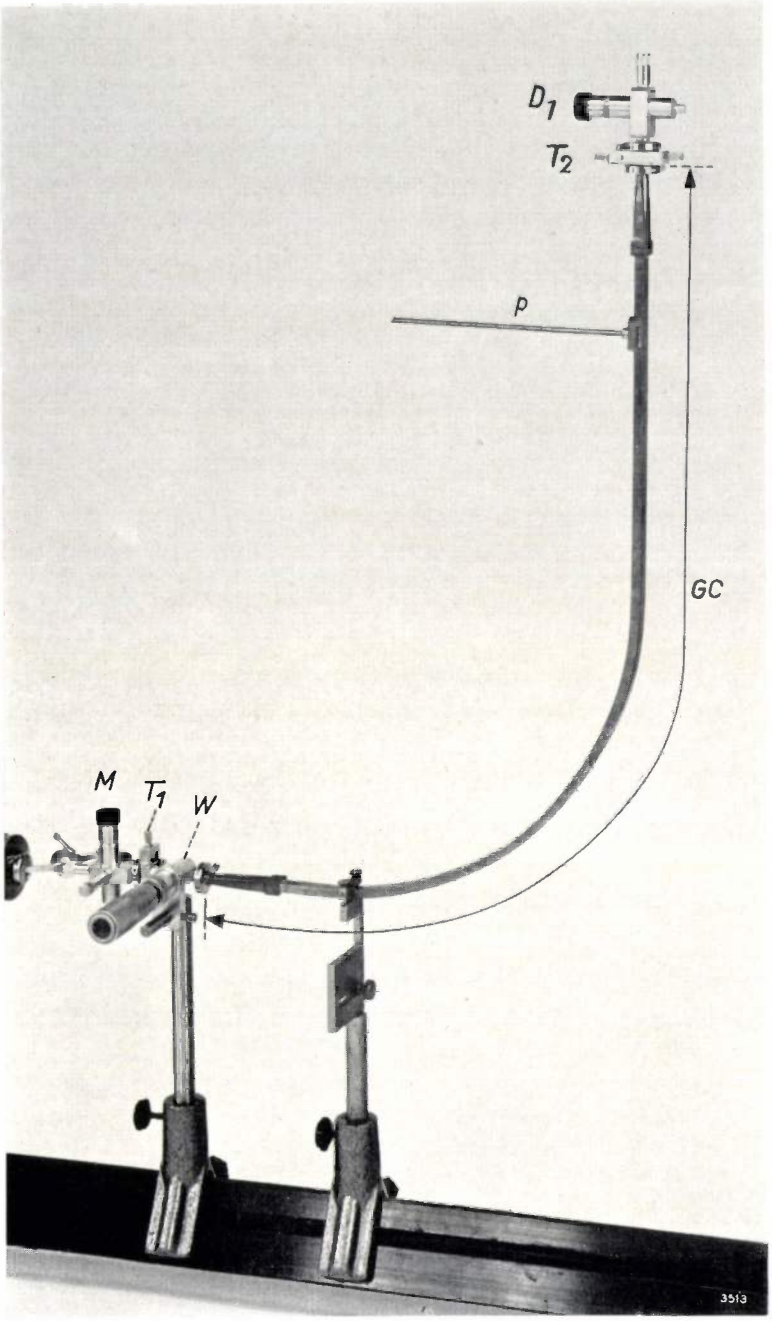


Fig. 10. The 2 mm part of the set-up, the diagram of which is reproduced in fig. 9. $p$ is the pump connection to the gas cell $GC$ (8 mm waveguide). Notation otherwise as in fig. 9. For the purpose of monitoring a wavemeter ($W$) ($TE_{010}$ mode) is included in the 2 mm part.

cell is made gas-tight by providing the transitions with mica windows ($c$ and $d$ in fig. 9). In fig. 10 the pump connection $p$ is visible, this being used to evacuate the gas cell before the admission of the gas sample.

One gas that exhibits an absorption line in the neighbourhood of 2 mm wavelength is carbonyl sulphide (COS). The COS molecule is a linear rotator, i.e. the three constituent atoms lie on a straight line (S = C = O), and the molecule rotates about an axis perpendicular to this line. The transitions between the rotational levels lie within the microwave region. Thus, for COS, the following transitions in the 2 mm region have been calculated:

transition from level 9 to level 10
at 121.625 Gc/s,
transition from level 11 to level 12
at 145.947 Gc/s,
transition from level 13 to level 14
at 170.267 Gc/s.

The method of measuring these absorption frequencies will now be described [7]). For this purpose, we choose the line in the neighbourhood of 146 Gc/s. The repeller voltage of the klystron is modulated by the sawtooth voltage which is used for the time-base of the oscilloscope $O$. The klystron oscillates only when the instantaneous value $v_r$ of the repeller voltage lies between the limits $V_{r1}$ and $V_{r2}$, and in this region the frequency $f$ varies more or less linearly with $v_r$ (*fig. 11*). In this way, a certain frequency region

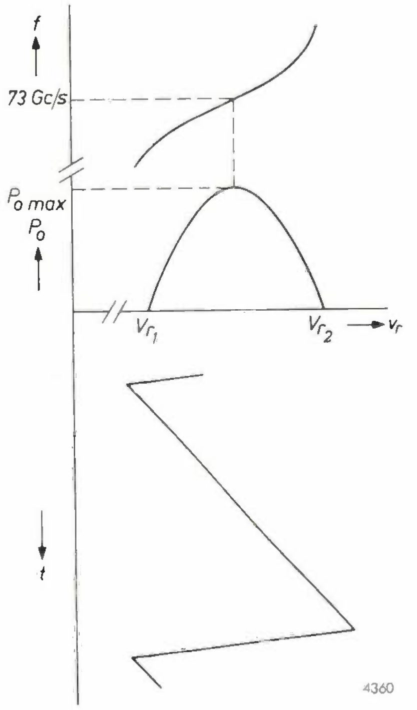


Fig. 11. The repeller voltage $v_r$ of the klystron in fig. 9 has a sawtooth waveform. The klystron oscillates between the limits $V_{r1}$ and $V_{r2}$. The output power $P_o$ and frequency $f$ vary as shown. At $P_o = P_{o\,max}$ the value of $f$ must be approximately 73 Gc/s.

[7]) C. G. Montgomery, Technique of microwave measurements, M.I.T. Radiation Lab. Ser., Vol. 11, McGraw-Hill, New York 1947, p. 24-33.

is covered [8]). The frequency sweep has a value of, for example, 0.1 Gc/s and the sweep must take place about the value 73 Gc/s approximately; this frequency must occur in the neighbourhood of the maximum output power $P_{o\,\max}$.

A check on whether the latter conditions are fulfilled can be made by switching off the generator $G$ and connecting the oscilloscope to the frequency multiplier $M$, which now acts temporarily as a 4 mm detector. This will give an oscillogram of the type shown in *fig. 12*, which represents the power characteristic of the klystron — $P_o$ as a function of $v_r$ (and therefore of $f$) — with a superimposed dip caused by the absorption in the 4 mm wavemeter. The latter is tuned to 73 Gc/s. If the klystron is properly tuned, then the peak of the curve occurs at the same frequency as the dip. If this is not so, then the klystron must be mechanically adjusted until the maximum *does* coincide with the dip.

Using the circuit of fig. 9, when *no* gas is present in the cell, an oscillogram of the type shown in *fig. 13a* is obtained. When the cell contains COS gas, the picture obtained is as shown in fig. 13*b*, where the absorption line is visible. The sharper this line is, the more accurately the frequency can be determined. At a given cell volume, the line width increases as the gas pressure rises; this is the result of the interaction between the molecules. An excessive microwave power has a similar detrimental effect because of the saturation which then occurs at the higher level.

The determination of the frequency is achieved in the first instance with the aid of a 4 mm or 2 mm wavemeter. With, for example, a carefully calibrated $TE_{01n}$ wavemeter of good construction, the measurement can be made with an accuracy of up to 0.01%. For greater accuracy, more elaborate equipment is necessary [9]), including a frequency standard which is regularly checked. However, the discussion of this does not fall within the scope of this article.

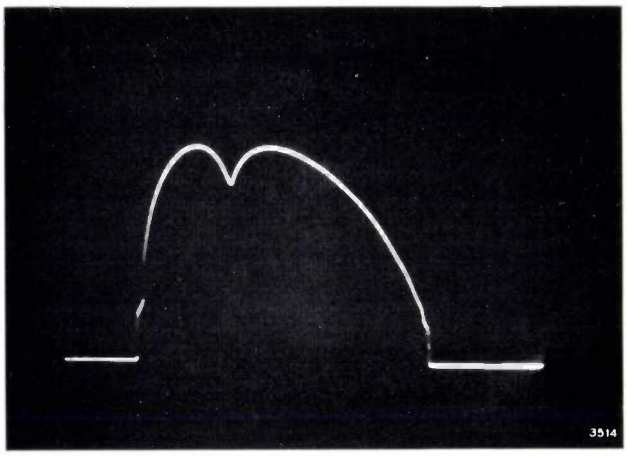


Fig. 12. Oscillogram of the detected output power $P_o$ from the klystron; the abscissa represents both time and frequency. The dip marks the frequency to which the wavemeter is tuned (73 Gc/s), the klystron being mechanically adjusted so that the peak of $P_o$ occurs at this frequency.

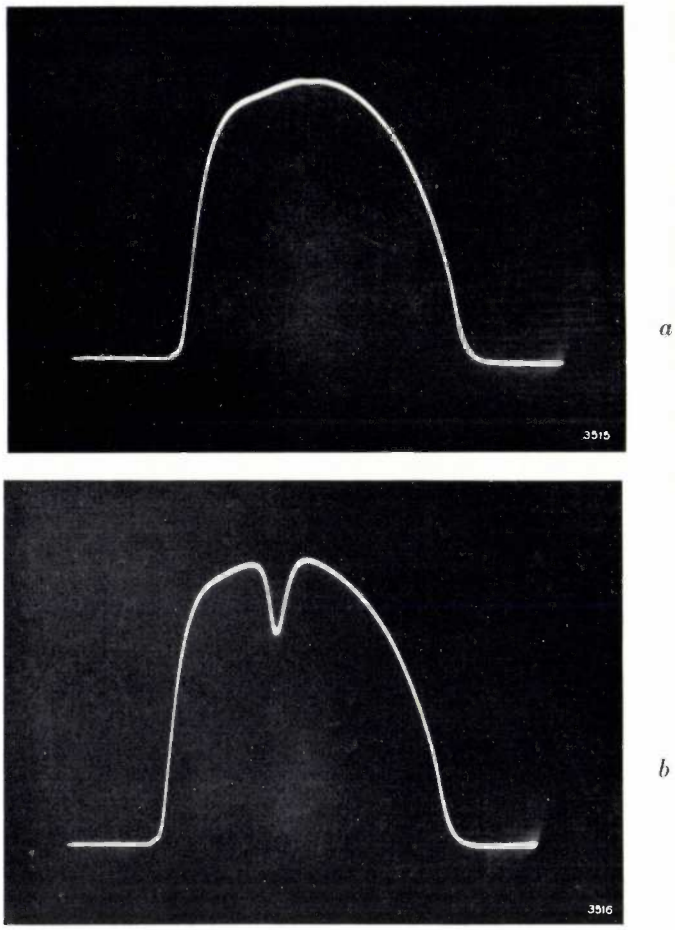


Fig. 13. Oscillograms obtained from the oscilloscope $O$ of fig. 9, *a*) evacuated gas cell, *b*) gas cell filled with COS gas. The absorption line in (*b*) occurs at a frequency of approximately 146 Gc/s.

[8]) Use is thus made of the "electronic tuning range" of the reflex klystron; see Philips tech. Rev. **21**, 224 (fig. 6, Note [4])), 1959/60 (No. 8).

[9]) O. R. Gilliam, Ch. M. Johnson and W. Gordy, Microwave spectroscopy in the region from two to three millimeters, Phys. Rev. **78**, 140-144, 1950.

**Summary.** In continuation of Part I of this article, in which waveguide components for 2 mm wavelength were discussed, Part II describes three measuring set-ups. The first of these is for measuring dissipative losses of the order of 0.1 to 3 dB in microwave components. In the second set-up, the reflection coefficients of unknown impedances are determined by means of a bridge circuit using a hybrid T. With certain precautions, voltage reflection coefficients lower than 0.01 can be measured with reasonable accuracy. The third set-up is used for the experimental determination of absorption lines in gases. The gas carbonyl sulphide (COS), which exhibits an absorption line at about 146 Gc/s, is chosen as an example. Some 2 mm components are also discussed which were not treated in Part I: a crystal detector, a waveguide switch and an isolator.

# SOLID-STATE RESEARCH AT LOW TEMPERATURES

## I. INTRODUCTION

576.48

by J. VOLGER.

*In the first decade of this century, the last "permanent gas", helium, was liquefied by Kamerlingh Onnes in the physics laboratory at Leiden University. Even at that time, however, interest was not confined exclusively to the thermodynamic properties of gases difficult to liquefy; investigations were already being made into the properties of solids in the new temperature range opened up. It was in the course of low-temperature research on the electrical conductivity of pure metals, for example, that superconductivity was discovered. Kamerlingh Onnes — whose name the laboratory now bears — rendered a valuable service to science in that he opened up the field of very-low-temperature research by devising equipment for the routine liquefaction of helium and other not readily condensable gases. Laboratory facilities of this kind were then virtually unique in the world.*

*Since then, solid-state research has expanded enormously and investigations at low temperatures are made into the most diverse properties. The author has chosen a number of instances of the work being done in this field, at the Philips laboratories and elsewhere, which will be presented in the form of three articles. It will be the aim in each subject discussed to explain why it is so important that the properties in question should be studied at low temperature.*

*The first article printed below, which is by way of an introduction to the subject, begins by considering what exactly is meant by "low temperature".*

In recent decades, cooling to extremely low temperatures has become a valuable research tool. The properties of solids, for instance — we shall not be concerned with liquids and gases in these articles — often exhibit characteristic changes at low temperature, from which inferences may be drawn regarding the structure of the substances investigated. Structure in this sense refers not merely to the geometry of the crystal lattice, whether or not perturbed by lattice defects, but also, for example, to the spectrum of possible lattice vibrations and the location and structure of the bands of energy levels occupied by electrons.

In this series of articles we shall endeavour to illustrate with a number of examples the scientific importance of low-temperature research, and also, where relevant, its technological importance. First, however, it will be useful to consider what is meant by "low" temperature. We shall see that the answer differs from case to case, and that in some instances even room temperature may be regarded as very low. Our considerations will be confined, however, to those properties of solids which show characteristic changes only after cooling to temperatures lower than about 50 °K.

### *The concept "low temperature"*

The criterion as to whether a temperature is to be regarded as "low" is to be found in the phenomenon in which we are interested. We may, for example, take the transition from the liquid to the solid phase, a phenomenon shown in principle by all substances at a sufficiently low temperature (with the exception of helium if the pressure is too low). It is found that in substances of corresponding crystal structure the melting point $T_S$ — which we regard in this case as the upper limit of the region where the temperature may be called low — is approximately proportional to the heat of fusion $\Delta U$. In *Table I* it is seen that $\Delta U/T_S$ is indeed roughly constant for the simple substances chosen, even though the values of $\Delta U$ and $T_S$ themselves differ considerably.

Table I. Comparison of the heat of fusion $\Delta U$ per gram-molecule and the melting point $T_S$ of various substances. For substances of simple crystal structure the ratios $\Delta U/T_S$ have much about the same value. In the case of water (ice), whose structure is more complicated, the value of $\Delta U/T_S$ differs considerably from that of the simpler structures.

| Substance | $H_2$ | $N_2$ | Ar | Ag | $H_2O$ |
|---|---|---|---|---|---|
| Heat of fusion (J/mole) | 126 | 960 | 1170 | 11300 | 6040 |
| Melting point (°K) | 16 | 63 | 84 | 1230 | 273 |
| Ratio $\Delta U/T_S$ | 8.0 | 15.0 | 13.8 | 9.2 | 22 |

The provisional conclusion, then, is that the limit of what we may regard as the low-temperature region is determined by an energy that is characteristic of the phenomenon involved.

A description of the solidification or fusion process by the methods of thermodynamics involves, apart from energy and temperature, the concepts of order and disorder. The degree of disorder may be expressed thermodynamically in terms of the entropy $S$; a high value of $S$ corresponds to a greater degree of disorder.

In the fusion of a solid under constant pressure the heat supplied is not used for raising the temperature but solely for bringing about the fusion, i.e. for creating a state of greater disorder. According to thermodynamics the process of fusion involves no change in the *free enthalpy* $G$ — also known as the Gibbs free energy or the thermodynamic potential — which is defined by the equation $G = U - TS + pV$ (where $U$ is the internal energy, $T$ the absolute temperature, $S$ the entropy, $p$ the pressure and $V$ the volume). The transition is therefore given by:

$$\Delta U - T_S \Delta S + p\Delta V = 0\,, \quad \dots \quad (I.1)$$

or, since, for a liquid-solid transition, $p\Delta V$ may be neglected:

$$\Delta S \approx \frac{\Delta U}{T_S}. \quad \dots\dots\dots \quad (I.2)$$

The entropy change (per mole) upon fusion is thus equal to the heat of fusion divided by the melting point.

Substances of corresponding crystal structure may be expected to have roughly the same value of $\Delta S$, and hence a similar value of $\Delta U/T_S$. (In the gas-to-liquid transitions the spread in the values of $\Delta S$ is even smaller than in the solidification process. This finds expression in Trouton's rule.)

Looking at the situation now more from the *molecular* point of view, it should be noted first of all that temperature is a quantity we use to describe merely the *macroscopic* state of the substance; the temperature concept is not applicable to individual molecules. The relation between the temperature of a system—we shall see later that a "system" does not necessarily imply a "quantity of matter" — and the energy of the individual molecules, is given by *statistical mechanics*. Although we cannot possibly do justice to the principles of statistical mechanics [1]) within the scope of this article, we shall nevertheless try to show their significance for the subject under consideration. To do this we take the following example.

### *The specific heat of a quantized system*

Suppose that a substance consists of molecules whose energy can have only the values $E_1$, $E_2$, ... $E_n$, ... and which exhibit no marked interaction, so that the total energy of the substance is specified almost entirely by the sum of the energy contributions of each of the molecules. Now statistical mechanics tells us that, when such a system is in thermal equilibrium, the probability $W(E_n)$ that a molecule will have the energy $E_n$ is given by the formula:

$$W(E_n) = g_n \exp\left(-E_n/kT\right). \quad \dots \quad (I.3)$$

In this expression, $k$ is Boltzmann's constant ($1.38 \times 10^{-23}$ J °K$^{-1}$), $T$ the absolute temperature, and $g_n$ the degeneracy of the energy level of rank $n$.

An energy level has a degeneracy of, say, 3 if it corresponds to three distinct quantum states of the molecule (atom) all having the same energy. The degeneracy can be removed by an external perturbing influence, e.g. by the application of a strong electric or magnetic field. The energies of the three states are then generally altered by unequal amounts, so that the threefold degenerate level breaks up into three non-degenerate levels. The splitting-up of the spectrum lines of a gas discharge subjected to a magnetic field (Zeeman effect) is an example of this.

From (I.3) we can immediately derive an expression for the average energy $\overline{E}$ of the molecules:

$$\overline{E} = \frac{\Sigma\, E_n g_n \exp\left(-E_n/kT\right)}{\Sigma\, g_n \exp\left(-E_n/kT\right)}. \quad \dots \quad (I.4)$$

On the assumption that each molecule of our substance may be treated as a (quantized) harmonic oscillator, the quantum theory states that the values which the energy may assume are equal to $\frac{1}{2}h\nu$, $1\frac{1}{2}h\nu$, $2\frac{1}{2}h\nu$, ... $(n - \frac{1}{2})h\nu$, ..., where $h$ is Planck's constant ($6.6 \times 10^{-34}$ J sec), $\nu$ the frequency of the oscillator, and $n$ is again the rank number. Substituting this in (I.4) we find, after some manipulation [2]), the Planck-Einstein formula:

$$\overline{E} = \tfrac{1}{2}h\nu + \frac{h\nu}{\exp\left(h\nu/kT\right) - 1}. \quad \dots \quad (I.5)$$

If the temperature is so high that $kT \gg h\nu$, we may write as a fair approximation:

$$\overline{E} = kT. \quad \dots\dots\dots \quad (I.6)$$

In this temperature range, then, the energy of the system is proportional to $T$. Its specific heat, which is equal to $N\,\mathrm{d}\overline{E}/\mathrm{d}T$, is therefore constant, viz. $Nk$ per gram-molecule, where $N$ is Avogadro's number. This is no longer true when the system is cooled down to temperatures where $kT$ becomes of the same order of magnitude as $h\nu$. At temperatures where $kT \ll h\nu$, the average energy of the molecules

[1]) An introduction to statistical mechanics is given by J. D. Fast, Entropy in science and technology, Philips tech. Rev. **16**, 258-269, 298-308 and 321-332, 1954/55. A more advanced treatment will be found in:
R. C. Tolman, Statistical mechanics with applications to physics and chemistry, New York 1927;
D. ter Haar, Elements of statistical mechanics, New York 1954;
L. D. Landau and E. M. Lifschitz, Statistical physics, London 1958, to name only a few.

[2]) See e.g. C. Kittel, Solid state physics, 2nd impression, p. 123, Wiley, New York 1957.

is given by

$$\overline{E} \approx \tfrac{1}{2}h\nu + h\nu \exp\,(-h\nu/kT)\,. \quad . \quad . \quad . \quad (I.7)$$

The specific heat $C$ is now no longer constant but depends on $T$. The variation of $\overline{E}$ and $C$ with temperature is shown in *fig. 1.*

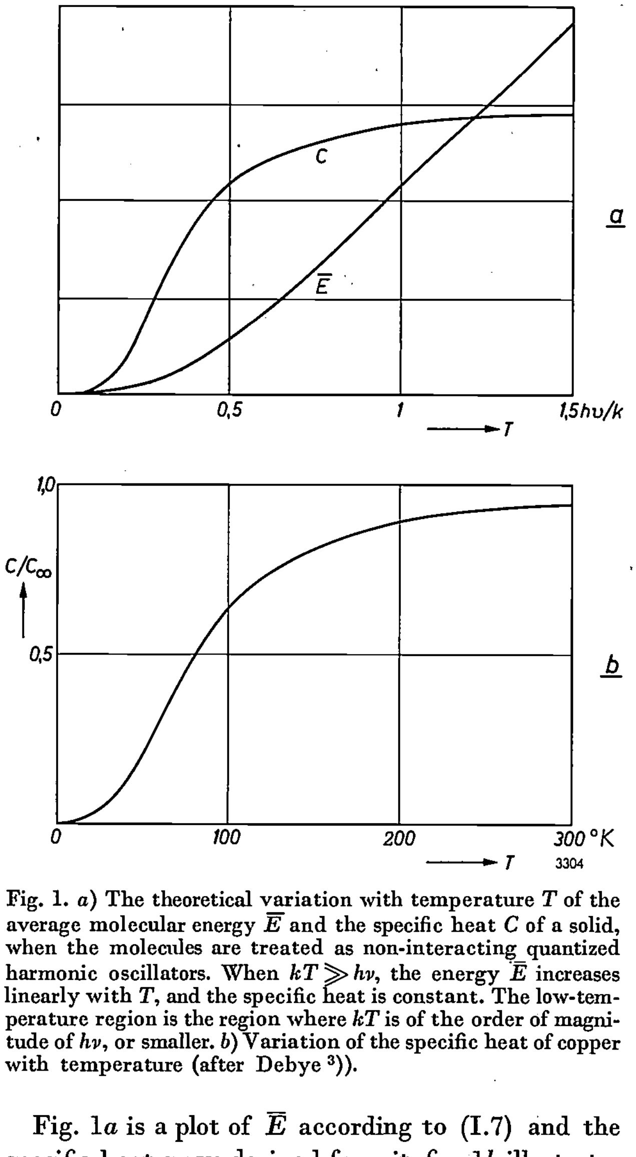


Fig. 1. *a*) The theoretical variation with temperature $T$ of the average molecular energy $\overline{E}$ and the specific heat $C$ of a solid, when the molecules are treated as non-interacting quantized harmonic oscillators. When $kT \gg h\nu$, the energy $\overline{E}$ increases linearly with $T$, and the specific heat is constant. The low-temperature region is the region where $kT$ is of the order of magnitude of $h\nu$, or smaller. *b*) Variation of the specific heat of copper with temperature (after Debye [3])).

Fig. 1*a* is a plot of $\overline{E}$ according to (I.7) and the specific-heat curve derived from it; fig. 1*b* illustrates how the specific heat of copper varies as a function of temperature. In spite of the sweeping simplifications in our example, the two curves show a remarkable resemblance.

We see that, according to statistical mechanics, the temperature in the above example may be qualified as "low" in the region where $kT$ is smaller than the energy $h\nu$. *As a general rule, low-temperature physics is concerned with the temperature region where kT is smaller than a certain characteristic molecular (or atomic) energy.*

Reverting to the example of the liquid-to-solid transition, we note that there the ratio $\Delta U/T_S$, calculated per mole, was approximately equal to $Nk$ ($Nk$ = 8.32 J/mole °K). Although in itself a macroscopic quantity, the heat of fusion may also be regarded in a sense as a molecular quantity, since its value depends closely on the energy with which a molecule (atom, ion) is bound to its site in the crystal lattice.

Finally, it should be pointed out that the quantity $kT$, which constantly occurs in statistical mechanics, does not always emerge from a distribution function of the type (I.3). The latter (due to Boltzmann) applies only to particles which obey the laws of "classical" statistics. The free electrons in a metal, for example, which collectively form a kind of "gas" about $10^4$ times as dense as an ordinary gas at normal temperature and pressure, and whose mass is roughly $10^4$ times smaller than that of ordinary gas molecules, are described by the Fermi-Dirac distribution function; to the "gas" of light quanta (photons) in a closed space the Bose-Einstein distribution function applies.

The existence of various different distribution functions is due to differences in the methods of calculating the number of states in which the system may be found. In the classical theory this is done on the underlying assumption that each molecule of a gas is distinguishable and in principle could be followed on its way. In quantum statistics identical particles are regarded as indistinguishable, and moreover states which can be derived from each other simply by the exchange of two identical particles are treated as the same state (one might thus say that the particles are not ouly indistinguishable but have no individuality). This change leads to the Bose-Einstein distribution. Where particles are concerned to which the Pauli exclusion principle applies — "no two particles can be in exactly the same quantum state" — (e.g. electrons), we come to the Fermi-Dirac distribution. Both quantum distributions reduce to the Boltzmann relation in the case of "gases" which are so rarefied that the distance between the particles is large compared with their De Broglie wavelength.

### *Paramagnetism*

As a third example of the general rule that the low-temperature region can be found by comparing $kT$ with a certain energy, we shall consider *paramagnetism.* By contrast with the two preceding examples, which served purely for illustration, the considerations here may serve as an introduction to the section on paramagnetic phenomena, which will appear in the third article of this series.

It will be known that paramagnetism arises from the fact that the electrons which together form the electron cloud of an atom, or ion, possess a magnetic (spin and orbital) moment. In many substances the electron clouds of the ions that make up the crystal lattice have what is called the "inert-gas structure", and the total moment is zero. There are

[3]) P. Debye, Ann. Physik **39**, 789, 1912.

some ions, however — e.g. those of the iron group — whose electron shells are not completely filled, and these do show a resultant moment. When a paramagnetic substance is placed in a magnetic field $H$, the component of the magnetic moment in the direction of $H$ can only assume certain discrete values, and the energy-level diagram shows correspondingly discrete levels. Generally, the distance between these levels is primarily determined by the strength of the applied field and further by the interaction of the magnetic atoms with neighbouring atoms (which may or may not be magnetic).

In the simplest case the magnetic moment can take up only two positions, with components of magnitude $\mu_B$ parallel or antiparallel to $H$ ($\mu_B$ being the Bohr magneton). The energy difference between the two levels is then $2\mu_B H$. Here again, the occupation of the levels follows from the Boltzmann equation (I.3), and by a method similar to that used for deriving (I.4) we find that the average magnetic moment $\overline{\mu}$ is given by:

$$\overline{\mu} = \mu_B \frac{\exp(\mu_B H/kT) - \exp(-\mu_B H/kT)}{\exp(\mu_B H/kT) + \exp(-\mu_B H/kT)} = \mu_B \tanh\frac{\mu_B H}{kT} . \quad . \quad . \quad (I.8)$$

For $kT \gg \mu_B H$, this yields the Langevin-Curie formula for the susceptibility $\chi$, the ratio between the resultant moment $\overline{\mu}N$ of a gram-atom and the magnetic field $H$:

$$\chi = N{\mu_B}^2/kT . \quad . \quad . \quad . \quad . \quad (I.9)$$

The results obtained with less simple energy-level diagrams are not essentially different from the above, provided $kT$ is again greater than the energy range in which the levels are situated.

The variation of $\overline{\mu}$ with $T$, as expressed in (I.8), is plotted in *fig. 2*. The transition between the region of high temperature, where $\overline{\mu}$ is small and proportional to $H$, and the region of low temperature, where saturation occurs, is seen to lie between 1 and 2 °K for an applied magnetic field of $10^6$ A/m (12 500 oersteds). Virtually complete saturation occurs only below 1 °K. The fact that the upper limit of the low-temperature region must be very low in this case may be directly inferred from (I.8), bearing in mind that $N\mu_B H$ at $10^6$ A/m is equal to only 7.0 J/mole, and that $NkT$ at $T = 1$ °K is as much as 8.3 J/mole.

### *Free electrons in metals and semiconductors*

As an introduction to the second article of this series, we shall now touch on some aspects of electrical conduction in metals and semiconductors.

It is common knowledge that, in most metals, every atom loses a certain number of electrons (usually one), and these electrons seem to move as free particles through the lattice; theoretically they may be regarded as independent. The energy levels they may occupy are so close together as to form effectively a continuous band. Broadly speaking, this band of energy levels is filled only up to a certain "height". As mentioned earlier, the occupation of energy levels in an "electron gas" is given not by the Boltzmann distribution but by application of Fermi-Dirac statistics. The occupation $f(E)dE$ of a narrow strip $dE$ of the energy band in which the density of the quantum states [4]) has the value $g(E)$ is given by:

$$f(E)dE = \frac{g(E)dE}{1 + \exp(E - E_F)/kT} . \quad . \quad (I.10)$$

In this expression $E_F$ is the energy of the Fermi level, which may be regarded to a first approximation as independent of temperature. We may deduce from this formula that, at absolute zero, $f(E) = g(E)$ for particles of energy $E \leqq E_F$, and $f(E) = 0$ for energies $E > E_F$. This means that all levels are filled below the Fermi level ($E \leqq E_F$) — with only *one* particle in each quantum state, in accordance with the Pauli exclusion principle — and all other levels are empty. Since the value of $E_F$ for metals amounts to one or two electronvolts, and the value of $kT$ at room temperature is only about $\frac{1}{40}$ eV, a metal at room temperature is already to be regarded as very cold. Although some levels for which $E > E_F$ are in fact occupied at this

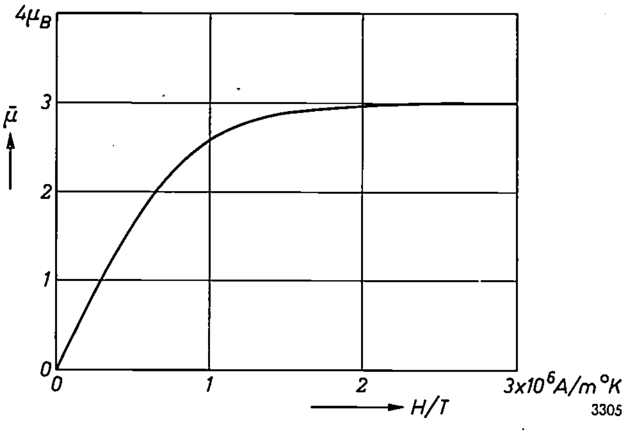


Fig. 2. The magnetization $\overline{\mu}$ of a paramagnetic substance is proportional to $H/T$ at high temperature (Langevin-Curie law), and virtually constant (saturation) even in weak fields at low temperature. For an applied field of $10^6$ A/m (12 500 oersteds), the transition region lies between 1 and 2 °K.

[4]) If there is no degeneracy at all, the density of the quantum states is equal to the density of the energy levels.

temperature, the occupation is nevertheless equal in good approximation to that obtaining at $T = 0$ (degenerate electron gas; see *fig. 3*). The reasons underlying the anomalous behaviour of the electrical conductivity of metals below 50 °K are therefore usually of an entirely different nature. That their study can contribute to our knowledge of the solid state will be explained later in these articles.

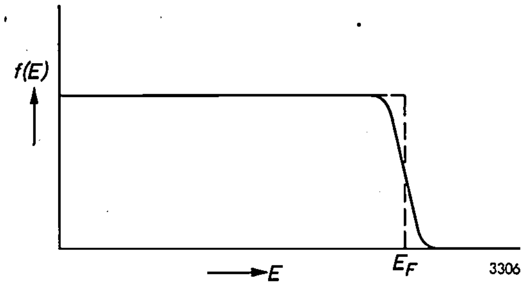


Fig. 3. In a metal the occupation f($E$) of the energy levels by conduction electrons is such that virtually all levels below the Fermi energy $E_F$ are filled and those above it empty. At room temperature (solid line) the occupation differs slightly from that which would exist at absolute zero (dashed line), by an amount of the order of $kT$ ($\approx \frac{1}{40}$ eV at room temperature). The value of $E_F$ is a few electron volts (1 eV $\approx k \times 10^4$ °K).

As may be inferred from the foregoing, the value of $E_F$ can be found directly by integrating formula (I.10) and equating $\int f(E)dE$ with the total number of electrons. The form of the function f($E$) will of course depend on the properties of the crystal lattice concerned, and will therefore differ from case to case. In the case of metals the value assumed by $E_F$ at absolute zero ($E_F$ is not, as we have seen, entirely independent of temperature) is found to be:

$$E_{F_0} = 4.2 \times 10^{-11} \times (m/m^*) n_0^{\frac{2}{3}} \text{ eV}. \quad . \quad . \quad (I.11)$$

Here $n_0$ is the concentration of the free electrons, $m$ the mass of the electron and $m^*$ its effective mass in the relevant lattice. The effective mass $m^*$ has the significance that the motion of an electron acted upon by external forces can be described as if the electron were a particle of mass $m^*$ moving in free space instead of through a crystal lattice.

One last remark. We have said that, because a metal is effectively very cold at room temperature, any anomalous effects it may show below that temperature are not connected with the occupation of the energy levels. Sometimes, however, this may not be entirely true, particularly if g($E$) has anomalous values at $E$ values close to $E_F$, as it has in sodium, for example.

In semiconductors the situation as a rule is quite different. Bands of energy levels are separated here by forbidden bands, i.e. zones of energy in which there are no levels at all. At low temperature there is very often an entirely filled band (valence band) and well above it an entirely empty one (conduction band). Only at a relatively high temperature can an electron acquire sufficient energy to reach the conduction band. In semiconductors the number of free electrons is therefore very much smaller than in metals. Frequently — in the case of "extrinsic semiconductors" — the free electrons are not supplied by the atoms or molecules of the substance itself but by foreign atoms (donor impurities); the concentration of the free electrons is then a function of the impurity concentration too. These electrons are released from the donor atoms by thermal agitation, at the cost of an ionization energy $\Delta E$.

At temperatures where $kT \ll \Delta E$, virtually all electrons will evidently be bound to the donors, and from the above considerations we should then expect these impure substances to be insulators, just as chemically pure semiconductors are at a sufficiently low temperature. However, even in this temperature range, conduction does take place in some of these substances, apparently in consequence of a different mechanism.

The idea underlying this "impurity band conduction" is that the electrons do not first have to be excited sufficiently to promote them directly into a level in the conduction band, but that they "jump" from one donor to another. Of course, between the region of high temperature, where normal conduction via the conduction band dominates, and the region of very low temperature, where impurity band conduction prevails, there is a transitional region where both mechanisms make a substantial contribution to the current. The effects observed in this region will be discussed in the second article of this series.

It may be useful to comment here briefly on an aspect of the behaviour of free electrons in semiconductors, which was not mentioned above because its consequences will not be discussed in these articles. As we have seen, the free electrons in a metal are described by the Fermi-Dirac statistics; a metal, statistically speaking, is always "very cold". This is not so in the case of semiconductors. It is possible in some semiconductors to control the nature of the impurity concentration in such a way that, in the temperature range of the experiments, the Maxwell-Boltzmann statistics apply at high temperatures and the Fermi-Dirac statistics at low temperatures. In the region of transition between the two statistics, particularly interesting effects are found. For most impurity semiconductors the Boltzmann distribution usually applies, and therefore their conductivity is much more temperature-dependent than that of metals.

In conclusion it should be mentioned that very many solid-state processes, such as the orientation of the ions (electron spins) in a paramagnetic salt subjected to a magnetic field, take place through the intermediary of lattice vibrations. When the temperature falls, these vibrations become weaker, thereby slowing down the processes concerned. If we regard the lattice vibrations as the vibrations of quantized oscillators, then the ratio of $kT$ to the magnitude of the energy quanta is the determining factor. This aspect of the low-temperature behaviour of solids will also be discussed in one of the following articles in this series.

**Summary.** This introductory article, the first of three on solid-state research at low temperatures, shows by a number of examples that a system may be said to be at a low temperature when $kT$ is smaller than a certain characteristic energy. This is the energy quantum $h\nu$ for the specific heat of a system of quantized harmonic oscillators, the heat of fusion (per mole) for the fusion of a solid, the energy $\mu_B H$ for the magnetization of a paramagnetic substance, and the Fermi energy $E_F$ for the electrical conductivity of metals and semiconductors. The article concludes with the remark that many solid-state processes, e.g. the orientation of electron spins in a paramagnetic substance, are slowed down at low temperature.

# AN OMEGATRON FOR THE QUANTITATIVE ANALYSIS OF GASES

by A. KLOPFER *) and W. SCHMIDT *).

621.384.8:621.039.343

*The present tendency towards high vacua of lower and lower pressures both in laboratory equipment and in electron tubes and other industrial products, makes it important to determine accurately the composition as well as the total pressure of the residual gas. Among the various kinds of mass spectrometer used for this purpose, the omegatron is particularly well suited — as this article describes — for determining, qualitatively and quantitatively, the composition of a gas at pressures lower than $10^{-5}$ mm Hg.*

For some years past, omegatrons have been developed and used in various laboratories for determining the composition of residual gases in high-vacuum systems. The first description of an omegatron was given by Sommer, Thomas and Hipple in 1951 [1]). The usefulness of the instrument was demonstrated by Alpert and Buritz in 1954 with a simplified arrangement [2]), which quickly found application in many laboratories.

The omegatron has been used at Philips since 1953 for *qualitative* analysis of the residual gas in sealed-off cathode-ray tubes [3]). With a view to making the instrument suitable for *quantitative* analyses, investigations were undertaken which have led to the omegatron described in this article.

## Principle of the omegatron

The operation of the omegatron was described in this journal some years ago. We shall briefly recapitulate here the underlying principle. A perspective sketch of the instrument is shown in *fig. 1*. A narrow beam of electrons from a cathode $K$ runs parallel to a uniform magnetic field $B$. Electrons colliding with gas molecules give rise to ions. If the latter have a velocity component perpendicular to the direction of the magnetic field, they describe circular paths perpendicular to the (static) magnetic field. The radii of these paths are given by the equation:

$$r = \frac{m}{e}\frac{v_0}{B}, \qquad \qquad (1)$$

where $m$ is the mass of the ion, $e$ its charge, $v_0$ the velocity component perpendicular to the direction of the magnetic field and $B$ the field-strength (or

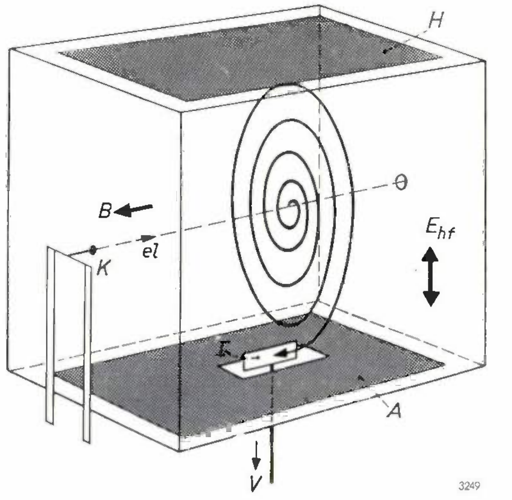


Fig. 1. Perspective sketch of a simplified omegatron. $B$ indicates the direction of the magnetic induction, $E_{hf}$ the direction of the RF field. The figure illustrates the spiral path described by an ion. $K$ cathode, which emits the ionizing beam of electrons *el*. $A$ and $H$ are the electrodes to which the RF voltage is applied. $I$ ion collector. $V$ connection to amplifier.

*) Philips Zentrallaboratorium GmbH, Aachen Laboratory.

[1]) H. Sommer, H. A. Thomas and J. A. Hipple, Phys. Rev. **82**, 697, 1951.

[2]) D. Alpert and R. S. Buritz, J. appl. Phys. **25**, 202, 1954.

[3]) J. Peper, Philips tech. Rev. **19**, 218, 1957/58.

temperature falls, these vibrations become weaker, thereby slowing down the processes concerned. If we regard the lattice vibrations as the vibrations of quantized oscillators, then the ratio of $kT$ to the magnitude of the energy quanta is the determining factor. This aspect of the low-temperature behaviour of solids will also be discussed in one of the following articles in this series.

**Summary.** This introductory article, the first of three on solid-state research at low temperatures, shows by a number of examples that a system may be said to be at a low temperature when $kT$ is smaller than a certain characteristic energy. This is the energy quantum $h\nu$ for the specific heat of a system of quantized harmonic oscillators, the heat of fusion (per mole) for the fusion of a solid, the energy $\mu_B H$ for the magnetization of a paramagnetic substance, and the Fermi energy $E_F$ for the electrical conductivity of metals and semiconductors. The article concludes with the remark that many solid-state processes, e.g. the orientation of electron spins in a paramagnetic substance, are slowed down at low temperature.

# AN OMEGATRON FOR THE QUANTITATIVE ANALYSIS OF GASES

by A. KLOPFER *) and W. SCHMIDT *).

621.384.8:621.039.343

*The present tendency towards high vacua of lower and lower pressures both in laboratory equipment and in electron tubes and other industrial products, makes it important to determine accurately the composition as well as the total pressure of the residual gas. Among the various kinds of mass spectrometer used for this purpose, the omegatron is particularly well suited — as this article describes —for determining, qualitatively and quantitatively, the composition of a gas at pressures lower than $10^{-5}$ mm Hg.*

For some years past, omegatrons have been developed and used in various laboratories for determining the composition of residual gases in high-vacuum systems. The first description of an omegatron was given by Sommer, Thomas and Hipple in 1951 [1]). The usefulness of the instrument was demonstrated by Alpert and Buritz in 1954 with a simplified arrangement [2]), which quickly found application in many laboratories.

The omegatron has been used at Philips since 1953 for *qualitative* analysis of the residual gas in sealed-off cathode-ray tubes [3]). With a view to making the instrument suitable for *quantitative* analyses, investigations were undertaken which have led to the omegatron described in this article.

## Principle of the omegatron

The operation of the omegatron was described in this journal some years ago. We shall briefly recapitulate here the underlying principle. A perspective sketch of the instrument is shown in *fig. 1*. A narrow beam of electrons from a cathode $K$ runs parallel to a uniform magnetic field $B$. Electrons colliding with gas molecules give rise to ions. If the latter have a velocity component perpendicular to the direction of the magnetic field, they describe circular paths perpendicular to the (static) magnetic field. The radii of these paths are given by the equation:

$$r = \frac{m}{e}\frac{v_0}{B}, \qquad (1)$$

where $m$ is the mass of the ion, $e$ its charge, $v_0$ the velocity component perpendicular to the direction of the magnetic field and $B$ the field-strength (or

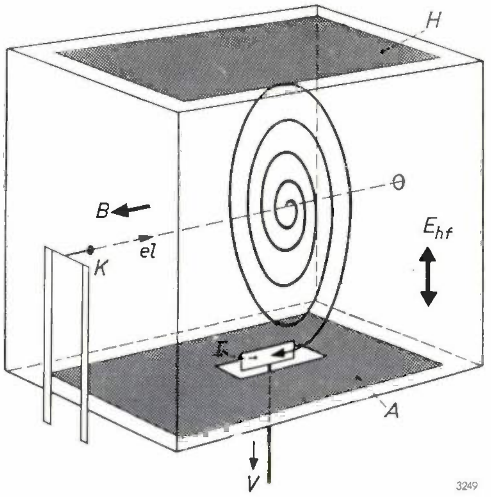


Fig. 1. Perspective sketch of a simplified omegatron. $B$ indicates the direction of the magnetic induction, $E_{hf}$ the direction of the RF field. The figure illustrates the spiral path described by an ion. $K$ cathode, which emits the ionizing beam of electrons *el*. $A$ and $H$ are the electrodes to which the RF voltage is applied. $I$ ion collector. $V$ connection to amplifier.

*) Philips Zentrallaboratorium GmbH, Aachen Laboratory.

[1]) H. Sommer, H. A. Thomas and J. A. Hipple, Phys. Rev. **82**, 697, 1951.

[2]) D. Alpert and R. S. Buritz, J. appl. Phys. **25**, 202, 1954.

[3]) J. Peper, Philips tech. Rev. **19**, 218, 1957/58.

rather the magnetic induction, using the rationalized Giorgi system). The angular frequency $\omega_c$ of the revolution of the ion is:

$$\omega_c = \frac{e}{m} B, \quad \ldots \ldots \quad (2)$$

and the period of revolution is therefore independent of the velocity $v_0$. This is the same situation as in a cyclotron, and just as in that case an RF field $\hat{E}_{hf} \sin \omega t$ is now applied perpendicular to the magnetic field. The effect of this is that the orbiting ions are accelerated or slowed down, depending on the value of $e/m$ (i.e. their angular frequency) and their phase. If we make the frequency $\omega$ of the alternating electric field equal to the angular frequency $\omega_c$ of a given kind of ion, the result is a condition of resonance. An ion in the correct phase then absorbs an equal quantity of energy upon each revolution, and thus describes a spiral of uniformly increasing radius (equiangular or Archimedes spiral). The resonating ions can be caught on a suitably placed collector electrode, and the ion current detected with the aid of a sensitive amplifier.

Ions having some other mass or charge (i.e. a different angular frequency) and which are thus not in resonance, also describe spiral paths, that is to say orbits whose radius varies with time. However, the maximum value of this radius generally remains smaller than the distance between the point of origin of the ions and the collector, and therefore such ions do not contribute to the measured current.

In order to determine the composition of a gas mixture, the frequency of the electric field applied to the omegatron must be varied either continuously or in steps. When a collector current, due to ions in resonance, is measured at a particular frequency, the $e/m$ ratio of these ions can be found from equation (2). It is then possible to determine the kind of gas concerned. The collisions between electrons and gas molecules give rise not only to singly ionized gas molecules, but also to doubly or multiply ionized particles. The gas molecules may also be broken up into ionized or neutral particles. In this way one finds for any type of gas and any given electron energy a complete "ionic spectrum", having a series of peaks corresponding to the respective masses of the particles, and with constant ratios between the heights of the peaks. For any gas these ratios, which correspond to the relative ion current measured at the respective resonant frequencies, can now be found by calibration; see *Table I*. When the omegatron contains an unknown gas, its nature can be deduced from the measured intensities for the various mass numbers by comparing the ratios between the ascertained values with the ionic spectra determined by calibration.

The pressure of each component of the gas mixture can in principle be found from the ion current, provided the relation between the pressure and the ion current is reproducible. With an electron current $i^-$ and a pressure $p$, the ion current $i_0^+$ is given by:

$$i_0^+ = s\,\sigma\,p\,i^-, \quad \ldots \ldots \quad (3)$$

where $s$ in the effective length of the electron beam through the gas, and $\sigma$ is the probability of ionization. But not all the resonant ions produced reach the ion collector. Some of them are lost by collision with gas molecules. Of the ions that do not collide with gas molecules, a fraction $a$ reaches the

Table I. Relative heights of the peaks due to ions of various masses for a number of gases (the peak of the most commonly occurring mass of each gas is taken as 100). In most cases the true mass of the gas will be that corresponding to the highest peak; see e.g. $H_2O$. This is not always so, however; see for example $C_3H_8$.

| Gas \ Mass | 2 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $H_2$ | 100 | — | — | — | — | — | — | — | — | — | — | — | — |
| Ar | — | — | — | — | — | — | — | | | 14.2 | — | — | — |
| $H_2O$ | 0.6-1.5 | — | — | — | — | 1.8 | 21 | 100 | 11-17 | 0.23 | — | — | — |
| $N_2$ | — | — | — | 7.4 | 0.03 | — | — | — | — | — | — | — | — |
| CO | — | 3.3 | 0.04 | 0.55 | — | 1.3 | — | — | — | — | — | — | — |
| $CO_2$ | — | 3.5 | 0.03 | 0.08 | — | 7.8 | — | — | — | — | — | — | — |
| $CH_4$ | | 1.8 | 5.7 | 12.5 | 81 | 100 | 2.7 | — | — | — | — | — | — |
| $C_2H_2$ | 3.5 | 1.4 | 4.0 | 0.3 | 0.04 | — | — | — | — | — | 5.1 | 19 | 100 |
| $C_2H_4$ | | 0.6 | 1.0 | 2.3 | 0.3 | 0.4 | — | — | — | — | 2.0 | 6.8 | 47 |
| $C_2H_6$ | | 0.2 | 0.55 | 2.0 | 3.1 | 0.15 | — | — | — | — | 0.5 | 2.7 | 18. |
| $C_3H_8$ | | 0.18 | 0.36 | 1.13 | 3.8 | 0.12 | — | — | — | — | 0.13 | 0.64 | 8.2 |

collector, and the remaining part (1—$a$) arrives on the other electrodes. The current $i^+$ incident on the ion collector is therefore:

$$i^+ = a\, i_0^+\, e^{-L/\lambda} = a\, s\, \sigma\, p\, i^-\, e^{-L/\lambda}, \quad . \; . \; . \quad (4)$$

where $L$ represents the total path length of the resonating ions and $\lambda$ the mean free path. In most cases, $\lambda$ will be very much greater than $L$. The requirement that the relation between $i^+$ and $p$ should be reproducible amounts to saying that, for given values of the parameters $s$, $\sigma$, $L$ and $\lambda$ and a given electron current $i^-$, the coefficient $a$ should be constant. It is particularly important that $a$ should not depend on the state of the instrument, nor on the presence of other gases whose ions are not in resonance. The ideal, of course, would be $a = 1$. Let us consider the effects that can make $a$ smaller than unity and upset the reproducibility.

As a result of their thermal energy, the ions have a tendency to escape from the plane of their spiral path, i.e. in the direction of the magnetic field. A weak electrostatic field is therefore applied which hinders this tendency. There are then in the middle of the omegatron components of the electric field which are perpendicular to the magnetic field. The effect of this is that the ions can describe cycloidal paths along the equipotentials. We shall return to this problem later. As a result of charged layers on the electrodes, the electrostatic field in the tube may undergo unpredictable variations, and thus affect the probability of ions escaping. Ions that are not in resonance may give rise to space-charge effects, which are dependent on the pressure and also influence the equipotential surfaces of the electrostatic field. Because of all these effects, $a$ is not constant with time and moreover differs from one tube to another. Since the fields produced by the effects referred to are of the same order of magnitude as the applied electrostatic field, widely divergent values may be found for $a$ (between 0 and 1). These variations of $a$ made it impossible to use the omegatron in its original form for quantitative analyses.

## Omegatron with side plates

In the Philips Aachen Laboratory an omegatron has now been developed whose sensitivity can be adjusted to give a constant value of $a = 1$, the sensitivity being the ion current divided by the product of electron current and pressure. The tube is shown schematically in *fig. 2*. The left half of the figure represents a cross-section parallel to the direction of the magnetic field. In this section the omegatron has its original form [1]). The cathode $K$ emits electrons which are accelerated through the omegatron parallel to the magnetic field and are finally caught by the electron collector $T$. To minimize the reactions of the gas with the hot cathode, a cathode is used which can operate at a relatively low temperature, namely a directly heated barium-oxide cathode [4]). The electrode $G_1$ may be used for stabilizing the electron current. The electrode $G_2$ has roughly the same electrical potential as the electron collector $T$. The omegatron must be aligned with extreme accuracy between the poles of the magnet, in order to prevent electrons striking the electrode $A$. This can be checked with a micro-ammeter. If the alignment is not accurate, the secondary electrons formed will affect the space charge and hence the sensitivity. The RF voltage

[4]) The use of this cathode was suggested by J. Peper. For various reactions of gases with cathodes, see the article by S. Garbe in Advances in vacuum science and technology, Proc. first int. Congress Vac. Tech., Namur 1958, Vol. I, p. 404, Pergamon Press, Oxford 1960.

| 27 | 28 | 29 | 30 | 31 | 32 | 36 | 37 | 38 | 39 | 40 | 41 | 43 | 44 | Mass / Gas |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| — | — | — | — | — | — | — | — | — | — | — | — | — | — | $H_2$ |
| — | — | — | — | — | — | 0.38 | — | 0.06 | — | 100 | — | — | — | Ar |
| — | — | — | — | — | 0.13 | — | — | — | — | — | — | — | — | $H_2O$ |
| — | 100 | 0.75 | — | — | — | — | — | — | — | — | — | — | — | $N_2$ |
| — | 100 | 0.88 | 0.2 | — | 0.02 | — | — | — | — | — | — | — | — | CO |
| — | 11.5 | 0.1 | — | — | 0.4 | — | — | — | — | — | — | — | 100 | $CO_2$ |
| — | — | 1 - 5 | — | — | — | — | — | — | — | — | — | — | — | $CH_4$ |
| .2 | — | — | — | — | — | — | — | — | — | — | — | — | — | $C_2H_2$ |
| 1.5 | 100 | 3.3 | — | — | — | — | — | — | — | — | — | — | — | $C_2H_4$ |
| 7.6 | 100 | 20.5 | 25.9 | 0.54 | — | — | — | — | — | — | — | — | — | $C_2H_6$ |
| 9.1 | 60.3 | 100 | 2.1 | — | — | 0.64 | 4.1 | 5.8 | 20 | — | 16 | 30.8 | 44.9 | $C_3H_8$ |

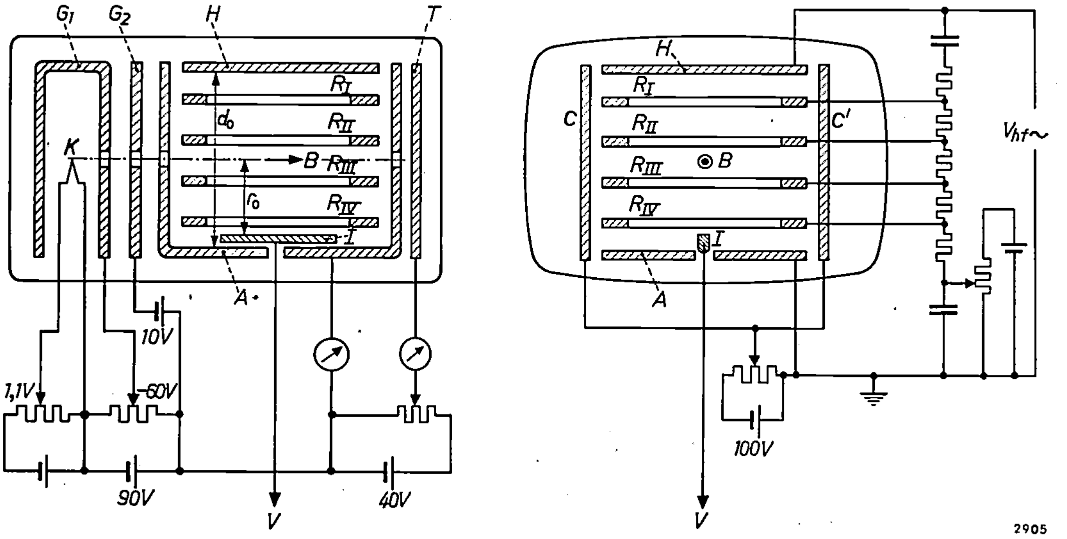


Fig. 2. Schematic representation of omegatron with rings $R_I$-$R_{IV}$ and side plates $C$-$C'$. The left part of the figure shows the section parallel to the magnetic field, the right half a section perpendicular to the magnetic field (induction $B$). $K$-$G_1$-$G_2$ electron gun. $T$ electron collector. $I$ ion collector. The RF voltage is applied between the electrodes $A$ and $H$. Other potentials are applied to the electrodes as indicated in the figure. The amplifier is connected at $V$.

is applied between the electrodes $A$ and $H$. The flat parallel rings $R$ are all held at a small positive electrostatic voltage; this serves to prevent the escape of ions in the direction of the magnetic field. In addition, RF voltage is applied to the rings $R$, a different amplitude for each ring (obtained by means of a voltage divider); this measure helps to establish a uniform RF field.

The right half of fig. 2 shows the cross-section perpendicular to the magnetic field. The main difference compared with the original form is the addition of two side plates $C$ and $C'$. By giving these plates a negative potential with respect to the electrodes $A$ and $H$, a suitable electrostatic field is created in the omegatron, that helps to make the sensitivity of the tube reproducible. We shall deal with this in more detail presently. The magnitude of this potential, required to obtain a maximum ion current on the ion collector, depends on the positive static potential $V_R$ on the rings $R$, as well as on their number. It is found that, when four rings are used, a value of approximately 1 : 100 is a suitable choice for the ratio $V_R : V_{CC'}$.

The arrangement is made clearer by the sketch of the cut-away electrode system shown in *fig. 3*. The photograph in *fig. 4* gives an idea of the dimensions of the omegatron in its glass envelope. The electrodes form a cube having edges 25 mm in length. Using no ceramic or mica, they are mounted on a glass foot and surrounded by a glass envelope. The whole tube can be degassed in an oven at 400 °C. The electrodes separately may be degassed at 900 °C by high-frequency heating. This high temperature sometimes proves to be necessary in order to remove disturbing surface layers on the electrodes. If the omegatron is not to be baked out in this way, the choice of electrode material is not critical, provided the electrodes have been thoroughly cleaned beforehand. A heat treatment is always desirable, however, in order to minimize the release of gases from glass and metal surfaces. To prevent the formation of oxide layers during the degassing process, considerable use is made of noble metals. The best results

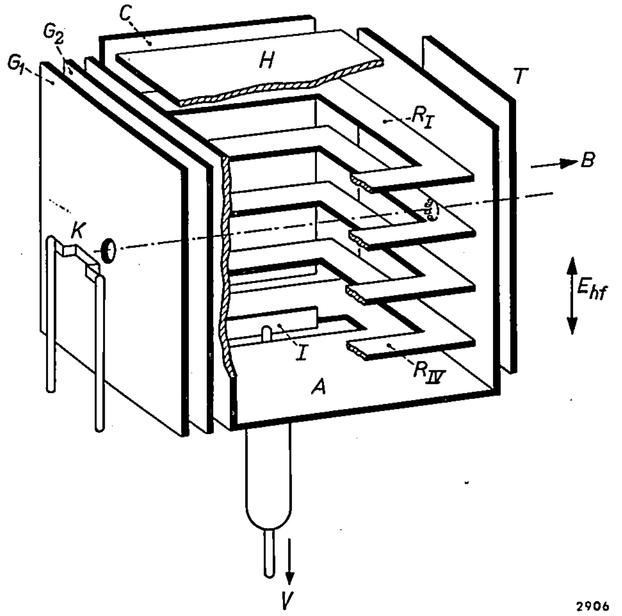


Fig. 3. Cut-away view of omegatron, showing rings and side plates (still somewhat schematic). Notation as in fig. 2.

were obtained with an alloy of platinum and iridium.

Little need be said about the auxiliary apparatus used with the omegatron. A permanent magnet or an electromagnet is used for producing the magnetic field. The permissible non-uniformity of the magnetic field over the omegatron is 2%. Any commercial signal generator may be used as the voltage source for the RF field, provided the output amounts to a few volts and is independent of frequency.

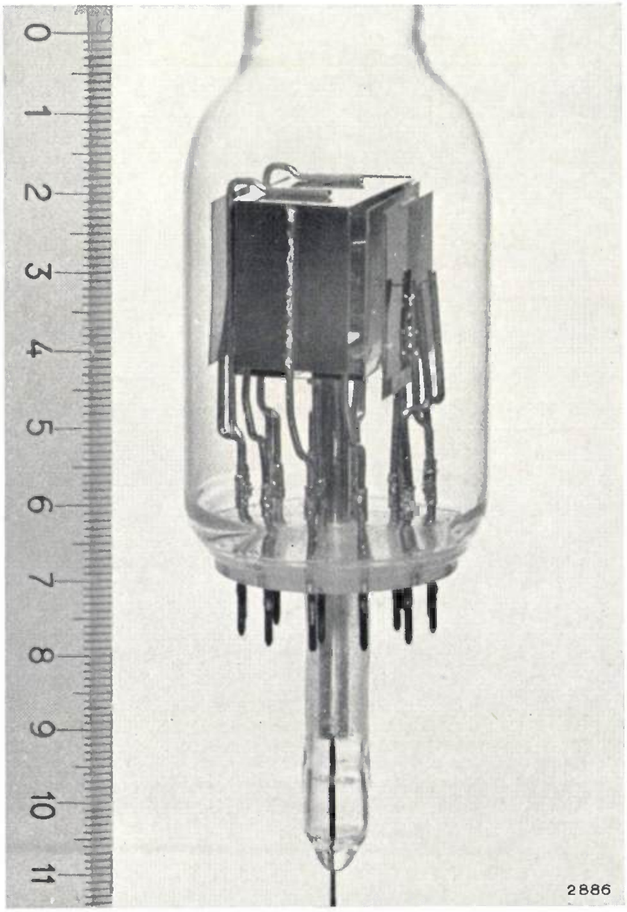


Fig. 4. The omegatron, with rings and side plates, in a glass envelope. The cathode is on the right, and on the extreme left is the collector for the electrons which have traversed the omegatron. The ions formed in the space inside the omegatron move to the ion collector. This is connected to the central metal pin fused into the glass tube.

The lowest measurable pressure of a given type of gas is determined by the lowest measurable direct current. A vibrating-reed electrometer is therefore particularly suitable for this measurement. Since the smallest currents which this can measure are of the order of $10^{-16}$ ampere, the connection between the ion collector and the electrometer input should be well screened.

Stabilization of the electron current proved to be essential. For this purpose a fixed potential is applied to the electrode $G_1$ and a resistance incorporated in the cathode lead. The electron current is then stabilized, via the space charge, by the voltage drop across the cathode resistor.

The mass spectrogram can be recorded on an oscilloscope or with a recording instrument as earlier described [3]).

## Operation of the omegatron with side plates

Consider the imaginary plane through the middle of the omegatron perpendicular to the magnetic field. The electrostatic equipotential lines in that plane are represented in *fig.* 5, where the rings are omitted for the sake of simplicity. The magnetic field is perpendicular to the plane of the drawing. The electrodes $C$ and $C'$ have a negative static potential of a few tens of volts with respect to $A$ and $H$. At this stage we shall disregard the effects of the RF field and the space charge in the omegatron. Electrolytic-tank measurements [5]) of the electrical field in the omegatron showed that at all points in the plane under consideration there exists a small negative potential, between a few tenths and one volt, in relation to the electrodes $A$, $H$ and $I$, except in the edge regions around the rings $R$. Depending on where they are formed, positive ions now pass along the equipotential lines in cycloidal paths outwards. The directions are indicated in the figure. To avoid ions escaping in the direction pointing away from the ion collector, the position $P$ where the ions are formed should lie on the ion-collector side of the electrical centre-point $M$ as shown in fig. 5. The

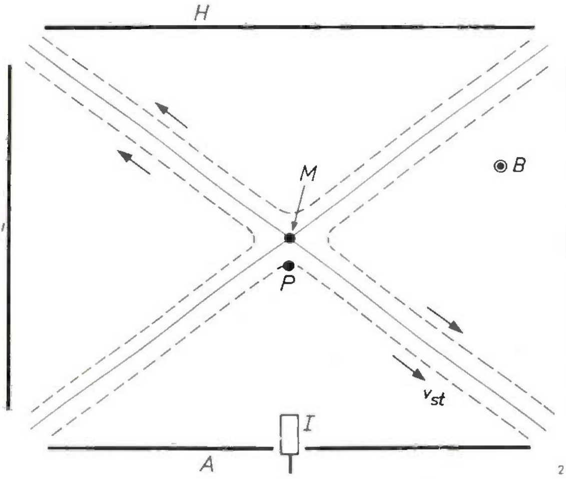


Fig. 5. Cross-section of omegatron perpendicular to the direction of the magnetic field. $M$ is the electrical centre, $P$ the position where the ions are formed. $I$ ion collector. $A$ and $H$ electrodes to which the RF voltage is applied. $C$ and $C'$ side plates, which are given a negative potential with respect to $A$ and $H$. The dashed lines represent equipotential lines.

[5]) These measurements were done by J. L. Verster and L. G. J. ter Haar at Eindhoven.

mere presence of the ion collector electrode $I$ contributes to the required asymmetry. The precise position required of $P$ with respect to $M$ (as in fig. 5) is obtained by careful choice of the position of the electrode $H$.

When a high-frequency alternating electric field is now applied, a spiral path is superposed on the cycloidal path already described by the ion. The resultant motion can be regarded as one in which the ions describe spiral paths whose centre moves continually outwards at a velocity $v_{\mathrm{st}}$. The result will be that the ions which are not in resonance are drawn outwards. This will have the effect of reducing the space charge. The ions which *are* in resonance describe a path that will make for an optimum ion capture at the ion collector. These highly intricate ionic movements lead to complicated considerations on the influence of the various parameters to be chosen. We shall try to deal with these considerations quite briefly.

To achieve optimum ion collection we must stipulate certain conditions regarding the magnitude of the electrostatic and RF fields. The velocity $v_{\mathrm{st}}$ is proportional to the electrostatic field-strength $E_{\mathrm{st}}$ and inversely proportional[6]) to the induction $B$:

$$v_{\mathrm{st}} \propto \frac{E_{\mathrm{st}}}{B}. \quad \ldots \ldots \quad (5)$$

The value $E_{\mathrm{st}}$ should be chosen large enough to ensure that the potential is not seriously affected by the space charge. An upper limit is set to $E_{\mathrm{st}}$, however, by the amplitude of the RF field. This may be understood as follows. The rate at which the radius of the spiral paths of the resonating ions increases is proportional to the RF field [1]):

$$v_{\mathrm{hf}} \propto \frac{E_{\mathrm{hf}}}{B}. \quad \ldots \ldots \quad (6)$$

If $E_{\mathrm{st}}$ is taken too large, and the RF field $E_{\mathrm{hf}}$ is too small, an ion in resonance is unable, in the time available, to acquire sufficient energy to reach the ion collector. If we raise $E_{\mathrm{hf}}$ we admittedly increase the energy which the ion can take up per revolution, but we adversely affect the resolution $(m/\Delta m)$. Consequently the region within which $E_{\mathrm{st}}$ may be chosen is somewhat limited.

Whilst the upper limit of the amplitude of the RF field $E_{\mathrm{hf}}$ is determined by the resolution, the lower limit of $E_{\mathrm{hf}}$ depends on the ion acquiring in the available time sufficient energy to reach the ion collector at the minimum value of $E_{\mathrm{st}}$. Otherwise the ion will not reach the collector and will be lost by collision with other electrodes.

Finally, it should be noted that $v_{\mathrm{st}}$ and $v_{\mathrm{hf}}$ are approximately independent of the mass of the ions. The values selected for maximizing the ion current at a given gas pressure are therefore valid for the whole mass spectrum of arbitrary gas mixtures.

When the ion capture is optimum, $a$ in equation (4) may be put equal to 1. The fact that this is permissible is to be seen from *Table II*. Here the values $\sigma$ for the probability of ionization, as reported in the literature, are compared with the values of $\sigma$ that may be calculated from our experiments on the assumption that $a = 1$. We proceed as follows. The ion current is measured at a known pressure and a given electron current. With the aid of equation (4) the value of $\sigma$ can then be determined, provided the effective length $s$ of the electron beam and the total length $L$ of the orbit of the resonating ions are known. The length $s$ is a constant and equal to 1.1 cm in the omegatron described. The length $L$ is a function of the pressure $p$, the induction $B$, the RF field $E_{\mathrm{hf}}$, the mass $m$ of the ion in resonance and the dimensions of the omegatron. The value of $L$ so derived is inserted in equation (4), yielding:

Table II. Values of the ionization probability $\sigma$ for various kinds of gas determined with the omegatron compared with values given in the literature. The unit is the number of ion pairs per electron and per centimetre path length, at a pressure of 1 mm Hg.

| Gas | Mass | $\sigma_{\mathrm{exp}}$ | $\sigma_{\mathrm{lit}}$ |
|---|---|---|---|
| $H_2$ | 2 | 3.6 | 3.6 |
| He | 4 | 1.63 | 1.36 |
| $N_2$ | 28 | 10 | 10 |
| CO | 28 | 10.7 | 10 |
| Ar | 40 | 11.8 | 12 |

$$i^{+} = a\, i^{-} p\, s\, \sigma \exp\left[-\frac{2.7\times10^{-5} r_0^{2} d_0}{\lambda_0 p_0}\, \frac{B^{2} p}{V_{\mathrm{hf}} m}\right], \quad (7)$$

where $\lambda_0 p_0$ is the pressure-independent product of the mean free path and pressure [7])[8]), $p$ is the pressure of the only type of gas assumed to be present in the omegatron, and $V_{\mathrm{hf}}$ is the r.m.s. RF voltage; the meaning of $r_0$ and $d_0$ is explained in fig. 2.

The values of $\sigma$ found from (7) correspond to the values reported in the literature if $a = 1$. In other words, all resonating ions that do not collide with

[6]) L. Spitzer jr., Physics of fully ionized gases, Interscience Publishers, New York 1956.

[7]) R. Jaeckel, Kleinste Drucke, ihre Messung und Erzeugung, Springer, Berlin 1950.

[8]) S. Dushman, Scientific foundations of vacuum technique, Wiley, New York 1949.

gas molecules are captured by the collector. *Fig. 6* shows the result for helium. From Table II it may be seen that the correspondence is even better for other gases.

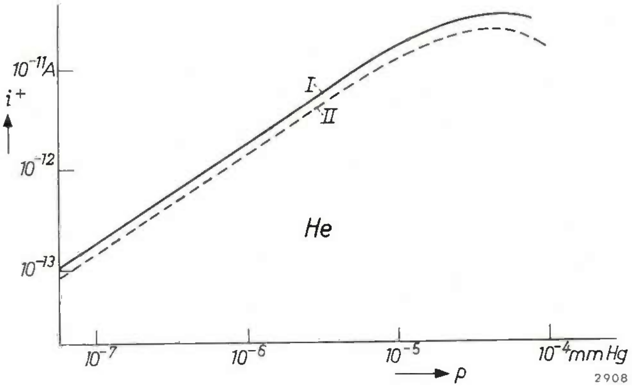


Fig. 6. The ion current $i^+$ as a function of pressure $p$, the omegatron containing helium only. Curve *I* gives the experimental values, curve *II* the theoretical values.

### Choice of parameters

Having shown in the foregoing that the omegatron can be used for quantitative analyses, we shall now discuss in more detail the choice of various parameters which affect the properties of the omegatron.

1) The *electrostatic field-strength* in the omegatron must be adjusted by the small positive potential $V_R$ and the negative potential $V_{CC'}$ on plates $C$ and $C'$ in such a way that the current of the ions in resonance will be maximum at a given pressure. The setting of these values is critical, since an ion current may flow even when no RF field has been applied. This residual current, as it is called, is proportional to the total pressure, i.e. the sum of the pressures of all gas components. It is attributable to the strong space charge that may be produced when there are low negative potentials on the side plates $C$ and $C'$. The potential at the point of origin of the ions can then be equal to the potential of the ion collector. The residual currents measured in such a case are represented by the dashed curves in *figs.* 7 and *8*. The solid curves represent the relation between the ion current and $V_{CC'}$ (fig. 7) and $V_R$ (fig. 8), when the amplitude of the RF voltage is 1 $V_{\text{rms}}$ and the magnetic induction 0.5 Wb/m². For the values of $V_R$ or $V_{CC'}$ where a residual current arises in the absence of an RF field, ions that are not in resonance will still reach the collector even when this field is present. Consequently it is possible that the ratio $i^+/i^+_{\text{max}}$ in the solid curves of fig. 8 will exceed the value of 1. To avoid interference from residual currents, it is best to choose a small positive voltage $V_R$ and large negative voltages $V_{CC'}$ (see the shaded regions in figs. 7 and 8). Owing to absorbed surface layers on the electrodes, the optimum values may vary from one tube to another.

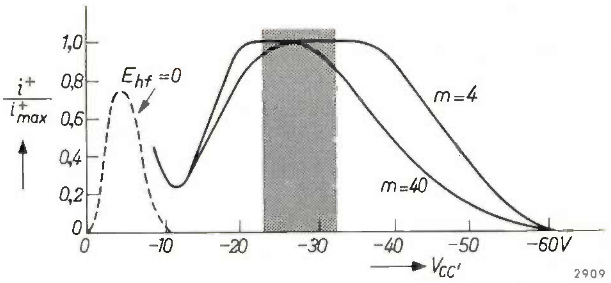


Fig. 7. Illustrating the effect which the negative potential $V_{CC'}$ (on the side plates $C$ and $C'$) has on the ion current $i^+$. $V_R = 0.15$ V; $V_{\text{hf}} = 1$ $V_{\text{rms}}$; $i^- = 1$ μA; $B = 0.5$ Wb/m².

The solid curves represent the relative ion current for the mass numbers 4 and 40. The dashed curve represents the ion current reaching the collector in the absence of an RF field (residual current). The shaded area is the operating region.

2) According to equations (4) and (7) the ion current at a given pressure is independent of the *RF voltage* $V_{\text{hf}}$ when $\lambda \gg L$. If this condition is no longer satisfied, the ion current decreases, and it does so more rapidly for light than for heavy particles. In the omegatron described, the ion current decreases faster than might be deduced from equation (7). This is due to the fact that, with electrostatic fields present in the omegatron, the resonance frequencies are a function of position. There is then a minimum value of the RF voltage at which the ions in resonance are still only just capable of reaching the ion collector [9]). Since the ion loss when the RF voltage is too low differs for each kind of ion, it is always necessary to operate the omegatron in the "saturation region"; in other words, the ion current must be independent of the RF voltage. Only then will the

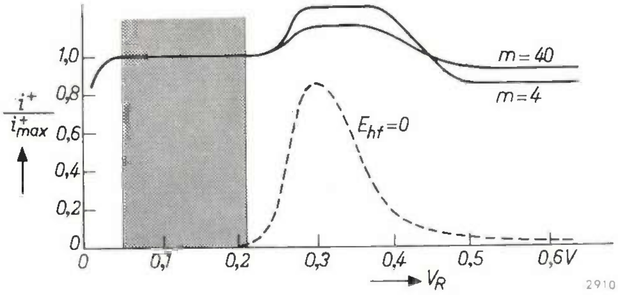


Fig. 8. Curves showing how the ion current $i^+$ depends on the positive potential $V_R$ on the rings $R$. $V_{CC'} = -25$ V; $V_{\text{hf}} = 1$ $V_{\text{rms}}$; $i^- = 1$ μA; $B = 0.5$ Wb/m². The dashed curve represents the ion current reaching the collector in the absence of an RF field. The shaded area is the operating region.

[9]) W. M. Brubaker and G. D. Perkins, Rev. sci. Instr. **27**, 720, 1956.

pressures of the various components of the gas mixture be measured in their correct relationships. Generally the RF voltage will be roughly 1 $V_{\text{rms}}$.

3) The *voltage $V_T$ of the electron collector* has little influence on the ion current provided this electrode is clean.

4) According to equation (4) the ion current at a given pressure is a linear function of the *electron current*. If the electron current is too high, the large numbers of ions formed will give rise to a strong space charge, which affects the electrostatic field. The ion current will then no longer be a linear function of the electron current. *Table III* shows for various pressure ranges the maximum permissible electron currents at which, irrespective of the mass, the relation between ion current and electron current is still linear.

Table III. The permissible electron current $i^-$ for various levels of the total pressure $p$. The higher the total pressure, the smaller the permissible electron current.

| $p$<br>mm Hg | $i^-$<br>μA |
|---|---|
| $\leq 10^{-8}$ | 30 |
| $\leq 10^{-7}$ | 10 |
| $\leq 10^{-5}$ | 1 |

### Pressure range, resolution and accuracy of measurement

From (7) the maximum permissible pressure is calculated to be $10^{-5}$ mm Hg where the induction $B$ is 0.5 Wb/m² and the amplitude of the RF voltage $V_{\text{hf}}$ corresponds to 1 $V_{\text{rms}}$. At this pressure the mean free path $\lambda$ is still $\gg L$, irrespective of the gas composition. At lower pressures than $10^{-5}$ mm Hg the relation between ion current and electron current is then linear provided the electron current is not greater than 1 μA (see Table III).

The lowest pressure $p_{\text{min}}$ measurable with the omegatron is determined by the smallest ion current which the electrometer is capable of measuring at the maximum permissible electron current:

$$p_{\text{min}} = \frac{i^+_{\text{min}}}{s\sigma i^-_{\text{max}}}. \quad . \quad . \quad . \quad . \quad . \quad (8)$$

Inserting $i^+_{\text{min}} = 10^{-16}$ A, $i^-_{\text{max}} = 3\times10^{-5}$ A, $s = 1.1$ cm and $\sigma = 10$, we find $p_{\text{min}} = 3\times10^{-13}$ mm Hg. The lowest pressure measured in our laboratory was approximately $p_{\text{min}} = 1\times10^{-12}$ mm Hg. As mentioned above, some gases can only be identified from their ion spectrum. If relative intensities of 1 : 100 are found between the ions (or agglomerates) of the various mass numbers, the lowest pressure that can be measured for any one type of gas is $p_{\text{min}} \approx 3\times10^{-11}$ mm Hg.

The resolution [1]) of the ideal omegatron, i.e. one in which the electrostatic field and the thermal energy of the ions in the direction of the magnetic field are zero, is given by the expression:

$$\frac{m}{\Delta m} = \frac{eS_0B^2}{2\hat{E}_{\text{hf}}m}. \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

This formula is arrived at on the assumption that, in the spectrogram, the peaks for the masses $m$ and $m \pm \Delta m$ are separated right down to the base. In the present instance, $S_0$ is the distance from the ion collector to the point on the equipotential line where the centre of the spiral path lies after the total time of flight of the resonant ion. In the ideal omegatron, $S_0 = r_0$. For an induction $B = 0.5$ Wb/m² and an RF field corresponding to $V_{\text{hf}} = 1$ $V_{\text{rms}}$, a resolution can usually be achieved such that at a mass number of 30 one mass number can still be resolved. In the ideal omegatron, according to equation (9), this would be possible up to the mass number 39. It is possible to increase the resolution by increasing the magnetic field. Reducing the RF voltage increases the resolution only very slightly, since this voltage may only be varied within the saturation region.

The sensitivity of the tube described remains constant up to 10% provided the electrode surfaces are clean. If the geometry of the various tubes is identical, so is their sensitivity. About 40 of these omegatrons were tested in our laboratory on their reproducibility. The presence of gases and vapours such as $H_2O$, $CH_4$, $C_2H_6$ and $CO_2$ up to maximum pressures of $10^{-5}$ mm Hg caused no change in the sensitivity, even after several weeks on test. However, the presence of Hg and also HCl on the electrodes causes marked changes in sensitivity. The accuracy with which the pressure of a pure gas can be determined amounts to 10%. In the case of a gas mixture, the error may be much greater, particularly if different components of the gas mixture contribute to the same mass peak (see Table I). All gases that contribute less than 10% to the ion current at this mass remain undetected.

The omegatron of the type described has been used at Philips for a variety of investigations, including the analysis of residual gases in vacuum equipment and in electron tubes, the investigation of reactions between getter and gas, and the determination of the release of gases from various kinds of glass.

*Fig. 9* shows a simple example of a mass spectrogram obtained in an investigation of the residual gases in a well-outgassed glass system containing a barium getter film.

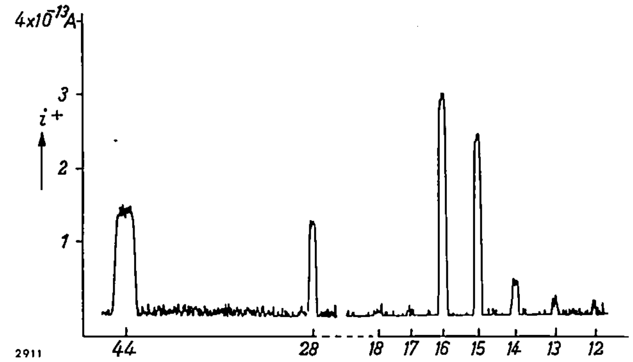


Fig. 9. Mass spectrogram of the residual gas in a vacuum system in which a barium getter has been flashed. The composition of the gas (expressed in partial pressures) is as follows.

| | | | |
|---|---|---|---|
| $CH_4$: | $2.8\times10^{-9}$ mm Hg | $N_2$ : | $2\times10^{-10}$ mm Hg |
| $CO_2$: | $1.0\times10^{-9}$ mm Hg | $C_2H_6$: | $2\times10^{-10}$ mm Hg |
| CO : | $8\times10^{-10}$ mm Hg | $H_2O$: | $1\times10^{-10}$ mm Hg |

**Summary.** In the manufacture of electron tubes and other industrial products, and also for high-vacuum work in laboratories, it is important to know not only the total pressure but also the composition of the residual gas. For this purpose, use is often made of the omegatron. In its original form, this instrument provided only qualitative results. This article describes a type of omegatron which can also be used for quantitative analyses. To make this possible a pair of side plates have been added to the original instrument. When a suitable potential is applied to these plates, the ion current can be brought to an optimum value at any mass which is in resonance.

For a magnetic induction $B$ of 0.5 Wb/m$^2$ and an RF voltage of 1 $V_{rms}$, the maximum permissible pressure is $10^{-5}$ mm Hg. The lowest measurable pressure of any one kind of gas is $p_{min} \approx 1\times10^{-12}$ mm Hg, when a DC amplifier is available capable of detecting a current of $10^{-16}$ A. The resolution is such that masses 30 and 31 can be distinguished. The accuracy of measurement is 10%.

# METHODS OF PRODUCING STABLE TRANSISTORS

by J. J. A. PLOOS van AMSTEL.

621.794:621.382.3

*One of the major problems in the manufacture of transistors is to make them stable, i.e. to produce transistors whose characteristics do not change in the course of time. Such changes are due to surface effects. It is therefore necessary to produce surface conditions that will minimize these changes.*

It was originally hoped that the characteristics of junction transistors, by their very nature, would change very little and that, if properly used, the transistors would have an almost unlimited life. Theoretical considerations show that the operation of these transistors is governed by the dimensions and the material properties of the constituent layers, i.e. of the emitter, base and collector [1]), and there is no reason to assume that these dimensions and properties alter in normal operation. At places where the *P-N* junctions reach the surface (see *fig. 1*) there must of course be no surface layer that might give rise to an undesired conductive path between the *P* and *N* regions. In the fabrication of transistors, cleaning the surface by etching is therefore one of the routine operations.

The optimistic expectations regarding the stability of transistors have not been borne out, however. It very soon appeared that, in spite of the etching, the state of the germanium surface has a very pronounced influence on the properties of the transistor. The electrical changes observed are very often due to changes in this surface — as, for example, the adsorption of foreign molecules and atoms — and these in their turn are bound up with the state of the ambient atmosphere. The obvious way to seek improvement was therefore to hermetically encapsulate the transistors. This, however, proved to be insufficient: life tests showed that the parameters — in particular the current amplification factor — continued to change. The production of stable transistors was more difficult than was at first thought.

In many laboratories the surface of germanium, and semiconductor surface phenomena in general, have consequently been the subject of much experimental and theoretical research [2]). One of the aims pursued was to find methods of achieving surface conditions that would result in good transistors and moreover ensure a high degree of stability. This article will describe some results of investigations

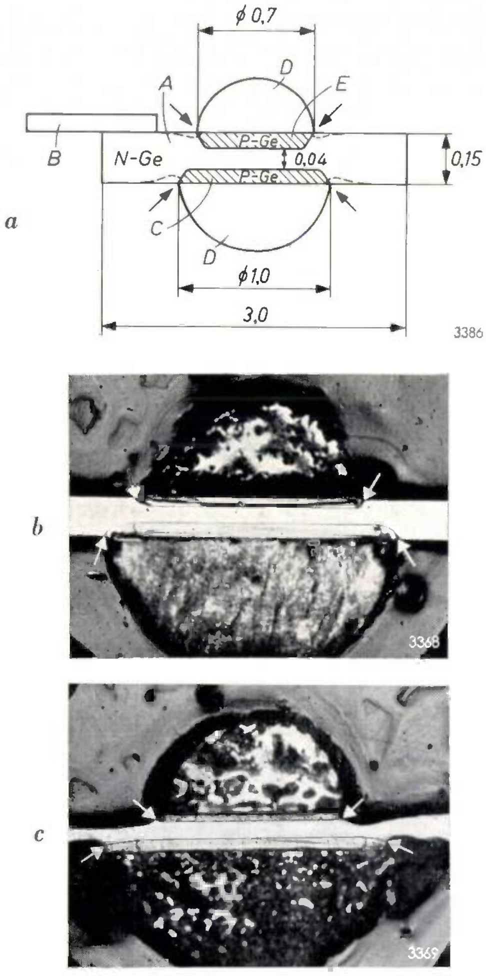


Fig. 1. *a*) Schematic cross-section of a *P-N-P* junction transistor, made by the alloying method. Dimensions in millimetres. *A* germanium single crystal of type *N* in the form of a wafer (3 × 2 × 0.15 mm). *B* base contact. *C* collector; this part of the germanium crystal is given *P*-type conductivity by alloying with the acceptor material *D*. *E* emitter, also of *P*-type germanium due to alloying with *D*. The arrows indicate where the *P-N* junctions reach the surface. After etching, the surface follows the dashed lines.
*b*) Polished cross-section, on which the various regions are made visible by etching. The arrows again indicate the *P-N* junctions at the surface. The transistor is here surrounded by a layer of shellac in connection with the preparation of the sample.
*c*) As *b*), but with etched surface.

[1]) See F. H. Stieltjes and L. J. Tummers, Simple theory of the junction transistor, Philips tech. Rev. **17**, 233-246, 1955/56, and F. H. Stieltjes and L. J. Tummers, Behaviour of the transistor at high current densities, Philips tech. Rev. **18**, 61-68, 1956/57.

[2]) R. H. Kingston, Review of germanium surface phenomena, J. appl. Phys. **27**, 101-114, 1956.

undertaken along these lines at Philips Research Laboratories at Eindhoven. The investigations concerned *N-P-N* and *P-N-P* germanium transistors for low frequencies and low power, made by the alloying method. A schematic cross-section of this type of transistor *(P-N-P)* is shown in fig. 1, together with the relevant dimensions.

The marked influence which the surface has on the operation of a transistor is accounted for by assuming that holes and electrons recombine at the surface to an extent that depends strongly on the state of the surface. To explain this effect, let us consider a *P-N-P* transistor. Here the emitter injects holes into the base, whilst the collector acts as a sink for holes from the base. If no recombination occurs in the base — in a stationary or quasi-stationary state — all injected holes will disappear to the collector, where they contribute to the collector current. That is the ideal situation. In fact, however, holes disappear in the base by recombining with electrons, and are thus lost to the collector current. This "base loss" is one of the two reasons why the ratio of the collector current to the emitter current — called the current amplification factor $\alpha$ — is less than unity [3]). It has been found that the base loss in transistors of the alloy type may largely be due to recombination at the surface. Changes in the velocity of recombination at the surface therefore have a considerable effect on the base loss and hence on the behaviour of the transistor.

We shall look a little closer at the reason for this effect. If we speak of a high surface recombination velocity, we mean that a hole in the neighbourhood of the surface is very likely to be lost by recombination with an electron. Whether there is in fact a high degree of recombination near the surface depends of course on whether there are holes there to be affected. To show that this will certainly be so, we may consider the case where no recombination at all takes place at the surface. *Fig. 2a* roughly illustrates the flow pattern of the hole current in that case, together with the lines of constant hole concentration. The two systems of lines are orthogonal, and the lines of constant concentration must moreover terminate perpendicular to the surface. The hole current must therefore run partly along the surface, so there will indeed be holes there, which may be destroyed by recombination at the surface. The recombination calls for a correction to the pattern in fig. 2*a*. The corrected pattern will look something like that in fig. 2*b*.

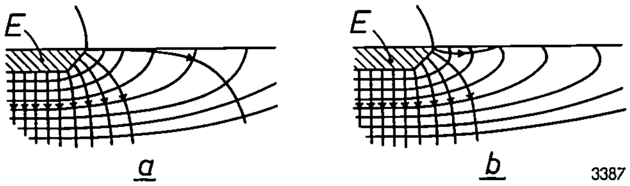


Fig. 2. Illustrating contours of constant hole concentration and the flow lines of the hole current (with arrows) in the base of a *P-N-P* junction transistor at the edge of the emitter *E*. The two sets of lines are orthogonal.
*a*) The pattern on the assumption that there is no recombination at the surface. The lines of constant hole concentration must then terminate at right-angles to the surface. The hole current passes along the surface; if there is any opportunity for surface recombination, holes will certainly be lost.
*b*) The pattern after correction for surface recombination. (Recombination inside the base is disregarded in both cases.)

To make a stable transistor it would be a great help if the surface could be treated in such a way as to make the surface recombination velocity insensitive to extraneous influences. Although work is being done in that direction, the results have so far been unsatisfactory. In this article we shall be solely concerned with the method whereby an attempt is made to reach the same objective by providing the transistor with a suitable *ambient atmosphere*. Obviously, this is possible only if the transistor is enclosed in a hermetically sealed container. We shall see presently why this air-tight enclosure is not in itself enough. There are certain substances that not only reduce the surface recombination velocity to an acceptable value but also, provided they are properly applied inside the enclosure, keep it constant over a long period of operation. Two outstanding representatives of these substances will be discussed here, namely *water* and *arsenic*, particular attention being given to the influence of water, i.e. water vapour.

A sensitive indication of the base loss, and hence of the surface recombination velocity, is the ratio of the collector current to the base current. We denote this "current amplification factor" by $\alpha'$, to distinguish it from the earlier mentioned $\alpha$, which is the ratio between collector and emitter currents. Between $\alpha$ and $\alpha'$ there exists the well-known relation

$$\alpha' = \frac{\alpha}{1-\alpha},$$

which follows directly from the fact that the sum of the emitter, base and collector currents is zero (provided these currents are counted positive when directed towards the transistor). Since $\alpha$ is not much smaller than unity — at least in a serviceable transistor — $\alpha'$ undergoes greater changes than $\alpha$. For the purpose of judging the surface effects of the various measures adopted, $\alpha'$ is therefore always measured.

## The effect of water

It is known that the surface recombination velocity depends on the surface occupation by water

[3]) The second reason is the emitter loss; see p. 239 *et seq.* of the first article quoted in reference [1]).

molecules[4]). When the surface is completely dry, the surface recombination velocity is high, and $\alpha'$ is correspondingly low. As the water "occupation" increases, the surface recombination velocity decreases and $\alpha'$ rises. The changes of $\alpha'$ observed on transistors are therefore certainly to a considerable extent attributable to changes in the water occupation of the surface. Even in hermetically encapsulated transistors, the surface water will almost certainly be affected by temperature variations, for example. The migration of the water may be so slow that one cannot always expect to find the same value of $\alpha'$ at the same temperature. Slow reactions involving water may also play a part.

The most obvious method of eliminating this undesired influence is to make sure that there is no water at all inside the transistor envelope. This is a method that is in fact used, but it has the drawback of resulting in a low value of $\alpha'$. It appears that the presence of only a minute trace of water is sufficient to make $\alpha'$ unstable.

In the methods discussed in this article an attempt is made to create conditions inside the transistor envelope such as to give the surface of the transistor a water occupation that will ensure a high value of $\alpha'$, and at the same time remain constant with time and temperature.

As we have seen, the surface recombination velocity decreases if the surface water increases. At constant temperature, a state of equilibrium will be reached inside the transistor envelope between the surface water and the water-vapour pressure: the greater the vapour pressure, the more densely will the surface be occupied by water molecules, and hence the greater will be the value of $\alpha'$. The water-vapour pressure must not, however, be unduly high, for if the surface is too wet, disturbances are caused by superficial ionic currents, one result of which may be the appearance of loops in the current-voltage characteristics of the $P$-$N$ junctions if measured by an AC method. It is known that a high surface water content reduces the sensitivity of the surface recombination velocity to fluctuations of that content[5]). Consequently, we may expect the curve of $\alpha'$ as a function of vapour pressure, the temperature being constant, to gradually flatten out (*fig. 3*). This tendency will be enhanced by the fact that the recombination near the surface gradually loses its importance in relation to the other factors governing $\alpha'$, namely the recombination inside the base and the emitter loss. In general, a favourable surface water occupation is found in air of room temperature with a relative humidity in the region of 60%, the value of $\alpha'$ then being high and little dependent on fluctuations of water-vapour pressure.

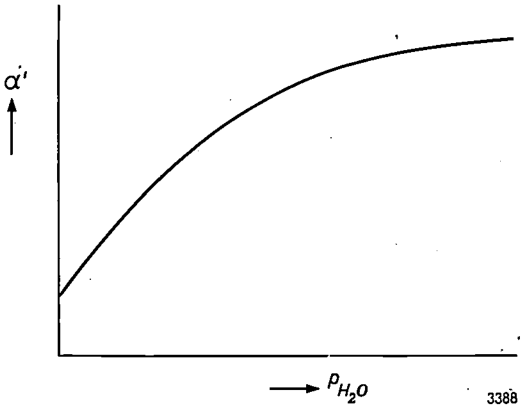


Fig. 3. Expected variation of current amplification factor $\alpha'$ with water-vapour pressure $p_{H_2O}$, assuming constant temperature and equilibrium between $p_{H_2O}$ and the surface occupation by water.

### *Stabilization with a water-vapour buffer*

It follows from the foregoing considerations that a stable transistor can be obtained by introducing a "stabilizer" inside the encapsulating envelope, i.e. a substance that fixes the water-vapour pressure at a favourable value and thus acts in that respect as a buffer. The stabilizer must ensure a favourable water-vapour pressure at all temperatures which the transistor is likely to reach in normal operation: if the temperature rises, the water-vapour pressure will have to increase in such a way that the existing surface water occupation is maintained. In a graph of temperature $t$ versus water-vapour pressure $p_{H_2O}$ there will be a region of favourable combinations of $t$ and $p_{H_2O}$ as shown by the hatching in *fig. 4*. The water-vapour pressure of the stabilizer as a function of temperature is required to have a curve that lies within this region in the whole temperature interval of interest for the transistor.

Since, in principle, the surface occupation by water is kept constant under all conditions, inertia effects due to changes in surface water no longer occur.

[4]) J. T. Wallmark and R. R. Johnson, Influence of hydration-dehydration of the germanium oxide layer on the characteristics of $P$-$N$-$P$ transistors, R. C. A. Review **18**, 512-524, 1957; also A. J. Wahl and J. J. Kleimack, Factors affecting reliability of alloy junction transistors, Proc. Inst. Radio Engrs. **44**, 494-502, 1956.

[5]) J. H. Forster and H. S. Veloric, Effect of variations in surface potential on junction characteristics, J. appl. Phys. **30**, 906-914, 1959; also J. R. A. Beale, D. E. Thomas and T. B. Watkins, A method of studying surface barrier height changes on transistors, Proc. Phys. Soc. **72**, 910-914, 1958, and G. Adam, Der Einfluss der Gasatmosphäre auf die Oberflächenrekombination bei Germanium, Z. Naturforschung **12a**, 574-582, 1957.

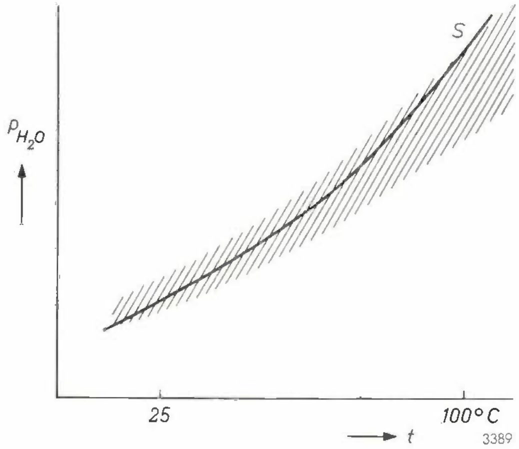


Fig. 4. Illustrating our picture of how the stabilization of transistors with the aid of a water-vapour buffer may occur. The temperature $t$ is plotted versus the water-vapour pressure $p_{\mathrm{H_2O}}$ inside the transistor envelope. The hatched region comprises combination of $t$ and $p_{\mathrm{H_2O}}$ that produce a favourable occupation of the transistor surface by water. The water-vapour pressure curve $S$ of the stabilizing buffer should lie fully within the hatched region, for the temperature interval of interest.

*Stabilization by "forming" the surface*

In the following discussion of experiments it will be shown that reasonably stable transistors can be made with the aid of a water-vapour buffer. After prolonged tests, e.g. after some thousands of operating hours, however, $\alpha'$ does usually begin to fall. This drawback can fortunately be overcome by slightly modifying the stabilizing method. A buffer is then used which, at room temperature, gives such a low water-vapour pressure that the surface water remains substantially below the region of favourable values ($I$ in *fig. 5*). A transistor provided with such a buffer therefore has a low $\alpha'$ until it has been subjected to a special treatment. The latter consists of baking the transistor at 140 °C for several days. During the baking process something of the nature of surface "forming" takes place [6]). Our hypothesis is that the region of favourable combinations of temperature and water-vapour pressure for the formed surface is now shifted so as to bring the $p_{\mathrm{H_2O}}$ curve for the buffer entirely into the favourable region (fig. 5).

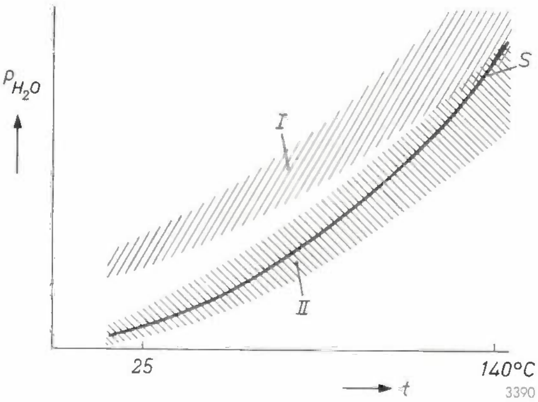


Fig. 5. Schematic representation of a working hypothesis concerning the stabilizing process where the surface of the transistor is "formed" with the aid of a buffer that gives a low water-vapour pressure. $t$ temperature. $p_{\mathrm{H_2O}}$ water-vapour pressure. Curve $S$ again represents the vapour pressure of the buffer. Region $I$ covers the combinations of pressure and temperature that correspond to a favourable surface occupation by water *before* forming. After the forming process (prolonged heating at 140 °C) the favourable region corresponds to $II$.

## Experiments

*Drying of transistors*

With the object of systematically investigating the influence of water vapour on transistors, a drying process was applied during which the change of $\alpha'$ was followed. The transistors were dried both in vacuum and in air. For the vacuum drying the transistors are sealed into a relatively large glass tube (*fig. 6*). The lead-in wires are fused to the glass so far

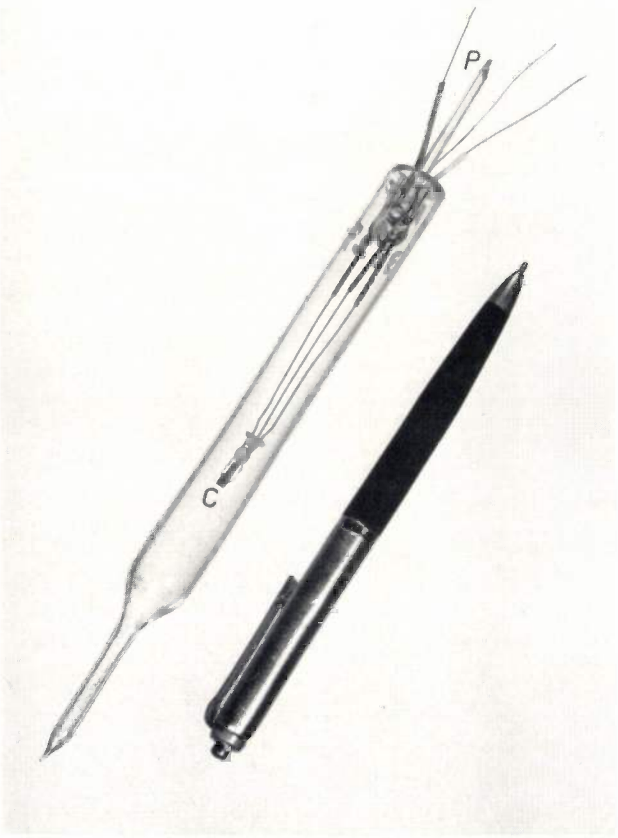


Fig. 6. In the water-vapour experiments the transistors were sealed inside glass tubes so designed that the temperature of the transistors is not significantly affected by the sealing operation. The transistor can be dried under vacuum. $P$ is the pump stem, $C$ the transistor.

[6]) J. T. Wallmark, Influence of surface oxidation on alpha c.b. of germanium *P-N-P* transistors, R.C.A. Review **18**, 255-271, 1957.

from the transistor that the transistor remains virtually unaffected. When the tube is now evacuated at room temperature, $\alpha'$ gradually falls, but so slowly that the final value is still not reached after days on the pump. Under these conditions the water bound to the germanium is released only very slowly. Evacuation at a higher temperature, e.g. at 100 °C, causes $\alpha'$ to drop faster, but here again, no final value is reached for several days. After pumping at 140 °C, however, $\alpha'$ generally drops in about six hours to a final value which is 10 to 15% of the initial value.

The rate at which $\alpha'$ decreases depends on the pretreatment of the transistors, particularly on the etching. For example, transistors that have been electrolytically etched in KOH show a much faster drop upon evacuation than transistors that have been etched in acid. However, the final value reached by $\alpha'$ after prolonged pumping at high temperature is always just about as low, whatever the method of etching adopted.

The subsequent admission of a dry atmosphere, e.g. dry air, dry oxygen, or dry nitrogen, has no or scarcely any effect on $\alpha'$ even after long storage. The admission of a humid atmosphere, however, sends $\alpha'$ up again. Sometimes it may rise very rapidly, often increasing in one second by a factor of 5 to 10, sometimes back to its original value. The rate at which $\alpha'$ recovers depends, like the rate at which it falls, on the pre-treatment, that is on the method of etching and on the baking temperature. It depends, too, on the humidity of the air admitted. A relative humidity of 60 % is found to be most effective. After the first steep rise, there is usually a slow increase to the final value.

The fact that the recovery of the transistor is attributable to the water vapour is confirmed by experiments where pure water vapour is admitted to transistors dried in a vacuum. They show that $\alpha'$ recovers in the same way as in humid air. The glass apparatus used for these experiments can be seen in *fig. 7*.

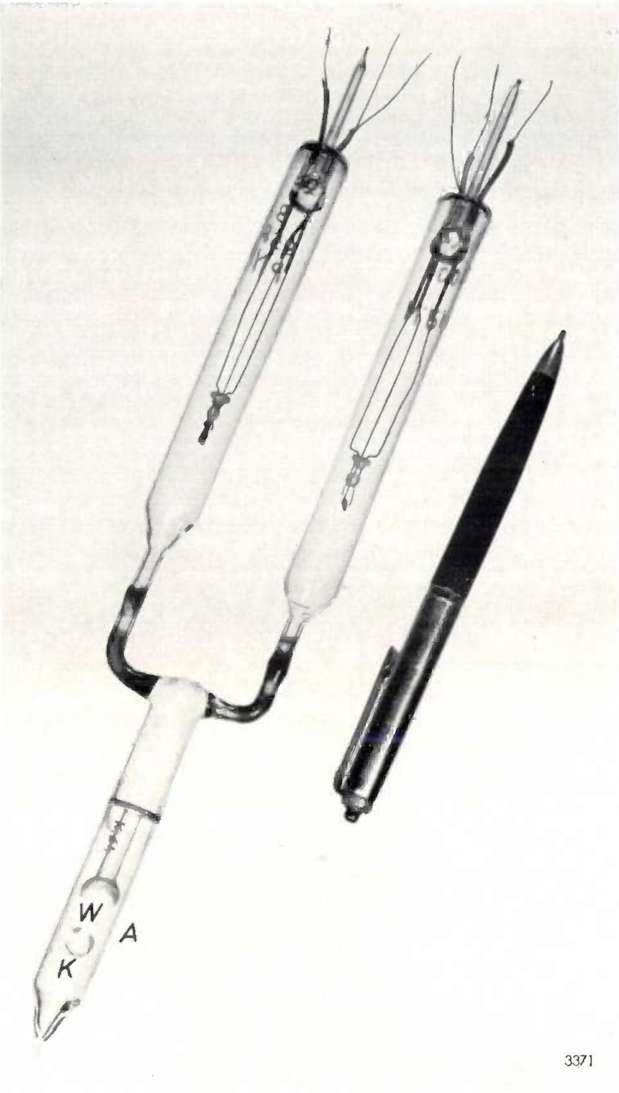


Fig. 7. Glass apparatus for admitting pure water vapour to vacuum-dried transistors. The marble *K* is used to break the spherical partition *W* which seals off the space *A* containing water. The water-vapour pressure is regulated by the temperature of space *A* (provided this temperature is lower than that of the space in which the transistor is situated). The apparatus here contains a *P-N-P* and an *N-P-N* transistor, for the purpose of comparing the behaviour of the two types.

In the drying experiments in air the transistors were heated in a small oven to which air had free access. The behaviour of transistors dried in air is broadly the same as that of transistors dried in vacuum: during the drying process, $\alpha'$ falls at a rate depending on how high the temperature is. Here again, $\alpha'$ falls faster (though not lower) in transistors electrolytically etched in KOH than in acid-etched transistors. Usually the decrease does not continue so far as when the transistors are dried in vacuum, where the drying process is more rigorous. When the dried transistors are again exposed to moist air, $\alpha'$ again recovers, but normally at a much slower rate than after drying in vacuum. In every case it appears that the state of the surface (i.e. the degree of oxidation) does not affect the rule that a certain surface occupation by water is necessary to obtain a high $\alpha'$, but the surface oxidation does affect certain details of the transistor's behaviour, as for example the rate with which the water makes its presence felt.

To check by other means whether the dehydration of the surface is in fact responsible for the decrease of $\alpha'$, transistors were heated at 140 °C, the water-vapour pressure being increased to maintain a certain surface water occupation. It was found that, under a water-vapour pressure of 300 to 400 mm Hg, the original values of $\alpha'$ remain virtually unchanged.

*Stabilization experiments with a water-vapour buffer*

Transistors are frequently sealed into small glass envelopes. It can be seen in *fig. 8* that the dimensions of such an envelope are still considerable compared with the size of the crystal. If this were not so, the crystal would be too near the seal and would be overheated during the sealing operation. Further protection is afforded the crystal by filling the envelope with an appropriate substance. Silicone grease, such as used for high-vacuum purposes, is a suitable and widely used filler.

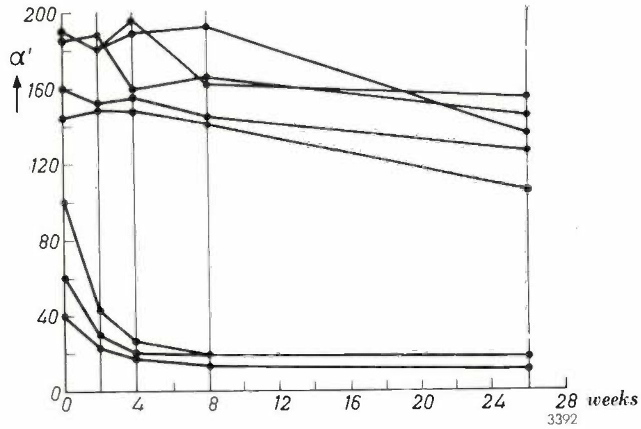


Fig. 9. Typical current amplification factor $\alpha'$ versus time after encapsulation in *dry* silicone grease (lower three curves) and in *moist* silicone grease (upper five curves). The ambient temperature during the experiments was 50 °C, and 50 mW was dissipated in the transistors. The temperature of the germanium crystal was 85 °C. The transistors used were *P-N-P* types, electrolytically etched in KOH. $\alpha'$ was measured at room temperature.

On page 206 it was argued that stable transistors may be expected if we introduce into the transistor enclosure a buffer which will keep the water-vapour pressure at a favourable value at any working temperature. A silicone grease which has absorbed some moisture, having for example been exposed for 24 hours to air of 30% relative humidity, performs this function reasonably well. *Fig. 9* shows that transistors with a filling of *dry* silicone grease exhibit low values of $\alpha'$ immediately after seal-off, which moreover drop appreciably in a few weeks. The same figure shows that, where a *moist* silicone grease is used, $\alpha'$ is much higher and, what is more, fairly constant.

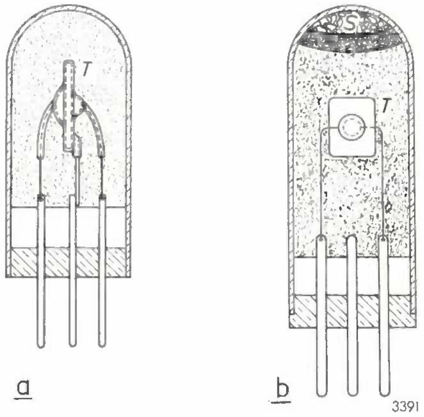
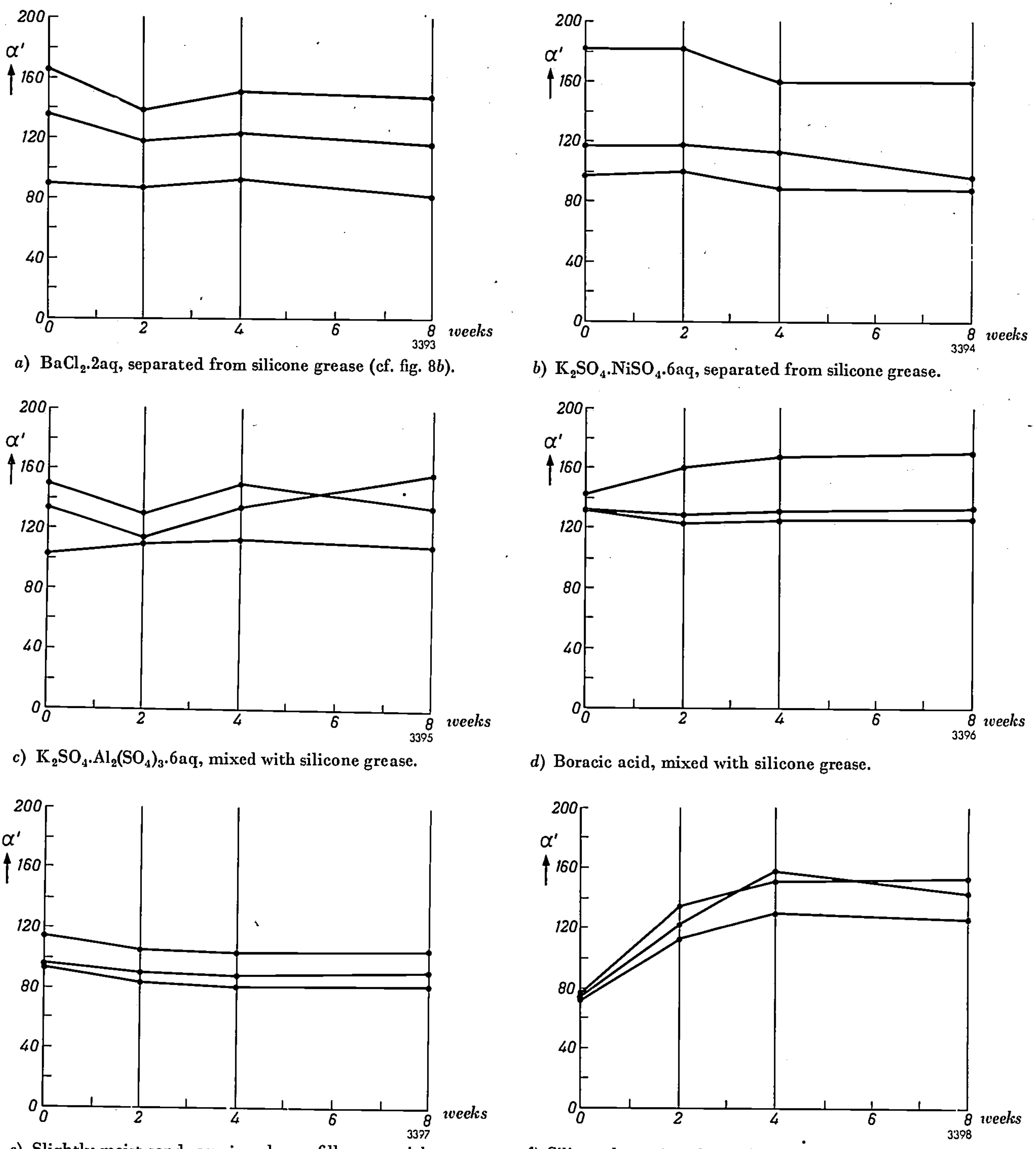


Fig. 8. *a*) Sketch of a transistor *T* in its envelope. The envelope is largely filled with a filler material (e.g. a silicone grease) to protect the germanium crystal during seal-off.
*b*) The envelope here contains a buffer or stabilizer *S*, separated from the silicone grease by a porous plug. At all temperatures the buffer produces a water-vapour pressure which keeps the water content of the germanium surface constantly favourable.

Numerous experiments were also done with transistors whose envelopes contained, in addition to the silicone grease, a substance separately introduced as a water-vapour buffer. The buffer was either kept distinct from the grease by means of a porous plug (fig. 8*b*) or it was mixed with the grease. *Figs. 10a, b, c* and *d* give some examples of the favourable effect produced by some of these buffers (mentioned in the captions). The results in figs. 10*e* and *f* relate to transistors which contained, instead of a silicone grease, slightly moistened sand or silica gel, both of which substances serve the dual purpose of filler and buffer. These two examples support the hypothesis that the behaviour of the transistor is governed mainly by the moisture inside the enclosure, and not by the silicone grease or the combination of water and silicone grease.

In fig. 10*f* it is noticeable that $\alpha'$ rises steeply during the first weeks, after which it remains roughly constant. We attribute this to the loss of water from the transistor surface when it is sealed into its envelope. Heat conduction through the electrodes then makes the crystal fairly hot, but the surroundings are kept cool, so that the water-vapour pressure remains low. After the seal-off, a low $\alpha'$ may therefore be expected. The surface water occupation is now out of equilibrium with the vapour pressure produced by the buffer, but at room temperature the equilibrium is restored only very slowly. If the transistor is operated, its temperature rises and the return to equilibrium is accelerated, which is apparent from changes in $\alpha'$. If the transistor is kept at 100 °C, $\alpha'$ usually reaches its stable value after a few days. The transistors to which fig. 10*f* relates were not subjected to this pre-heating.

The temperature must not be raised above 100 °C with the object of speeding-up the above process. Experience has shown that higher temperatures, in conjunction with the high water-vapour pressure then produced by the buffers, inflict damage to the germanium surface, one result of which is a particularly low $\alpha'$. Re-etching is then the only way to save the transistor.



*a*) $BaCl_2.2aq$, separated from silicone grease (cf. fig. 8*b*).

*b*) $K_2SO_4.NiSO_4.6aq$, separated from silicone grease.

*c*) $K_2SO_4.Al_2(SO_4)_3.6aq$, mixed with silicone grease.

*d*) Boracic acid, mixed with silicone grease.

*e*) Slightly moist sand, serving also as filler material.

*f*) Silica gel, serving also as filler material.

Fig. 10. Examples of the behaviour of transistors with a buffer incorporated in the encapsulant to stabilize the water-vapour pressure. Ambient temperature 50 °C; dissipation 50 mW; temperature of germanium crystal 85 °C. $\alpha'$ was measured at room temperature. The curves relate to *P-N-P* transistors, electrolytically etched in KOH. The buffer used is mentioned below each graph. The level of $\alpha'$ is not characteristic of the buffer used.

### *Stabilization experiments by "forming" the surface*

In order to stabilize transistors by "forming" their surface (p. 207), a buffer has to be introduced that gives a lower water-vapour pressure than is required for normal stabilizing. We have achieved successful results with transistor fillings consisting of a silicone grease mixed with a little boracic acid (say 5% by weight), from which water is expelled to the required degree by drying. If the encapsulated transistors are formed by heating them for three days at 140 °C, stable transistors are produced which possess favourable properties in every respect. *Fig. 11* demonstrates the stability of $\alpha'$ after weeks of continuous loading; *fig. 12* represents $\alpha'$ as a function of the time of storage at 100 °C.

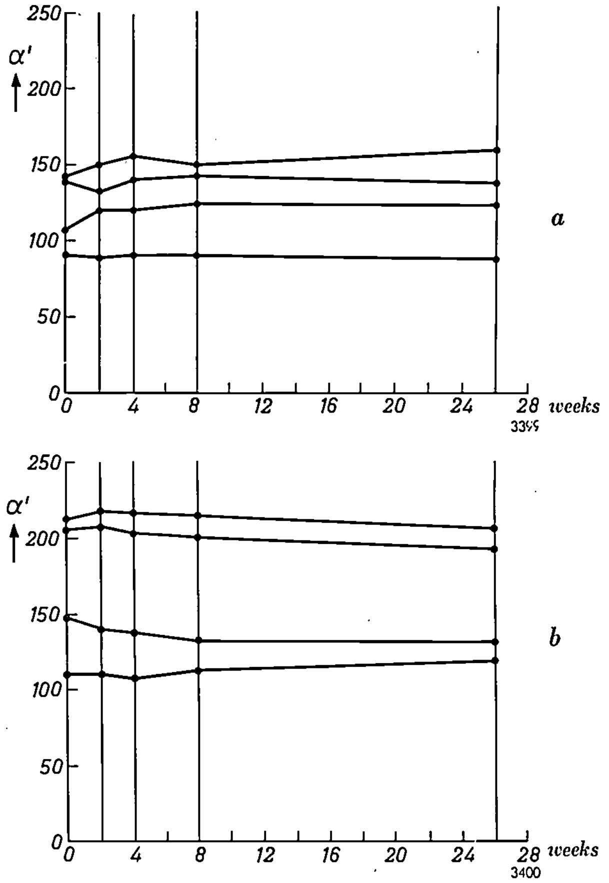


Fig. 11. Examples of the behaviour of $\alpha'$ in life tests on transistors stabilized by surface forming at 140 °C. Buffer 5% pre-heated boracic acid, mixed with silicone grease. Dissipation 50 mW (5 mA, 10 V). Ambient temperature 50 °C, the temperature of the crystal then being 85 °C. $\alpha'$ was measured at room temperature. (The level of $\alpha'$ is not characteristic of the stabilizing method or of the type of transistor.)
*a*) *P-N-P* transistors.
*b*) *N-P-N* transistors.

It is particularly important to extend the forming process over a sufficiently long period of time. It is seen from *fig. 13* that $\alpha'$, as expected, is low immediately after encapsulation. After heating at 140 °C for one day, its value has increased considerably, but there is again a sharp drop after storing the transistor for a day at room temperature. The forming process was too short. After prolonged heating at 140 °C (here 6 days), however, the improvement achieved is not lost again. But the forming should not be too prolonged, otherwise $\alpha'$ begins to fall once more.

Longer forming is required, or a higher forming temperature, the lower is the vapour-pressure curve of the buffer. This is illustrated in *fig. 14*, which relates to a transistor whose envelope was filled with a silicone grease mixed with boracic acid, which had been dried out more than in the cases earlier discussed. In practice the mixture of silicone grease and boracic acid will be chosen with a view to limiting the forming process to a few days of heating at 140 °C. A higher temperature is undesirable, since the indium in the collector and emitter melts at about 155 °C.

It is likely that during the forming process — and perhaps afterwards — the silicone grease in combination with the boracic acid has its own advantageous effect. At the forming temperature there may well be reactions between the grease and the boron compounds which favourably influence the transistor's characteristics. However, that water vapour plays the major role in the forming process appears from the fact that transistors with very low values of $\alpha'$ are obtained when a drying agent, e.g. barium oxide, is added inside the envelope.

The method of stabilization by forming the surface in a mixture of silicone grease and boracic acid not only produces high and stable values of $\alpha'$, but also benefits other important transistor characteristics (*Table I*). For example, the saturation leakage currents at the *P-N* junctions are small and do not drift. This is in contrast with the leakage currents in many commonly used transistors, which may gradually assume appreciable values, particularly if the temperature of the transistor is relatively

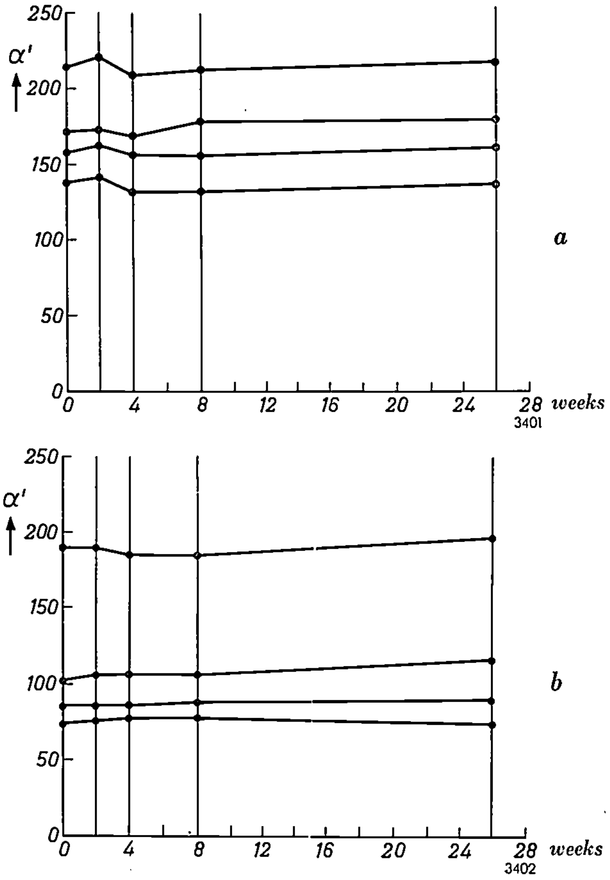


Fig. 12. Examples of the behaviour of $\alpha'$ in storage tests at 100 °C on transistors stabilized by forming. Buffer 5% preheated boracic acid, mixed with silicone grease. $\alpha'$ was measured at room temperature. (The level of $\alpha'$ is not characteristic of the stabilizing method or of the type of transistor.)
*a*) *P-N-P* transistors.
*b*) *N-P-N* transistors.

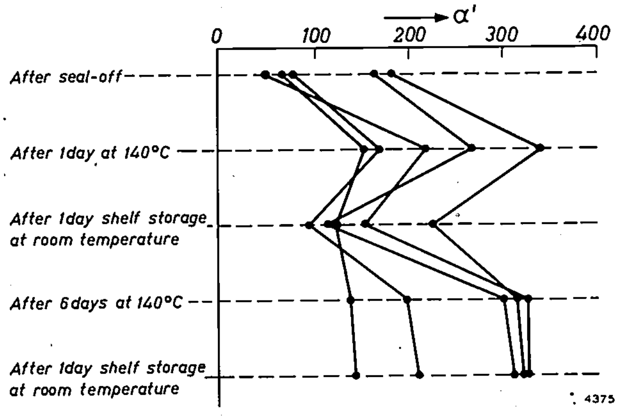


Fig. 13. Variation of $\alpha'$ of some $P$-$N$-$P$ transistors, electrolytically etched in KOH, which were initially formed for too short a time (one day) and subsequently for a long enough time (six days). $\alpha'$ was measured at room temperature.

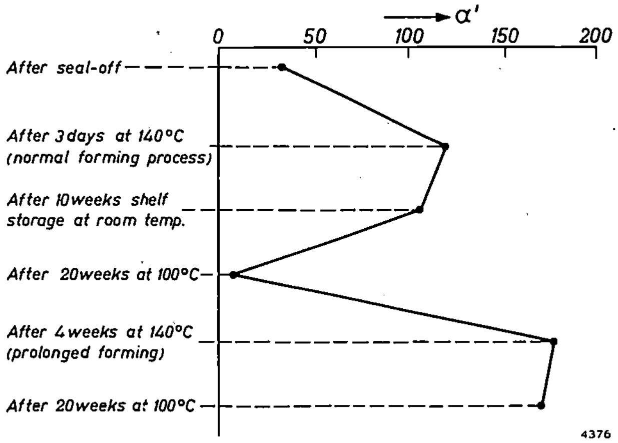


Fig. 14. Variation of $\alpha'$ of a transistor formed with a buffer giving a considerably lower water-vapour pressure than in the cases to which figs. 11, 12 and 13, and Table I, refer. $\alpha'$ was measured at room temperature.

Table I. The current amplification factor $\alpha'$, the saturation leakage currents $I_{C0}$ at the collector junction and $I_{E0}$ at the emitter junction, and the noise, of representative examples of $P$-$N$-$P$ and $N$-$P$-$N$ transistors, after various successive treatments. The transistor envelopes were filled with a silicone grease mixed with 5% boracic acid, dehydrated to a certain degree by pre-drying. The measurements were done at rpom temperature.

| | *P-N-P* | | | | *N-P-N* | | | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha'$ | $I_{C0}$ (μA) | $I_{E0}$ (μA) | Noise (dB) | $\alpha'$ | $I_{C0}$ (μA) | $I_{E0}$ (μA) | Noise (dB) |
| After electrolytic etching in KOH | 210 | 8 | 6 | — | 89 | 0.6 | 0.6 | — |
| After seal-off | 56 | 26 | 22 | — | 40 | 1.3 | 1.2 | — |
| After 1 day at 100 °C | 75 | — | — | — | 68 | — | — | — |
| After 24 hrs storage at room temperature | 60 | 2.8 | 2.7 | 4 | 53 | 4.4 | 3.5 | 6 |
| After 3 days at 140 °C (forming) | 158 | — | — | — | 164 | — | — | — |
| After 24 hrs storage at room temperature | 158 | 1.8 | 1.6 | 4 | 172 | 0.3 | 0.2 | 4 |

high (e.g. 60 °C). The table further shows that the noise level, which is often correlated with the leakage currents, is also favourable in formed transistors. In fact, the noise values found are just about the lowest yet measured on transistors.

In the case of $P$-$N$-$P$ transistors the breakdown potentials of the two $P$-$N$ junctions after forming are generally 20 to 30% lower than before. The breakdown potentials in $N$-$P$-$N$ transistors are not significantly affected.

Finally it should be noted that surface forming also has a fairly marked influence on the $I_E$-$V_{EB}$ characteristic (*fig. 15*).

## Stabilizing with arsenic

Water is not alone in its property of reducing the recombination of holes and electrons at the germanium surface. Another substance with which we have successfully experimented is arsenic. At first sight there would seem to be no relation between water and arsenic, and the reader may well wonder how it happened to be chosen for experiments at all. It is not really so odd, however. Extensive physicochemical investigations into the influence exerted by water on the surface of semiconductors have revealed that water adhering to such a surface tends to induce $N$-type surface conductivity. On $N$-type germanium, water thus makes the surface layer more strongly $N$-type than the interior, and on $P$-type germanium it makes the surface layer less strongly $P$-type than the interior. It may even result in an $N$-type surface layer on germanium that is only weakly $P$-type [7]. The question arose whether donor elements like arsenic, phosphorus, antimony and bismuth, with which germanium is doped to induce $N$-type conductivity, might have a similar effect on the germanium surface as water has. This proved indeed to be the case, particularly as regards arsenic and phosphorus, which have measurable vapour pressures at 140 °C.

[7] R. H. Kingston, Water-vapor-induced $N$-type surface conductivity on $P$-type germanium, Phys. Rev. **98**, 1766-1775, 1955.

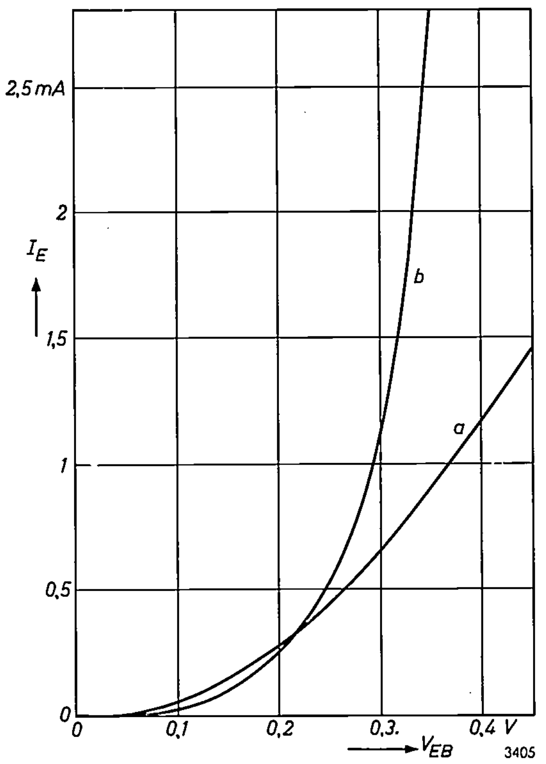


Fig. 15. Illustrating the effect of forming on the $I_E$-$V_{EB}$ characteristic of a *P-N-P* transistor; *a*) before forming, *b*) after forming.

We shall only comment briefly on the stabilizing method whereby a dried silicone grease mixed with a few percent by weight of arsenic powder is used as the filler substance in transistor envelopes. In its effect this mixture closely resembles the mixture of silicone grease and boracic acid used for the transistors stabilized by surface forming with water vapour. Again, a forming period of several days at 140 °C or higher is necessary. The resulting transistors have a high $\alpha'$ and excellent stability (*fig. 16*). In fact, in many aging tests, $\alpha'$ showed no change whatsoever.

It is not to be expected that the arsenic at 140 °C will really diffuse in the germanium surface layer and be incorporated in the germanium lattice as donor impurities normally are. For any significant diffusion to occur, the temperature would have to be at least 600 °C. Experiments have proved that the arsenic nevertheless gives rise to a surface layer which is strongly *N*-type. This layer persists as long as the transistor together with the arsenic is hermetically sealed. If the transistor is exposed to the ambient air, the surface layer changes and the transistor is no longer stable. The effect of the arsenic is entirely destroyed if the transistor is introduced into a space which is evacuated at high temperature. Evidently the arsenic is bound only very weakly to the germanium surface.

An advantage over the forming method using a silicone grease and boracic acid is that the transistors obtained are much more capable of withstanding high temperatures. Arsenically treated transistors can be held, for example, at a forming temperature of 140 °C for several months without deteriorating, whereas the same treatment with silicone grease and boracic acid would result in a considerable drop in $\alpha'$.

A complication is that transistors etched in KOH and subsequently formed with arsenic exhibit a marked "shelf after-effect". When such transistors, after a period of storage at room temperature, are raised to a higher temperature, $\alpha'$ gradually rises for a few days. When the transistors are then returned to room temperature, $\alpha'$ again declines, at first rapidly and then slowly (*fig. 17*). It may be a month before the original value is reached. This effect is not found on transistors formed with a silicone grease and boracic acid mixture. The effect is apparently bound up with the presence of traces

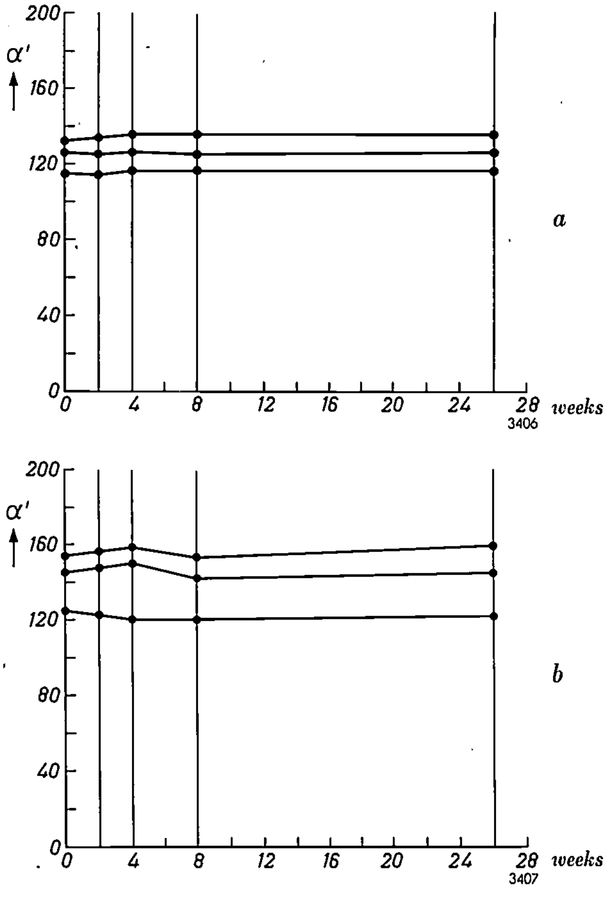


Fig. 16. Some examples of the constancy of $\alpha'$ of transistors formed with arsenic. Before each measurement of $\alpha'$, the transistor was shelf-stored at room temperature for 24 hours.
*a*) During operation; dissipation 50 mW, ambient temperature 50 °C.
*b*) Stored, at 140 °C.

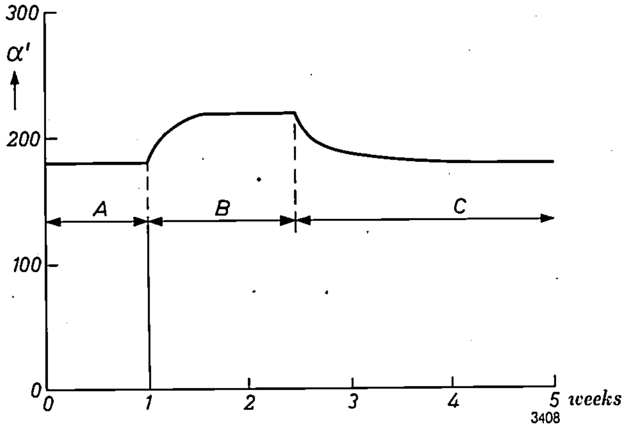


Fig. 17. Transistors formed at 140 °C, with powdered arsenic mixed in the silicone grease, often show a marked "shelf after-effect".
Period *A*: transistor at room temperature.
Period *B*: transistor at 100 °C.
Period *C*: transistor at room temperature.

of water in the transistor envelope, since if the transistors are rigorously dried after forming (by enclosing a drying agent in the envelope), the result is a high and stable $\alpha'$ without this shelf after-effect.

A simpler way of eliminating the effect is to mix boracic acid as well as arsenic with the grease. This is the procedure with which we have so far achieved the best results. The transistors so treated combine the advantages of boracic-acid and arsenic forming, namely no shelf after-effect, high and very constant value of $\alpha'$ and high temperature stability.

**Summary.** The surface state of germanium transistors has a very marked influence on their characteristics, in particular on the current amplification factor. The occupation of the surface by water molecules is shown to be exceptionally important, and is thought to have an optimum value. A favourable water occupation may be maintained in all operating conditions by incorporating in the encapsulant a buffer substance — a stabilizer — which provides the appropriate water-vapour pressure as a function of temperature. Whilst this method produces good transistors with reasonably stable characteristics, better results are obtained with a buffer whose water-vapour pressure is initially too low. After the transistors have been heated for several days, they are then found to have very favourable and stable characteristics. The process is described as "surface forming". There are other substances that resemble water in their effect on the surface. Results obtained with one such substance (arsenic) are briefly discussed.

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

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

**2780:** S. Duinker: General energy relations for parametric amplifying devices (T. Ned. Radiogenootschap **24,** 287-310, 1959, No. 5).

It is shown that the energy relations pertaining to parametric amplifying devices, as they have been derived by various authors, are a direct consequence of the invariance of the total-energy function of the parametric system under certain transformations. The theory is generalized so as to comprise arbitrary parametric systems. Some general properties of parametric systems, which can be deduced immediately from the energy relations, are discussed. A small number of typical examples are briefly treated to illustrate some fundamental principles following from the general theory.

**2781:** W. Albers and J. T. G. Overbeek: Stability of emulsions of water in oil, I. The correlation between electrokinetic potential and stability (J. Colloid Sci. **14,** 501-509, 1959, No. 5).

Experiments on water-in-oil emulsions of moderate concentration, stabilized with oil-soluble, ionizing stabilizers, show that in these emulsions no correlation exists between stability against flocculation and electrokinetic potential. Although, according to theoretical calculations, energy barriers of over 15 $kT$ are present if the radius of the dispersed globules is about 1 $\mu$ and the electrokinetic potential exceeds 25 mV, they apparently do not prevent lasting contact between particles. All the emulsions flocculate rapidly, even in the presence of a surface potential considerably higher than 25 mV. A rather pronounced anticorrelation exists between the zeta potential and coalescence. It is explained as a consequence of the free mobility of the stabilizing molecules in the interface. The good stabilization against coalescence caused by some oleates of polyvalent metals is due to the formation of a thick film of partial hydrolyzates in the interface.

**2782:** W. Albers and J. T. G. Overbeek: Stability of emulsions of water in oil, II. Charge as a factor of stabilization against flocculation (J. Colloid Sci. **14,** 510-518, 1959, No. 5).

It is shown by theoretical calculations that the energy barrier between charged droplets in water-in-oil emulsions is strongly diminished when the concentration of the emulsion is not extremely low. This is a consequence of the great extension of the diffuse electrical double layer in oil. The high concentration in the sediment (or cream) therefore strongly promotes flocculation. Gravity also promotes flocculation directly in all but the most dilute water-in-oil emulsions because the weight of the particles in higher layers transmitted by the extended double layers presses on those in the lower layers and forces them together.

**2783:** F. N. Hooge: Influence of bond character on the relation between dielectric constant and refraction index (Z. phys. Chemie (Frankfurt a.M.) **21**, 298-301, 1959, No. 3/4).

In the relation given by Szigeti relating the dielectric constant of cubic ionic compounds with the refractive index, it is shown that the value of the experimental constant *s* depends on the bond character obtaining in the crystal. Using Pauling's model it is shown that *s* may be calculated from quantities that determine the bond character of a crystal.

**2784:** G. Diemer and J. G. van Santen: Nieuwe toepassingen van elektro-optische verschijnselen (Ned. T. Natuurk. **25**, 265-287, 1959, No. 10). (New applications of electro-optical phenomena; in Dutch.)

A survey is given of the fundamental properties and possible applications of electroluminescence, photoconduction and combinations in the fields of image intensification, image display, radiation detection, electric amplification and logic, switch and memory devices.

**2785:** T. J. de Man and J. R. Roborgh: The hypercalcemic activity of dihydrotachysterol$_2$ and dihydrotachysterol$_3$ and of the vitamins $D_2$ and $D_3$; comparative experiments on rats (Biochem. Pharmacology **2**, 1-6, 1959, No. 1).

The hypercalcemic activities of dihydrotachysterol$_2$, dihydrotachysterol$_3$, vitamin $D_2$ and vitamin $D_3$ have been compared at different intervals (2, 4, 7 and in some cases 1 and 10 days) after the administration to rats of one oral dose of the crystalline compounds in pure peanut oil. The maximal serum calcium levels appear to be obtained 2-4 days after the administration. From the results it appears that dihydrotachysterol$_3$ is the most active hypercalcemic agent, followed, in decreasing order, by dihydrotachysterol$_2$, vitamin $D_3$ and vitamin $D_2$. The activity ratios proved to be highly dependent on the time interval. As to the hypercalcemic activity, the potency ratio vitamin $D_3$/vitamin $D_2$ has been found to be about 1.6, while, with respect to their antirachitic activities, both vitamins are equipotent in rats. The antirachitic activity of dihydrotachysterol$_3$ was shown to be nearly twice as high as the potency of dihydrotachysterol$_2$, the antirachitic activity of dihydrotachysterol$_2$ being about 0.5 per cent of the vitamin $D_3$ activity.

**2786:** J. H. Uhlenbroek and J. D. Bijloo: Isolation and structure of a nematicidal principle occurring in Tagetes roots (Proc. 4th int. Congr. Crop Protection, Hamburg 1957, Vol. 1, pp. 579-581; published 1959).

Note and discussion concerning extraction and structure of a nematicidal principle from Tagetes roots.

**2787:** J. Mcltzcr and F. C. Dietvorst: Rclation between chemical structure and ovicidal and leaf-penetrating properties of some new acaricides (Proc. 4th int. Congr. Crop Protection, Hamburg 1957, Vol 1, pp. 669-673; published 1959).

Sulphur-linked diphenyl compounds possess high acaricidal activity. The sulphone ("Tedion"), sulphoxide and sulphide homologues of the 2,4,5,4'-tetrachloro-derivatives do not differ greatly in ovicidal and larvicidal activity. Direct contact action on the eggs, however, is stronger with the sulphone and least with the sulphide. Leaf penetration is fastest for the sulphide and slowest for the sulphone. As regards residual action, the sulphone is appreciably longer lasting than the sulphoxide and the sulphide. p-chlorophenyl benzenesulphonate and chlorbenside show a faster leaf penetration than Tedion, and in consequence a shorter residual action.

**2788:** J. Rodrigues de Miranda and H. van den Kerckhoff: Designing a multi-purpose stereo pre-amplifier (J. Audio Engng. Soc. **7**, 75-80, 1959, No. 2).

In between the required sensitivity and the required output voltage the stereo pre-amplifier should be designed with a good tone and volume control system. Logical design leads to a pre-amplifier with distortion as low as 0.03% for all frequencies. Included in the design is a simple but effective rumble filter.

**2789:** J. L. Ooms and C. R. Bastiaans: Some thoughts on geometric conditions in the cutting and playing of stereodiscs and their influence on the final sound picture (J. Audio Engng. Soc. **7**, 115-121, 1959, No. 3).

The influence of various geometric conditions in the cutting and reproducing process is investigated. It is found that stylus contour plays no important role. Axis orientation does play a role; any deviation herein causes cross modulation, ultimately leading to distortion of the original stereophonic sound picture. Slant and rotation of planes of axes may be disregarded since the resultant difference angles are small enough to be neglected. Finally, the nature of sound-picture distortion caused by such cross modulation is investigated. Amplitude distortion leads to angular shift of the sound sources (panorama distortion), phase relationship of cross-modulation components causing increase or decrease in width of the sound picture (basis distortion).

**2790:** J. L. Meijering: Thermodynamical calculation of phase diagrams (The physical chemistry of metallic solutions and intermetallic compounds, Proc. Symp. Nat. Phys. Lab., June 1958, Vol. 2, paper No. 5 A).

Survey article outlining the methods and limitations of thermodynamics for the prediction and extrapolation of phase diagrams. The subject matter is treated under the following headings: binary miscibility gaps, solid-liquid equilibria, ferrite-austenite equilibria, ternary miscibility gaps. About 50 references are given to the original literature.

**2791:** G. Meijer: Photomorphogenesis in different spectral regions (Proc. Conf. Photoperiodism and related phenomena in plants and animals, Gatlinburg, Tennessee, Oct.-Nov. 1957, edited by R. Withrow, pp. 101-109; published 1959).

In experiments with some photoperiodically-sensitive plants it is shown that for obtaining a long-day effect two different photo reactions are involved, a red-sensitive one which can be reversed by near infrared, and a blue/near-infrared sensitive reaction. The elongation of internodes is inhibited by light. Two inhibiting processes were distinguished, one occurring in red light and being antagonized by near infrared, the other one occurring in blue light. With a certain intensity it was found that it depends on the plant species which process predominates.

**2792:** W. Verweij: Probe measurements in the positive column of low-pressure mercury-argon discharges (Physica **25**, 980-987, 1959, No. 10).

In the positive column of electric discharges in mixtures of argon and mercury vapour at low pressure, electron concentration, electron temperature and axial field strength are determined with the aid of Langmuir probes. The mercury pressure is varied from $0.50 \times 10^{-3}$ mm Hg to $90 \times 10^{-3}$ mm Hg, the argon pressure from 0 to 20 mm Hg and the mean current density from 10 mA/cm$^2$ to 80 mA/cm$^2$. If very thin cylindrical probes (20 $\mu$ diameter) are used, the measurements of the electron concentration based on plasma potential and those found from the characteristic at positive probe voltage are in very good agreement. For the discharges under examination, the mobility of the electrons is evaluated from the electron concentration, the electron gradient and the tube current.

VOLUME 22, 1960/61, No. 7 pp. 217-244 Published 5th April 1961

Philips Technical Review

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

# PROPERTIES AND APPLICATIONS OF INDIUM ANTIMONIDE

by R. E. J. KING*) and B. E. BARTLETT*).

546.682.'86

*The increasing demand for photoconductive cells which extend their sensitivity range into the far infra-red, has led to the construction of two new photocells based on crystals of indium antimonide.*

*In this article the preparation of InSb crystals and the construction and performance of photocells based on InSb are described. Mention is also made of the use of InSb in Hall generators. These and other devices will be discussed in a forthcoming article in this journal.*

In recent years a search has been made for semiconductors to supplement silicon and germanium for use in electronic applications. Since no other elements appear particularly useful, interest has been focussed on compounds. Most work has been done on compounds of elements from group III and V of the periodic table, such as GaAs, GaP, InSb, etc. Of these compounds, indium antimonide has been studied most intensively [1]), because its properties make it particularly suitable for a number of applications. In addition, the problems involved in the preparation of InSb are more tractable than those encountered with other, similar, compounds. Photocells, Hall generators and other devices based on InSb are already being manufactured by a number of factories, amongst others the Mullard division at Southampton.

## Properties of InSb

### *Optical and electrical properties*

In *Table I* is shown part of the periodic table of the elements. The elements within the columns of the table have roughly similar chemical properties because they have the same number of outer shell valency electrons.

The elements in column IVA, with the exception of lead, are found to be semiconductors and crystallise in the diamond configuration by forming four tetrahedral bonds. InSb forms the same type of structure utilising the three valence electrons of indium and the five of antimony to give the four bonds. Consequently the electrical and optical properties of the material are qualitatively similar to those of germanium and silicon. However, the magnitudes of the important parameters underlying the physical properties of the bulk material are different.

Table I. Groups IIIa, IVa and Va of the periodic table of elements.

| IIIA | IVA | VA |
|---|---|---|
| *B*<br>5 | *C*<br>6 | *N*<br>7 |
| *Al*<br>13 | *Si*<br>14 | *P*<br>15 |
| *Ga*<br>31 | *Ge*<br>32 | *As*<br>33 |
| *In*<br>49 | *Sn*<br>50 | *Sb*<br>51 |
| *Tl*<br>81 | *Pb*<br>82 | *Bi*<br>83 |

4114

Two important quantities which affect the optical and electrical properties of the material are the *energy gap* and the *mobility of the charge carriers*. These will now be considered.

In solids there are *bands* of permissible energies for electrons in contrast to the discrete electron energy *levels* permissible in separate atoms. These bands

*) Mullard Radio Valve Co., Southampton Works.

[1]) A. N. Blum, N. P. Mokrovski and A. R. Regel, J. tech. Phys. (Moscow) **21**, 237, 1951; H. Welker, Z. Naturf. **7a**, 744, 1952 and **8a**, 248, 1953; R. Gremmelmaier and O. Madelung, Z. Naturf. **8a**, 333, 1953; H. Weiss, Z. Naturf. **8a**, 463, 1953.

are separated by forbidden ranges of energy. Now in a semiconductor the highest occupied energy band is exactly filled by the valence electrons available. Energy must be supplied — e.g. thermal, electrical or optical energy — to excite an electron from this band (the valence band) to the next (empty) band of higher electron energy (the conduction band). The energy separating these two bands is termed the "energy gap". (The greater this gap, the more the semiconductor approximates to an insulator.) Electrons with energies within a full band cannot carry any current but after such an excitation both the electron in the conduction band and the "hole" left in the valence band can conduct. When excitation can be achieved by optical energy, the material has the characteristics of a photoconductor.

In *Table II* the energy gaps of Ge, Si and InSb are compared. It is seen that the energy gap of the latter material is small, much smaller than that of Gc and Si. A consequence of this small energy gap is that radiation of quite long wavelengths is absorbed: the optical absorption edge of InSb is at 7.5 μ, as compared with 1.7 μ for Ge and 1.2 μ for Si.

Thus photoconductive cells covering the wavelength range from the visible to 7.5 μ can be made from InSb, utilising conduction changes due to carriers generated optically by excitation from the valence to the conduction band.

Table II. Some properties of germauium, silicon and indium antimonide.

| | Ge | Si | InSb |
|---|---|---|---|
| Band gap (eV) | 0.72 | 1.1 | 0.18 |
| Intrinsic carrier concentration at 300 °K (cm$^{-3}$) | $2.5\times10^{13}$ | $6.8\times10^{10}$ | $2\times10^{16}$ |
| Electron mobility at 300 °K (cm$^2$/Vs) | 3600 | 1300 | 70 000 |
| Hole mobility at 300 °K (cm$^2$/Vs) | 1700 | 500 | 1000 |

The *mobility* of charge carriers is the average drift velocity acquired per unit applied electric field intensity in the direction of the field. This property is important for two reasons:

1) It is a parameter determining performance in both photocells and Hall-effect devices.

2) It is relevant to the investigation and control of the purity of material produced.

In a perfectly periodic crystal lattice at zero absolute temperature the mobility of charge carriers would be infinite. However, at higher temperatures, thermal vibrations of the crystal lattice reduce the mobility. In addition, the forces associated with impurities interfere with the motion of the carriers and this "scattering" can also be important in determining the mobility in the material. In order to ensure that the InSb produced meets the design desiderata for photocells and other devices, it is essential that the purity of crystals be investigated and controlled.

Two types of impurities are of particular importance, namely donors and acceptors, which give rise to electrons (*N*-type semiconductor) and holes (*P*-type), respectively. In the presence of both types of impurities simultaneously, compensation occurs and it is the difference between the concentrations of the two types which is operative in determining the carrier concentration. The determination of this difference in concentrations (or net concentration) is complicated by the presence of carriers due to purely thermal excitation across the energy gap (*intrinsic* carriers). The concentration of such intrinsic carriers in a pure InSb crystal at room temperature is about $2\times10^{16}$ cm$^{-3}$, as compared with only $2.5\times10^{13}$ cm$^{-3}$ and $7\times10^{10}$ cm$^{-3}$ for germanium and silicon, respectively. Therefore, in order to determine impurity concentrations of this order or less, it is necessary to reduce the intrinsic concentration by cooling.

On the other hand, there is a certain (small) amount of activation energy required to excite an electron from the energy levels associated with the impurities which in fact is the very means of detecting the presence of the impurities. The temperature at which measurements are made must therefore be sufficiently high for the impurities to be all ionised, but sufficiently low for the intrinsic carrier concentration to be considerably less than the impurity concentration. In practice for InSb, at present levels of purity, 77 °K (boiling point of liquid nitrogen) is a suitable temperature.

To obtain the necessary information about the impurity concentrations and the mobilities, both the conductivity and the Hall effect are measured [2]).

The electrical conductivity $\sigma$ for an *N*- or *P*-type semiconductor is given by

$$\left.\begin{aligned} \sigma &= ne\mu_n \\ \text{or}\quad \sigma &= pe\mu_p, \end{aligned}\right\} \quad \dots\dots\dots \quad (1)$$

*e* being the electronic charge, *n* and *p* the respective concentrations of electrons and holes and $\mu_n$ and $\mu_p$ their respective mobilities.

To evaluate separately the carrier concentrations and mobilities, a Hall-effect measurement is made.

[2]) See, for example, C. Kittel, Solid-state physics, 2nd Edn. p. 296, Wiley, New York 1957.

The Hall effect may be illustrated with the aid of *fig. 1.*

If a current $I$ is passed along a rectangular specimen and a magnetic field $B$ is applied at right-angles to the direction of current flow, then a potential difference is established in a direction perpendicular to both the field and the current directions. If contacts are made on the specimen at $A_1$ and $A_2$ this potential difference known as the Hall voltage, can be measured. The open-circuit Hall voltage $V_{\mathrm{H}0}$ is given by the equation,

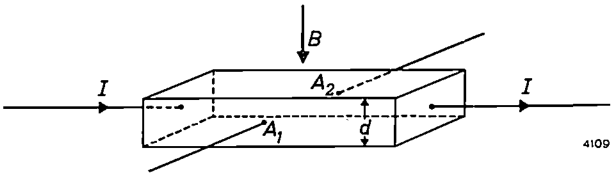


Fig. 1. The Hall effect. $B$ is the induction of the applied magnetic field, $I$ the current through the specimen of thickness $d$. $A_1$ and $A_2$ are the Hall-voltage contacts.

$$V_{\mathrm{H}0} = \frac{R_{\mathrm{H}} I B}{d}, \quad . \; . \; . \; . \; . \quad (2)$$

where $d$ is the thickness of the specimen in the direction of the magnetic field. For a particular temperature $R_{\mathrm{H}}$ is a constant of the material and is known as the Hall coefficient. The above equation assumes that the ratio of length to breadth of the specimen is considerable greater than unity.

For an $N$- or $P$-type semiconductor, the Hall coefficient is:

$$\left.\begin{array}{l} R_{\mathrm{H}} = -\dfrac{3\pi}{8}\Big/ne \\ \text{or} \\ R_{\mathrm{H}} = +\dfrac{3\pi}{8}\Big/pe\,. \end{array}\right\} \quad . \; . \; . \; . \; . \; . \quad (3)$$

In *figs. 2* and *3* are shown the variation of the Hall coefficient and conductivity with temperature for two typical samples of InSb, the one being $P$-type and the other $N$-type, at low temperatures. Eq. (3) gives the electron or hole concentration, giving information on the difference between the concentrations of donors and acceptors. From the values of $n$ and $p$ and using eq. (1), the mobility of the carriers can be found.

In many cases, the mobility of the carriers may be wholly or almost wholly determined by their scattering by impurity atoms. Under these conditions the *total* impurity concentration may be determined by comparison with scattering theory [3]).

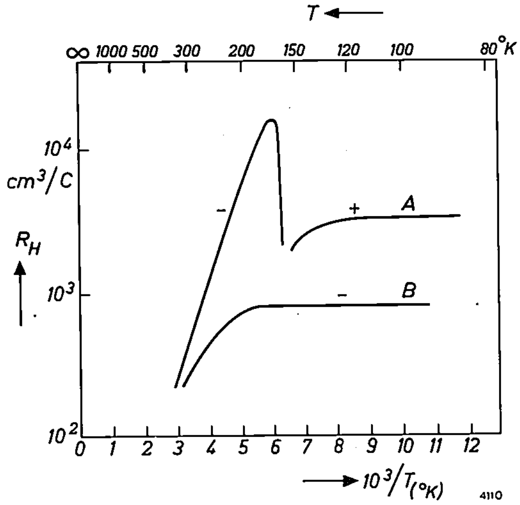


Fig. 2. Variation of Hall coefficient of ($A$) $P$-type and ($B$) $N$-type InSb with temperature. At temperatures below about 160 °K the Hall coefficient of sample ($A$) is positive. The high ratio of electron to hole mobility in indium antimonide causes the Hall coefficient to be negative at temperatures above 160 °K when it is a "mixed" conductor i.e. both electrons and holes are making a significant contribution to the conduction process. At higher temperatures both curves $A$ and $B$ approximate to the same straight line, which represents the intrinsic material.

From the Hall measurements, and using eq. (3), the *difference* in impurity concentrations is found. Hence the individual concentration of both acceptors and donors may be separately estimated.

From Table II it can be seen that the electron mobilities encountered in InSb are high compared with Ge and Si, being typically 70 000 cm$^2$/Vs at room temperature. At 77 °K values as high as 650 000 cm$^2$/Vs have been observed. The highest hole mobility observed at 77 °K is 10 000 cm$^2$/Vs, and values deduced for room temperature may often reach 700-1000 cm$^2$/Vs.

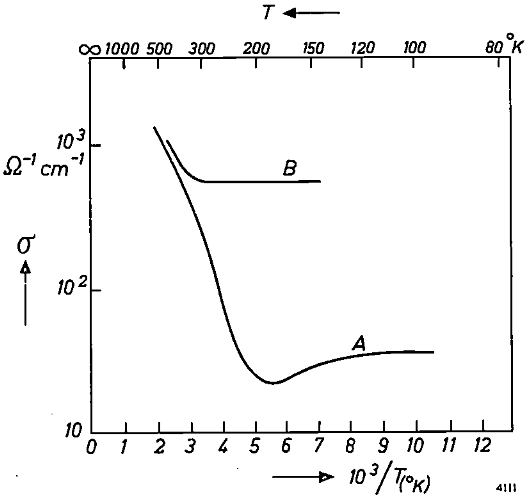


Fig. 3. Variation of conductivity of ($A$) $P$-type and ($B$) $N$-type InSb with temperature. For higher temperatures the material becomes intrinsic which is indicated by the fact (as in fig. 2) that curves $A$ and $B$ approximate to the same line.

[3]) R. B. Dingle, Phil. Mag. 46, 831, 1955.

A consequence of these high mobilities is that InSb displays large magnetoresistance effects [4]). A further consequence is that the power efficiency of Hall generators made from InSb is high compared with similar devices manufactured from Ge or Si.

*Recombination and trapping*

Non-equilibrium carrier concentrations in the bulk of semiconductor materials can be achieved by injection of carriers at contacts or by irradiation of the material with light of such a wavelength or wavelengths that valence electrons can be excited into the conduction band (leaving holes in the valence band) after acquiring energy from a photon-electron interaction. The first-mentioned process — injection at contacts — is made use of in diodes and transistors. The second process is made use of, as stated earlier, in photoconductors (photo-resistors).

After these processes the bulk of the material remains electrically neutral (no space charge is built up). When the process of injection or radiation is stopped, the concentrations return to equilibrium in a time which is in general long compared with the space-charge relaxation time when electrical forces accelerate the process of returning to equilibrium. The time for return to equilibrium is related to the *lifetime* of the charge carrier, i.e. the time that an electron remains in the conduction band, or a hole in the valence band, after excitation, and is, in general, determined by the probabilities of *recombination* and *trapping*. The first is the process of returning of an electron to the valence band by recombining with a hole. The second is the process of being "trapped" by a trapping centre and thus becoming immobile. As with transistors, the operation of photo-resistors depends on the existence of a *finite* lifetime of excess injected carriers in the material. The signal obtained from a photoconductive cell of given geometry is directly proportional to this lifetime.

In InSb, in contrast to Ge and Si, it is not at present possible to obtain non-equilibrium carrier concentrations in the bulk material, at room temperature or higher temperatures, by injection of carriers at contacts. To obtain injection it is necessary that a potential barrier is present at thc contact to prevent the flow of one type of carrier into or out of the material. Consequently, it is not possible at present to make transistors or diodes from InSb for operation at room-temperature and above.

The room-temperature lifetime of excess carriers in InSb is *not* governed by traps and recombination centres, as in Ge or Si, and it is, at present, uncertain whether the recombining electron loses its energy in the form of radiation (radiative recombination) or by the Auger effect [5]) in which it loses its energy to another electron in the conduction band (see *fig. 4*). At 300 °K, typical lifetime values for InSb are about $5 \times 10^{-8}$ sec.

At lower temperatures other mechanisms determine the recombination [6]). In particular, the presence of traps in the bulk or on the surface is important. These mechanisms can be studied by measurements of the photoconductive and photomagneto-electric type.

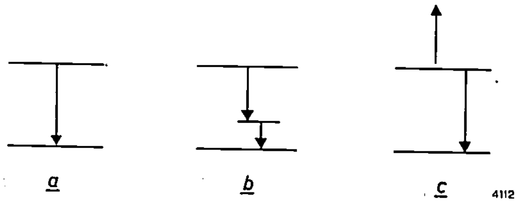


Fig. 4. Recombination processes.
*a*) Radiative recombination. An electron in the conduction band recombines directly with a hole in the valence band, accompanied by the emission of a photon.
*b*) Recombination via a trap. An electron is "trapped" in an energy level in the forbidden zone, and after some time recombines again with a hole in the valence band.
*c*) Auger effect. An electron in the conduction band falls back to the valence band and the energy which comes free is transferred to another conduction electron, lifting the latter to a higher energy level in the conduction band.

## Preparation of InSb

As InSb is a compound, its preparation differs somewhat from the preparation of the elemental semiconductors. It has the advantage that the two constituent elements may be purified before the compound is prepared and impurities difficult to remove from the compound may be removed from the elements themselves, but there is the possibility that non-stoichiometry (excess of one of the elements in the compound) will occur. In fact, there is no evidence for the solubility of significant amounts of either excess In or excess Sb in solid InSb, although it is not certain whether this is still true for very low concentrations of the order of $10^{14}$ cm$^{-3}$.

For the efficient production of reproducible devices from InSb, there are three requirements to be mct in the material production process. These are: 1) high degrees of purity, 2) single crystals and 3) uniformity of material over useful working volumes.

[4]) H. P. R. Frederikse and W. R. Hosler, Phys. Rev. **108**, 1136, 1957.

[5]) A. R. Beattie and P. T. Landsberg, Proc. Roy. Soc. A **249**, 16, 1959.

[6]) D. W. Goodwin, Report of the meeting on semiconductors, Physical Society and British Thomson-Houston Ltd. (Rugby, April 1956), p. 137.

The fulfilment of the first of these requirements ensures that the high mobilities realisable in InSb can be utilised in devices such as the Hall generator. Also doping the material to levels set by design considerations for photocells can be readily accomplished if pure starting material is employed.

Single crystals are necessary if thin foils or filaments are to be prepared by anodic etching (polycrystalline material is preferentially etched at grain boundaries) and uniformity is essential for the attainment of uniform photoconductive response along a filament of InSb.

To meet the above-mentioned requirements, the preparation of InSb, as with Ge and Si, is carried out in two stages: the production of high purity polycrystalline ingots and the growth of uniform single crystals from this material. For the purification, use is made of the principle of zone refining developed by Pfann [7]). In this process a molten zone is repeatedly passed in the same direction along a bar of the material to be purified. Impurities tend to be either more soluble in the solid than in the melt or vice versa and are swept to the one or the other end of the bar. The property of differing solubility in the solid and melt can be put on a quantitative basis by defining the distribution coefficient for a given impurity:

$$k_0 = \frac{\text{concentration of impurity in the solid}}{\text{concentration of impurity in the melt}}$$

for thermodynamic equilibrium conditions between the solid and melt at the molten zone. If $k_0$ is less than 1, impurities concentrate in the molten zone and are swept to the end of the bar which freezes last, while if $k_0$ is greater than 1 they tend to remain in the solid and the molten zone contains less impurities than the starting material. In this case, the end which freezes last contains the purest material. Zone refining is impossible if $k_0$ is unity.

### *Polycrystalline ingot preparation*

The starting materials for the preparation are commercial high purity indium and antimony, each containing about one part per million of impurities. Experiment has shown that, for the production of indium antimonide of the highest purity, the commercially pure *antimony* must be purified further by zone refining in a hydrogen atmosphere before the compound is prepared. During this zone refining, impurities (probably S or Se) which are not readily zoned out of InSb, are removed.

As regards the *indium*, zinc and cadmium are present in the commercially pure metal and must be removed as they are acceptors in InSb [8]). Both have $k_0$'s so near to 1 in InSb that zone refining is inefficient, but fortunatly they both have sufficiently high vapour pressures to allow them to be removed by evaporation. Most of the zinc and cadmium is removed by baking the indium under vacuum at 800 °C in the crucible in which the compound is prepared. The chemically equivalent quantity of antimony, correct to about 1%, is then added. The crucible is sealed off under vacuum and the contents fused together at 750 °C for some hours. After freezing, the compound is further purified by giving the ingot thirty zone passes through an eddy-current heater. During this process, not only are a large number of impurities concentrated at the two ends of the bar but also Zn and Cd are condensed on the upper part of the crucible which remains relatively cool.

Hall measurements show that 65% of the ingot has a difference of donor and acceptor impurity concentrations approximately equal to $10^{14}$ cm$^{-3}$. The remaining impurity has not been identified. Harman [9]) has suggested that it is tellurium originating in the indium but this has not been confirmed.

### *Preparation of single crystals*

Single crystals are pulled by the Czochralski method [10]), in which a rotating seed crystal is slowly withdrawn from a melt of the material. A typical crystal puller is shown in *fig. 5*. Crystals are grown under vacuum at a rate of 2.5 cm/hour and a rotation rate of 120 rpm. It has been observed that donor concentrations are markedly lower under vacuum than when pulling in a gaseous ambient. During pulling there is a relatively large loss of antimony (about 0.1%), but the solubility of indium in indium antimonide is evidently sufficiently low that the stoichiometry of the crystal is not appreciably affected. Undoped crystals, which are *N*-type, due to residual impurities, have an electron concentration of approximately $10^{14}$ cm$^{-3}$ and can have mobilities up to 650 000 cm$^2$/Vs at 77 °K as mentioned earlier.

Germanium is an acceptor in InSb and is used for doping in preference to Zn and Cd, which are volatile in vacuum. It has a $k_0$ of 0.02, giving an acceptor concentration varying by a factor of 2 during the pulling of the first 50% of the melt.

[7]) W. G. Pfann, J. Metals **4**, 347, 1952. See also J. Goorissen, Philips tech. Rev. **21**, 185, 1959/60 (No. 7).

[8]) J. B. Mullin, J. Electronics and Control **4**, 358, 1958.

[9]) T. C. Harman, J. Electrochem. Soc. **103**, 128, 1956.

[10]) J. Czochralski, Z. phys. Chem. **92**, 219, 1917. See also B. Okkerse, Philips tech. Rev. **21**, 340, 1959/60 (No. 11).

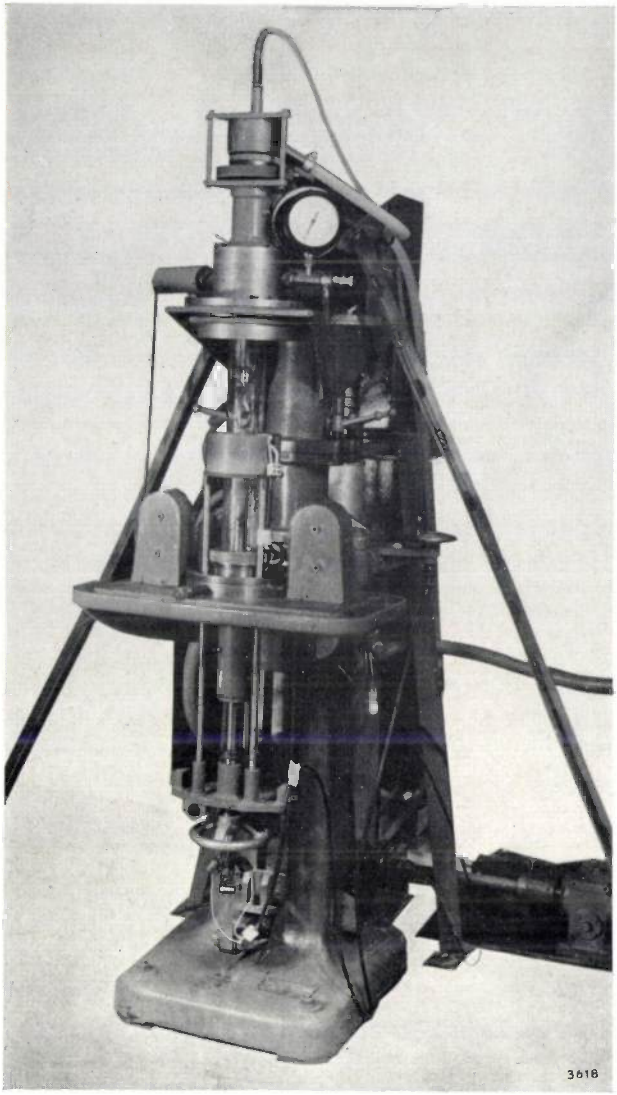


Fig. 5. Crystal puller used for production of single crystal InSb. The crucible is heated by radiation from a resistance coil not in contact with the crucible.

At doping levels greater than $10^{15}$ cm$^{-3}$ the resistivity across crystal slices is uniform to within about 25% on crystals pulled on the (111) plane, i.e. when a (111) face of the crystal is in contact with the melt. At lower doping levels, larger variations are observed, although sufficient uniformity can be maintained to produce material with a *P*-type resistivity at 77 °K of 10 Ωcm (i.e. hole concentration of $10^{14}$ cm$^{-3}$). More uniform crystals are, however, easily produced by pulling on the (211) plane, when resistivities up to 100 Ωcm with 10% variation across slices can be obtained. Hulme and Mullin [11]) have shown that the variation in resistivity in crystals pulled on the (111) plane is caused by the $k_0$ of the residual *N*-type impurity varying across the (curved) growing interface due to the presence of (111) facets. If the crystal is pulled on the (211) plane, (111) facets and consequent changes in resistivity do not occur.

*Fig. 6* shows a single crystal of InSb grown on a (211) plane.

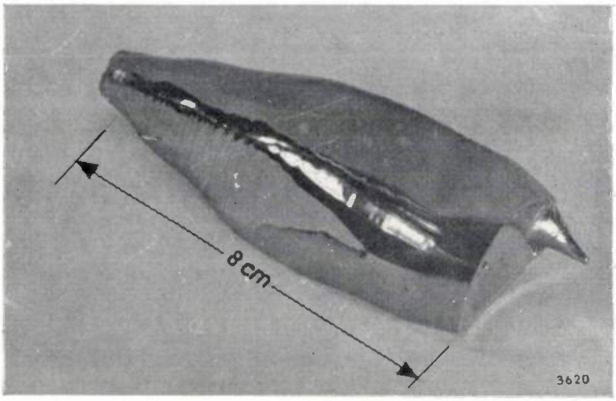


Fig. 6. Single crystal of InSb grown on the (211) plane. It can be seen that the cross-section is not circular. This is due to the different growth velocities in the different crystallographic directions.

## Applications of InSb

### *Photoconductive cells; general*

As mentioned above, the small energy gap of InSb makes it particularly suitable for use in infrared detectors. Photocells using InSb have been made utilising photoconduction, the photomagnetoelectric effect or the photovoltaic effect. The photoconductive cell is the simplest of these.

Two photoconductive cells will be discussed. The first of these, the ORP 10, is a detector which is operated at or near room temperature. This cell can be used to detect infrared radiation up to a wavelength of 7.5 μ. This extends the sensitivity range of the existing set of infrared detectors, the lead sulphide (PbS), lead selenide (PbSe) and lead telluride (PbTe) cells with their respective limits of 3.5 μ, 5 μ and 6 μ.

The second InSb cell, the ORP 13, is designed for operation at liquid-nitrogen temperature. This is a much more sensitive cell but it has a reduced spectral sensitivity, operating to 5.5 μ, which is comparable with the PbTe photocell.

The InSb cells have shorter time constants than those made from other materials. Thus they are particularly suited for applications where a fast response is required, e.g. military infra-red systems or fast automatic recording instruments.

Photoconductive detectors are usually operated by passing direct current through the sensitive filament or layer. The change in current resulting

[11]) K. F. Hulme and J. B. Mullin, Phil. Mag. 4, 1286, 1959.

from the increase in conductance during illumination is then amplified and measured. In order to facilitate amplification, the incident radiation is usually chopped at a suitable frequency, often near 1 kc/s.

When the performances of detectors at a given wavelength and in a particular system are compared, there are two cell characteristics which are important. These determine the overall signal-to-noise ratio of the two cases, (*a*) system noise large compared with cell noise, and (*b*) system noise small compared with cell noise.

For case (*a*) the performance is determined by the *responsivity*, defined as the detector output voltage per unit incident signal power.

In case (*b*) the *noise equivalent power* (N.E.P.) is important. This is sometimes termed minimum detectable energy. The N.E.P. is the incident radiation power for which the signal equals the cell noise. The N.E.P. is referred to a particular bandwidth, usually 1 c/s.

*The ORP 10 detector*

This infra-red photocell is illustrated in *fig. 7*. It consists of a 10 μ thick strip of InSb attached to a copper mount drilled to facilitate mounting on a heat sink. The sensitive area is a rectangle 6 mm× 0.5 mm. The dark resistance of the cell is 100 Ω, which is suitable for use with transistor amplifiers.

If, at room temperature, $I$ photons per unit area per second are incident on a filament of InSb of width $b$ and resistance $R$, the steady electric field in the filament being $E$, then the open-circuit signal voltage $V_s$ is given by

$$V_s \propto I\, e\,(\mu_n + \mu_p)\, \tau E R b\,, \qquad (4)$$

where $\tau$ is the lifetime of the charge carriers. If the lifetime in the material was independent of the carrier concentration the signal could be increased by increasing the resistivity, e.g. by doping the material with acceptors. However, it has been found that there is a decrease in lifetime which tends to offset any increase in resistance obtained by doping.

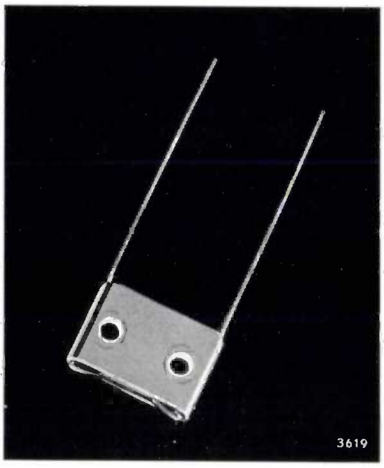


Fig. 7. The Mullard photoconductive cell ORP 10. The InSb element is visible on the edge of the drilled copper block. The leads to the InSb run though the copper block which thus serves as a heat sink during the soldering of connections.

Doping with acceptors increases the resistivity, which is given by the equation $\varrho = (ne\mu_n + pe\mu_p)^{-1}$ when both electrons and holes participate in the conduction process. Other conditions which hold are $np = n_i^2$ and $p - n = n_A$, where $n_i$ is the intrinsic carrier concentration and $n_A$ the acceptor concentration. With the aid of these relationships it can be shown that $\varrho_{\max}$ occurs when

$$n_A = n_i \left( \frac{\mu_n - \mu_p}{\sqrt{\mu_n \mu_p}} \right)$$

and *not* for $n_A = 0$, i.e. intrinsic material.

The noise is found to be always less than twice the Johnson noise whose voltage is denoted $V_n$. Now

$$V_s / V_n \propto \frac{(\mu_n + \mu_p)\,\tau}{\sqrt{\sigma}} \qquad . \quad . \quad . \quad (5)$$

for a given power dissipation in the filament. This quantity decreases with increased doping: intrinsic or near intrinsic material is therefore used for the cell.

In *Table III*, the characteristics of this cell are compared with those of other photoconductive cells.

The ORP 10 detector possesses the following special features:

1) Rapid measurements can be made of radiations with wavelengths up to 7.5 μ. In this wavelength range the cell therefore supplants thermal bolometer detectors, which are rather slow.

2) The cell is particularly suitable for *spectrometer* applications. The form of the InSb element — a narrow strip — and its mounting permit an array to be used and simultaneous observation of several bands seems feasible. Observations for extended periods of time may be made without complicated cooling arrangements. The wide wavelength range, which includes the main atmospheric absorption bands, permits the study of fundamental absorption bands of many chemical groups without the use of thermal bolometer detectors. The latter are not conveniently incorporated into automatic equipment.

3) The cell is sensitive to the thermal radiation from bodies at relatively low temperatures, e.g. the radiation from 1 cm² of a black body at 40 °C can be easily measured at a range of 40 cm.

**Table III.** Comparison between some photoconductive cells.

| Cell category | Type No. | Effective sensitive area | Spectral range | Peak response | Time constant | Dark resistance | Sensitivity, noise and figure of merit | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | (mm$^2$) | ($\mu$) | ($\mu$) | ($\mu$sec) | (k$\Omega$) | Radiation | Sensitivity | Noise equivalent power per unit bandwidth ($10^{-9}$ W) | Figure of merit $D^*$ (cm/$\mu$W) |
| PbS | 61 SV | 36 | 0.3 - 3.5 | 1.8 - 2.8 | 75 | 1000-4000 | Tungsten lamp at 2700 °K | 3mA/lumen | 0.055 | 11000 |
| | | | | | | | Black body radiation at 200 °C | 180 $\mu$V/$\mu$W | 5 | 120 |
| PbSe | 61 RV | 6 | 1.0 - 5.0 | 2.0 - 4.3 | $<$1.0 | 15 - 100 | Monochromatic at 4 $\mu$ | 15 $\mu$V/$\mu$W | $<$8.5 | $>$29 |
| InSb | ORP 10 | 3 | 0.6 - 7.5 | 5.0 - 7.2 | $<$1.0 | 0.1 | Monochromatic at 6 $\mu$ | 0.3 $\mu$V/$\mu$W | $<$4 | $>$43 |
| | | | | | | | Black body radiation at 200 °C | 0.36 $\mu$V/$\mu$W | $<$10 | $>$17 |
| InSb liquid-nitrogen cooled | ORP 13 | 3.5 | 0.6 - 5.5 | 4.5 - 5.0 | $<$10 | 20 - 40 | Monochromatic at 4 $\mu$ | 14 mV/$\mu$W | $<$0.02 | $>$9000 |
| | | | | | | | Black body radiation at 200 °C | 2.4 mV/$\mu$W | $<$0.12 | $>$1500 |

$D^* = (\mathrm{NEP})^{-1} \times (\mathrm{area})^{\frac{1}{2}}$ and is a figure of merit for photocells when used for detecting low-level radiation [12]). $D^*$, unlike NEP, is independent of photocell sensitive area and represents a sound basis for comparison of photocells of different area.

### *The ORP 13 cooled photoconducting cell*

Detectors with the same light-sensitive area as the ORP10 have been developed for operation at liquid nitrogen temperature, see *figs. 8* and *9.*

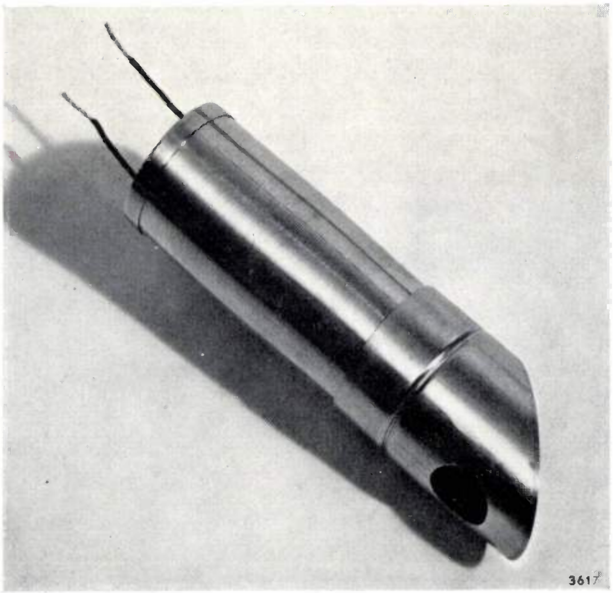


Fig. 8. The Mullard cooled InSb cell ORP 13, in its metal housing.

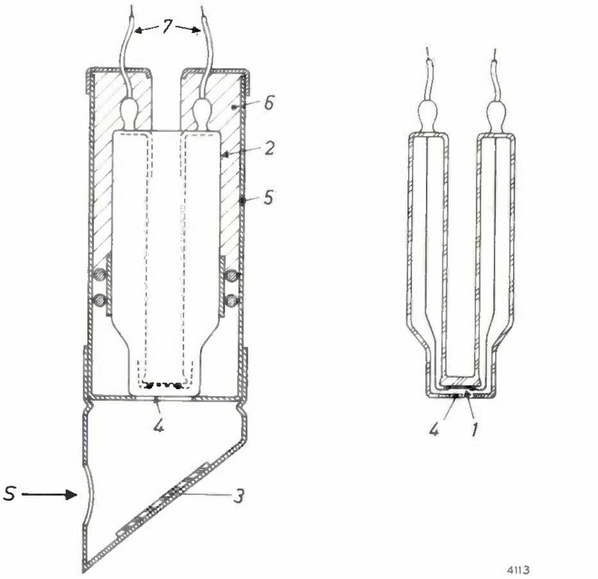


Fig. 9. Construction of the cooled InSb cell. The InSb element *1* is cooled by filling the Dewar flask *2* with liquid nitrogen. A demountable mirror *3* facilitates the measurement of horizontally incident radiation. A sapphire window *4* is sealed to the Dewar flask just in front of the InSb strip. The Dewar is housed in a metal tube *5* containing a resilient filling *6*. The leads *7* of the cell pass through glass-metal seals in the Dewar flask.

The values of the parameters in equation (4) are modified at this temperature. *P*-type material is used because the hole mobility is much less than the electron mobility — this gives high resistivity,

[12]) R. C. Jones, Proc. Inst. Radio Engrs. 47, 1495, 1959.

about 30 times the room temperature value, and therefore high cell resistance and large signals. The process determining the lifetime of excess holes and electrons is *electron* trapping. The signal voltage $V_s$ is given by

$$V_s \propto I\, e\, \mu_p\, \tau_p\, R\, E\, b, \quad . \; . \; . \; . \quad (6)$$

where $\tau_p$ is the hole lifetime.

At this temperature the resistivity would be *decreased* by doping because there are a negligible number of intrinsic carriers present. The product $\mu_p \tau_p$ would also be decreased by doping. The purest possible *P*-type material is therefore used.

In this cell the noise appears to be semiconductor "fluctuation noise" (fluctuations in the recombination process). Using *P*-type material of resistivity at 77 °K of up to 10 Ωcm the characteristics given in Table III are obtained.

After the cell has been exposed to visible radiation while cooled, a quasi-permanent change in the dark resistance takes place. The original resistance value may be re-attained by allowing the cell to warm to room temperature and then cooling it once more.

Owing to its great sensitivity this cooled detector is a useful addition to the present range of photoconductive cells.

### *Hall generators*

Indium antimonide has been quite widely used as the basis of Hall generators. These are devices in which the output signal is proportional to the product of two currents, either of which may be steady or variable. One of the currents is passed through a plate of the material (as in the measurement of the Hall effect, see page 219) and the other is fed into the winding of an electromagnet producing the magnetic field on the plate. The output voltage from the Hall probes is then proportional to the product of the two currents. A load may be inserted between the Hall probes and power drawn in the output circuit.

Many applications and refinements of these devices have been described in the literature. Considerations concerning this type of device will form the subject of a forthcoming article in this journal.

The large magnetoresistance effect in InSb has likewise led to a number of applications, including displacement gauges and tiny measuring probes for high intensity magnetic fields. In these devices a disc geometry is often used for the InSb element; this arrangement gives a large change in resistance when a magnetic field is applied. The magnetoresistance effect is proportional to the square of the carrier mobility, thus InSb is particularly suitable for such applications.

**Summary.** Indium antimonide is a compound with semiconducting properties. The small energy gap makes it a good photoconductor. This has led to the construction of photoconductive cells with either long-wavelength response (the Mullard ORP 10 to 7.5 μ) or high sensitivity (the Mullard ORP 13, cooled with liquid nitrogen, with a sensitivity of 14 mV/μW at 4 μ). Moreover the lifetime of the free charge carriers is very short, so that InSb cells are very fast: the time constant of the ORP 10 is <1 μsec, that of the ORP 13 is <10 μsec.

The electron mobility is observed to be very high (650 000 cm$^2$/Vs at 77 °K), making InSb also particularly suited for use in Hall generators. The large magnetoresistance effect in InSb has likewise led to a number of applications, including displacement gauges and tiny measuring probes for high intensity magnetic fields.

Very pure single crystals of InSb are prepared from previously purified indium and antimony. The compound is further purified by zone melting, after which single crystals are formed by pulling from the melt on the (211) plane. Doping with Ge (for example) gives *P*-type material, while without doping it is *N*-type because of residual impurities.

# SOLID-STATE RESEARCH AT LOW TEMPERATURES

## II. ELECTRON CONDUCTION IN METALS AND SEMICONDUCTORS [1]

by J. VOLGER.

536.48

*Following the previous article in this series, which was mainly introductory, the article below deals with various recent investigations on solids at low temperatures. Except for the first one, all these investigations were carried out at Philips Research Laboratories in Eindhoven.*

### Electrical conduction in metals at low temperature

When a metal is cooled from room temperature, its resistance at first decreases more or less linearly with temperature. However, upon reaching the temperature region below about 50 °K — the actual temperature differs from one metal to another — the resistance curve bends over and, at very low temperatures, becomes virtually horizontal (*fig. 1*). Near absolute zero the value of the resistance is not zero, but has a finite value known as the residual resistance. The magnitude of this residual resistance depends on the concentration of the impurities in the metal — the purer the metal, the lower the residual resistance — and also on the physical lattice imperfections. Theoretically, the resistivity of a metal is regarded as the sum of a component $\varrho'$, caused by thermal vibrations of the crystal lattice, and a component $\varrho''$, due to scattering of the electrons (regarded as wave packets) by the foreign atoms. The magnitude of $\varrho'$ obviously depends on the temperature, and becomes zero when $T$ approaches the absolute zero point of temperature. The residual resistance is thus entirely determined by $\varrho''$. Provided the impurity concentration is not unduly large, $\varrho''$ is independent of temperature (Matthiessen's rule). The value of $\varrho'$, apart from being temperature-dependent, is almost entirely governed by the nature of the metal and is not significantly affected by the impurity concentration.

We shall now discuss some examples of investigations concerning the contributions made to the resistivity by physical lattice imperfections (in aluminium and in copper) and by impurities (carbon in iron). To yield results, these investigations had to be carried out at such a low temperature that the value of the resistivity was primarily determined by $\varrho''$ (viz. at 20.4 °K, the boiling point of hydrogen at 1 atm).

### *The influence of physical lattice imperfections on the electrical resistivity of highly purified aluminium*

As our first example we shall briefly describe an investigation concerning aluminium [2]. The results obtained give a clear picture of the separate influences of the point defects (vacancies, interstitial atoms) and linear lattice defects (dislocations).

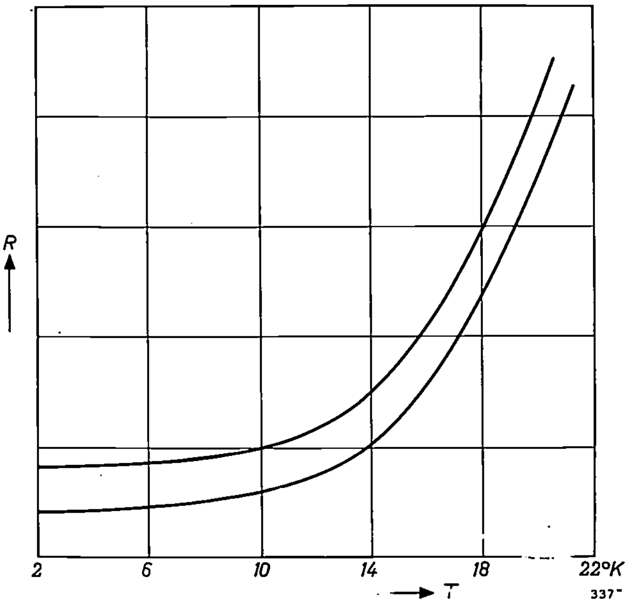


Fig. 1. The variation of the electrical resistance $R$ with temperature of two sodium samples of differing purity. Below the relatively low temperature of 10 °K, there is hardly any further decrease in $R$. The purer sample has the lower residual resistance (after D.K.C. MacDonald and K. Mendelssohn, Proc. Roy. Soc. A **202**, 103, 1950).

Three experiments were done. The first consisted of plastically stretching by 10% an annealed aluminium wire. (Annealing removes a considerable proportion of the lattice imperfections produced during the drawing process or during a previous elongation [3].)

[1] Sequel to the article: J. Volger, Solid-state research at low temperatures, I. Introduction, Philips tech. Rev. **22**, 190-195, 1960/61 (No. 6). This article is further referred to as I.

[2] M. Wintenberger, C. R. Acad. Sci. Paris **242**, 128, 1956 and **244**, 2800, 1957.

[3] For further particulars of this subject, see H. G. van Bueren, Lattice imperfections and plastic deformation in metals, Philips tech. Rev. **15**, 246-257 and 286-295, 1953/54.

The resistivity, measured at 20 °K, was found to have increased. After about two hours, part of the excess resistivity had disappeared, the other part proved to be permanent. From the two other experiments it was concluded that the temporary part of the excess resistivity was due to the point defects caused by stretching, which gradually disappear, and that the remaining part must be attributed to an increase in the number of dislocations. When the number of point defects (vacancies) was raised, *without* appreciably increasing the dislocation density — this can be done by heating the metal almost to the melting point and then quenching it rapidly in air — the excess was found after some time to have almost completely disappeared, even though it was originally roughly five times greater than in the first experiment. The fact that the *rate* at which the excess resistivity disappeared was much less here than in the first experiment is also in agreement with the explanation given above, since the "sinks" into which the point defects can vanish are, of course, the dislocations [4]), and the rate at which they vanish is clearly less for a smaller dislocation density. Confirmation of all this was obtained from a third kind of experiment, in which the wire was again first heated and quenched but afterwards stretched. The excess now proved to be somewhat greater than in the previous experiment — additional vacancies had now been created in *two* ways — but it decreased at the same rate as in the first experiment. Here again part of the excess was permanent, and it was roughly of the same magnitude as in the first case. Both the rate of decrease of the temporary part and the magnitude of the permanent part can be related to the dislocation density.

### *Magnetoresistance of copper*

It has been known for some time that the electrical resistance of a metal undergoes a slight change when the metal is subjected to a magnetic field (magnetoresistance). In the case of well-annealed samples of the same metal the fractional change in resistance $\Delta\varrho/\varrho_0$ plotted against $H/\varrho_0$ always yields the same curve, irrespective of the temperature and of the value of the residual resistance ($\Delta\varrho$ is the change in the resistivity, $\varrho_0$ is the resistivity in the absence of a magnetic field, and $H$ the magnetic field-strength). It has been found that, although this rule (Kohler's rule) is valid for the residual resistance due to impurities and other point defects, it does not hold for the contribution from the dislocations [5]). The results obtained from measurements of this effect are collected in *fig. 2*. The curves show the relation between the above-mentioned ratios as found for copper wire. It can be seen that the points for pure and impure copper indeed lie on the same curve, provided the wires have been annealed beforehand and then slowly cooled. After stretching (at 20 °K), however, the curve is seen to have shifted. On heating the wire to 220 °C — at which temperature the (physical) point defects are removed — no reduction in this effect is found. The effect can only be eliminated by *annealing*, that is to say by again reducing the dislocation density to the original value. This result, combined with that found from measurements of the recovery of the electrical resistance after stretching (in the absence of a magnetic field) led to the conclusion that Kohler's rule is valid only for the point-defect contribution to the resistance. The cause of this anomalous behaviour is thought to be the anisotropy of the scattering

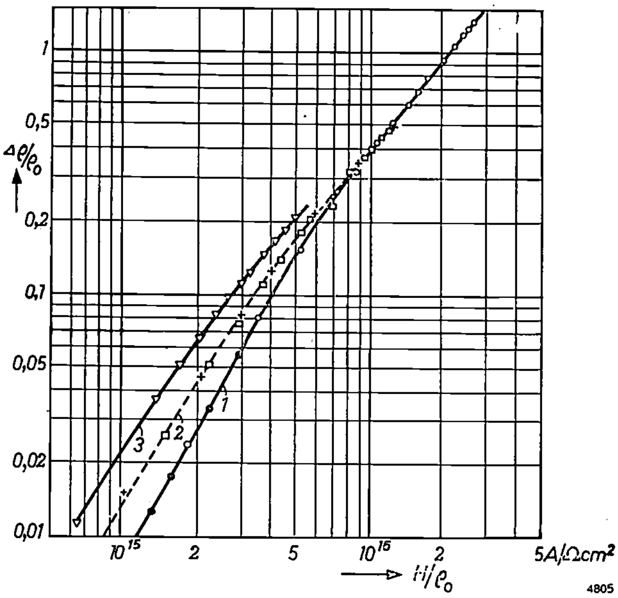


Fig. 2. The way in which the magnetoresistance $\Delta\varrho/\varrho_0$ of a metal depends on $H/\varrho_0$ ($H$ being the magnetic field) is not affected by point defects in the lattice, but it *is* affected by dislocations. This is illustrated here for copper. Curve *1* relates to annealed copper wire. The points for pure (●) and impure (○) material are seen to lie on the same curve. Curve *2* relates to pure copper wire which was plastically stretched 9.6% at 20 °K (points □) and also after subsequent heating at 220 °C (points +). It can be seen that curve *2* differs appreciably from curve *1*, particularly at low values of $H/\varrho_0$, and that this deviation is not eliminated by heating at 220 °C; to bring the wire back to resistance of curve *1* it has to be reannealed. Curve *3* relates to a wire of 0.5 mm diameter which, after annealing, was stretched to a diameter of 0.2 mm.

[4]) See e.g. B. Okkerse, A method of growing dislocation-free germanium crystals, Philips tech. Rev. 21, 340-345, 1959/60 (No. 11).

[5]) P. Jongenburger, Ned. Tijdschr. Natuurk. 22, 297, 1956 (in Dutch).

power of dislocations. A theory based on this supposition[6]) has in fact yielded an effect of the right order of magnitude.

Finally, a remark on the fact that, where the plastic deformations are not excessive, the anomalous effect disappears when $H$ is stronger than roughly $10^6$ A/m. The radius of curvature of the orbits described by the electrons then decreases to about $10^{-3}$ cm, which is smaller than their mean free path. In a strong magnetic field, then, an electron may thus make one or more complete revolutions between two collisions, and it is conceivable that this will attenuate or completely eliminate the anisotropic effect of the dislocations [7]).

### *Electrical resistivity of iron contaminated with carbon*

The way in which the residual resistance is affected by the impurity concentration is elegantly illustrated by the results of an investigation made on iron [8]). All resistivity measurements were done at 20 °K and at 20 °C (293 °K). First of all the ratio $\varrho_{293}/\varrho_{20}$ was determined for very pure iron [9]); the value found was 147 ± 2. If we compare this with the value 187 ± 5, found on a similar iron wire which was subsequently further purified by zone melting (11 ×), we get some idea of the considerable effect which the impurities have on the value of $\varrho$ at low temperature. It is to be hoped that means will be found of using this effect conversely for the quantitative chemical analysis of traces of impurities at concentrations which cannot be detected, or only with great difficulty, by other methods.

The way in which $\varrho''$ depends on the carbon concentration $C$ was found by measuring the resistivity of wires of different concentrations at the temperatures mentioned. Denoting the ratio of these resistivities, $(\varrho'_{20} + \varrho'')/(\varrho'_{293} + \varrho'')$, by $p_c$, and the ratio $\varrho'_{20}/\varrho'_{293}$ by $p_z$, we have that $\varrho''/\varrho'_{293}$ is equal to $(p_c—p_z)/(1—p_c)$. *Fig. 3* shows the results of the measurements. It is seen that the relation between $\varrho''$ and the concentration $C$ can be represented by a straight line passing through the origin. This leads to the conclusion that the contributions of the carbon atoms to the scattering of the conduction electrons are additive; there is apparently no mutual interaction between the neighbouring carbon atoms. In view of the small carbon concentration and the correspondingly large average distance between the carbon atoms, this result is not surprising.

It may further be deduced from fig. 3 that the proportionality factor between $\varrho''/\varrho_{293}$ and the carbon concentration (expressed in percentage by weight) is approximately 2.1 to 2.2. Measurements at a higher temperature, done by other workers [10]), yielded higher values. Matthiessen's rule, according to which $\varrho''$ is not dependent on temperature, is apparently valid only in a restricted temperature range. This was confirmed by measurements at 77 °K on the same iron wires used in the above investigation. These measurements resulted in a value of 2.55.

### *Superconductivity*

Certain metals and alloys, and also certain metallic nitrides and carbides, exhibit the effect of superconductivity when their temperature is reduced to an extremely low value. Although this was discovered half a century ago (by Holst and Kamerlingh Onnes in 1911), progress towards a complete fundamental theory of the phenomenon has only been made in recent years [11]).

A superconductive material shows two characteristic properties: 1) the electrical resistance is

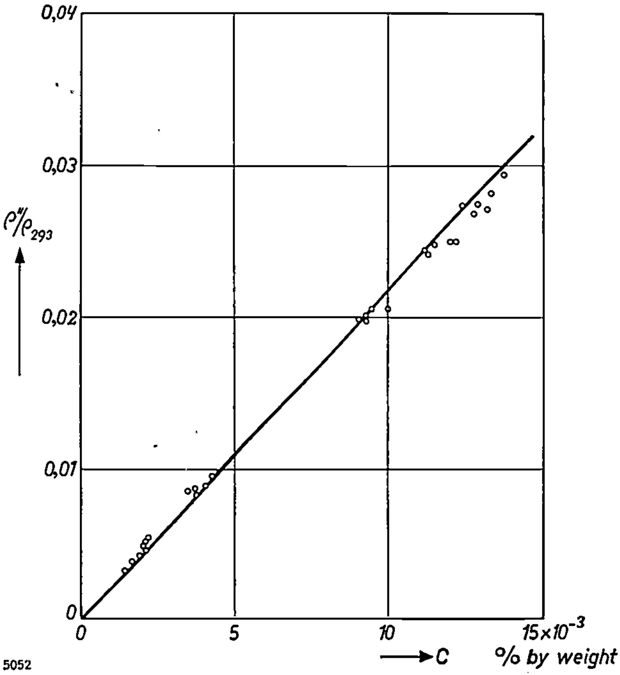


Fig. 3. The variation of the residual resistivity $\varrho''$ (expressed as a fraction of $\varrho_{293}$) with the concentration $C$ of carbon dissolved in iron. The relation between $\varrho''$ and $C$ is seen to be given by a straight line passing through the origin.

[6]) H. G. van Bueren, Philips Res. Repts. **12**, 1 and 190, 1957. See also reference [5]).

[7]) More recent measurements by Prof. Jongenburger will be published shortly in Acta Metallurgica.

[8]) Carried out by G. Baas of this laboratory.

[9]) For the method of preparation see J. D. Fast, A. I. Luteijn and E. Overbosch, Philips tech. Rev. **15**, 114, 1953/54.

[10]) W. Köster, in Arch. Eisenhüttenw. **2**, 503, 1928/29, and L. J. Dijkstra, in Philips Res. Repts. **2**, 357, 1947, found a value of 2.5 at 25 °C, W. Pitsch and K. Lücke, in Arch. Eisenhüttenw. **27**, 45, 1956, found a value of 2.75 at 22 °C.

[11]) See e.g. C. G. Kuper, Adv. Phys. **8**, 1, 1959 (No. 29) and I. M. Khalatnikov and A. A. Abrikosov, ibid. page 45.

zero (except in the case of alternating currents whose frequency exceeds a certain value), and 2) an external magnetic field penetrates only a very thin surface layer and is excluded from the bulk (the Meissner effect). The temperature below which a substance becomes a superconductor, the *transition temperature*, is changed by the application of an external magnetic field. The curve showing the relation between the transition temperature and the external field is shown in *fig. 4*.

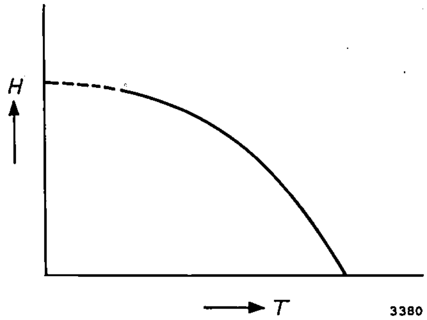


Fig. 4. Curve showing the transition point of a superconductor At a certain temperature $T$ the material is superconductive only if the strength of the external magnetic field $H$ is lower than the corresponding ordinate value of the curve.

Because of the complete absence of electrical resistance, there can be no electrical field inside a superconductor, for such a field would evidently give rise to an infinitely strong current. If a ring of superconducting material, placed in a magnetic field, is cooled to below the transition temperature corresponding to that field, and if the magnetic field is then switched off, a current is induced in the ring of such magnitude that the magnetic flux enclosed by the ring remains as it was. If the temperature is maintained below the transition point, this current persists for an unlimited time without becoming noticeably weaker. This persistent-current phenomenon has attracted attention in recent years, since it can be turned to use for storage elements in electronic computers.

The fact that the superconductivity can be removed and restored again by means of an external magnetic field is important in the design of storage elements, in connection with the write-in and read-out processes. But the phenomenon is of technical interest in itself, in that it can be used to make an electrical switching device which functions very roughly in much the same way as a triode [12]). The principle is illustrated in *fig. 5*. The "resistance" *a*, which is connected to the terminals *2* via the conductors *b*, is surrounded by a solenoid *c*. The temperature of *a* is kept just below the transition point, so that superconductivity occurs and the resistance between the terminals *2* is entirely determined by that of the leads *b*. If a current source is now connected to the terminals *1* which causes a current flow in *c* such that the resultant magnetic field removes the superconductivity of *a*, there will be an increase in the resistance between the terminals *2*. If the resistance of *a* is high in relation to that of *b*, the arrangement may be regarded as a switch.

The most suitable form of *a* is that of a straight strip of very small thickness. Its resistance $R$ must obviously be high, but at the same time its inductance $L$ as low as possible, since the speed with which the superconductivity state changes to the normal state, and vice versa, is greater the larger is the value of $R/L$. A long thin wire, which might be coiled to save space, is therefore out of the question.

In view of the extremely restricted choice of basic materials — there are not many superconductive materials, and boiling helium is the only suitable cooling bath — it was a long time before a strip could be made that adequately fulfilled the requirements. A particularly suitable material from the electrical viewpoint is tantalum. The transition point of bulk tantalum lies at 4.4 °K, which is only 0.2 °K above the boiling point of helium; this means that its superconductivity can be removed with a relatively weak magnetic field. If one tries, however, to make a thin strip of this material by vapour-deposition in vacuum, the strip usually acquires electrical properties so different from those of the bulk metal that it cannot be used. The reason is that tantalum is a good getter, so that the strip becomes strongly contaminated by gas absorption. Preparation of tantalum strips having virtually the same properties as the bulk metal only recently became possible by working at a pressure of about

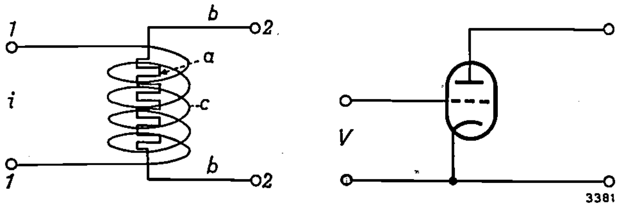


Fig. 5. Making use of the fact that the superconductivity can be removed by the application of a magnetic field, a switching device can be designed whose characteristic resembles that of a triode. An essential difference is that, whereas the triode is driven by a voltage, the device in question (known as a "cryotron") is driven by a current. The figure shows: *a* the superconductor, *b* supply leads and *c* the solenoid providing the magnetic field.

[12]) The first experiments in this field were done by D. A. Buck, Proc. Inst. Radio Engrs. **44**, 482, 1956.

$10^{-11}$ mm Hg [13]), achieved with a vacuum-pump system of the type recently described in this journal [14]). The use of tantalum is attractive not only because of the favourable situation of its transition point, but also because the vapour-deposited strips are quite hard and of good composition.

It should be emphasized that a superconductor differs essentially from a normal conductor of zero resistance. The fact that a magnetic field penetrates only very superficially into a superconductor is not simply to be explained from the disappearance of the resistance. According to the London-London phenomenological theory, it is necessary in order to describe the behaviour of superconductors to add to the Maxwell equations the expression:

$$\text{curl}\ \Lambda \mathbf{J}_s = -\mathbf{H}. \quad \ldots\ldots\ldots \quad (\text{II}, 1)$$

Here $\Lambda$ is a constant, H the magnetic field, and $J_s$ the current in the superconductor. For a superconductor this equation takes the place of Ohm's law. If we replace $J_s$ in (II, 1) by curl H (in accordance with one of the Maxwell equations), we arrive, after some manipulation, at the equation:

$$\Lambda\, \nabla^2 \mathbf{H} = \mathbf{H}. \quad \ldots\ldots \quad (\text{II}, 2)$$

For a superconductor with a plane surface, with H varying only in the direction perpendicular to that plane, (II, 2) becomes:

$$\frac{\partial^2 \mathbf{H}}{\partial x^2} = \frac{\mathbf{H}}{\Lambda}. \quad \ldots\ldots\ldots \quad (\text{II}, 3)$$

The variation of H is therefore given by:

$$\mathbf{H}_z(x) = \mathbf{H}_0 \exp\left(-x/\sqrt{\Lambda}\right), \quad \ldots\ldots \quad (\text{II}, 4)$$

where $H_0$ is the value of H on the surface ($x = 0$). Since the quantity $\sqrt{\Lambda}$, called the penetration depth, amounts to only about $10^{-5}$ cm, it follows from (II, 4) that the magnetic field scarcely penetrates at all into the superconducting material.

### Electron conduction in semiconductors at low temperature

As mentioned in I, the conduction which still occurs in some extrinsic semiconductors when the temperature is so low that there can be no electrons at all in the conduction band, has been attributed to a mechanism called "impurity band conduction", where the electrons are assumed to jump directly from one donor to the other. In this section we shall examine the considerations which led to the discovery of the nature of this remarkable effect — one of the many shown by semiconductors at low temperature. First we shall briefly recall the general experimental method of investigating electron conduction, starting from various theoretical aspects. To begin with the simplest case, we shall return for a moment to metals.

As is well known, electron conduction in metals may be described by the formula:

$$\sigma = ne\mu, \quad \ldots\ldots \quad (\text{II}, 5)$$

where $\sigma$ is the conductivity, $n$ the concentration of the electrons, $e$ their charge and $\mu$ their mobility. The latter quantity is the mean drift velocity of the electrons in unit electrical field, and is related to the mean free path $l$, the arithmetic mean velocity $v$ of the electrons and the electron mass $m$, as given by:

$$\mu = el/mv. \quad \ldots\ldots \quad (\text{II}, 6)$$

The conductivity can of course be found directly from the resistance and the geometry. To find the concentration $n$, and hence indirectly to arrive at the value of $\mu$, use is made of the Hall effect. This effect is a consequence of the force acting on the charge carriers (here electrons) moving in the conductor when it is subjected to a transverse magnetic field. This gives rise to a transverse potential gradient in a direction perpendicular both to the current and to the magnetic field. Between two points on the surface of the conductor, which would otherwise have the same potential, a potential difference is thus measured, whose magnitude $V_H$ in the case of a small bar of rectangular cross-section is given by:

$$V_H = A_H \frac{iH}{d}. \quad \ldots\ldots \quad (\text{II}, 7)$$

Here $H$ is the strength of the magnetic field, $i$ the current, $d$ the thickness of the bar in the direction of $H$, and $A_H$ is a constant, called the Hall coefficient. It can be shown that $A_H$ is equal to $1/ne$; by determining the value of $A_H$ we can thus find indirectly the value of $n$.

As stated, (II, 5) and (II, 6) apply only to metals, that is to substances where the mobility is governed entirely by the electrons with energies roughly equal to the Fermi energy. Returning to semiconductors, we see that this condition in their case is not fulfilled. The electron concentration is lower, the energy distribution is often described by Boltzmann statistics instead of by Fermi-Dirac statistics, and it may be necessary to regard the mean free path no longer as a constant but to take into account the way in which it depends on the energy. Having regard to all these considerations, we find that $A_H$ still satisfies an equation of the above-mentioned form, but that a numerical factor $f$ must be added, the value of which — between 1 and 2 — may differ somewhat depending on the circumstances. It is evident that,

[13]) J. F. Marchand and A. Venema, Philips Res. Repts. 14, 427, 1959.

[14]) A. Venema and M. Bandringa, Philips tech. Rev. 20, 145, 1958/59.

in order properly to interpret a measurement of the Hall effect, reliable assumptions are needed regarding the mechanism by which the electrons are scattered and regarding the statistics to be applied.

To approach a little closer to our goal, we shall now consider what changes the Hall effect may undergo where two conduction mechanisms are responsible for the motions of the electrons. It is then as if two currents were flowing in the semiconductor. The electrons of the one current possess the mobility $\mu_1$, those of the other the mobility $\mu_2$. The manner in which the total electron concentration $n$ is built up from the fractions $n_1$ and $n_2$, corresponding to the two currents, will depend on the temperature. (The value of $n$ itself is also temperature-dependent, but that is not relevant for our present purposes.) We may express the above mathematically as follows:

$$\sigma = \sigma_1 + \sigma_2 = n_1 e \mu_1 + n_2 e \mu_2 \quad . \ . \quad \text{(II, 8)}$$

and

$$n_1/n_2 = a(T). \quad . \ . \ . \ . \ . \quad \text{(II, 9)}$$

The densities of the two component currents are:

$$i_1 = \frac{\sigma_1}{\sigma_1 + \sigma_2}\, i \quad \text{and} \quad i_2 = \frac{\sigma_2}{\sigma_1 + \sigma_2}\, i \, .$$

It can be calculated that the Hall coefficient in this case is given by:

$$A_\mathrm{H} = \frac{\sigma_1\mu_1 + \sigma_2\mu_2}{(\sigma_1 + \sigma_2)^2}. \quad . \ . \quad \text{(II, 10)}$$

Putting $\mu_1/\mu_2 = b$, we can write (II, 10) as:

$$A_\mathrm{H} = \frac{1}{(n_1 + n_2)e} \, \frac{(ab^2 + 1)\,(a + 1)}{(ab + 1)^2} \quad . \quad \text{(II, 11)}$$

$$= \frac{1}{ne}\, \mathrm{f}(a,b). \quad . \ . \ . \quad \text{(II, 12)}$$

The variation of $A_\mathrm{H}$ with temperature is thus determined by the way in which $n$ and $\mathrm{f}(a,b)$ vary with temperature. Now, $n$ is of course a monotonic function of $T$, but $\mathrm{f}(a,b)$ is not. At a certain temperature the latter function shows a maximum, and consequently the same holds for $A_\mathrm{H}$; where a combination of conduction mechanisms occurs as in the model just described, the Hall coefficient in a particular temperature range may often be appreciably larger than outside that range.

If we assume that $b$ is not temperature-dependent, it follows from (II, 11) that the maximum value of $\mathrm{f}(a,b)$ is equal to $(b + 1)^2/4b$. The temperature at which this maximum occurs is then governed only by the way in which $a = n_1/n_2$ varies with temperature.

This remarkable behaviour of $A_\mathrm{H}$ has indeed been found in certain cases, including germanium [15]) and cadmium sulphide [16]). The impossibility of explaining this behaviour with any other model gave rise to the hypothesis that two conduction mechanisms must be operative in these semiconductors. In germanium the maximum value of $A_\mathrm{H}$ was found in a particular case to be about 100 times greater than the values found at higher and lower temperatures. From considerations that cannot be discussed here (see reference [16])), and the fact that the second mechanism, although characterized by a small electron mobility, makes a relatively large contribution to the current in the temperature range where the conduction band is virtually empty, it was concluded that this mechanism must be that understood by "impurity band conduction".

[15]) C. S. Hung, Phys. Rev. **79**, 727, 1950.

[16]) F. A. Kröger, H. J. Vink and J. Volger, Philips Res. Repts. **10**, 39, 1955.

**Summary.** The residual resistance shown by a metal at extremely low temperature is due to the scattering of the electrons by lattice imperfections. Plastic-deformation experiments on highly purified aluminium give a picture of the individual influence of point defects and dislocations. The magnetoresistance of copper is found to obey Kohler's rule only in so far as the residual resistance is due to point defects in the lattice: dislocations cause a deviation from this rule. The residual resistance of iron contaminated with carbon is proportional to the carbon concentration but not entirely independent of temperature (as it should be to obey Matthiessen's rule). The fact that the superconducting transition temperature depends on an external magnetic field is turned to use in a switching element made of vapour-deposited tantalum. In some extrinsic semiconductors, conduction occurs at low temperature as a result of the mechanism whereby the electrons jump directly from one donor to the other (impurity band conduction).

# A MAGNETIC JOURNAL BEARING

by F. T. BACKERS.

621.822.824:538.12:621.318.124

*The supporting of a rotating shaft in such a way that no material contact is made with the shaft can be of importance for technical applications. Cases in point are where the friction or wear must be particularly small, where contamination by lubricants is inadmissible or where the lubricating oil would decompose under the influence of radiation (in a nuclear reactor, for example). The article below considers the theory of a shaft which is held in suspension by magnetic fields, and compares the theory with the results of measurements which have been made on "magnetic bearings" of various dimensions.*

Rotating shafts are supported by bearings, and the necessary reactions are provided by material contact between shaft and bearing. The friction and wear which are the unavoidable consequence of this contact, are limited as much as possible by using lubricants.

The idea of reducing the friction and wear to nil by making the shaft "float" is attractive. One approach to this problem is based on levitation by the application of magnetic fields [1]). Those methods which rely on the use of a diamagnetic body or of a superconductor [1]) are for practical reasons not very useful. Also methods in which the supporting force comes from electromagnets are open to the objection that a continuous supply of energy is necessary for levitation. The only remaining alternative possibility, using permanent magnets to provide the support, has been described in a patent application [2]) but, as far as we know, has never been realized. In the Philips laboratory at Eindhoven, a study has been made of such magnetic bearings. A number have been constructed and measurements on them have verified the theory.

## Description of the magnetic bearing

The principle of the magnetic bearing considered here is illustrated in *fig. 1.* On the shaft $A$ are fixed a number of radially magnetized rings $B$, made of the ceramic permanent-magnet material ferroxdure I. Adjacent rings have opposite polarity: if the magnetization in the $p^{\text{th}}$ ring is directed *towards* the shaft, the magnetization in the $(p-1)^{\text{th}}$ ring and the $(p+1)^{\text{th}}$ ring is directed *away* from the shaft. In fig. 1, each of the two bearings has four of these shaft rings. A greater number can also be used; the theory is based upon an extremely large number.

The shaft, with its rings $B$, is placed in the field of a like number of larger rings $C$ of the same thickness, which are fixed in the bearing housing. These $C$ rings are also made of ferroxdure I and have alternate polarities which, however, are opposed to those of the corresponding shaft rings. That the resulting *radial* equilibrium is stable is seen from the following. When, as a result of external forces upon the shaft, the concentricity of the rings $B$ and $C$ is disturbed, the magnetic force at the place where the rings are closest is always larger than that at the diametrically opposite position. If each shaft ring is of opposite polarity to the corresponding outer ring, the resulting force tends to return the shaft towards the concentric position. Thus, as far as radial deviations are concerned, the shaft is in stable equilibrium. If opposing shaft and outer rings had the same polarity, there would still be an equilibrium possible, but it would be unstable.

During an *axial* deviation, a force develops which tends to make the deviation still greater, so that in an axial direction the equilibrium is unstable. This is in accordance with a theorem due to Earnshaw [3]), which can be formulated in the following way: a permanent magnet placed in the field of other permanent magnets cannot remain in stable equilibrium. However, the axial instability can be kept within bounds by letting the shaft abut a stop when the deviation reaches a certain value. Another method, in which material contact is excluded, will be discussed later (see Notes, *c*, page 237).

The theory will show that it is better to construct a bearing from a number of adjacent rings of opposed polarities instead of from one ring having the same polarity throughout.

[1]) A. H. Boerdijk, Levitation by static magnetic fields, Philips tech. Rev. **18**, 125-127, 1956/57.

[2]) German Patent Application B 30 042 dated 1954 (German Specification 1 017 871 dated 1957) by M. Baermann.

[3]) S. Earnshaw, Trans. Cambr. Phil. Soc. 7, 97-112, 1842. See also: J. C. Maxwell, A treatise on electricity and magnetism Clarendon, Oxford 1873, Part 1, pp. 139-141.

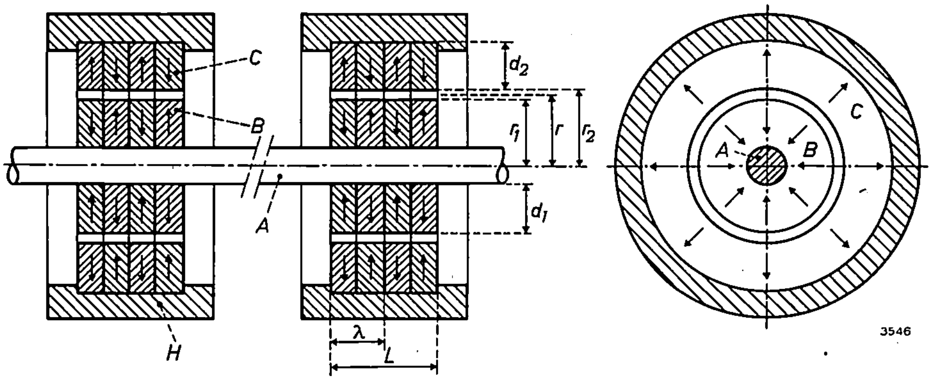


Fig. 1. Diagram showing the principle of a shaft with two magnetic bearings. The inner rings $B$ are fixed on the shaft $A$. This assembly can rotate in the field of the stationary outer rings $C$. All the rings are magnetized radially. Adjacent rings are magnetized in opposite directions; likewise corresponding inner and outer rings. $H$ housing.

## The theory of the magnetic bearing

In the magnetic bearing, two characteristic quantities occur: the radial carrying capacity $F_0$, being the external force upon the shaft which is necessary to bring the shaft rings into contact with the outer rings; and the radial stiffness $S$, denoting the force per unit displacement in a radial direction. These two quantities are calculated below.

We consider an infinitely large flat plate $C$ of ferroxdure having a thickness $d$ (*fig. 2*). The plate is so orientated in the system of coordinates shown that the upper surface lies in the plane $y=0$. The plate is magnetized in the $y$ direction, the strength of the magnetization $I$ ($= B - \mu_0 H$) being solely a function of $z$ and, as we will assume for the moment, a cosine function:

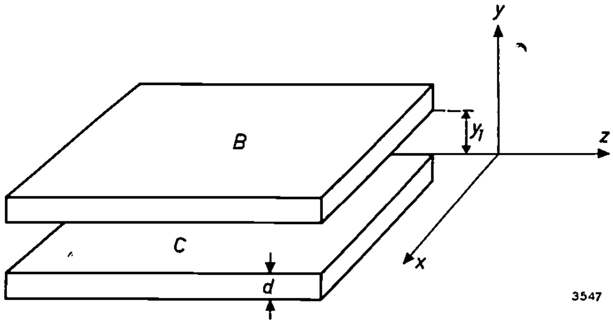


Fig. 2. Two plates $B$ and $C$ regarded as being infinitely large, both magnetized in the $y$ direction. The magnetization of plate $C$ is $I = I_0 \cos 2\pi z/\lambda$ and that of plate $B$ is $I = I_0 \cos 2\pi(z + z_0)/\lambda$.

$$I = I_0 \cos 2\pi z/\lambda .$$

This expression defines a periodically recurring distance $\lambda$ (the "wavelength") in the $z$ direction. Later, we shall consider the case when $I$ is another function of $z$.

An important quantity is the magnetic potential, i.e. the quantity from which the components $H_x$, $H_y$ and $H_z$ of the magnetic field are obtained by differentiating with respect to $x, y$ and $z$ respectively. The magnetic potential $U$ at a point $(x,y,z)$ above the plate can be calculated, albeit not simply, by means of the formula:

$$U(x,y,z) = \frac{1}{4\pi\mu_0} \int_V \frac{\mathbf{I} \cdot \mathbf{r}}{r^3} \, \mathrm{d}V .$$

Here, $\mu_0$ is the permeability of free space (equal to $4\pi \times 10^{-7}$ H/m), $\mathbf{r}$ the radius vector of the volume element $\mathrm{d}V$ at the point $(x,y,z)$, and $\mathbf{I}$ the magnetization vector. Integration over the volume of the plate yields:

$$U(x,y,z) = \frac{I_0 \lambda}{4\pi\mu_0} \{1 - \exp(-2\pi d/\lambda)\} \{\exp(-2\pi y/\lambda)\} \cos 2\pi z/\lambda .$$

By differentiation, the $y$ component of the magnetic field-strength is given by: $H_y = -\partial U/\partial y$.

Let $B$ be a second plate of ferroxdure parallel to the plate $C$ at a distance $y_1$ away from it (fig. 2). The plates $B$ and $C$ are assumed to be identical, except that the magnetization of $B$ is spatially displaced in phase with respect to that of $C$:

$$I = I_0 \cos 2\pi(z + z_0)/\lambda .$$

The potential energy of $B$ in the field of $C$ is now:

$$E = -\int_V \mathbf{I} \cdot \mathbf{H} \, \mathrm{d}V = -\int_V I \, H_y \, \mathrm{d}V ,$$

where the vector $\mathbf{I}$ is the magnetization of the plate $B$ and the vector $\mathbf{H}$ the field of the plate $C$. If we write $E(\lambda,w)$ to denote the energy of an element of plate $B$ of length $\lambda$ in the $z$ direction, of breadth $w$ in the $x$ direction and of thickness $d$, then the repulsive force to which $B$ is subjected in the field of $C$ is, per unit area:

$$\sigma_y = -\frac{1}{w\lambda}\frac{\partial E(\lambda,w)}{\partial y}.$$

In this way it is found that:

$$\sigma_y = -\frac{I_0^2}{4\mu_0}\{1-\exp(-2\pi d/\lambda)\}^2 \{\exp(-2\pi y_1/\lambda)\}\cos 2\pi z_0/\lambda.$$

If only $\lambda$ is varied in this expression, it appears that $\sigma_y$ approaches zero in the two limiting cases $\lambda \rightarrow 0$ and $\lambda \rightarrow \infty$. In the first case, the opposed polarities approach infinitely closely to one another so that they neutralize each other's effect. In the second case, the magnetic field is contained totally within the ferroxdure and there is thus no external field present; this can be seen when it is realized that $H$ in the infinitely large plate must be constant (lines of force parallel) and can therefore only have the value zero, since the lines of force outside the plate are infinitely long. The existence of two limiting cases suggests that somewhere in between there should be an optimum value of $\lambda$.

Since $\sigma_y$ is positive in the direction of the positive $y$ axis, a maximum *attractive* force will be obtained when the phase displacement $z_0$ is equal to zero, i.e. the magnetization of plate $B$ is spatially in phase with that of $C$ (this can be seen by inspection). Since, in the case of the magnetic bearing, a maximum *repulsive* force is required, $z_0$ is chosen equal to $\frac{1}{2}\lambda$; the magnetization of $B$ is then displaced by a half wavelength from that of $C$. Again, the thickness $d$ is assumed to be sufficiently large in order to make $\exp(-2\pi d/\lambda)$ negligible with respect to unity; $d = \lambda$, for example, makes this term approximately 0.002. These considerations lead to:

$$\sigma_y = \frac{I_0^2}{4\mu_0}\exp(-2\pi y_1/\lambda). \quad . \quad . \quad . \quad (1)$$

In order to transform the flat plates $B$ and $C$ into the bearing, we cut pieces of suitable dimensions out of the plates and bend these pieces into the form of hollow cylinders, the axes of which are parallel with the $z$ axis and whose radial direction corresponds to the original $y$ direction. The distance $y_1$ is now represented by the distance between the shaft rings $B$ and the outer rings $C$ of the bearing described earlier, and the length (dimension in axial direction) of a ring is $\frac{1}{2}\lambda$. It is assumed throughout that the repulsive force, acting upon an element of $B$, predominantly emanates from a limited area of $C$ in the immediate vicinity of that element of $B$. If the radius of curvature of the cylinder is large compared with $y_1$, then this small area of $C$ can be regarded as flat. The above theory for the "flat case" can then be applied to the cylindrical case, i.e. that of the bearing.

The radial force acting upon a surface element $Lr\,\mathrm{d}\Theta$, where $L$ is the length of the bearing and $r$ the mean radius (fig. 1), is now:

$$\mathrm{d}F_y = \sigma_y Lr\,\mathrm{d}\Theta. \quad . \quad . \quad . \quad . \quad . \quad (2)$$

Here, $\sigma_y$ is given by (1); the $y_1$ which occurs in (1) can be approximated (see *fig. 3*) by

$$y_1 = c + e\cos\Theta, \quad . \quad . \quad . \quad . \quad . \quad (3)$$

where $e$ is the eccentricity of the shaft and $c$ its *maximum* radial displacement which is equivalent to the difference of the radii of the opposing surfaces of the rings — in the following called the "clearance". The angle $\Theta$ is measured in fig. 3 from the point where $y_1$ is a maximum.

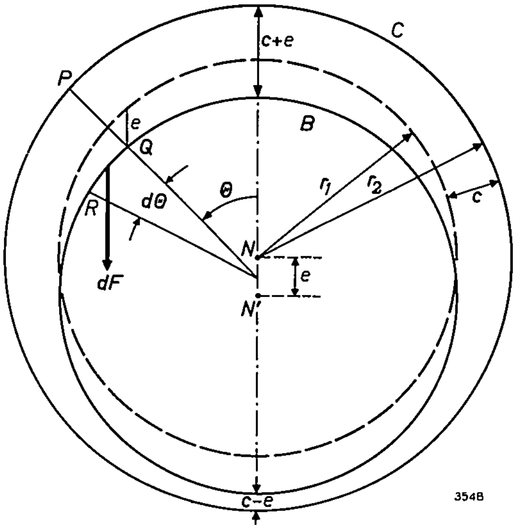


Fig. 3. Shaft eccentric with respect to bearing. $N$ is the centre of the outer rings, $N'$ the centre of the inner rings. The eccentricity is $e$. In the concentric position (dashed line) $N'$ coincides with $N$. If $r_1$ does not differ greatly from $r_2$, then, approximately:

$$y_1 = PQ = c + e\cos\Theta \text{ and } RQ = r\,\mathrm{d}\Theta,$$

where $r$ is the average of $r_1$ and $r_2$. $\mathrm{d}F$ is the component of the force acting upon the element $rL\,\mathrm{d}\Theta$, parallel to the displacement.

From considerations of symmetry it can be seen that, upon radial displacement, the resulting force cannot have any component perpendicular to the displacement. The contribution which the force acting on the element $Lr\mathrm{d}\Theta$ makes to the total restoring force, opposing the displacement, is therefore:

$$\mathrm{d}F = -\cos\Theta\,\mathrm{d}F_y. \quad . \quad . \quad . \quad . \quad . \quad (4)$$

Thus, for $-\frac{1}{2}\pi < \Theta < \frac{1}{2}\pi$ this contribution is negative, in other words it tends to increase the displacement, while the part of the bearing for which $\frac{1}{2}\pi < \Theta < \frac{3}{2}\pi$ makes a positive contribution. After integrating equation (4) over $\Theta$, and substituting from equations (1), (2) and (3), the total restoring force $F$ is found, this force being a maximum when the eccentricity $e$ is equal to the clearance $c$, and being then equivalent to the radial carrying capacity $F_0$ which was what we set out to calculate. We now have:

$$F = -A \exp(-b) \int_0^\pi \cos\Theta \exp(-b' \cos\Theta)\,\mathrm{d}\Theta,$$

where $A = LrI_0^2/2\mu_0$, $b = 2\pi c/\lambda$ and $b' = 2\pi e/\lambda$. The maximum value of $F$, i.e. the radial carrying capacity $F_0$, follows immediately by putting $b' = b$, i.e. $c = e$.

From a numerical calculation of $F_0$ as a function of $b$, it appears that $F_0$ has a fairly flat maximum for values of $b$ in the neighbourhood of 1. From this it follows that the relation

$$b = 2\pi c/\lambda = 1 \quad . \; . \; . \; . \; . \quad (5a)$$

defines an optimally dimensioned bearing. The axial length $\frac{1}{2}\lambda$ of each ring should thus be approximately three times the clearance $c$. The magnitude of the carrying capacity is then 0.655 $A$, i.e.

$$(F_0)_{\max} = 0.655\, LrI_0^2/2\mu_0 \, . \quad . \; . \quad (5b)$$

In *fig. 4* the restoring force $F$, as a percentage of $(F_0)_{\max}$, is reproduced as a function of $b'$ for various values of $b$. These curves represent the bearing characteristics. For values of $b$ approximately equal to 1 the characteristic is virtually linear, so that the radial stiffness $S$ can be represented by:

$$S = (F_0)_{\max}/c\,.$$

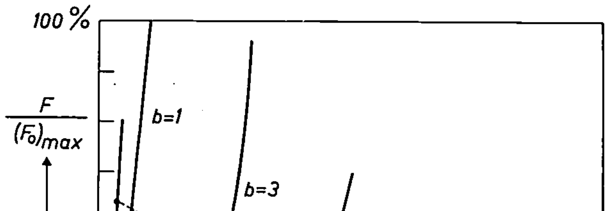


Fig. 4. Calculated bearing characteristics where the magnetization is assumed to vary according to a cosine function. The force $F$ (restoring force towards the concentric position) is represented as a percentage of the carrying capacity $(F_0)_{\max}$ and reproduced as a function of the eccentricity $e$ for values of the parameter $b$ ($= 2\pi c/\lambda$). For a given $b$, the largest possible value of $F$ is attained when $e$ is equal to the radial clearance $c$, i.e. when $b' = b$. In that case $F$ represents the load capacity $F_0$ of the bearing. The most favourable value of $b$ is unity, this giving a maximum value for $F_0$.

In the above, it is assumed that the magnetization varies according to a cosine function. The calculation can also be carried out for a discontinuously changing radial magnetization, i.e. for a square waveform:

$$I = +I_0 \quad \text{for} \quad 0 < z < \tfrac{1}{2}\lambda,$$
$$I = -I_0 \quad \text{for} \quad \tfrac{1}{2}\lambda < z < \lambda, \quad \text{and so on.}$$

In this case, the maximum carrying capacity is:

$$(F_0)_{\max} = 1.20\, LrI_0^2/2\mu_0, \quad . \; . \; . \quad (6)$$

in other words, about 1.8 times as large as in the case of (5b).

## Experimental results

A bearing was made in which $\lambda$ and $c$ could be varied, with a view to testing certain theoretical relationships. The fact that $b = 2\pi c/\lambda$ should optimally be equal to about 1 can easily be tested, there being no difficulty in measuring $c$ and $\lambda$. It is not so simple to check $(F_0)_{\max}$, however, since to do so we must know the magnitude of $I_0$, which is not easy to measure with the required accuracy. Furthermore, it is necessary to know the "type" of magnetization, e.g. whether it varies as a cosine or a square wave function. And lastly there is the question of whether $I_0$ is constant in the ferroxdure in a radial direction. Plainly, then, the experiments can do no more than give a rough check of the formula for the carrying capacity.

Three types of rings were made (80 of each type), comprising one type of outer ring $C$ and two types of shaft rings $B$ and $B'$ (*fig. 5*). At both ends of a shaft, which also carried a wheel, 40 shaft rings were mounted; two groups of outer rings $C$, each consisting of 40 rings, formed the fixed part of the bearings. The axial length of the rings is 1.5 mm; this defines the

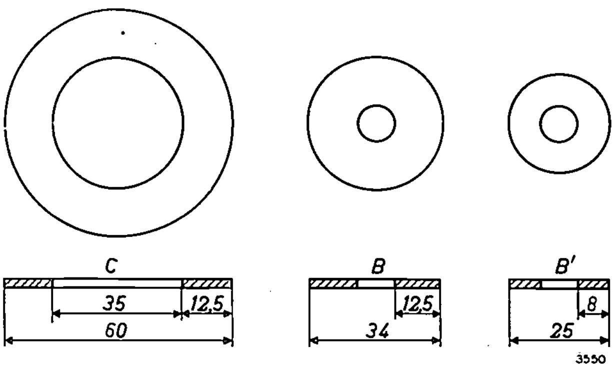


Fig. 5. Dimensions of the outer ring $C$ and two kinds of shaft rings $B$ and $B'$ used in the tests.

minimum value of $\frac{1}{2}\lambda$. If $n$ rings of like polarity are placed together, then $\frac{1}{2}\lambda = n \times 1.5$ mm.

The following combinations of rings were tried:

1) Rings $C$ and $B$ with alternate polarity, so that $\lambda = 3$ mm. The dimensions of the rings are such that the clearance $c$ is 0.5 mm, from which it follows that $b$ ($= 2\pi c/\lambda$) is $\sim 1$.
2) Rings $C$ and $B$ in batches of ten adjacent rings of like polarity: $\lambda = 30$ mm, $b \approx 0.1$.
3) Rings $C$ and $B'$. The $B'$ rings are of smaller diameter than the $B$ rings, such that $c = 5$ mm. Here, $\lambda$ is again 3 mm, so that $b$ is $\sim 10$.

In the experiments (1) and (2) the radius $r$ was about $17\frac{1}{2}$ mm; in all three cases the combined length of the two bearings was $L = 120$ mm.

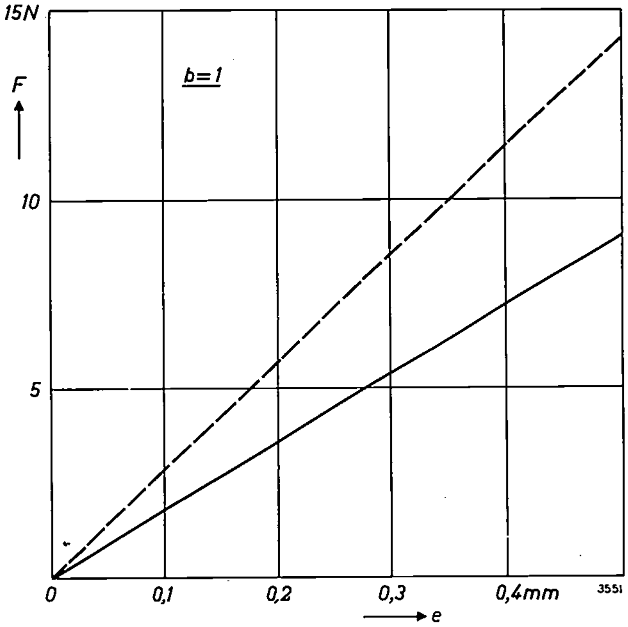


Fig. 6. Characteristic of an optimally dimensioned bearing ($b = 1$). The dashed characteristic was calculated for sinusoidal magnetization with $I_0 = 0.17$ Wb/m$^2$. The full line is the measured characteristic.

The characteristics of the three types of bearings (force $F$ in newtons as a function of the eccentricity $e$ in mm) are given in *figs. 6, 7, 8* and *9*; the full curves represent the experimental results and the dashed curves the calculated values. In fig. 9 particularly, it is clear that the radial carrying capacity and the radial stiffness are greatest for $b = 1$.

As far as the type of magnetization is concerned, a good approximation to a square form can be expected in case (2), where $\lambda$ is large, whilst in the cases (1) and (3) a cosine function would approximate more closely to the actual form. For case (2), fig. 7, the theoretical curves for both types of magnetization are given.

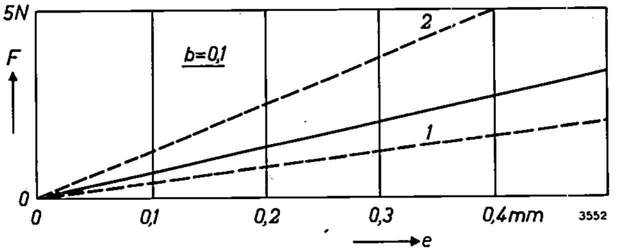


Fig. 7. Characteristic of bearing where $b = 0.1$. The dashed lines *1* and *2* are calculated characteristics with $I_0 = 0.17$ Wb/m$^2$ (*1* for sinusoidal magnetization, *2* for discontinuously changing magnetization). The full line is the measured characteristic.

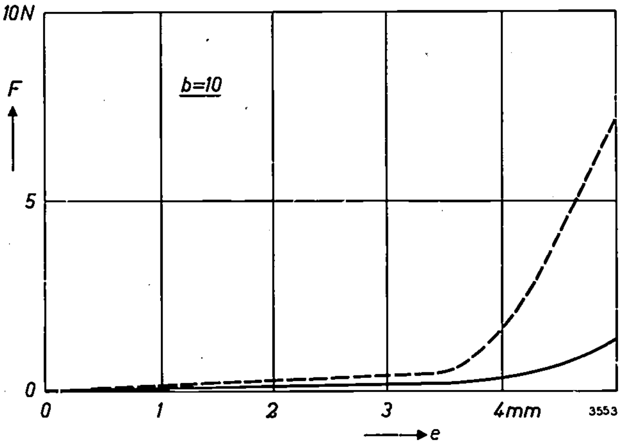


Fig. 8. Characteristic of a bearing where $b = 10$. The dashed line is the calculated characteristic for sinusoidal magnetization with $I_0 = 0.17$ Wb/m$^2$. The full line is the measured characteristic.

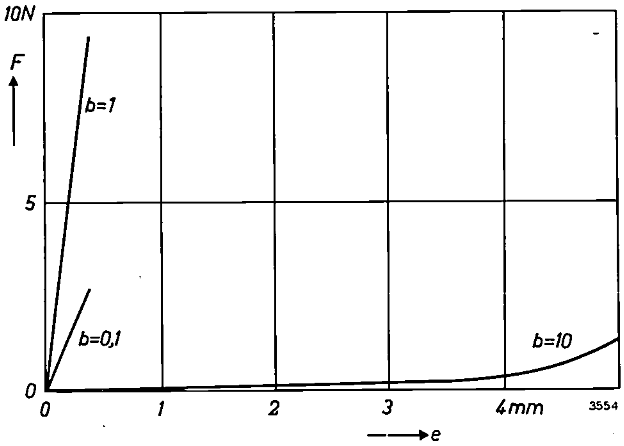


Fig. 9. The three measured bearing characteristics in one diagram. The bearing is optimally dimensioned when $b = 1$.

In the case of the rings $B$ and $C$ the radial thickness $d$ is approximately 12.5 mm and in the case of rings $B'$ approximately 8 mm. It follows from this that we are amply justified in neglecting exp ($-2\pi d/\lambda$) with respect to unity in the cases (1) and (3), but not in case (2), where $d$ is approximately $\frac{1}{2}\lambda$ and $\exp(-2\pi \times 12.5/30) \approx 0.07$. In the calculation for case (2) the term exp ($-2\pi d/\lambda$) is therefore not neglected.

If we now compare the theoretical and the measured characteristics with each other, we see first of all that in case (3), fig. 8, the divergence between the theoretical and the measured value of $F_0$ is very considerable. This is not surprising, since the theory supposes the clearance $c$ to be small compared with the average radius $r$, and this requirement is not fulfilled here. On the other hand, in cases (1) and (2), it appears that the differences between the theoretical and the measured curves are probably due to an error in the measurement of the magnetization. The latter was measured in various ways. The average of the measurements was: $I_0 = 0.17$ Wb/m$^2$ with a possible error of approximately 20%. Since $(F_0)_{\max}$ is proportional to $I_0^2$, this error can lead to one of approximately 40% in the final result. Assuming that the theory is correct and that in case (1) the magnetization followed a cosine function, it follows from the measured curve in fig. 6 that: $I_0 = 0.13$ Wb/m$^2$. This value does not deviate more than about 20% from the average measured value. If we choose this "corrected" value for calculating the theoretical curves in case (2), fig. 7, then the curve for a discontinuously changing magnetization more or less coincides with the measured curve.

It should be added that the measurements of $I_0$ are, in fact, measurements of the remanence or residual magnetization. This is only equal to the magnetization in the present situation (magnet in the field of another magnet) if the relative permeability $\mu_r$ is equal to 1, since an external field $H$ increases the magnetization by the amount $\mu_0(\mu_r—1)H$, which is zero only when $\mu_r$ is unity. This is approximately so in the case of ferroxdure, and therefore it may be assumed that the measurements give the approximate magnetization of the ferroxdure in the assembled bearing.

Another practical advantage of ferroxdure is its high electrical resistivity, so that very low eddy-current losses occur during the rotation of the shaft.

## Notes

*a*) Both the weight of the shaft ring and $(F_0)_{\max}$ increase linearly with the length $L$; however, as the radius $r$ of the bearing is increased, the weight increases quadratically whilst $(F_0)_{\max}$ increases only linearly. Now, in order to avoid using up the whole radial carrying capacity to maintain the weight of the shaft itself in the case of large horizontal shaft assemblies, we make use of the fact that although the radial thickness $d$ must be large enough (compared to $\lambda$), it does not have to be so much larger than $\lambda$. It is therefore possible to make use of hollow shafts of large diameter, fitted with rings of relatively small radial thickness. In this way, advantage is taken of the linear increase of $(F_0)_{\max}$ with respect to $r$, while the weight now increases much less than proportionally to the square of $r$.

With the types of bearing used in our experiments, the weight of the shaft complete with rings and wheel was just under 1 kg, i.e. almost as much as the carrying capacity of the bearing with optimum clearance ($b = 1$, fig. 6). By means of the measures just mentioned, the design can be improved to fulfil practical requirements. Again, a better material can be used (see the following note).

*b*) In the foregoing we have been concerned solely with radially oriented magnetization. The same formulae also apply, however, to axial magnetization of the same rings, although this will not be further amplified here. One advantage of axial magnetization is that the direction of $I$ is the same throughout each ring, making it possible to use crystal-oriented ferroxdure II, which has a higher remanence [4]); this is a great advantage since $(F_0)_{\max}$ is proportional to $I_0^2$. The preferred direction of magnetization in the ferroxdure II should then coincide with the axial direction of the rings.

*c*) It is assumed that the shaft and outer rings are directly opposite each other (see theory; $z_0 = \frac{1}{2}\lambda$). If this is not the case, the carrying capacity must be multiplied by cos $2\pi z_0/\lambda$, which indicates that

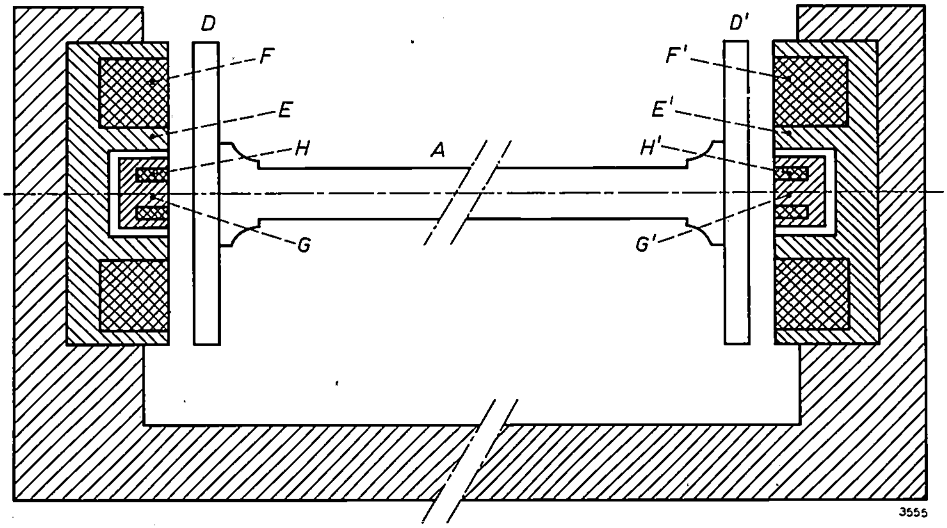


Fig. 10. System for stabilizing the shaft in an axial direction. $D$ and $D'$ are ferroxcube disks which can be attracted by electromagnets $E$ and $E'$, excited by coils $F$ and $F'$. The self-inductance of the coils $H$ and $H'$ on the cores $G$ and $G'$ is dependent upon the distance to the disks $D$ and $D'$. The magnetic bearings themselves are omitted here.

[4]) A. L. Stuijts, G. W. Rathenau and G. H. Weber, Ferroxdure II and III, anisotropic permanent magnet materials, Philips tech. Rev. 16, 141-147, 1954/55.

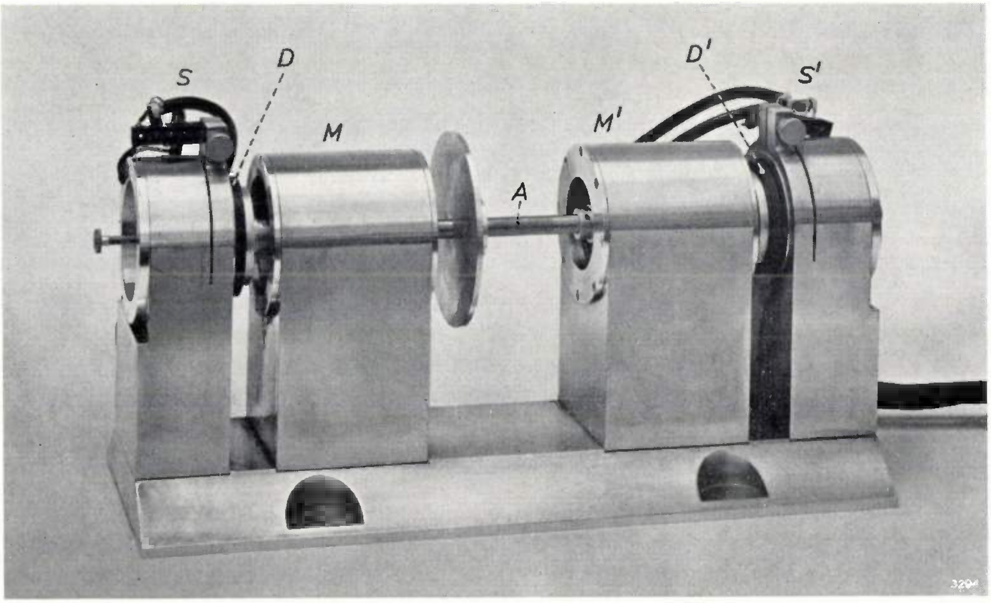


Fig. 11. Shaft $A$ with magnetic bearings $M$ and $M'$, ferroxcube disks $D$ and $D'$ and axial stabilizers $S$ and $S'$.

good positioning in the axial direction is important, especially when $\lambda$ is small. A method has been developed for locating the shaft in an axial direction without material contact. Use is made of an electro-mechanical servo-mechanism, the principle of which can be seen in *fig. 10* (the bearing itself being excluded). A photograph of the whole bearing with servo-mechanism is shown in *fig. 11*. $D$ and $D'$ are disks of ferroxcube, $E$ and $E'$ are electromagnets excited by coils $F$ and $F'$, while $H$ and $H'$ are coils having cores $G$ and $G'$. The self-inductance of each of the coils $H$ and $H'$ decreases as the distance to its opposing disk increases, which makes it possible to electrically "measure" the axial position of the shaft. $H$ and $H'$ form two arms of a bridge circuit (*fig. 12*); an error in the position of the shaft disturbs the balance of the bridge. Depending on the direction of displacement of the shaft from the central position, a phase-sensitive detector $D$ energizes the coil $F$ or $F'$ in such a way that the shaft is drawn back to the equilibrium position. A suitable network (not drawn) between the detector and the coils $F$ and $F'$ provides the necessary stability in the servo-system.

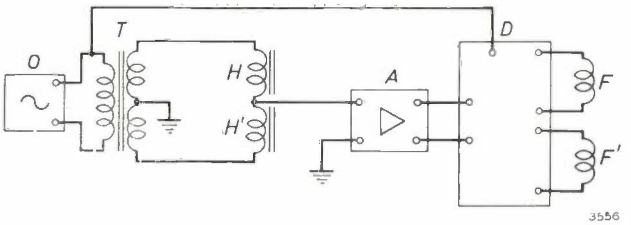


Fig. 12. The coils $H$ and $H'$ (fig. 10) form two arms of a bridge circuit which is fed via a transformer $T$ by an oscillator $O$. When the self-inductance of $H$ is equal to that of $H'$, the bridge is in balance. An axial displacement of the shaft unbalances the bridge; this causes the phase-sensitive detector $D$ to energize whichever of the coils $F$ and $F'$ is required to bring the shaft back to the central position. $A$ amplifier.

**Summary.** Bearing wear and friction of a rotating shaft can be avoided by avoiding all material contact between shaft and bearing. This may be done by letting the shaft "float" in a magnetostatic field. For this purpose "magnetic bearings" have been constructed, consisting of ring magnets fixed to the shaft situated within a set of outer stationary ring magnets. The rings consist of ferroxdure I and are radially magnetized. The direction of magnetization of adjacent rings is alternately directed towards and away from the shaft, and opposing inner and outer rings have mutually opposed directions of magnetization. In an axial direction, the alternating direction of magnetization defines a "wavelength" $\lambda$. It is found theoretically that the bearing is optimally dimensioned when the clearance between the inner and outer rings is equal to $\lambda/2\pi$; experiments confirm this result. The bearing is stable in a radial direction; it is stabilized in an axial direction by a simple electromechanical servo-system.

# VISUAL INSPECTION OF MOVING LAMP FILAMENTS ON A COILING MACHINE

by F. EINRAMHOF *).

621.397.331.2:621.326.032.321

In the mechanical production of coiled filaments for incandescent lamps the quality of the coils must be regularly inspected. It is not necessary to examine each coil separately, a sampling inspection being sufficient. The number of samples to be taken is quite considerable, however, for example about 1 in 30. Since the primary object of the inspection is to ensure that the machine is properly lined up and that the wire used is of good quality, the coil should preferably be examined immediately after it has been wound, i.e. before it is cut into separate filaments. This means that either the coils must be observed while they are moving, or the machine must be stopped for a moment for each inspection. The latter is obviously undesirable, and so the coils are examined in motion. In view of the minute details involved, a microscope is indispensable for this purpose.

Until recently the time taken by one coil to traverse the field of view of the microscope was long enough for a reliable appraisal. The latest coiling machines work so fast, however, that this is no longer the case. It was therefore necessary to find some means of artificially extending the observation time without depriving the inspection of its instantaneous character. This can be done in several ways. The choice ultimately fell on a method using closed-circuit television equipment. The principle will be illustrated with reference to *fig. 1*.

On the left the coil *1*, still on the mandrel, can be illuminated by a flash-tube *2* (circular or U-shaped) and a cylindrical mirror *3*. The microscope *4* is focused on the coil. Behind the microscope is mounted a television camera *6*, focused on infinity and connected via a power supply and control circuit *7* to a television receiver *8*. The cathode-ray tube screen is shielded from daylight by a visor *9* and is fitted with a yellow filter *10*. The observer presses a push-button switch *11* to actuate the circuit *12* for triggering the flash-tube. There then appears on the screen a picture of the coil which is under the microscope at the moment of the flash. In the absence of a flash, the picture tube is biased just below the threshold of illumination.

The picture remains visible for about 30 seconds, and it is bright enough during the first 10 to 15 seconds to be closely examined. On the other hand the light-flash is short enough to preclude movement blur. These advantages are due to the following features of the set-up. In the first place, the television camera is equipped with a vidicon, i.e. a photoconductive camera tube [1]). Owing to the slow decay of the charge pattern produced in the photoconduct-

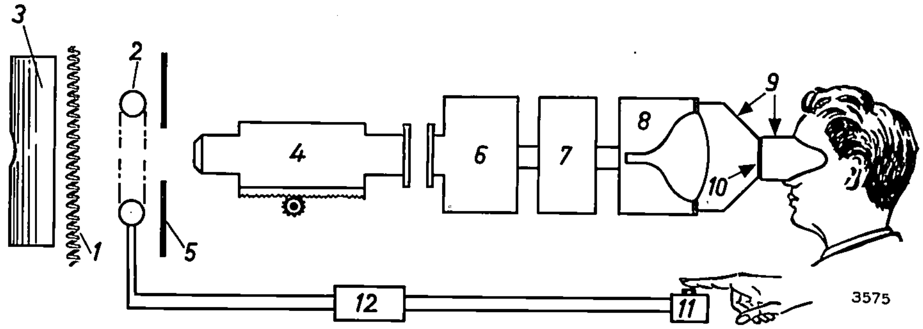


Fig. 1. Block-diagram of set-up for inspecting lamp filaments on the point of leaving the coiling machine. *1* coil on winding mandrel (not shown). *2* flash-tube (duration of flash 10 μsec). *3* mirror for concentrating the illumination. *4* microscope, focused on *1*. *5* screen preventing direct incidence of light on microscope objective (for the same reason the mirror has a hole in the middle). *6* vidicon television camera and pre-amplifier. 7 electronic circuit for power supply and vidicon control. *8* television receiver with long-persistence screen (30 sec afterglow). *9* visor to shield the screen from daylight. *10* yellow filter. *11* switch for actuating the flash-tube triggering circuit *12*.

*) Philips Lighting Division, Eindhoven.

[1]) P. K. Weimer, S. V. Forgue and R. R. Goodrich, R. C. A. Rev. **12**, 306, 1951. See also: L. Heijne, P. Schagen and H. Bruining, Philips tech. Rev. **16**, 23, 1954/55.

ing layer upon exposure, scanning is possible in the normal time of 1/25th sec. The flash does not have to be synchronized with the movement of the scanning beam, the position of the beam at the moment of the flash being unimportant. The extension of the observation time to the duration mentioned above is due to the use of a cathode-ray tube screen having an exceptionally long afterglow. This screen — a normal radar type — consists of two luminophor layers, one fluorescent and the other phosphorescent. The first is exicted into blue fluorescence by the electron beam. The second, excited by the light from the first, emits yellow light, which persists for a time after the fluorescence has ceased. It will now be clear why the observer looks at the screen through a yellow filter: the short-lived but intense blue light would otherwise dazzle him, making further observation very difficult.

Now a word about the illumination of the object. With a suitable flash-tube, the duration of the flash can be varied from a few microseconds to a few milliseconds by varying the inductance in the flash-tube circuit. If we choose a flash duration of 10 μsec, for example, and allow a movement blur of 0.5 mm on the screen — a better definition would be pointless — an object reproduced at its true size may then admissibly travel at a speed of up to 50 metres per second, or 180 km/h. Where a magnification of × 100 is required, as it is for the inspection of coiled filaments, the maximum permissible speed of the

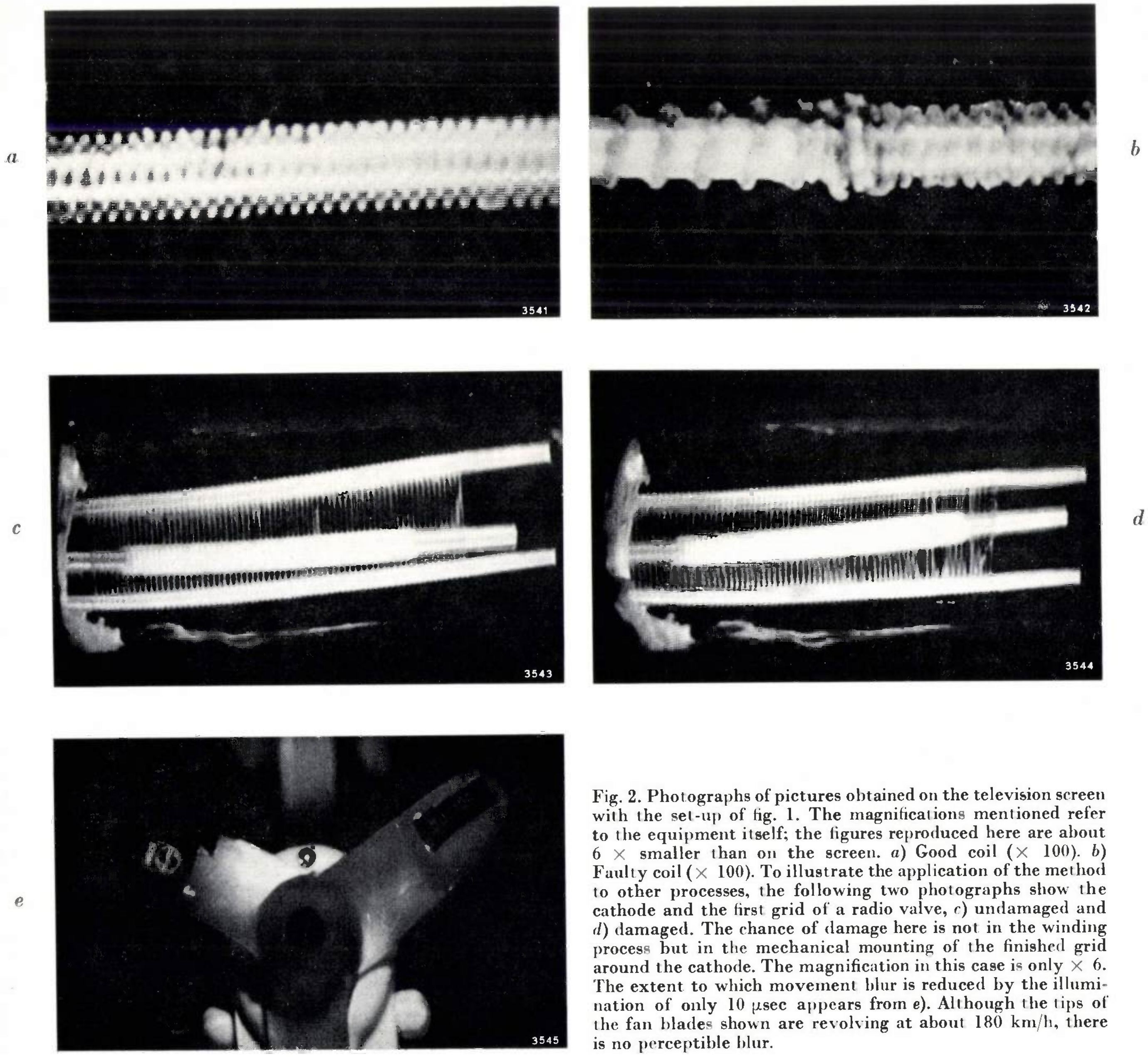


Fig. 2. Photographs of pictures obtained on the television screen with the set-up of fig. 1. The magnifications mentioned refer to the equipment itself; the figures reproduced here are about 6 × smaller than on the screen. *a*) Good coil (× 100). *b*) Faulty coil (× 100). To illustrate the application of the method to other processes, the following two photographs show the cathode and the first grid of a radio valve, *c*) undamaged and *d*) damaged. The chance of damage here is not in the winding process but in the mechanical mounting of the finished grid around the cathode. The magnification in this case is only × 6. The extent to which movement blur is reduced by the illumination of only 10 μsec appears from *e*). Although the tips of the fan blades shown are revolving at about 180 km/h, there is no perceptible blur.

object must be 100 times smaller, i.e. 50 centimetres per second. This is still quite a considerable speed.

The illumination of the object must obviously not be so intense as to overload the vidicon and to risk overdriving the amplifier behind it. A wide range of illumination intensities can be covered, however, if an iris diaphragm is incorporated in the optical system and the amplifier suitably adjusted.

It will be evident that the method described here can be used for examining all kinds of other rapid transients (periodic processes can best be observed stroboscopically). The microscope may then need to be replaced by some other optical system. As regards its general usefulness, the new method is certainly superior to the classic method, which uses a synchronously revolving mirror and where the object is continuously illuminated. Here, the problem of the synchronous movement of object and mirror has to be solved afresh for every new application. The photographic method compares unfavourably with the present technique because of its slowness — it takes at least a minute to develop a negative — and because of the high costs it entails in photographic material.

There are two other methods based on television techniques, one using a magnetic wheel store on which the picture information is recorded [2]), and the other a storage tube [3]). Compared with our method, however, they both call for much more expensive equipment. Elegant and universally applicable though these systems may be, their use is necessary and justified only in those cases where an inspection time of about 10 seconds is inadequate and where prolonged examination or storage of the pictures is required.

The photographs in *fig. 2* give some idea of the quality of the pictures obtained. Further particulars are mentioned in the caption.

[2]) See J. H. Wessels, A magnetic wheel store for recording television signals, Philips tech. Rev. **22**, 1-10, 1960/61 (No. 1).

[3]) A description of existing types of storage tubes is given by H. G. Lubszynski, in J. sci. Instr. **34**, 81, 1957.

Summary. A set-up is described for examining lamp filaments at the moment they leave the coiling machine, and are thus still in motion. A microscope, focused on the plane in which the coils are moving, produces an enlarged image which is viewed by a vidicon television camera connected to a television receiver. The object is illuminated by a flash-tube. The very short duration of the flash (10 μsec) precludes movement blur. Since the charge pattern on the photoconductive layer of a vidicon decays slowly, it can still be scanned in the normal time (1/25th sec). By using a cathode-ray tube screen of very long afterglow, the picture remains bright enough to be observed for 10 or 15 seconds. The method is potentially applicable to the observation of other moving objects provided an observation time of this order is sufficient.

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

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution.

H 2*: G. Schulten: Novel method for measuring impedances on surface wave transmission lines (Proc. Inst. Radio Engrs. **47**, 76-77, 1959, No. 1).

Brief description of methods of measuring the reflection coefficient of millimetre and sub-millimetre impedances using a microwave reflectometer.

H 3*: H. Severin: Zur Analogie akustischer und elektromagnetischer Randwertprobleme (Acustica **9**, 270-274, 1959, Akust. Beihefte, No. 1). (On the analogy of acoustic and electromagnetic boundary-value problems; in German.)

The conditions for the mathematical analogy of electromagnetic and corresponding acoustic boundary-value problems are investigated. Examples of complete and partial analogy are given. Finally some scalar potential functions are compiled, which are solutions of acoustic boundary-value problems and may be applied to the treatment of corresponding electromagnetic problems.

H 4*: D. Gossel: Die Korrektur des Phasenfehlers von *RC*-Gliedern in der Umgebung ihrer Grenzfrequenz (Arch. elektr. Übertr. **13**, 525-529, 1959, No. 12). (The correction of phase distortion of *RC* networks in the neighbourhood of their cutoff frequency; in German.)

*RC* networks lose much of their high-pass or low-pass characteristics if they have to be designed for small phase distortion in thc pass band. In this case the distance between transmission-band limit and cutoff frequency becomes very large. A large attenuation results if the networks are used for differentiation or integration and if small phase distortion is required at that. This disadvantage can be greatly reduced if a larger phase distortion is permitted and the latter is then wide-band-compensated by a phase-correcting network. Examples of such phase-correcting networks and the conditions for optimum design are given. For the optimum corrected network and under the condition that flat basic attenuation of 6 dB is tolerated, it is shown that the distance between transmission-band limit and cutoff frequency (high-pass or low-pass filter) or the damping factor (differentiator or integrator), respectively, can be reduced by a factor $2\sqrt[3]{2\psi^2}$, where $\psi$ is the overall phase distortion permitted.

H 5*: F. Karstensen: Über die Diffusion in Germaniumkristallen, die eine Korngrenze enthalten (Z. Naturf. **14a**, 1031-1039, 1959, No. 12). (On diffusion in germanium crystals containing grain boundaries; in German.)

Investigation into the diffusion of donors and acceptors along low-angle tilt boundaries in germanium. The diffusion is examined by measuring the displacement of the *P-N* junction which marks the position of equal donor and acceptor concentrations. The dislocations forming a low-angle tilt boundary act as "diffusion pipes", whereby the diffusion is much faster in the direction of the dislocations than in the normal lattice. Perpendicular to the dislocations this is not so. Diffusion along the dislocations is investigated for As and Sb, for various times and temperatures. It appears that more than one diffusion coefficient is required to describe the diffusion along the grain boundary. From the measurements, and a very rough calculation based on Whipple's formulae, it appears that diffusion along the dislocations is about $10^5$-$10^6$ times faster than in the normal lattice. The diameter of the dislocation pipe is assumed to be six lattice spacings.

H 6*: H. Severin: Sommerfeld- und Harms-Goubau-Wellenleiter im Bereich der Zentimeter- und Millimeterwellen (Arch. elektr. Übertr. **14**, 155-162, 1960, No. 4). (Sommerfeld and Harms-Goubau guides in the cm and mm wave region; in German.)

In addition to well-known numerical results for Sommerfeld and Harms-Goubau guides in the region of decimetre and metre waves the numerical evaluation is extended to centimetre and millimetre waves. Field extent and attenuation as functions of frequency and line data (wire radius, thickness and permittivity of the dielectric coating) are discussed by reference to numerous examples. With tolerable values of field extent, attenuation factors are found that are much smaller than those of hollow metal waveguides of the same frequency range. However, below 5 cm wavelength Sommerfeld and Harms-Goubau guides cannot be used for long-distance transmission if a maximum attenuation of 3.5 dB/km is allowed.

H 7*: G. Schulten: Messung der Eigenschaften von dielektrischen Leitungen bei Millimeterwellen in einem optisch angekoppelten Resonator (Arch. elektr. Übertr. **14**, 163-166, 1960, No. 4). (Measurement of the properties of dielectric rods in the mm-wave region in an optically coupled resonator; in German.)

Dispersion and attenuation of the $HE_{11}$-mode of the dielectric rod have been measured using the resonator method. The coupling of the resonator has been effected according to optical principles. The coupling element is a nearly transparent mirror consisting of a grid of dielectric threads. Measurements have been made at wavelengths in the 5 and 8 mm region. The dielectric guides were polyethylene threads of various diameters. The deviations of the guide wavelength from the free space wavelength were between $10^{-1}$ and $10^{-3}$%, the attenuation constant in the order of 0.1 dB/m while the radial field extent was about 70 mm.

H 8*: H. Severin: Neuere Entwicklungen der Mikrowellenphysik (Naturwiss. **47**, 217-221, 1960, No. 10). (New developments in microwave physics; in German.)

Discussion of two recent developments in microwave technique. The first concerns the use of ferrites in waveguides for microwave isolators. The phenomena occurring are discussed qualitatively. The author refers to the spin waves that can occur in ferrite materials and to their use for measuring the energy of exchange interactions. The second development concerns parametric amplification which, compared to other methods of microwave amplification, offers the advantage of a lower noise. The principle of parametric amplification is explained and a practical form of such an amplifier using a semiconductor diode is described. Other possibilities of realizing parametric amplifiers are mentioned.

**A 19:** A. Klopfer: Das Omegatron als Partialdruckmesser (Advances in vacuum science and technology, Proc. 1st int. congress on vacuum techniques, Namur, June 1958, edited by E. Thomas, Vol. 1, pp. 397-400, Pergamon, Oxford 1960). (The omegatron for the measurement of partial pressures; in German.)

A description is given of an omegatron with noble-metal electrodes, which when applying a suitable electrostatic field and with the right kind of cathode enables partial pressures of gases and vapours in the pressure region below $10^{-5}$ mm Hg to be measured with an accuracy of 10%. The sensitivity of the tube remains constant with time even after prolonged exposure to chemically active gases and vapours such as $H_2O$, $CO_2$ and $CH_4$, and does not depend on the particular tube used as long as the geometrical dimensions are maintained. A comparison of the ionization probabilities taken from the literature with the values calculated from the calibration curves of the omegatron shows that practically all the resonance ions produced by the electron beam are caught by the ion collector. The adjustment of the operating parameters to achieve this is in general independent of the mass.

**A 20:** E. Baronetzky and A. Klopfer: Einfluss von Gasreaktionen in Vakuumsystemen auf die Zusammensetzung des Restgases (as **A 19**; pp. 401-403). (The influence of gas reactions in vacuum systems on the composition of the residual gas; in German.)

Residual gases in vacuum systems may change their composition either on account of decomposition or because of reaction with components or impurities. The measuring methods often have a marked influence. Care must therefore be exercised when assessing such results. By means of some examples it is shown how the effects of pressure and temperature manifest themselves. Measurements on the kinetics of the decomposition of methane with various cathodes are reported.

**A 21:** Š. Garbe: Restgasanalysen mit dem Omegatron (as **A 19**; pp. 404-409). ( Residual-gas analysis with the aid of the omegatron; in German.)

The advantage of the omegatron as compared to other mass spectrometers lies in the possibility of performing gas analyses at very low pressures in a relatively small volume that is cut off from the pump. A description is given of the construction of a glass high-vacuum apparatus, with metal taps, for measurements of gas desorption and for partial-pressure analyses. In the case of permanent gases it was possible to measure gas-desorption rates of a few $10^{-10}$ torr l/min. The difficulty of determining small amounts of strongly adsorbing gases is explained by the example of water vapour. On account of the chemical reactions at hot surfaces the kind of cathode used in the omegatron and also in ionization manometers has a decisive effect on the result of the analysis. As an example of a residual-gas analysis it is shown how the residual gas above the barium-getter layer of a vacuum tube with an L-cathode alters during operation. (See also Philips tech. Rev. **22**, 195-203, 1960/61, No. 6.)

**A 22:** A. Klopfer and W. Ermrich: Erfahrungen mit Titan-Ionenpumpen (as **A 19**; pp. 427-429). (Experiences with titanium ion pumps; in German.)

By means of gas analyses with the omegatron, the suitability of the titanium ion pump for use as a high-vacuum pump was investigated. To obtain very low pressures, it became apparent that careful degassing is essential, as for ion and Hg diffusion pumps. Effects that determine the final pressure attainable and the pumping time are reported. (See also Philips tech. Rev. **22**, 260-265, 1960-61 (No. 8.)

**A 23:** E. Baronetzky: Ein neuartiger, metallischer Getterstoff (Vol. 2 of book mentioned under **A 19**, pp. 646-647). (A new metallic getter material; in German.)

The addition of silver and other noble metals enables the gettering properties of thorium-aluminium alloys at room temperature, in particular the autocatalytic gettering of hydrogen after preliminary oxidation with pure oxygen, and the adsorption of carbon monoxide to be considerably increased. In the $Th_2(Al,Ag)$ system the $Th_2Al$ and $Th_2Ag$ form a continuous series of mixed crystals.

**A 24:** P. Eckerlin and A. Rabenau: Die Struktur einer neuen Modifikation von $Be_3N_2$ (Z. anorg. allgem. Chemie **304**, 218-229, 1960, No. 3/4). (The structure of a new modification of $Be_3N_2$; in German.)

A new hexagonal modification of $Be_3N_2$ is formed by heating the known cubic form to temperatures above 1400 °C. The transformation is influenced by silicon compounds. The crystal structure of the new modification has been determined by single-crystal X-ray photographs. The space group is P $6_3$/mmc. The dimensions of the unit cell containing 2 formula units are $a = 2.841$ Å and $c = 9.693$ Å. There are two kinds of coordination for the Be atoms, viz.

triangular and tetrahedral, the N atoms being then surrounded by five and six atoms, respectively.

A 25: H. G. Grimmeiss, R. Groth and J. Maak: Lumineszenz- und Photoleitungseigenschaften von dotiertem GaN (Z. Naturf. **15a**, 799-806, 1960, No. 9). (Luminescence and photoconductive properties of doped GaN; in German.)

A description is given of a method for the preparation of GaN, which offers the advantage of low working temperature and facilitates doping with a large varicty of elements. The luminescence properties of such GaN preparations have been investigated as a function of the preparative conditions, and the emission bands produced by the doping have been determined. In the case of doping with Zn, Cd and Li a level scheme is proposed based on infra-red quenching of fluorescence and the maximum of the emission bands. Glow curves permit of an explanation of the short-wave emissions (so-called satellites) as a trap emission of impurity-containing GaN. In addition a method is described for producing single crystals of GaN, whose photoconductivity was investigated.

**Now available:**

P. A. Neeteson: Junction transistors in pulse circuits (Philips Technical Library, 1959, pp. viii + 139, 105 figures and 4 plates).

This book forms a companion to the earlier book by the same author, "Vacuum tubes in pulse technique" (Philips Technical Library 1955, second edition 1959). In pulse circuits tubes or transistors are used as switches. Since the transistor is a better approximation than a tube to the ideal switch, transistor pulse circuits are much simpler than the corresponding tube circuits. To keep the treatment brief and simple, the physical background of transistor operation is omitted. The seven chapters of the book are entitled: 1. Introduction; 2. Survey of fundamental pulse circuits; 3. Pulse generators; 4. Pulse shapers; 5. Frequency divider and voltage-level switch; 6. Some auxiliary pulse circuits; 7. Some logic circuits.

The book has also appeared in German.

Harley Carter: An introduction to the cathode ray oscilloscope, second edition (Philips Technical Library, popular series, 1960, pp. 121, 99 figures).

This book explains the operation and design of the cathode ray oscilloscope in non-mathematical language. It is addressed to technicians and shop engineers who are not experts in electronics, and will also appeal to the serious amateur and hobbyist. The chapter headings are as follows: 1. Introduction; 2. The cathode ray tube; 3. The time base; 4. Amplifiers for vertical deflection and pick-ups for converting non-electrical phenomena into electrical magnitudes; 5. Power supply for cathode ray oscilloscopes; 6. Practical applications of the oscilloscope; 7. Standard cathode ray tubes for oscillography; 8. Some complete oscilloscope circuits.

E. Rodenhuis: Hi-Fi amplifier circuits (Philips Technical Library, popular series, 1960, pp. x + 105, 64 figures).

Until a few years ago the high-fidelity reproduction of sound was an ideal attainable only by enthusiasts with expensive equipment. The situation is now considerably improved in that the price of quality has dropped very considerably, which has brought Hi-Fi within the reach of a much wider circle.

The book noticed here is a companion to "Electron tubes for A.F. amplifiers" by the same author (Philips Technical Library, popular series, 1960). It deals in detail with a number of amplifier circuits of very high quality which can be built at a reasonable cost. The three chapters of the book are: 1. General considerations on the design of Hi-Fi amplifiers; 2. Power amplifier circuits; 3. Pre-amplifiers.

P. van der Ploeg: Industrial electronics apparatus — steps in design and maintenance (Philips Technical Library, popular series, 1960, pp. xi + 97, 20 figures, 33 plates).

The object of this book is to show how the trouble-free operation of electronic equipment is dependent on details of its design, manufacture, use and maintenance. The book includes many practical tips and hints of value to both designers and service engineers. The chapters are: 1. The function of the equipment; 2. The laboratory test; 3. The prototype; 4. Production; 5. Installing the equipment; 6. The purpose of maintenance; 7. Maintenance; 8. Fault finding. A supplement, "Electronic tube data", at the end of the book forms a brief guide to the use of tube characteristics, operating data and limiting values.

VOLUME 22, 1960/61, No. 8 pp. 245-280 Published 10th May 1961

Philips Technical Review

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

# AN APPARATUS FOR AUTOMATICALLY PLOTTING ELECTRON TRAJECTORIES

by J. L. VERSTER.

537.533.3:621.317.729.1

*The electrolytic tank is a versatile tool for determining the paths of electrons in an electric field (e.g. in an electron tube or an electrostatic lens). Generally, the procedure is to use the tank to determine the equipotential surfaces, from which the electron trajectories are then construed step by step. The apparatus described here, which is based on a principle put forward by Gabor and Langmuir, traces out the trajectories automatically. A trolley, riding on a drawing board over the tank, is controlled by the voltages from a number of probes (in this case four) in such a way as to cause a stylus to plot the trajectory on the board.*

When designing the electrode system of an electrostatic lens, for example, or an electron gun for a cathode-ray tube, it is important to be able to predict the motion of the electrons in the system.

For this purpose one can determine the potential distribution in the system (e.g. with an electrolytic tank or a resistance network) and from this construe the electron trajectories. This method, however, is very cumbersome. The apparatus described in this article traces out the electron trajectories automatically. The field strength and the voltage, quantities that are needed for calculating the trajectories, are measured with an electrolytic tank. A trolley to which a stylus is attached rides on a drawing board above the tank. From the measured quantities at each position the apparatus computes the curvature which the track of the trolley must have at that position in order for the stylus to trace out the trajectory of an electron.

## The electrolytic tank

We shall first briefly describe the electrolytic tank as used for determining the potential distribution in an electrode system [1]). The tank is filled with a weakly conducting liquid — the electrolyte (usually water) — in which a model of the electrode system is placed. When voltages are applied to the electrodes of the model, a potential distribution is produced in the electrolyte which is independent of the nature of the dielectric, and thus identical with the distribution that would be found in a vacuum.

Use is now made of a similarity rule which states that the shape and relative potential of the equipotential surfaces do not change when the voltages between the electrodes are multiplied by an arbitrary factor, or when the electrode system is made proportionally larger or smaller. Making use of this, it is possible to work with conveniently low voltages, lower by a suitable factor than in the actual system. The factor need not remain constant, that is to say it is possible to use alternating voltage. This is in fact done in most cases, with the object of avoiding polarization at the electrodes. The dimensions of the electrodes are generally chosen larger than in the actual system, which improves the accuracy of the plot.

For practical reasons, the electrolytic tank is used almost exclusively for measurements in symmetry planes of an electrode system (see the first of the two articles mentioned under reference [1]). To this end the model is mounted in the tank in such

[1]) See e.g. K. F. Sander and J. G. Yates, The accurate mapping of electric fields in an electrolytic tank, Proc. Instn. Electr. Engrs. **100 II**, 167-175, 1953. See also: G. Hepp, Measurements of potential by means of the electrolytic tank, Philips tech. Rev. **4**, 223-230, 1939; N. Warmoltz, Potential distribution at the igniter of a relay valve with mercury cathode, Philips tech. Rev. **8**, 346-352, 1946; F. Reiniger, The study of thermal conductivity problems by means of the electrolytic tank, Philips tech. Rev. **18**, 52-60, 1956/57.

a way that its plane of symmetry coincides with the surface of the electrolyte (*fig. 1*). The potentials in this plane are measured with a pin probe dipped in the electrolyte.

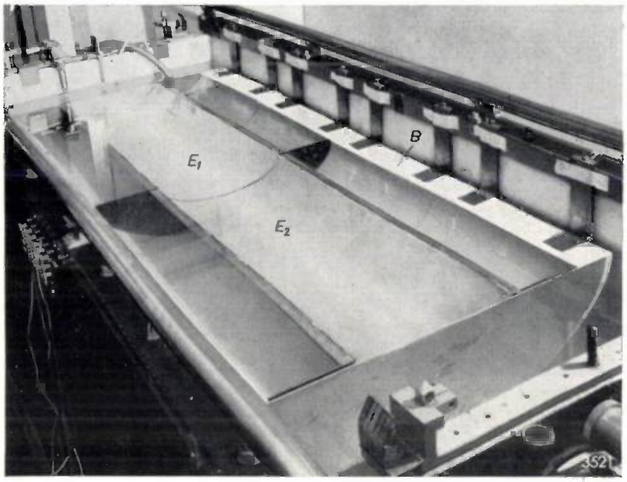


Fig. 1. The electrolytic tank containing a model of an electrostatic lens consisting of two electrodes. The system possesses rotational symmetry; the measurement is made in a plane through the axis of the system. The model must be mounted in such a way that this plane coincides with the surface of the liquid. For this purpose the electrodes $E_1$ and $E_2$ rest in a semicylindrical plastic trough *B*, which is suspended from the edge of the tank and can be adjusted vertically by means of screws.

### Principle of the automatic plotter

When a particle of mass *m* and charge $-e$ is accelerated from rest by a potential difference *V* to a velocity *v*, its kinetic energy is given by $\frac{1}{2}mv^2 = eV$. The curvature of the trajectory is determined at every point by the instantaneous velocity of the particle and the force $F_n$ acting upon it in a direction perpendicular to the trajectory. The radius of curvature $\varrho$ is:

$$\varrho = \frac{mv^2}{F_n} = \frac{2eV}{eE_n} = \frac{2V}{E_n}, \quad \ldots \quad (1)$$

where *V* is the potential difference between the point of interest and the place where the velocity of the particle was zero, and $E_n$ is the field-strength component perpendicular to the trajectory at the position of the particle.

The trajectory is plotted with the aid of a three-wheel trolley which rides on a drawing board above the tank. A sketch of the trolley is given in *fig. 2*. The stylus is located under point *O* on the rear axle. The trolley is propelled by an electric motor which drives the front wheel. The angular position of the front wheel is changed by turning the steering shaft *St* (perpendicular to the plane of the drawing). If the distance from the steering shaft to the rear axle is *h*, and the angle between front and rear axles is $\alpha$, the stylus describes a circle whose radius is

$$\varrho = \frac{h}{\tan \alpha}. \quad \ldots \ldots \quad (2)$$

The stylus thus traces the trajectory of the particle, provided the following condition is satisfied:

$$\frac{h}{\tan \alpha} = \frac{2V}{E_n},$$

or, otherwise expressed,

$$\alpha = \tan^{-1}\left(\frac{hE_n}{2V}\right). \quad \ldots \quad (3)$$

At any given point, then, $\alpha$ must be adjusted in accordance with the values of $E_n$ and *V* at that point.

This principle was independently described by D. Gabor [2]) and D. B. Langmuir [3]) as early as 1937. They determined the potential difference *V* and the field strength $E_n$ from the potentials of two probes. In our case, use is made of four probes in line, which results in greater accuracy, the curve which the potential variation describes along the line through the probes now being approximated by a third-degree function instead of by a linear function. The probes are mechanically linked with the trolley in such a way that they always remain vertically below the rear axle. They are mounted symmetrically with respect to the stylus at equal distances apart (see fig. 2). Voltages proportional to *V* and $E_n$ are applied to a servo system driving a servo motor

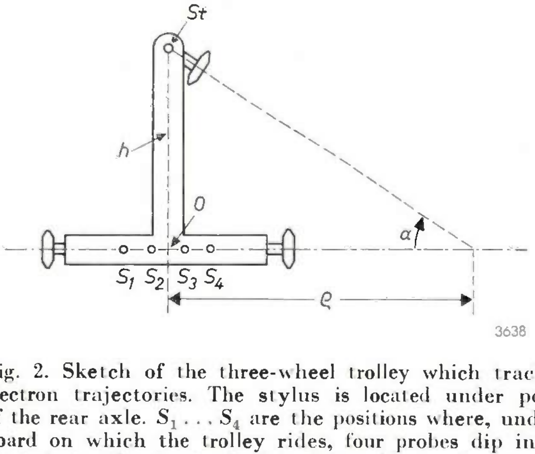


Fig. 2. Sketch of the three-wheel trolley which tracks the electron trajectories. The stylus is located under point *O* of the rear axle. $S_1 \ldots S_4$ are the positions where, under the board on which the trolley rides, four probes dip into the electrolyte. The front wheel is driven by an electric motor. The steering shaft *St* (vertical shaft which turns the front axle) is controlled by a servo motor so as to cause the stylus to trace out the trajectory.

[2]) D. Gabor, Mechanical tracer for electron trajectories. Nature **139**, 373, 1937.

[3]) D. B. Langmuir, Automatic plotting of electron trajectories, Nature **139**, 1066-1067, 1937, and An automatic plotter for electron trajectories, R.C.A. Rev. **11**, 143-154, 1950.

which controls the angular position of the steering shaft of the trolley.

It is possible to start plotting a trajectory at any arbitrary point. The stylus of the trolley is then simply placed above that point. The direction of the initial velocity at which the electron travels at the starting point can also be arbitrarily chosen by facing the trolley in a particular direction. (The tangent of the trajectory coincides with the line *h* in fig. 2, which thus gives the direction in which the electron is moving.) The speed of the trolley — which is of course in no way related to the velocity of the electron — may be freely chosen; the trolley can also be stopped and reversed for the purpose of checking the trajectory described.

### General description of the apparatus

A schematic diagram of the complete equipment is shown in *fig. 3*. An *RC* generator delivers a sinusoidal alternating voltage of about 500 c/s. This is supplied via a phase shifter *FD*, whose function will presently be explained, to a power amplifier *VV*, which delivers a square-wave voltage to an output transformer, the secondary of which is floating. Connected to this secondary are four potentiometers, so that from the total voltage, which is applied across two electrodes of the model, four variable intermediate voltages are available for applying to up to four other electrodes of the model. The voltages of the four probes *S* are supplied via cathode followers *KV* to two computing circuits. Circuit *VR* determines from these four voltages the potential *V* midway between the probes (i.e. at the position of the stylus). It further delivers the voltage —*V*, which is also required for the servo system. Circuit *ER* determines from the four probe voltages the field strength $E_n$ at the position of the stylus.

The servo system consists of a potentiometer circuit, an amplifier and a two-phase servo motor. The potentiometer circuit is mounted on the trolley *W*, and consists of two potentiometers and some fixed resistors. This circuit is supplied with the voltages *V*, —*V* and $gE_n$ (*g* is a constant having the dimensions of length, and thus $gE_n$ has the dimensions of voltage; the method of calculating *g* is discussed below). The wipers of the potentiometers are mechanically coupled to the steering shaft on the trolley. The circuit is so designed as to give no output signal when the steering shaft is correctly aligned. If there is a deviation, the potentiometer circuit delivers a voltage $V_p$, which, via the amplifier *SV*, actuates the servo motor *SM*. For the sake of accuracy it is necessary (see later) to use the voltage $V_p$ only during a small part of each cycle. The phase of this part is adjusted with the aid of the second output voltage from the phase shifter *FD*.

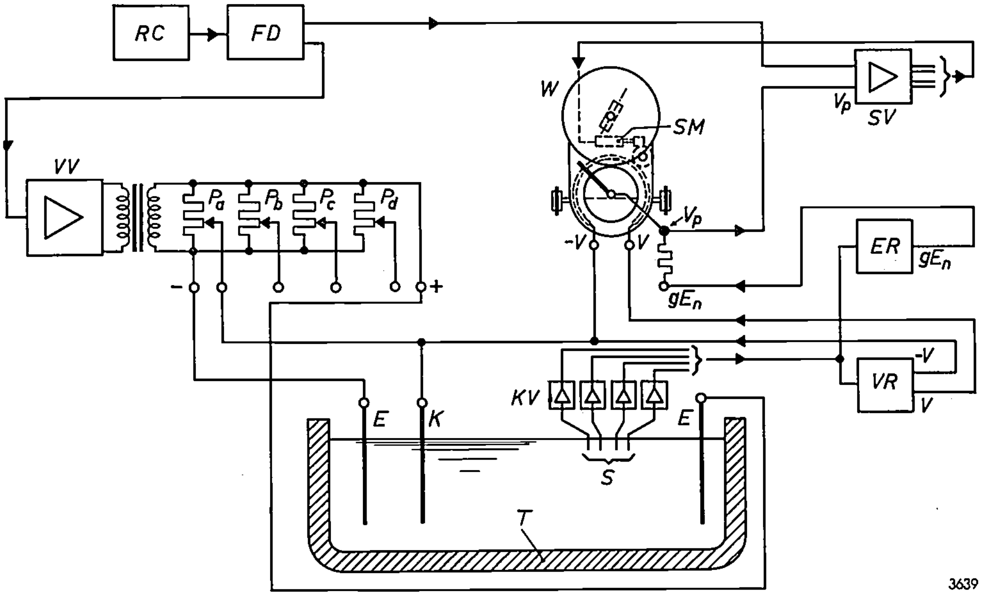


Fig. 3. Schematic diagram of the automatic plotter. *RC* generator for 500 c/s. *FD* phase shifter. *VV* power amplifier. $P_a \ldots P_d$ potentiometers for adjusting the supply voltages for the electrodes *E*. *T* electrolytic tank. *K* cathode of electrode system, i.e. the electrode at which the electrons have zero velocity. *S* probes. *KV* cathode followers. *VR* computing circuit for determining the average probe potential *V*. *ER* computing circuit for determining the local field strength $E_n$. *W* trolley with potentiometer circuit and servo motor *SM*. *SV* servo amplifier.

*Fig. 4* shows a general view of the apparatus, in the process of plotting a trajectory in an electron gun for a television picture tube. The electronic equipment is not visible in this photograph.

### Tank and electrodes

The electrolytic tank is made of reinforced concrete, 192×78 cm and 50 cm deep. The walls are clad with a layer of polyethylene 3 mm thick, to prevent ions from the concrete dissolving in the water.

The surface of the probes and electrodes are subject to effects which may adversely influence the accuracy. When a metal is brought into contact with

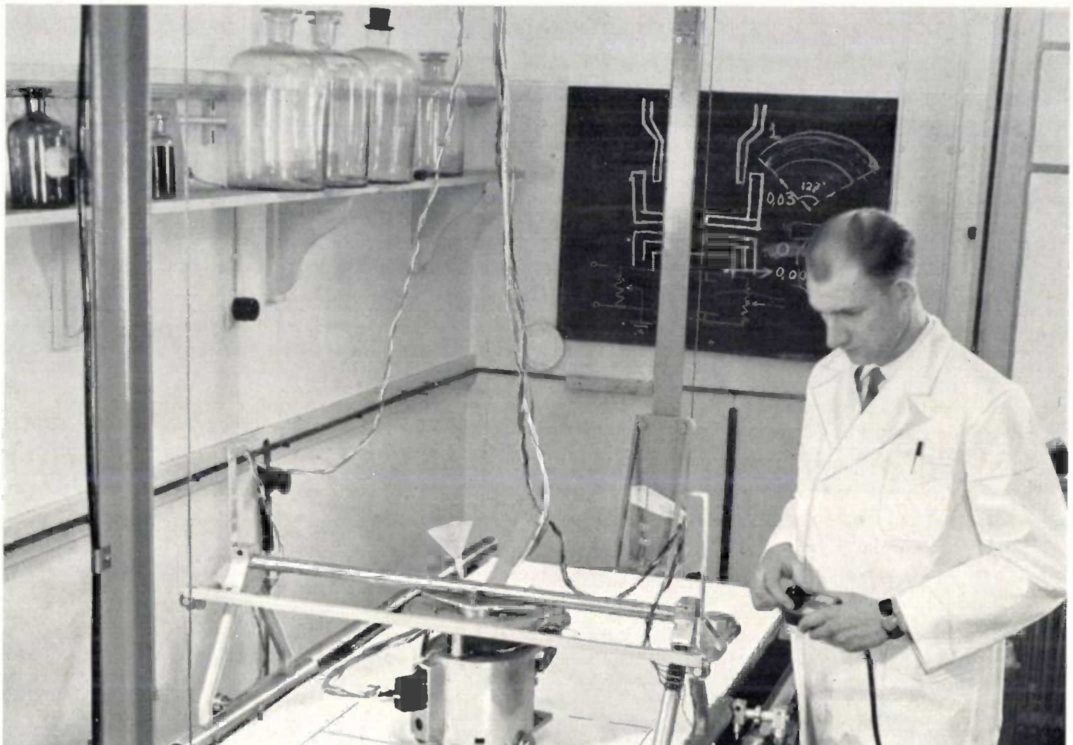


Fig. 4. General view of the automatic plotter. The control box held by the laboratory assistant contains a switch for reversing the direction of the trolley, and a potentiometer for controlling the speed of the trolley. A trajectory is here being plotted in an electron gun for a television picture tube (to make the trace visible in the photograph the trajectory plotted has been traced over by hand). The form of the electrodes is marked out on the drawing board. When all the trajectories have been plotted, the board is photographed and then cleaned. To enable the model to be set up in the tank the whole board and appendages are hoisted up the vertical beams.

an electrolyte, a potential difference arises between the metal and the electrolyte, and if a current flows between them an additional potential difference, due to polarization, may arise. This effect depends on the current density. The polarization of the probes and electrodes must be kept to a minimum; the constant potential difference then remaining is harmless, owing to the use of alternating voltage. Polarization can be reduced to negligible proportions by covering the metal surface with platinum black, i.e. a porous layer of very finely divided platinum. This is in fact done in the case of the probes, but the electrodes are too large to make such a coating practicable. In their case, silver plating was found to result in a sufficiently low polarization, and the electrodes are therefore made of silver-plated brass or copper.

The effect of polarization on the potential measurements can be further reduced by using an electrolyte of high resistivity. In our case, use is made of deionized water, whose resistivity is roughly 30 kΩ cm as against the 2 kΩ cm of main water.

To achieve maximum accuracy, two further conditions must be satisfied.

1) The electrolyte must be uniformly conductive. This can be assured simply by stirring the electrolyte shortly before the measurement.
2) The probes must not draw current from the electrolyte, as this would disturb the potential distribution. For this reason the probes are connected to cathode followers having an extremely high input resistance.

As a result of polarization of the electrodes, a phase shift occurs between the voltage on the electrodes and the current; consequently, when a sinusoidal voltage is used there is no clearly defined zero point in the output voltage $V_p$ of the potentiometer circuit. $V_p$ is a linear combination of $gE_n$ and $V$. The voltage $gE_n$ is calculated from the potential difference between the probes, that is between points in the electrolyte, whereas $V$ is the potential difference between a point in the electrolyte and an electrode of the model (the cathode of the system). The phase shift is therefore not felt in $gE_n$, but it does affect $V$. These two voltages do not therefore have the same phase, and so they can never be zero at the same time and $V_p$ can never be zero.

As a method of obtaining a sharp zero-point setting, Sander and Yates[4]) have proposed the use of a square-wave voltage for the electrodes of an electrolytic tank. This is done in the present equipment; the voltage $V_p$ then has the waveform shown in *fig. 5a*, *b* and *c*, depending on the position of the wiper of the potentiometer on the trolley. It is now possible to make a sharp adjustment to the equilibrium setting (fig. 5*b*) by using a circuit that responds only to the flat portions (plateaus) of the waveform.

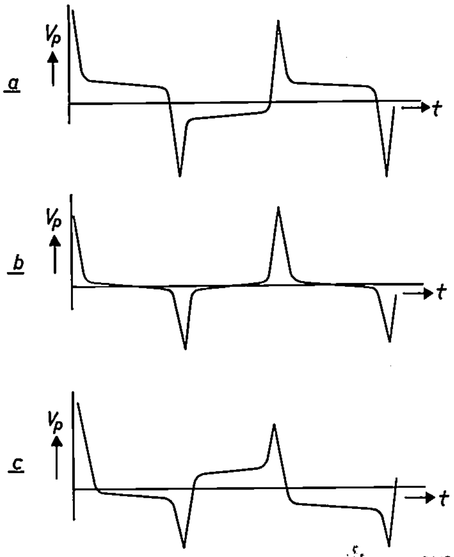


Fig. 5. Output voltage $V_p$ of the potentiometer circuit when the electrodes are supplied with a square-wave voltage.
*b*) The steering shaft of the trolley is in the correct position. The middle of the "flat" portion (plateau) of the curve is zero.
*a*) and *c*) The steering shaft is not in the correct position. The voltage of the middle of the plateau is amplified and fed to the servo motor, which corrects the position of the steering shaft.

### Excitation of the field in the tank

As we have seen, the electrodes are fed with a square-wave alternating voltage, whilst in the actual system the electrodes are at positive or negative DC potentials. When the electrodes are connected to the output of the power amplifier $VV$ (fig. 3), one of the output terminals is regarded as positive and the other as negative. Interchanging these terminals would imply that the phase of the voltages on the electrodes (and therefore also the phases of $V$, $E_n$ and $V_p$) would be shifted 180°. It will be shown when dealing with the servo amplifier that this would make the servo system unstable.

As an example of adjusting the voltages on the electrodes, we shall consider a triode where $V_c = 0$ V, $V_g = -10$ V, and $V_a \doteq +200$ V. The grid is connected to the "negative" terminal of the power amplifier, and the anode to the "positive" terminal. The cathode must now be given a potential, relative to the negative terminal, amounting to 10/210 times the output voltage of the amplifier. For this purpose the cathode is connected to the wiper of potentiometer $P_a$ in fig. 3. For pre-setting the wiper to obtain the required potential, a bridge circuit

[4]) See the first article quoted under reference [1]).

is used as shown in *fig. 6*. An accurately calibrated decade potentiometer $P_{\text{dec}}$ is connected, in parallel with $P_{\text{a}}$, to the amplifier. The wiper of $P_{\text{dec}}$ is adjusted until the resistances of the parts into which the potentiometer is divided are in the ratio of 10 : 210. The wiper of $P_{\text{dec}}$ is earthed, and that of $P_{\text{a}}$ is adjusted with the aid of an oscilloscope until its potential is also zero.

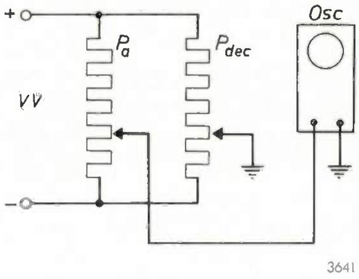


Fig. 6. Bridge circuit for adjusting the potentiometers from which the electrode potentials are tapped. *VV* output of power amplifier. $P_{\text{a}}$ potentiometer to which the electrode will be connected. $P_{\text{dec}}$ calibrated decade potentiometer, pre-set to the required voltage ratio. *Osc* oscilloscope.

For reasons which will be discussed in connection with the computing circuit for $E_{\text{n}}$, the potential midway between the probes must be zero. This is achieved by deducting the voltage $V$ from all electrode voltages, the method being to connect the output $-V$ of the $V$-computing circuit to the electrode of the system for which the electron velocity is zero, as illustrated in fig. 3. This electrode is referred to as the cathode, irrespective of its function in the actual system. If the system contains no such electrode, one of the potentiometer terminals must nevertheless be set at zero potential and connected to the output $-V$. Take, for example, an electron lens system with two electrodes where $V_2 > V_1 > 0$. The zero voltage is taken from the negative terminal, and electrode *2* is connected to the positive terminal. Electrode *1* is connected to the wiper of $P_{\text{a}}$, which is adjusted to the ratio $V_1/V_2$.

### Method of determining $E_{\text{n}}$ and $V$

The field-strength component $E_{\text{n}}$ and the potential $V$ may in principle be determined, as Gabor and Langmuir have done, from the voltages of two probes at either side of the measuring point, along

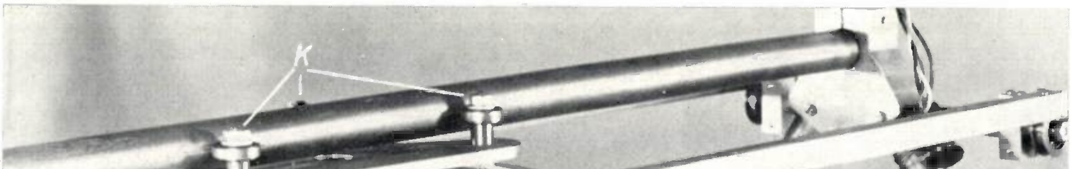


Fig. 7. The trolley on the drawing board. *M* motor which drives the front wheel for propelling the trolley. *SM* servo motor which controls the angular position of the steering shaft via a worm and pinion and gear $T_2$. $P_1$ and $P_2$ potentiometers of servo system. $T_1$ and $T_2$ gear wheels which mechanically couple the potentiometer wiper arms to the steering shaft. *K* ball bearings which transmit the length-wise movement of the trolley to a cross-frame that carries along the probes under the drawing board. (The cover-plates protecting the potentiometers in fig. 4 have here been removed.)

a line perpendicular to the direction of the trajectory. If the distance between the probes is $a$ and their voltages are $V_1$ and $V_2$, then $E_\mathrm{n} = (V_1 - V_2)/a$ and $V = \frac{1}{2}(V_1 + V_2)$. The voltages might be determined by means of a difference amplifier and a sum amplifier. However, at points where the field strength depends markedly on position, the values found in this way would not be accurate. The error can be reduced, of course, by reducing the distance between the probes, but in that case the difference between the voltages is smaller, and therefore higher demands must be made on the accuracy with which the voltages are measured. Moreover, the local disturbance which the probes cause in the potential distribution is then greater.

As we have said, the instrument here described uses four probes, making it possible to determine $E_\mathrm{n}$ and $V$ with greater accuracy without reducing the distance between the probes. Of course, the formulae defining $E_\mathrm{n}$ and $V$ are more complicated, and more elaborate electronic equipment is needed for determining these voltages.

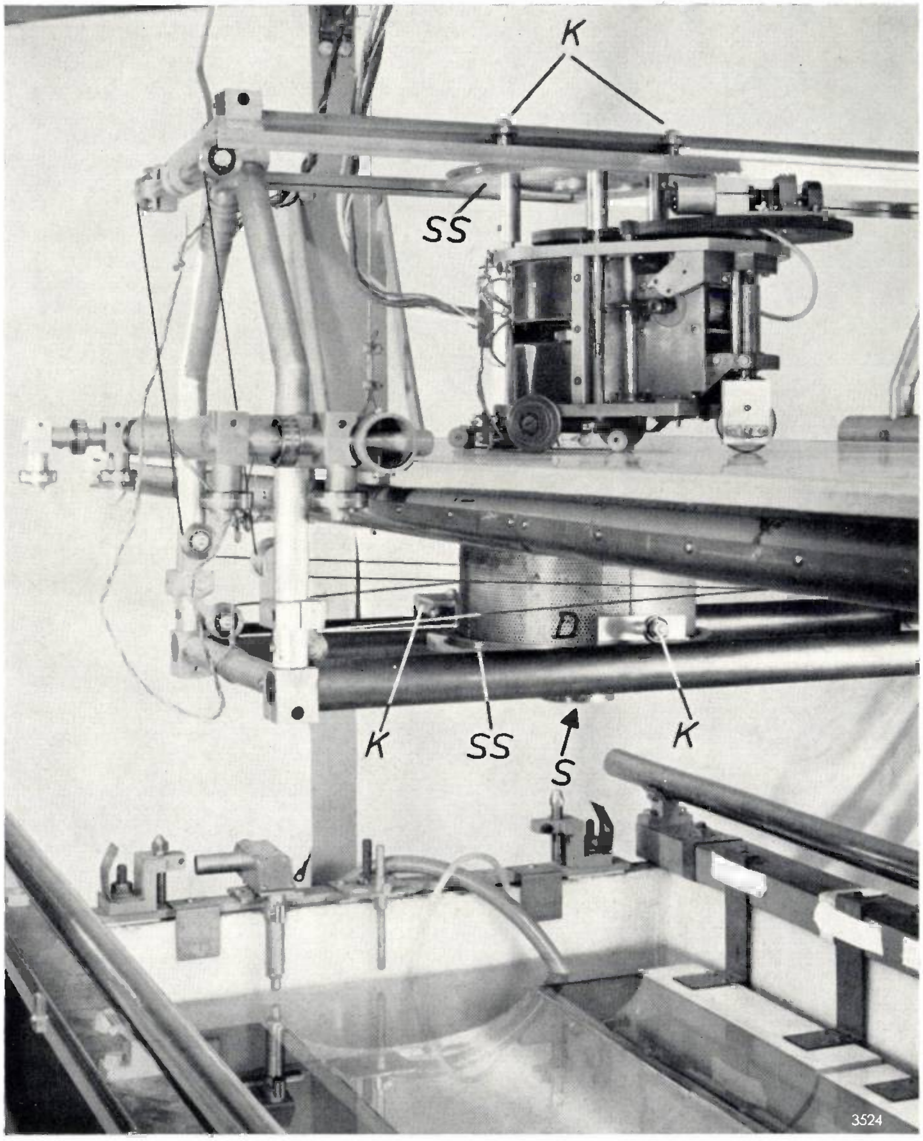


Fig. 8. In this photograph the drawing board has been raised to show the cross-frame and components underneath the drawing board. The box $D$ can be seen, which contains the four cathode followers, and under it the probe holder with the four probes $S$. These protrude underneath, but are too small (0.5 mm diameter) to be visible.

The movement of the trolley along the length of the tank is transmitted by ball bearings $K$ to the cross-frame, which, again by ball bearings, transmits this movement to the box $D$. The lateral and rotational movements of the trolley are transmitted by cords passing over the large pulleys $SS$. Underneath the trolley can be seen the tracing mechanism, consisting of a relay, two rollers over which a typewriter ribbon runs and the wheel which draws the line.

Before dealing at length with the computing circuits and the derivation of the formulae, we shall touch briefly on the components of the servo system.

### The trolley

A photograph of the three-wheel trolley with which the electron trajectories are plotted is shown in *fig. 7*. The trolley is propelled by an electric motor $M$ which drives the front wheel. A servo motor $SM$ turns, via a wormwheel and pinion, a gearwheel $T_2$ fixed to the steering shaft. Mounted on the trolley are the two potentiometers, $P_1$ and $P_2$, of the potentiometer circuit. With the aid of gearwheels $T_1$ and $T_2$ the two wiper arms are mechanically coupled to the steering shaft. In *fig. 8* the drawing board has been raised to reveal various parts situated under it. Here are located the probes $S$ in the probe-holder, which is mounted on a box $D$ containing the four cathode followers. The probes must at all times follow the trolley in both position and orientation; all the movements of the trolley must therefore be transmitted to them. Along the main axis of the tank this is done by means of a cross-frame which travels along the rim of the tank, and against which the trolley and the box abut via ball bearings $K$. The trolley further carries a large pulley $SS$, over which two cords run via small pulleys to an identical large pulley fixed to the box underneath the drawing board. This mechanism, then, is responsible for transmitting the lateral and rotary movements of the trolley to the probes.

The recording mechanism is visible under the trolley in fig. 8. It comprises an endless typewriter ribbon which passes over two rollers. A fairly sharp-rimmed wheel, attached to the armature of a relay, presses the ribbon on to the drawing board when this relay is energized. Being stationary relative to the drawing board, the ribbon rolls continuously as the trolley moves.

### Potentiometer circuit

The principle of the potentiometer circuit mounted on the trolley is illustrated in *fig. 9a*. The deviation of the wiper of potentiometer $P_1$ from the centre point, denoted $x$, may vary from $-\frac{1}{2}$ to $+\frac{1}{2}$. Since $\alpha$ (the angle between the front and rear axles of the trolley) must be able to run from $-\frac{1}{2}\pi$ to $+\frac{1}{2}\pi$, the transmission between the potentiometer spindle and the steering shaft must be chosen such that $\alpha = \pi x$.

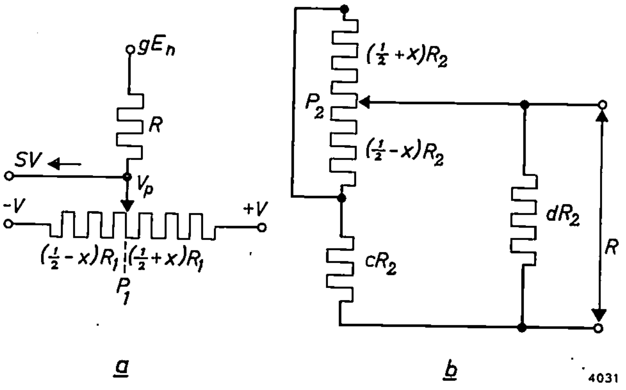


Fig. 9. *a*) Principle of the steering circuit mounted on the trolley. The potentiometer $P_1$ is supplied with the voltages $V$ and $-V$, and the resistor $R$ with the voltage $gE_n$. The position of the potentiometer wiper arm is determined by the angular position of the steering shaft. The voltage $V_p$ of the wiper is applied to the servo amplifier $SV$.
*b*) Circuit which ensures that $R$ depends according to a certain function on the position of the steering shaft, whereby $V_p$ is zero when the steering shaft is in the correct position. The wiper arm of potentiometer $P_2$ is coupled mechanically to that of potentiometer $P_1$ in (*a*).

When the wiper of the potentiometer is in either of its extreme positions, a small resistance still remains between the wiper contact and the connection terminal, so that $x$ cannot become precisely $\frac{1}{2}$ (or $-\frac{1}{2}$). In the extreme positions, then, the angle $\alpha$ remains somewhat smaller than $\frac{1}{2}\pi$, and the radius of curvature of the trajectory cannot therefore be zero. The minimum radius of curvature that can be described by the trolley is 3 mm. For most practical purposes this is quite sufficient. If an even stronger curvature is required, part of the model can be made on a larger scale and the trajectory can then be described in that particular part.

From Kirchhoff's first law, $\Sigma i = 0$, it follows that:

$$V_p = \frac{(\frac{1}{4} - x^2)R_1 g E_n - 2xRV}{(\frac{1}{4} - x^2)R_1 + R}. \quad . \quad . \quad (4)$$

This voltage is used as the output voltage of the circuit and must therefore be zero when the steering shaft is in the correct angular position. In that position, $\alpha = \tan^{-1}(hE_n/2V)$, and therefore $V_p$ must be zero when $\pi x = \tan^{-1}(hE_n/2V)$. The resistance $R$ must then be:

$$R = R_1 \frac{g}{h}\left[\frac{\frac{1}{4} - x^2}{x}\tan \pi x\right]. \quad . \quad . \quad . \quad (5)$$

If the factor between brackets is plotted as a function of $x$, a curve resembling a parabola with a flattened top is produced. A similar variation as a function of $x$ is shown by the resistance between the output terminals of the circuit in fig. 9*b*. This contains a second potentiometer, $P_2$, whose wiper arm is coupled mechanically to that of $P_1$. The resistance of this circuit is:

$$R = d\,\frac{c + \frac{1}{4} - x^2}{c + d + \frac{1}{4} - x^2}\,R_2, \quad . \quad . \quad . \quad (6)$$

where $c$ and $d$ are constants. This function does not exactly correspond to the function required, but the values of $R_1$, $R_2$, $c$, $d$ and $h/g$ may be chosen in such a way as to keep the relative error of the radius of curvature smaller than $3 \times 10^{-5}$ over the whole range $|x| \leqq \frac{1}{2}$ (this error is thus negligible compared with that caused by the normal non-linearity of the potentiometers $P_1$ and $P_2$). The distance $h$ is fixed by the construction of the trolley, so that the value of the factor $g$ can be calculated from $h/g$.

### The servo amplifier

Since no current may flow in the circuit of the potentiometer tap, the voltage $V_p$ is applied to an amplifier capable of delivering the power needed to energize the servo motor. The output voltage of this servo amplifier is an alternating voltage of 50 c/s, whose amplitude is proportional to that of $V_p$.

The waveform of $V_p$ is shown in fig. 5. In the equilibrium position (the steering shaft is then correctly aligned) the middle of the plateau is at zero level (*b*, fig. 5). The servo amplifier is therefore required to amplify the voltage only during the midplateau part of each cycle ($t_1$, see *fig. 10*). This is achieved in the following way.

The servo amplifier contains a stage which amplifies the voltage $V_\mathrm{p}$ to $A\,V_\mathrm{p}$. This voltage is then added to a direct voltage $V_\mathrm{B}$ and also subtracted from it, thus producing two voltages

$$V_\mathrm{C} = V_\mathrm{B} + A\,V_\mathrm{p} \quad \text{and}$$

$$V_\mathrm{D} = V_\mathrm{B} - A\,V_\mathrm{p}$$

(fig. 10). $V_\mathrm{C}$ and $V_\mathrm{D}$ are fed to a gate circuit which, via the phase shifter *FD* (fig. 3), is controlled by the same alternating voltage that is applied, via the power amplifier, to the electrodes of the model. The phase difference between the two output voltages of *FD* is so adjusted that the voltages $V_\mathrm{C}$ and $V_\mathrm{D}$ are passed by the gate circuit to the next stage only during the intervals $t_1$. This next stage is a modulator, which delivers an alternating voltage of 50 c/s whose amplitude is proportional to the difference $V_\mathrm{C} - V_\mathrm{D}$, and thus proportional to the value of $V_\mathrm{p}$ during the interval $t_1$. This also applies when $V_\mathrm{p}$ is negative, in which case the phase is shifted 180°.

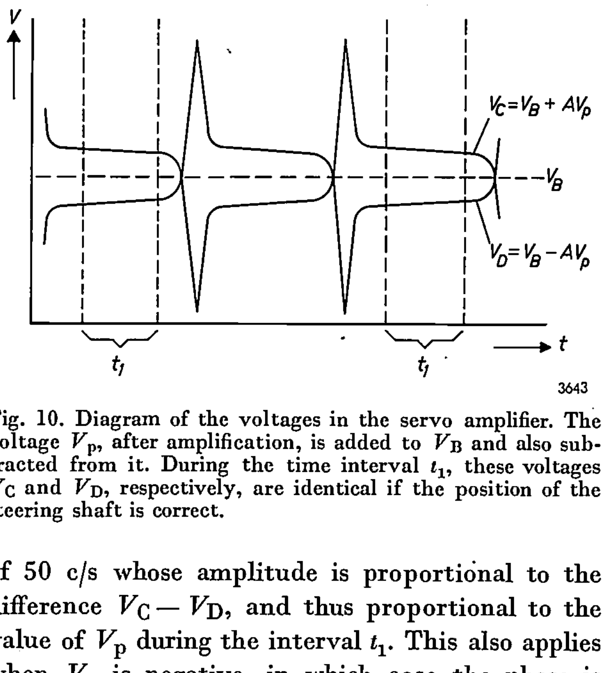


Fig. 10. Diagram of the voltages in the servo amplifier. The voltage $V_\mathrm{p}$, after amplification, is added to $V_\mathrm{B}$ and also subtracted from it. During the time interval $t_1$, these voltages $V_\mathrm{C}$ and $V_\mathrm{D}$, respectively, are identical if the position of the steering shaft is correct.

The servo motor is a two-phase motor, one winding of which is fed via a transformer from the mains, the other being supplied with the output voltage from the servo amplifier. Depending on the phase of the latter voltage, the motor turns clockwise or anti-clockwise until $V_\mathrm{p}$ is zero.

It is now clear why the terminals of the power amplifier must not — as pointed out earlier — be interchanged. The phase of $V_\mathrm{p}$ would then be shifted 180°, and in the interval $t_1$, therefore, $V_\mathrm{p}$ would have the opposite sign and the motor would rotate in the wrong direction.

The entire servo system is required to respond fast enough to enable the trolley to follow properly all the curvatures of the trajectory. This can be checked by, for example, letting the trajectory be plotted in both directions. Any discrepancy between the two trajectories indicates that the system has not responded rapidly enough, or that the trolley was moving too fast. The response time and the damping of the system depend on the moment of inertia of the rotating parts, on the friction, and on the ratio $M/\Delta a$, where $M$ is the torque acting on the steering shaft as a result of a deviation $\Delta a$ from its correct angular position. The response of the system is faster, and at the same time the damping smaller, the greater the value of $M/\Delta a$. The ratio $M/\Delta a$ is proportional to the gain of the servo amplifier and to the voltage $V$, and also depends to some extent on the radius of curvature. Since $V$ is proportional to the electrode supply voltage, $M/\Delta a$ can be adjusted either with the servo amplifier or with the power amplifier.

The system works best, that is to say the trolley speed can be greatest, when there is hardly any damping and the amplitude of the trace remains small. The trolley speed can be about 2 cm/sec if the trajectory curvature radius $\varrho$ varies only slightly, but it must be substantially less if the variations of $\varrho$ are considerable. For this purpose the supply voltage of the drive motor can be varied with a potentiometer whilst the trajectory is being plotted.

## Derivation of the formulae for $E_\mathrm{n}$ and $V$

The field strength $E_\mathrm{n}$ and the potential $V$ must be derived by the computing circuits from the voltages of the four probes. We call the line through the probes, which is perpendicular to the direction of the trajectory, the $n$ axis (*fig. 11*). The $n$-coordinates of the probes $S_1 \ldots S_4$ are: $-\frac{3}{2}a$, $-\frac{1}{2}a$, $\frac{1}{2}a$ and $\frac{3}{2}a$; the distance between the probes is $a$. Four values are known of the function $V(n)$ that gives the potential variation along the $n$ axis. We assume that $V(n)$

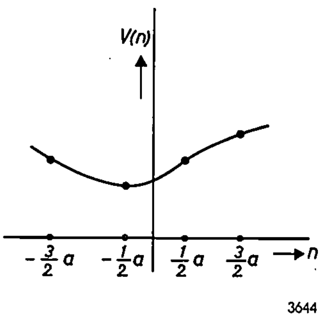


Fig. 11. Potential distribution along the $n$ axis. The probes are situated at points $n = -\frac{3}{2}a$, $n = -\frac{1}{2}a$, etc. The function $V(n)$, which describes the potential variation, is approximated by a third-degree curve through the points representing the potentials of the four probes.

is a third-degree polynomial in $n$, for this is the most general polynomial determined by four values. We can now calculate the function $V(n)$, from which $V = V(0)$ and $E_n = -V'(0)$ can be evaluated [5]). We find:

$$E_n = \frac{1}{24a}\left[27(V_2-V_3) + (V_4-V_1)\right] \quad . \quad (7)$$

and
$$V = \frac{1}{16}\left[9(V_2+V_3)-(V_1+V_4)\right]. \quad . \quad (8)$$

### The effect of induced dipoles

A complication arises in that the probes disturb the potential field in the electrolyte. In each probe a dipole is induced which affects the field near the other probes. The voltages $V_1 \ldots V_4$ measured on the probes are consequently not the same as the voltages $V_{01} \ldots V_{04}$ prevailing in the electrolyte when the probes have not yet been immersed. The latter voltages must be used instead of $V_1 \ldots V_4$ in formulae (7) and (8), and therefore we must ascertain how these voltages are derived from the measured $V_1 \ldots V_4$.

The dipole moments of the four probes are proportional to the local field strengths and have the same direction. Since the components perpendicular to the $n$ direction have no influence on the field near the other probes, we shall consider only the components in the $n$ direction, $M_1 \ldots M_4$. These are proportional to the $n$ components of the local field strengths, $E_1 \ldots E_4$:

$$M_i = \varepsilon\, k\, E_i \quad (i = 1 \ldots 4), \quad . \quad . \quad . \quad (9)$$

where $\varepsilon$ is the dielectric constant of the electrolyte, and the factor $k$ has the dimension (length)$^3$.

Between the voltages $V_{0i}$ and $V_i$ the following relations exist:

$$\left.\begin{aligned}
V_{01} &= V_1 + \frac{k}{(6a)^3}(71V_1+126V_2-171V_3-26V_4),\\
V_{02} &= V_2 + \frac{k}{(6a)^3}(-414V_1+783V_2-270V_3-99V_4),\\
V_{03} &= V_3 + \frac{k}{(6a)^3}(-99V_1+270V_2-783V_3-414V_4),\\
V_{04} &= V_4 + \frac{k}{(6a)^3}(-26V_1+171V_2+126V_3-71V_4).
\end{aligned}\right\} \quad . \quad . \quad . \quad (10)$$

In formula (7) for $E_n$ we must replace $V_1 \ldots V_4$ by $V_{01} \ldots V_{04}$. Substituting the expressions (10) for $V_{01} \ldots V_{04}$ we obtain:

$$E_n = \frac{1}{24a}\left[27(V_2-V_3)+(V_4-V_1) + \right.$$
$$\left. + \frac{k}{a^3}\left\{130(V_2-V_3) + 40(V_4-V_1)\right\}\right]. \quad (11)$$

The same manipulations in respect of formula (8) for $V$ results in:

$$V = \tfrac{1}{2}(V_2+V_3) + \frac{1}{16}(V_2+V_3-V_1-V_4)(1+21.6\frac{k}{a^3}). \quad . \quad . \quad . \quad (12)$$

The second term is already small in relation to the first, so there is no point in applying the correction factor to this term. As regards the voltage $V$, therefore, the correction for the dipole moments of the probes need not be applied and (12) reduces to equation (8).

The equations in (10) are derived from various formulae of electrostatics, confined in the present case to the above-mentioned components in the $n$ direction and to points on the $n$ axis.

If a dipole, the $n$ component of whose moment is $M$, is situated at the origin of the $n$ axis, the field strength at the point $n = a$ is given by $E = 2M/|a|^3$ and the potential at that point by $V = Ma/|a|^3$. Bearing in mind that the field near a probe is built up from the original field and the contribution from the dipoles of the other probes, we find, where $E_{0i}$ is the $n$ component of the field strength for non-immersed probes:

$$\left.\begin{aligned}
E_1 &= E_{01} + \frac{2k}{a^3}E_2 + \frac{2k}{(2a)^3}E_3 + \frac{2k}{(3a)^3}E_4,\\
E_2 &= E_{02} + \frac{2k}{a^3}E_1 + \frac{2k}{a^3}E_3 + \frac{2k}{(2a)^3}E_4,
\end{aligned}\right\} \quad . \quad . \quad (13)$$

$$\left.\begin{aligned}
V_1 &= V_{01} - \frac{k}{a^2}E_2 - \frac{k}{(2a)^2}E_3 - \frac{k}{(3a)^2}E_4,\\
V_2 &= V_{02} + \frac{k}{a^2}E_1 - \frac{k}{a^2}E_3 - \frac{k}{(2a)^2}E_4.
\end{aligned}\right\} \quad . \quad . \quad (14)$$

Similar relations hold for $E_3$, $E_4$, $V_3$ and $V_4$.

We further assume, for the reason already mentioned, that the function giving the potential variation $V_0(n)$ along the $n$ axis when the probes are not immersed in the water, is a third-degree polynomial in $n$. The first derivatives at points $-\frac{3}{2}a$, $-\frac{1}{2}a$, $\frac{1}{2}a$ and $\frac{3}{2}a$ are the field strengths $E_{01} \ldots E_{04}$. This gives another four equations between the quantities. We now have 12 equations available, containing 16 variables. After elimination (matrix algebra is the simplest way of doing this) we can express $V_{01} \ldots V_{04}$ in terms of $V_1 \ldots V_4$, which yields the above result.

The value of $k/a^3$ can be calculated from the formula for the dipole moment $M$ induced in a cylinder of radius $R$ and height $h$, where the field strength $E$ is normal to the axis: $M = \frac{1}{3}\varepsilon R^2 hE$. Comparison with (9) yields:

$$k = \tfrac{1}{3}R^2h.$$

The diameter of the probes is 0.5 mm and they project about 1 mm into the electrolyte. The distance $a$ is 4 mm, and hence $k/a^3 \approx 0.0005$.

[5]) For a more detailed treatment see: J. L. Verster, On the use of gauzes in electron optics, dissertation Delft, to be published shortly by Centrex Publishing Co., Eindhoven.

### The computing circuit for $E_n$

It follows from (11) that $E_n$ is given by:

$$E_n = \frac{1}{24a}\left[(27+130\frac{k}{a^3})(V_2-V_3)+(1+40\frac{k}{a^3})(V_4-V_1)\right]. \quad \ldots \quad (15)$$

The circuit required to deliver a voltage proportional to $E_n$ must be a kind of difference amplifier. The output voltage $V_0$ of an ideal difference amplifier is proportional to the difference of two input voltages $V_{i1}$ and $V_{i2}$. It is difficult, however, to make a difference amplifier whose output voltage is not at the same time to some extent dependent on $V_{i1} + V_{i2}$. In that case:

$$V_0 = A\left\{(V_{i1}-V_{i2}) + \frac{1}{H}(V_{i1}+V_{i2})\right\}, \quad (16)$$

where $H$ is the so-called rejection factor [6]). For an ideal difference amplifier, $H = \infty$. The second term in eq. (16) can still be minimized, however, by ensuring that $V_{i1} + V_{i2}$ is small. To this end, a voltage is deducted from both input voltages which has roughly the mean value $\frac{1}{2}(V_{i1} + V_{i2})$.

This operation is done on the input voltage to the difference amplifier for computing $E_n$. It is for this reason, as mentioned on page 250, that the voltage $V$ is subtracted from the voltages $V_1 \ldots V_4$, the value of $V$ being close to the average voltages for both subtractions in eq. (15). A further incidental advantage of this method is that it considerably reduces the influence of differences between the cathode followers, which might give rise to the same error as a finite rejection factor.

The principle of the circuit for computing $E_n$ is illustrated in *fig. 12.* Amplifier *1* has a high gain, so that the potential at point $P$ is zero. The resistances of the network are chosen such that $V_S$ at point $S$ is proportional to $E_n$, and only $R_5$ is dependent on the constants $k$ and $a$ of the apparatus. No current may be drawn from point $S$, and for this reason amplifier *2* is of a type that has a high input impedance and whose output may therefore permissibly be loaded. The gain is adjusted to give the output voltage $gE_n$.

High demands are made on the accuracy of the computing circuit, which means that strong negative feedback is required from amplifier *1*. In view of the high gain which the latter must have, this makes it difficult to achieve a stable circuit. A solution was found by utilizing a new principle termed "addition of the complement" [7]). A main amplifier raises the input signal to roughly the required output signal. A combination of input and output signal is amplified by a correction amplifier, and this combination vanishes when the output signal has the correct value. The output signal from the correction amplifier is added to that of the main amplifier. With this arrangement an accurately defined gain is obtained, without the gain of the individual amplifiers having to be particularly high. The negative feedback in these amplifiers is relatively small, so that no complicated feedback network is needed to make the circuit stable.

Since the gain of amplifier *2* need not be high (about 60), it is sufficient to use a conventional amplifier consisting of two pentodes with negative current feedback to the cathode of the first tube.

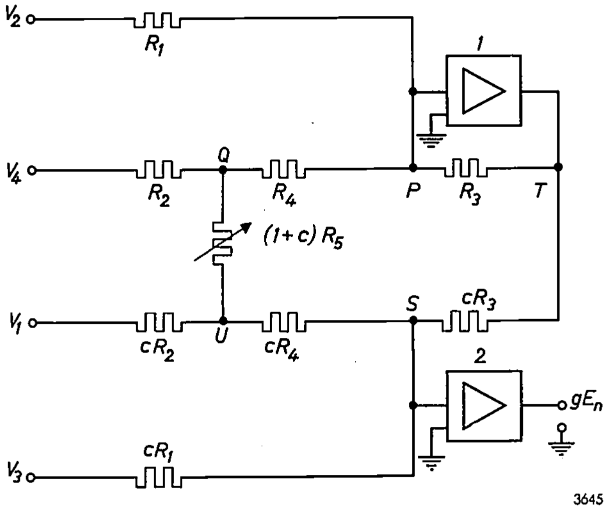


Fig. 12. Block diagram of the computing circuit for $E_n$. The probe potentials are applied to the terminals on the left. Amplifier *1* has a high gain, so that the potential of point $P$ is effectively zero. The resistance values are such as to make $V_S$ proportional to $E_n$. Since the network must not carry current, amplifier *2* is given a high input impedance. Only $R_5$ is dependent on the dimensions of the probes.

### The computing circuit for $V$

The computing circuit for $V$ is required to supply the voltages $V$ and $-V$ to the trolley. At the same time the voltage $-V$ must be applied to the "cathode" output of the power amplifier in order to make the potential at the point midway between the probes equal to zero. We shall explain how the latter is achieved in this way by considering a

[6]) See G. Klein, Rejection factor of difference amplifiers, Philips Res. Repts. **10**, 241-259, 1955; see also Philips tech. Rev. **21**, 32-33, 1959/60.

[7]) J. J. Zaalberg van Zelst, Stabilised amplifiers, Philips tech. Rev. **9**, 25-32, 1947/48.

simple case as represented in *fig. 13*. Here there is only one probe, which has a potential $V$ relative to the cathode. If $V_\mathrm{k}$ is the potential of the cathode relative to earth, the probe voltage with respect to earth is:

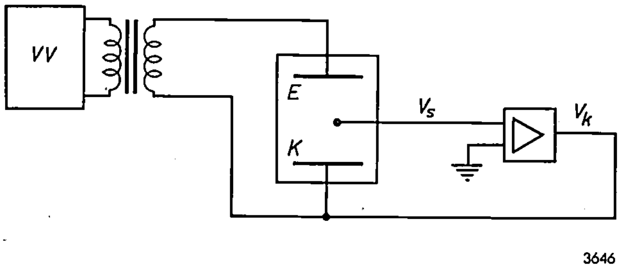


Fig. 13. Circuit for keeping the probe at earth potential, for the case of only one probe. $VV$ power amplifier for supplying the electrodes $E$ and $K$ ($K$ is the cathode of the system). The voltage between the probe and the cathode is $V$. If the gain $A$ is very much greater than unity, $V_\mathrm{s} = 0$ and $V_\mathrm{k} = -V$.

$$V_\mathrm{s} = V_\mathrm{k} + V. \quad . \quad . \quad . \quad . \quad . \quad (17)$$

The figure shows further that:

$$V_\mathrm{k} = -AV_\mathrm{s}. \quad . \quad . \quad . \quad . \quad . \quad (18)$$

It follows from eq. (17) and (18) that:

$$V_\mathrm{k} = -\frac{V}{1+1/A} \quad \text{and} \quad V_\mathrm{s} = \frac{V}{A+1}.$$

If the gain $A$ is sufficiently high, we can thus put $V_\mathrm{s} = 0$ and $V_\mathrm{k} = -V$. The voltage $V$ can be obtained by using an inverting stage.

When four probes are used, the output voltage must satisfy eq. (8). The principle of the circuit used in this case is illustrated in *fig. 14*. For the input voltage $V_\mathrm{p1}$ a voltage divider of high resistance is used to obtain the mean of $V_2$ and $V_3$, so that $V_\mathrm{p1}$ with respect to earth is:

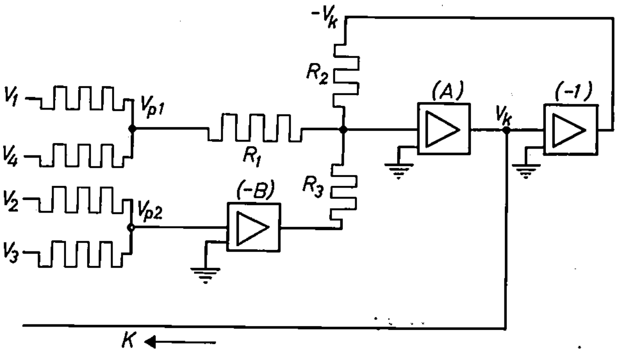


Fig. 14. Principle of the computing circuit for $V$. This circuit ensures that the potential at the point midway between the probes always remains zero. The resistances and gain figures $A$ and $B$ are chosen such that $V_\mathrm{k} = -V$; the inverting amplifier delivers the voltage $V$.

$$V_\mathrm{p1} = \tfrac{1}{2}(V_1 + V_4) + V_\mathrm{k}. \quad . \quad . \quad . \quad (19)$$

Likewise:

$$V_\mathrm{p2} = \tfrac{1}{2}(V_2 + V_3) + V_\mathrm{k}. \quad . \quad . \quad . \quad (20)$$

By a suitable choice of the resistances $R_1$, $R_2$ and $R_3$ and the gains $A$ and $B$, we can obtain (see eq. (8)):

$$V_\mathrm{k} = -\frac{9}{16}(V_2 + V_3) + \frac{1}{16}(V_1 + V_4) = -V. \quad (21)$$

As the cathode now has a potential $-V$, the potential midway between the probes is zero. The circuit produces both $V$ and $-V$. The gains $A$ and $B$ need not be high, so that simple amplifiers can be used. The inverting amplifier is again designed on the principle of "addition of the complement".

### The cathode followers

The probes, regarded as voltage sources, have a high internal resistance owing to the fact that the part projecting into the electrolyte is small and the conductivity of the electrolyte is low. To prevent the probes drawing current from the electrolyte, they are connected to cathode followers which pass the voltages on to the computing circuits.

These cathode followers must meet three requirements:

*a*) The input impedance must be high compared with the probe impedance.

*b*) The ratio of the output voltage $V_0$ to the input voltage $V_\mathrm{i}$ must be highly constant and identical for all cathode followers.

*c*) The output impedance must be small compared with the input impedance of the two computing circuits in parallel.

The latter requirement is necessary because the cathode followers are not equally loaded by the computing circuits. A cathode follower consisting of one valve does not simultaneously fulfil the second and third requirements. If the output impedance is made small, the output voltage is then too dependent on the valve characteristics and therefore can no longer be identical for all four. A cascade arrangement of two valves (*fig. 15*) satisfies the above conditions (see reference [5]) for the calculation), provided the following measures are taken:

*a*) To achieve a high input impedance, the grid of valve *2* is directly coupled with the probe, thus dispensing with the need for a grid leak. This grid now receives, in addition to the alternating voltage, the direct voltage from the probe, which is approximately zero.

*b*) To minimize the dependence of $V_0/V_\mathrm{i}$ on the valve characteristics, the amplification factor of valve *2* must be high. To minimize the load, the internal resistance of valve *1* must be high. Both valves used are therefore pentodes.

*c*) A low output impedance is achieved by using valves having a fairly steep slope and large $R_a$.

Measurements yielded the following values for all cathode followers:

$$V_o/V_i = 0.9998 \pm 0.0001 \text{ where } R_l = \infty ,$$
$$V_o/V_i = 0.9996 \pm 0.0001 \text{ where } R_l = 4.7 \text{ k}\Omega.$$

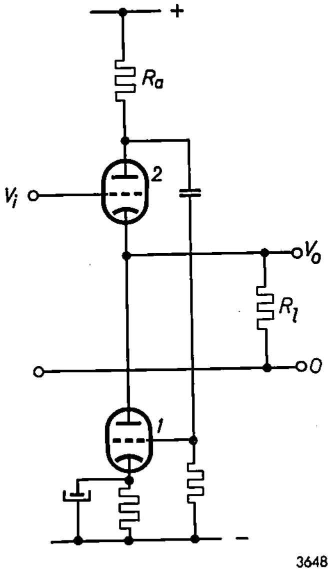


Fig. 15. Basic diagram of a cathode follower consisting of two valves in cascade. If the two valves are pentodes, the gain differs very little from unity and is largely independent of the valve characteristics. Moreover, the input impedance is very high since no grid leak is needed, the cathode follower being directly coupled to its probe and thus biased to the potential of the probe, which differs little from zero.

## Applications and performance

In electrostatic fields with no space charge the apparatus described is capable of plotting electron trajectories lying in planes of symmetry. For electrostatic lenses having rotational symmetry, these are the planes through the axis. By plotting a series of trajectories from different starting points and with different initial directions, it is possible to determine the focal planes, principal planes, spherical aberration, distortion and field curvature of such lenses. Certain forms of misalignment and axial astigmatism (lens errors caused by the electrodes not being centered on the axis or being non-circular) can also be investigated.

As our first example we shall consider the trajectories in a planar triode having an intersecting-"wire" gauze grid, where the grid is at a positive potential, and the average field strength between grid and anode is zero. *Fig. 16a* shows the trajectories in a perpendicular section which cuts the gauze parallel and perpendicular to the directions of the "wires", and fig. 16*b* in a section that cuts the gauze diagonally. We see that in the first case the trajectories become more divergent the closer they pass by the grid "wires". In the second case the trajectories are again divergent, but their divergence shows a maximum, that is to say the trajectories passing

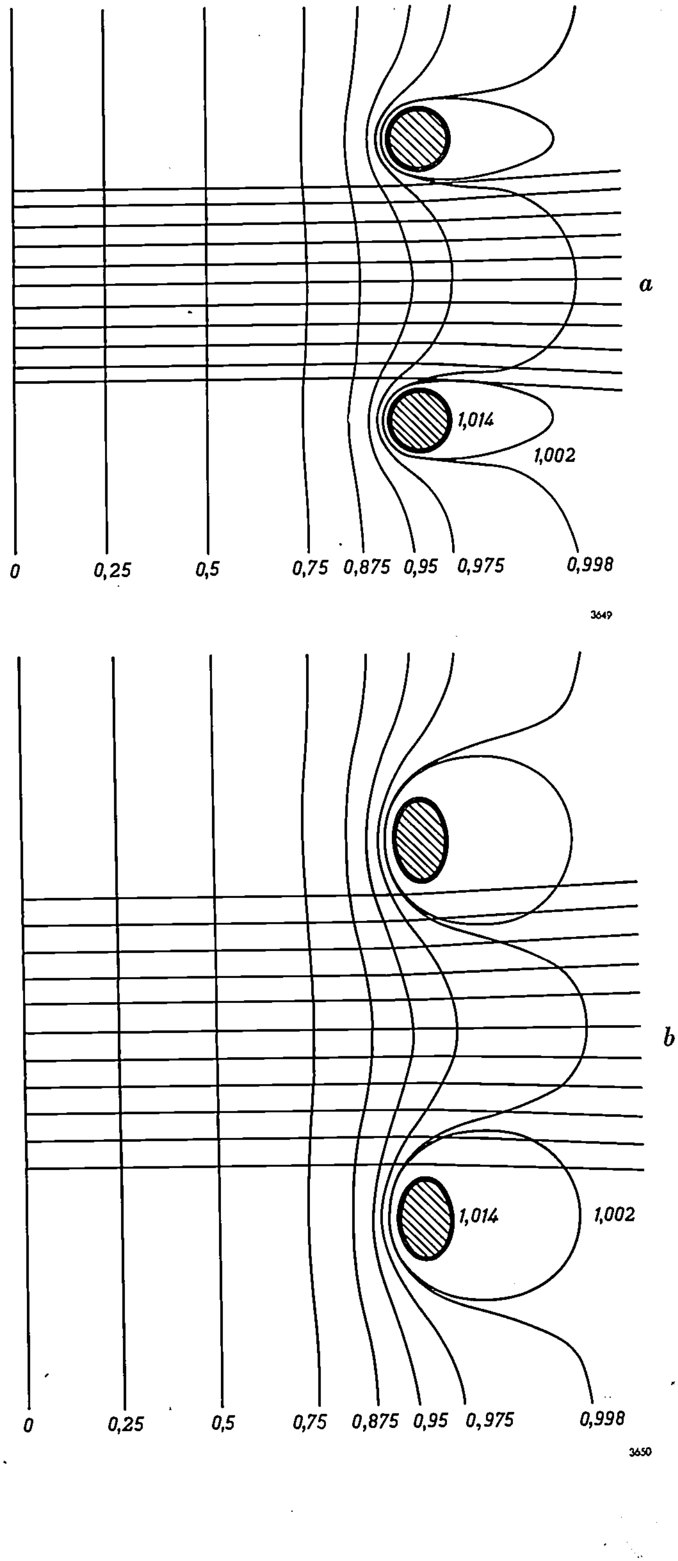


Fig. 16. The electron trajectories and equipotential lines in the field of a planar triode with an intersecting-"wire" gauze grid. (These are drawings copied from the plotted trajectories.)
*a*) The plane in which the trajectories are drawn cuts the gauze parallel and perpendicular to the directions of the "wires".
*b*) The plane cuts the gauze diagonally.

In both drawings the cathode, which has zero potential, is on the left, and the grid, which has a positive potential, is on the right. The anode is not shown. Between grid and anode there exists an equipotential space. The potential of this space is put equal to 1, and the relative potential with respect to this value is marked beside the equipotential lines.

close by the points of intersection of the grid wires are deflected less than those that pass somewhat farther away. Also drawn in fig. 16*a* and *b* are the equipotential lines.

The second example (*fig. 17a*) concerns the electron trajectories in an accelerating lens consisting of two adjacent coaxial cylinders of identical diameter. Both cylinders have a positive potential with respect to the cathode, the right one being 20 times as high as the left one. The trajectories drawn

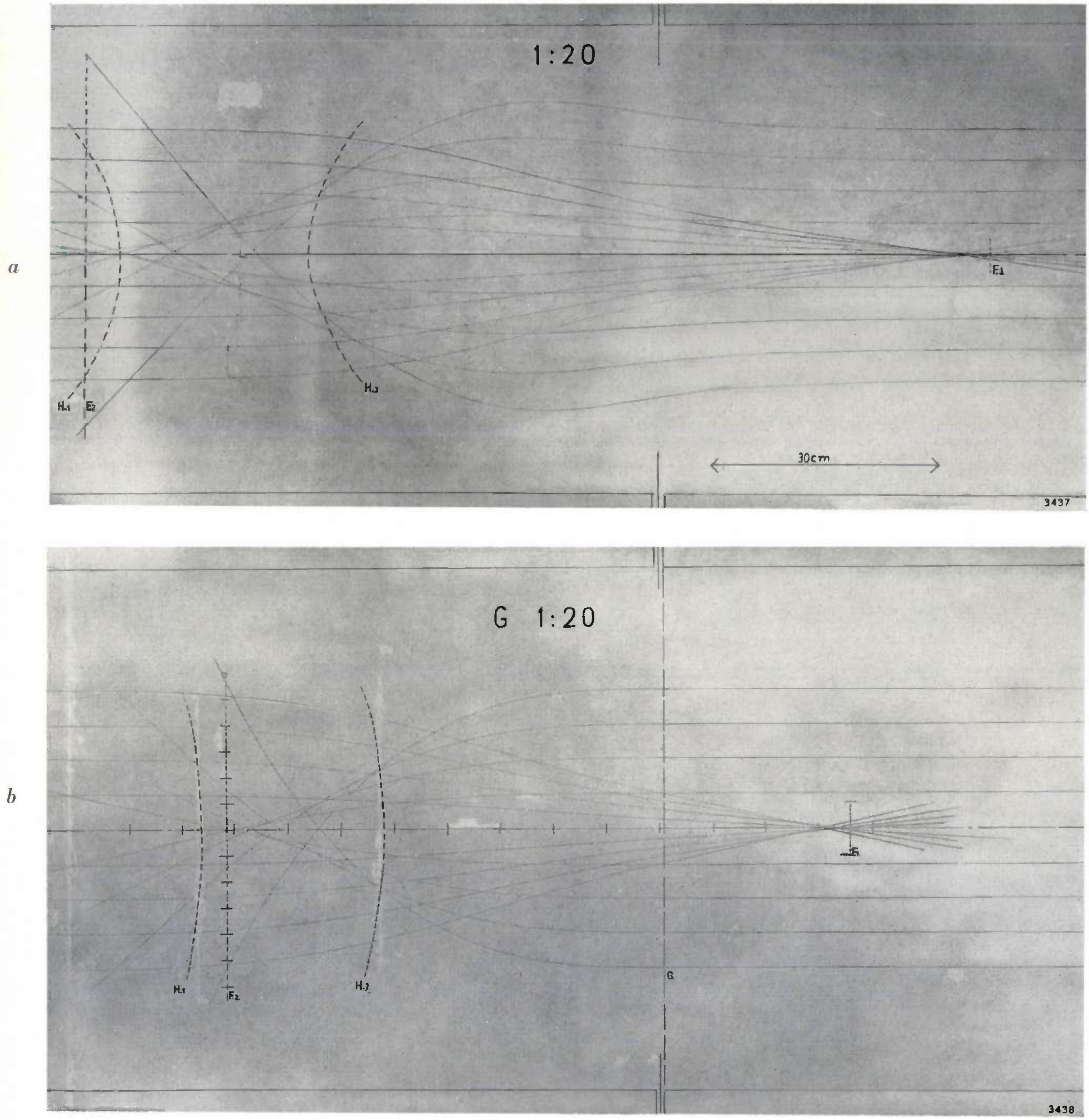


Fig. 17. *a*) Electron trajectories as traced out by the automatic plotter for the model of an accelerating lens. (The electrodes of this model are shown in fig. 1.) The lens consists of two adjacent coaxial cylinders of the same diameter; the potential with respect to earth of the right cylinder is 20 times that of the left one. The trajectories are drawn in a plane through the axis of the lens. The electrons enter parallel to the axis, from the left and also from the right. The principal planes and focal planes are indicated, together with the situation of the electrodes.
*b*) The trajectories in the same accelerating lens, but now with a plane gauze *G* at the inner end of the right cylinder, having the same potential as the cylinder. It can be seen that this makes the lens more powerful and results in a considerably smaller curvature of the focal planes.

are those of electrons entering parallel to the axis, both from the left and from the right. From these the focal planes and principal planes can be determined. Fig. 17*b* shows the trajectories in the same lens when a flat gauze is fixed on the inner end of the right cylinder, the gauze having the same potential as the cylinder. It can be seen that the power of the lens is now considerably greater. (For a detailed treatment, see reference [5].)

Potential distributions can also be measured with this instrument. The determination of a potential at a particular point in a field is greatly simplified by making use of various components that are also used for plotting electron trajectories. The ends of the decade potentiometer $P_{\text{dec}}$ are then connected to the electrodes which are respectively at the lowest and highest potential. The voltage on the wiper of $P_{\text{dec}}$ is applied to an oscilloscope. The stylus on the trolley is placed over the measuring point, and $P_{\text{dec}}$ is adjusted until the voltage on its wiper is zero. Sincc the computing circuit for $V$ keeps the potential at the point midway between the probes equal to zero, the wiper of $P_{\text{dec}}$ and the measuring point now have the same potential with respect to the electrodes. This potential can be derived from the position of $P_{\text{dec}}$.

In order to plot a complete equipotential line, $P_{\text{dec}}$ is again connected up in the same way and its wiper is adjusted to the potential of the equipotential line in question. When the stylus is above a point that possesses this potential with respect to the electrodes, the voltage of the wiper with respect to earth will then, for the same reason as above, be zero. The circuit is now modified as follows. The steering circuit and the servo motor are disconnected, and the wiper of $P_{\text{dec}}$ is connected to the input of the servo amplifier. The output of this amplifier energizes a relay. A make contact of this relay is connected in parallel with the stylus relay, so that the latter is energized only when thc voltage from the servo amplifier, and thus the potential of the wiper, drops to zero. When the trolley is now moved to and fro by hand over the place where the equipotential line is presumed to be, successive dashes are recorded. These dashes are later joined up to make a continuous line.

The automatic plotter has an accuracy of 0.2%, which means that the relative error in the radius of curvature of the trajectory is smaller at every point than 0.2%. All circuits and components of the plotter must be optimally adjusted, and the speed of the trolley must be properly adapted to the curvature of the path described. The average time taken to plot a trajectory is about 2 to 3 minutes per meter.

The accuracy of the system should be checked now and then by plotting a number of paths in the field between two parallel plates (the paths should be parabolae) and in the field of a spherical capacitor. In both these cases the form of the paths can be derived exactly. Any components or circuits requiring re-adjustment can be identified from the deviations shown by the paths thus described.

**Summary.** Electron trajectories in an electrostatic field can be automatically plotted by an apparatus designed round an electrolytic tank. The tank contains a model of the electrode system in which the trajectories are to be determined. On a board above the tank rides a three-wheel trolley which is mechanically coupled to four closely-spaced probes, mounted in line and dipping in the electrolyte. The radius of curvature of the trajectory at the point midway between the probes is derived from the voltages of the four probes. The apparatus does this by means of two computing circuits. A servo system ensures that the trolley — propelled at a suitable speed of, say, half a meter per minute — describes a path which everywhere has the correct radius of curvature. The trajectory is traced by a stylus fixed underneath the trolley.

The electrodes of the model are fed with a square-wave alternating voltage of 500 c/s with the object of minimizing polarization. For the same purpose the electrodes are silver-plated. The accuracy achieved is 0.2%. Practical examples mentioned are the plotting of electron trajectories in an accelerating lens, with and without a gauze, and the trajectories in a planar triode. The instrument can also be used as a particularly simple means of determining equipotential lines.

# A SMALL GETTER ION-PUMP

by A. KLOPFER *) and W. ERMRICH *).

621.528.5/.6

*The advances made in vacuum technique in the last ten years have given a new lease of life to the apparently outmoded idea of continuously pumping a vacuum tube during operation.*

The getter ion-pump forms a compact, convenient vacuum pump, immediately set in operation and capable of pumping small quantities of gas to maintain a low pressure for a long period of time.

The principle of this pump has been known for some years [1]: a getter material, such as zirconium or titanium, is continuously or intermittently evaporated, and the gas molecules are ionized in the same way as in a hot-cathode ionization gauge [2]) or a Penning vacuum gauge [3]). The ions formed are then trapped in the getter film deposited on the inside of the envelope.

In this article a getter ion-pump will be described in which a certain quantity of titanium is evaporated, and where the titanium vapour itself simultaneously sustains a discharge of the type occurring in a Penning vacuum gauge. In conclusion, various examples of the pump's application will be discussed.

## Construction and operation of the getter ion-pump

The pump consists of a glass envelope containing an electrode system. The electrode assembly (*fig. 1*) is similar to that in a Penning gauge. The Penning system is formed by the loop-shaped anode $A$, made of molybdenum, the cathodes $K_1$-$K_4$, and a magnetic field. The cathodes carry a supply of getter material. The direction of the magnetic field, with an induction of 0.04 Wb/m$^2$, is perpendicular to the plane of the anode. The electrodes are mounted on a glass base. The cylindrical glass bulb measures 33 mm in diameter and 60 mm in length. The weight of the pump (40 g) together with the permanent magnet is 450 g. The complete assembly is therefore conveniently light and easy to connect up with the object to be evacuated.

The pump is set in operation by applying a DC potential of 2 kV between anode and cathodes, the latter being heated by an alternating current. It is also possible to heat only one cathode. The cathodes consist of stranded tungsten and titanium wires, and are heated to a temperature at which the titanium evaporates. The rate of evaporation is governed by the temperature and the number of cathodes heated. Part of the evaporated getter material settles directly on the glass walls of the pump; another part contributes to the Penning discharge, in that, during evaporation, titanium atoms are ionized and electrons are released. If the total pressure of the gases present is less than $10^{-3}$ torr ($10^{-3}$ mm Hg), the discharge is sustained not by these gases but almost entirely by the titanium vapour. The molybdenum shield $P$, together with the protective caps $T$, serves to prevent short-circuiting between anode and cathode if the titanium were to

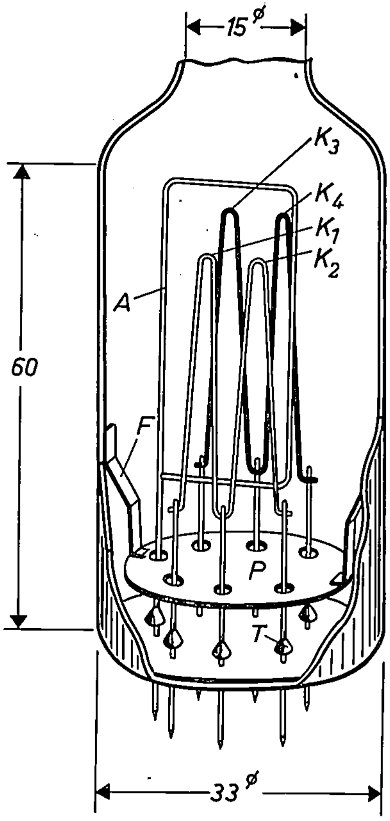


Fig. 1. Getter ion-pump in glass envelope. $K_1$-$K_4$ cathodes. $A$ anode. $P$ shield. $T$ protective caps. $F$ contact spring.

*) Philips Zentrallaboratorium GmbH, Aachen Laboratory.

[1] R. G. Herb, R. H. Davis, A. S. Divatia and D. Saxon, Phys. Rev. **89**, 897, 1953.

[2] A. Venema and M. Bandringa, The production and measurement of ultra-high vacua, Philips tech. Rev. **20**, 145-157, 1958/59.

[3] F. M. Penning, Physica **4**, 71, 1937; see also Philips tech. Rev. **11**, 116, 1949/50.

deposit on the glass base between the lead-ins. That can only happen, however, if the titanium is evaporated at excessive pressures, causing the mean free path of the titanium atoms to become shorter than the distance to the glass wall. In that case, as a result of collisions in the gas, titanium atoms may arrive on the base. If the pump is to be used for producing very high vacua, so that the titanium is only evaporated at low pressures, the shield *P* can better be removed. The reason for this is that, for such low pressures, all metal parts must be degassed by heating; this can be done very simply for the electrodes by passing a current through them, but this is not practicable in the case of the shield. (The anode loop would be led out through the base via two pins to enable current to be passed through it.)

The power supply for the pump consists merely of a DC voltage source of 2 kV, 2 mA, and an AC voltage source of 10 V, 10 A.

The pump operates as follows.

1) The getter material deposited on the glass wall adsorbs the molecules of the chemically active gases impinging on the surface, such as oxygen, hydrogen, nitrogen, etc.

2) The Penning discharge contains a vast number of ions and electrons. The ions, but especially the electrons, which can describe a long path owing to the presence of the magnetic field, have a considerable chance of colliding with gas molecules entering the discharge from the space to be evacuated. The chemically active as well as the inert gas molecules are thereby ionized. The gas ions formed are then, under the action of the electric field, shot into the getter film — which is kept at cathode potential by the contact spring *F* — and there they are trapped.

As a result of continuously evaporating the titanium, the getter ion-pump, as its name implies, has a double action in that a fresh surface is continuously created for trapping ions, and a fresh chemically active surface for gettering.

### Properties of the pump

The characteristic properties determining the performance of a getter ion-pump are:

*a*) the maximum permissible initial pressure,
*b*) the pumping capacity,
*c*) the pumping speed, and
*d*) the lowest attainable pressure (ultimate pressure).

The maximum permissible initial pressure, which is determined by the method of evaporation chosen, amounts in the pump described here to a few tenths of a torr.

The pumping capacity for chemically active gases, such as CO and $H_2$, i.e. the total quantity that can be extracted, is given by the getter supply present in the cathode. For chemically active gases the pumping capacity is equal to the gettering capacity. This differs according to the gas, the maximum value being achieved when each getter atom traps as many gas molecules as corresponds to the chemical reaction equation. In practice this situation seldom obtains. For each cathode we generally use 12.5 mg of titanium in the form of three wires, each of 100 microns diameter, stranded with the tungsten heater wire. The four cathodes together thus contain 50 mg of titanium. The quantity of gas that can be removed under favourable conditions with this quantity of getter material is about 2.5 torr.litre in the case of CO, calculated for a maximum specific gettering capacity of 0.05 torr.litre/mg. Since a vacuum system may contain a mixture of the most diverse chemically active gases, one can generally reckon only with an average specific gettering capacity of 0.01 torr.litre/mg.

For inert gases, such as He and $CH_4$, the pumping capacity is considerably less. Measurements have shown, however, that in the case of He it is certainly more than 0.05 torr.litre.

In principle the cathodes could accommodate a larger supply of getter material.

The quantity of gas removed per second in the steady state is given for a chemically active gas by the expression:

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = C\frac{\mathrm{d}M}{\mathrm{d}t} = Sp. \quad . \; . \; . \; . \; . \quad (1)$$

Here $\mathrm{d}M/\mathrm{d}t$ represents the rate of evaporation of the getter material, *C* the specific gettering capacity, *S* the pumping speed (governed by the speed with which the getter takes up gas and the geometry of the glass envelope of the pump) and *p* the resultant pressure at the pump mouth, i.e. the place where the getter ion-pump is connected to the vessel to be evacuated. *Fig. 2* shows the measured pumping speeds *S* in l/sec as a function of the rate of evaporation $\mathrm{d}M/\mathrm{d}t$ in mg/sec, for various quantities of CO supplied per unit time in torr.litre/sec. The maximum value for *S*, in this case 62 l/sec, is determined by the orifice of the pump mouth. For other chemically active gases the pumping speed *S* is of the same order of magnitude. The curves in fig. 2 were obtained irrespective of whether the anode was under high tension or not. This means that, for CO, the contribution of ion trapping to the total pumping speed in the measuring range investigated is small compared with the contribution of the gettering action alone.

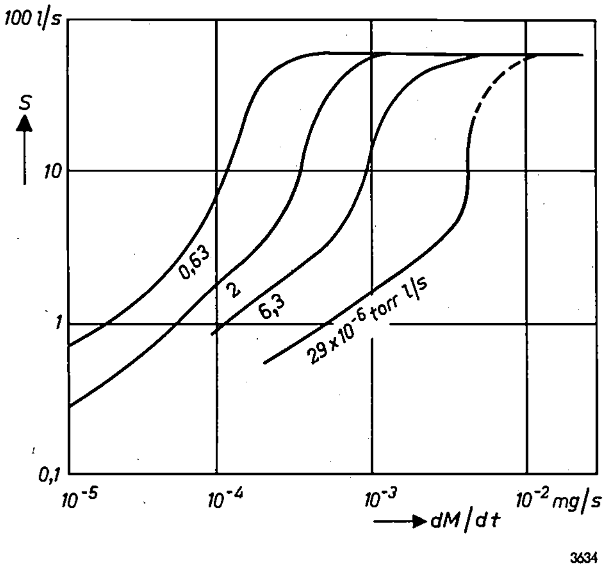


Fig. 2. Pumping speed $S$ in l/sec as a function of evaporation rate $\mathrm{d}M/\mathrm{d}t$ in mg/sec, for various influx rates of CO in torr.litre/sec.

In order to use equation (1) it is desirable to determine the relation between the specific gettering capacity $C$ and the pumping speed $S$. From the curves in fig. 2 we can calculate for any value of $S$ the corresponding values of $C$ for the various quantities of CO supplied. The surprising result is that the same value of $C$ is always found for a given $S$. A plot of $S$ versus $C$ is shown in *fig. 3*. Along the abscissa at the top the specific gettering capacity is setforth in torr.litre/mg, and at the bottom the quantity $C$ as the ratio between the number of CO molecules trapped and the number of titanium atoms incident on the getter surface. Only at large values of $S$ is the relation between $S$ and $C$ no longer certain, for it is here, as we have seen, that the size of the pump orifice imposes an upper limit on the pumping speed. Apart from this, the pumping speed is established for any given value of $C$. We may therefore deduce that a constant $S$, independent of the influx rate of gas, can only be obtained if the rate of evaporation varies in proportion to the pressure in the pump, since it follows from (1) that

$$S = C\left(\frac{\mathrm{d}M}{\mathrm{d}t} / p\right). \quad . \quad . \quad . \quad . \quad . \quad (2)$$

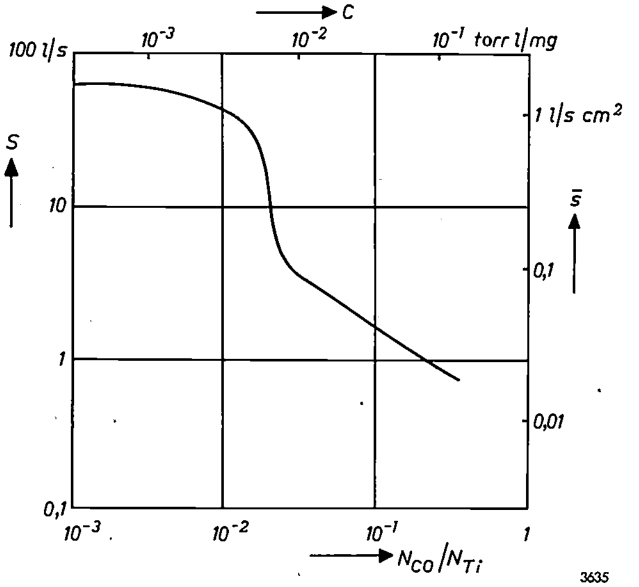


Fig. 3. Pumping speed $S$, in l/sec as a function of specific gettering capacity $C$ in torr.litre/mg, calculated from the data in fig. 2; or, which is equivalent, the ratio $N_{\mathrm{CO}}/N_{\mathrm{Ti}}$ between the number of gettered CO molecules and the number of titauium atoms incident on the getter surface, as a function of $\bar{s}$, the mean pumping speed (in l/sec) of the getter film per cm$^2$.

A fairly good approximation to the rate of evaporation of the hot cathode is arrived at by deducting from the discharge current $I$ in the pump the discharge current when the cathode is cold. The relation between the rate of evaporation and the discharge current is represented by the curve in *fig. 4* (corrected for residual gas).

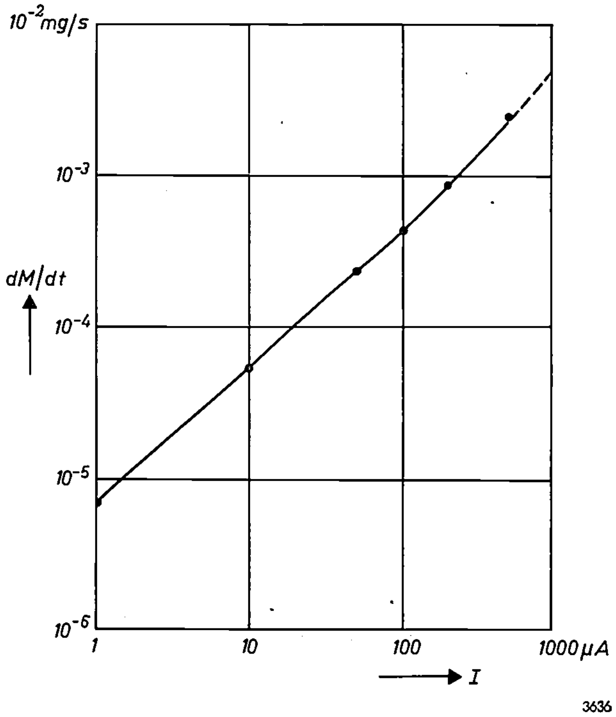


Fig. 4. Evaporation rate $\mathrm{d}M/\mathrm{d}t$ in mg/sec of one cathode in the getter ion-pump, as a function of the discharge current $I$ in µA.

The above considerations make it evident that the discharge current when the cathode is hot is no measure of the gas pressure. It *is*, however, when the cathode is cold, and — which is sometimes convenient — the pump may then be used for pressure measurements in a certain range of pressures, as shown by the curves in *fig. 5*. These curves relate to both $N_2$ and CO at a magnetic induction of 0.04 Wb/m$^2$ and an anode voltage of 1 and 2 kV, respectively, with an anode resistor of 1 MΩ.

The pumping speeds found for chemically inactive gases, such as the rare gases and methane, are much

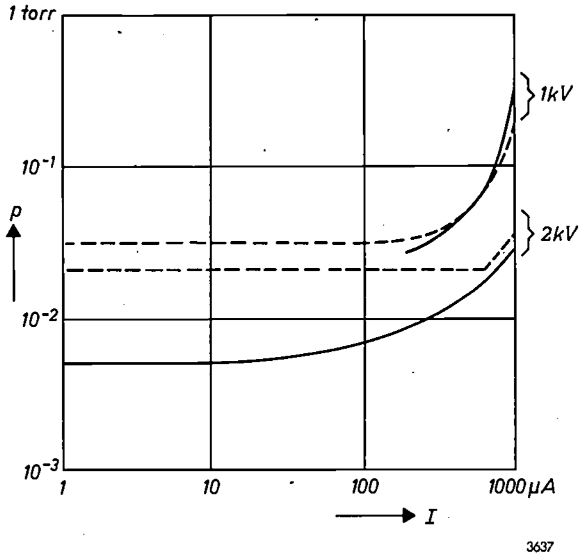


Fig. 5. The getter ion-pump as a pressure gauge: the pressure $p$ in torr as a function of the discharge current $I$ in µA, with cold cathode, for two anode voltages. Full curves: with getter film; dashed curves: without getter film.

lower (see *Table I*). It can be seen from the table that the pumping speed varies only slightly with the discharge current. The rare gases are not chemically bound by the layer of evaporated titanium, and in this case the pumping action relies entirely on the ionization in the gas. The ions formed are shot into the titanium layer and trapped there. How long they remain trapped will depend entirely on the rate of diffusion in the titanium layer. As the rate of diffusion is temperature-dependent, it is advisable to cool the pump wall with a fan. Where methane is concerned, thermal decomposition at the hot cathode also makes an essential contribution to the pumping speed.

The lowest pressure obtainable with this getter ion-pump depends very much on the way in which the pump itself and the vessel are degassed. Degassing is necessary for various reasons, mainly to reduce the hydrogen content of the getter reserve and to prevent the unwanted formation of non-active gases such as $CH_4$ as a result of reactions between getter and gas [4]). The formation of chemically inactive gases is undesirable because of the low pumping speed for such gases.

On a small vacuum system having a volume of 0.5 l, which consisted of an omegatron, an ion gauge of the Bayard-Alpert type and a getter ion-pump of the type here described, we measured with the omegatron the partial pressures of the various gas components during the process of evacuation. Four such systems, in various degassed states, were investigated.

A system which had not been degassed at all was found to contain, after evaporation of the getter, mainly water vapour, carbon dioxide and hydrocarbons. The pressure of $H_2O$ and $CO_2$ is governed by the release of gas from the glass surfaces and by the effective pumping speed at the position of the omegatron. The presence of hydrocarbons is attributable to two reactions: in the first place, during the evaporation, hydrogen and carbon present as impurities in the getter may react with one another and form hydrocarbons with up to five carbon atoms. In the second place, hydrocarbons may be formed from water and carbon at temperatures as low as room temperature, as we were able to demonstrate. As a result of the above-mentioned gas desorption and gas reactions, the lowest pressure in a system that has not been degassed lies between $10^{-7}$ and $10^{-8}$ torr [5]).

Table I. Pumping speed $S$ for various non-active gases.

| Gas | Discharge current $I$ µA | Pumping speed $S$ $10^{-3}$ l/sec | Pressure $p$ $10^{-3}$ torr |
|---|---|---|---|
| He | 100-1000 | 2 | 0.01-0.03 |
| Ne | 500 | 5.5 | ~ 1 |
| Ar | 100 | 3.5 | ~ 1 |
| | 500 | 7.5 | ~ 1 |
| $CH_4$ | 100 | 74 | ~ 1 |
| | 500 | 210 | ~ 1 |

Even slight preliminary degassing causes a marked drop in the release of $H_2O$ and $CO_2$. The pumping time and the final pressure attainable are then entirely governed by the formation of hydrocarbons from $H_2$ and C during the evaporation of the getter, and by the pumping speed of the ion pump in respect of hydrocarbons. The pumping times are shorter the more thoroughly the vacuum system and the getter wires have been degassed. Pressures of about $1 \times 10^{-9}$ torr can be achieved after relatively little preliminary degassing. The residual gas is principally composed of $CH_4$, CO and $H_2$. Thorough degassing, as described by Alpert [6]), makes it possible to reach extremely low pressures ($< 10^{-10}$ torr). The final pressure is then determined by the equilibrium pressure of the hydrogen dissolved in the titanium, and by the diffusion of atmospheric helium through the glass wall [4])[7]). Heating of the getter film by the hot cathode promotes the formation of $CH_4$ and raises the equilibrium pressure of the hydrogen.

[4]) A. Klopfer and W. Ermrich, Vakuum-Technik **8**, 162, 1959.

[5]) G. Reich and H. G. Nöller, Vacuum Tech. Trans. **4**, 97, 1959.

[6]) D. Alpert and R. S. Buritz, J. appl. Phys. **25**, 202, 1954.

[7]) F. Norton, J. appl. Phys. **28**, 34, 1957.

For this reason too it is advisable to cool the pump wall with a fan.

### Some applications

In order to produce vacuum required in electron tubes, the tubes are subjected during manufacture to two different evacuation operations. First of all, the electron tube is degassed and evacuated at suitable temperatures for a sufficiently long period with the aid of a good high-vacuum pump, usually a diffusion pump. Because of the bulk of the pumping equipment, the tube during this period is virtually immovable. At the end of this major evacuation process, the system consisting of electron tube and getter is sealed off. Further degassing can be effected by letting the tube operate normally or by even overloading it, whereby the evaporated getter material together with the ionizing electron current in the tube is responsible for reducing the residual gas pressure. The loading of the tube must not be increased so fast as to cause the pressure to assume values at which components, e.g. a photo-cathode, might be damaged.

The use of the small getter ion-pump described has considerable advantages for the evacuation of special electron tubes, the degassing of which must often be very protracted. For example, the outgassing times during evacuation by the diffusion pump can be shortened, and the baking-out process and the degassing under initial load can be carried out whilst the getter ion-pump takes over the task of the diffusion pump. Because of the small weight and dimensions it is now possible to transfer the tube together with the pump to the test equipment. Since the getter ion-pump is inexpensive, one can be fitted to each electron tube, and in certain circumstances it is possible for the pump to remain connected to the tube for life. Whereas it was formerly the normal practice to work with tubes on the pump, it is here rather a question of working with the pump on the tube.

Generally speaking, it is cheaper to use a getter ion pump in cases where long periods of evacuation would be necessary on the diffusion pump. It is also better to use a getter ion-pump when lower pressures than are normally required are needed for certain experiments.

We shall now consider three examples to demonstrate the application of our getter ion-pump.

In various experiments with a type of oscillator tube related to the magnetron [8]), designed for high-power pulsed operation at decimetre wavelengths, the object was to maintain a pressure of $10^{-6}$ torr during operation. The tubes had a volume of about 2 l and were all-metal, with a copper envelope and molybdenum electrodes. In the first place the tubes, to which an ion gauge and a getter ion-pump were connected, were evacuated on an oil-diffusion pump and degassed for 18 hours. The impregnated cathodes [9]) of the tubes were heated up to about 1250 °C, and the cathodes of the getter ion-pump to about 900 °C at the end of the degassing period. Between the tube and the oil-diffusion pump two cold traps were mounted, one cooled with water and the other with liquid air. The pressure in the tubes, after seal-off from the diffusion pump, amounted to roughly $5 \times 10^{-7}$ torr. When the tubes were afterwards put into operation, the pressure *without* the getter ion-pump in operation rose to $10^{-4}$ torr. With the getter ion-pump in operation, however, it proved possible, after an initial rise in pressure, to pump off the released gas rapidly and to maintain a pressure lower than $10^{-6}$ torr during the further operation of the tube.

As a second example we shall mention the application of the pump in various experiments on low-noise travelling-wave tubes [10]). To keep the noise level low, the condition imposed on the vacuum in these tubes is that the pressure under load should not exceed $10^{-9}$ torr. The travelling-wave tubes, which deliver an RF power of a few microwatts at 4 Gc/s, have a glass envelope measuring 40 cm in length and 2 cm in diameter (approximate volume 0.15 l); the electron source is a barium-oxide coated cathode, and the helix is a molybdenum wire. The vacuum system, which consisted of the travelling-wave tube, an ion gauge and a getter ion-pump of the type described (*fig. 6*), was evacuated with diffusion-pump equipment described in this journal some time ago [2]). During the last hours of the degassing period, which lasted 48 hours and during which the tube was baked out at a temperature of 400 to 450 °C, the electrodes of the ion gauge and the getter ion-pump were glowed. The oxide cathodes of the tubes were activated shortly before seal-off, while the whole system was still at a temperature of 200 °C. After seal-off the pressure amounted to $10^{-7}$ torr. The getter ion-pump reduced the pressure to $10^{-9}$ torr, and it was possible to maintain this pressure throughout operation of the tube.

In the examples given, the getter ion-pump was used to maintain the vacuum in the tube during operation.

[8]) Investigated, using the getter ion-pump, by T. I. Sprenger, Philips Research Laboratories, Eindhoven.

[9]) R. Levi, Philips tech. Rev. **19**, 186, 1957/58.

[10]) Investigation, using the getter ion-pump, by A. Versnel. Philips Research Laboratories, Eindhoven.

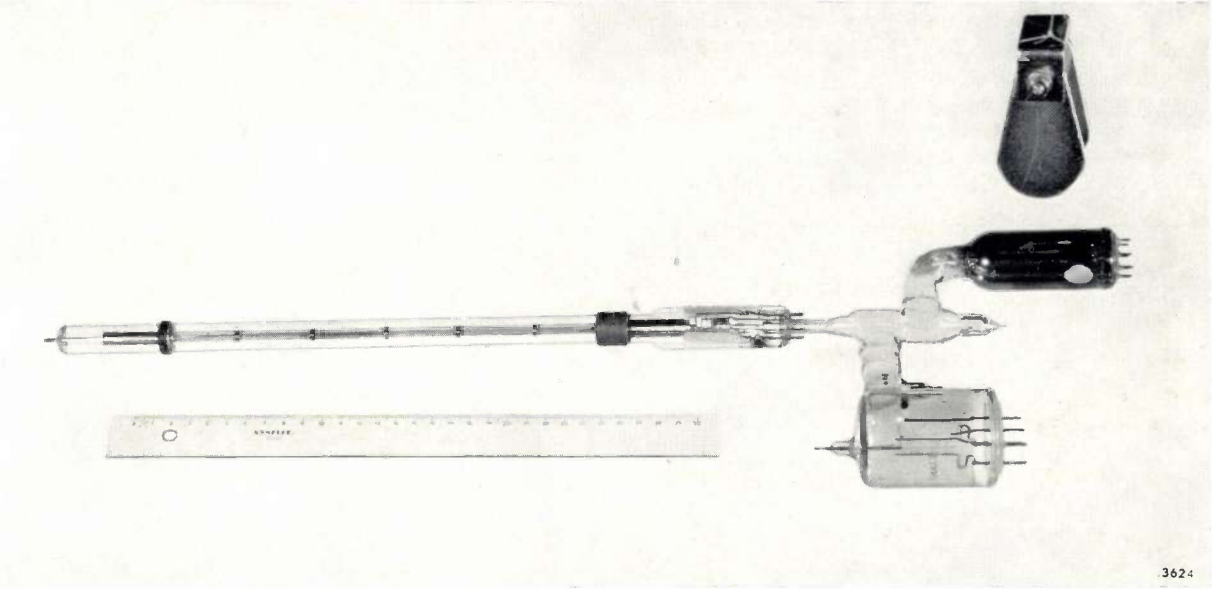


Fig. 6. Travelling-wave tube with getter ion-pump (above) and ion gauge (below) fused to one end. Top right, above the pump, can be seen the magnet for the getter ion-pump.

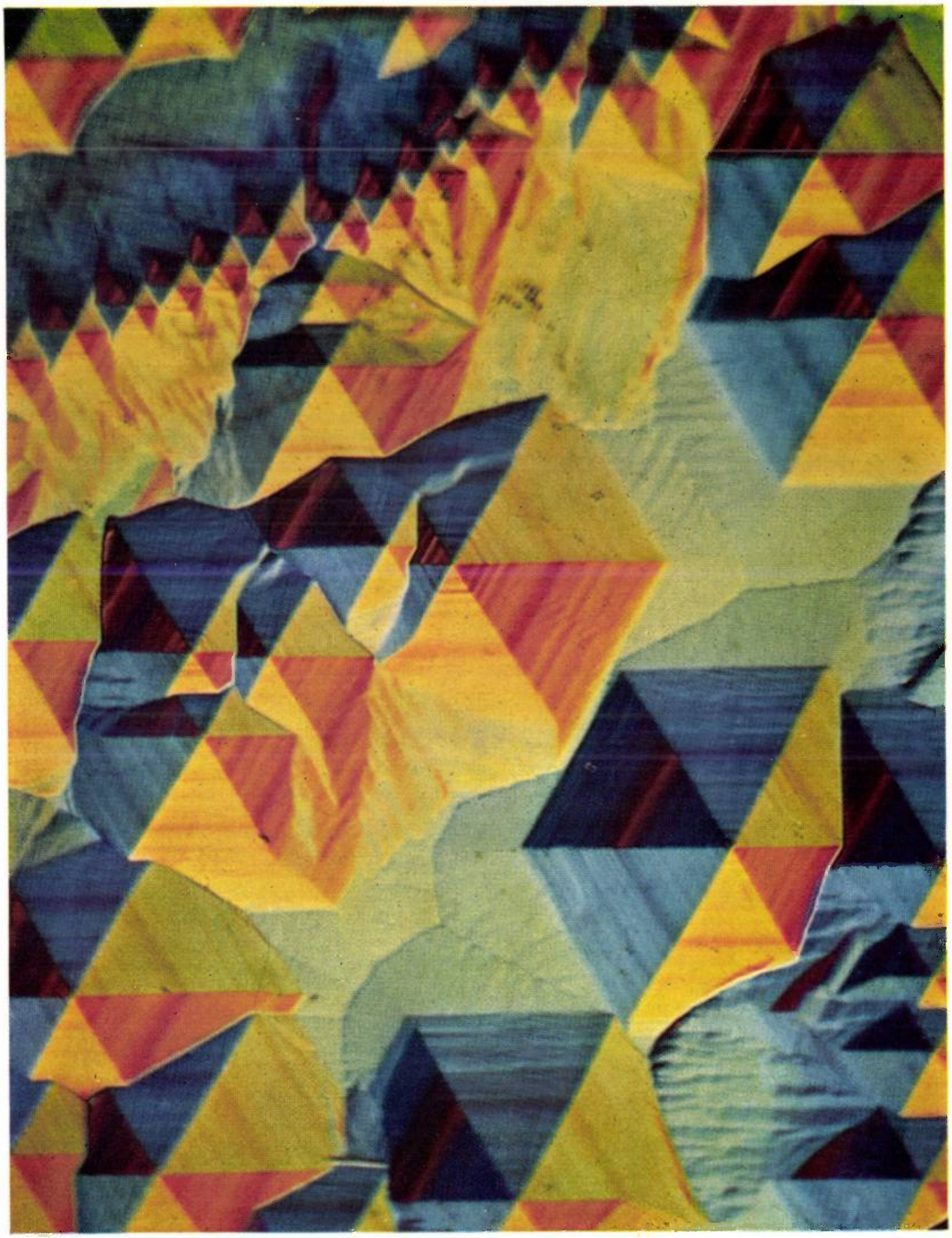
As a third example of the use that may in principle be made of a getter ion-pump, we mention in conclusion the evacuation of television picture tubes during manufacture [11]). Tubes having screen diagonals of both 43 and 53 cm were evacuated by using solely a two-stage rotary backing-pump and a getter ion-pump. This method proved to be highly satisfactory. As it happens, however, its application in mass production offers no special advantages over the usual method of pumping, which is already very economical.

**Summary.** For evacuating certain types of electron tubes, use can be made of a getter ion-pump. Because of its compactness, light weight and low cost, many special types of electron tubes can retain their own pump throughout their working life. This article describes a getter ion-pump using titanium as getter material. Ionization takes place as in a Penning vacuum gauge; the titanium itself sustains the Penning discharge. With its reserve of 50 mg of titanium the pump can remove a total of 2.5 torr.litre of CO. The maximum pumping speed is 62 l/sec. The lowest pressure achieved in small vacuum systems is roughly $10^{-10}$ torr.

[11]) These experiments were done partly by C. J. W. Panis and J. van der Waal in the Development Laboratory of the Electron Tubes Division, Eindhoven, and partly in the laboratory at Aachen.

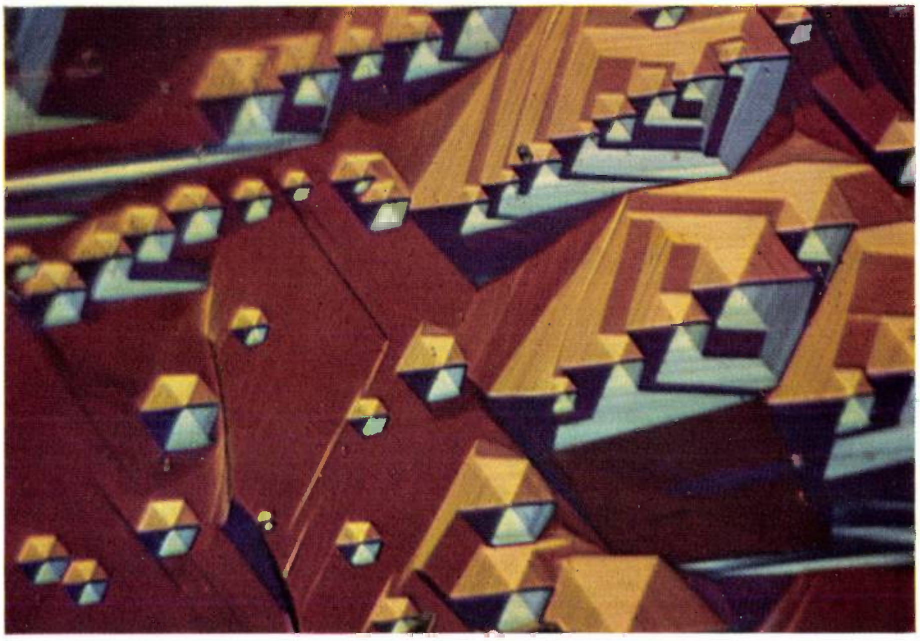
# ETCH PITS ON CADMIUM-SULPHIDE CRYSTALS

548.4:546.48'22



The properties of cadmium sulphide for use in photoresistors, electroluminescent panels and other solid-state applications depend in large measure on the chemical impurities in the crystallites — a fact which has been the subject of much research. They also depend, however, on the physical imperfections of the crystal, and to investigate their influence, attempts are made to grow CdS crystals as nearly perfect as possible. In recent times various laboratories have succeeded in growing good single crystals of CdS having dimensions up to a few centimetres. CdS can occur in two crystal modifications — cubic and (polar) hexagonal. Under the present conditions of growth (sublimation at relatively high temperature) the hexagonal modification is obtained.

A crystal grown in this laboratory was etched on the basal plane in hydrochloric-acid vapour for about 1 minute (after wetting with water). Observed



under a microscope, using interference contrast [1]), the etched surface shows colourful patterns of etch pits, two photographs of which are reproduced here (overall magnification about ×250). The symmetry of the CdS crystal comes out beautifully in the hexagonal-pyramidal etch pits. It is not yet entirely certain how the formation of the steps visible in the pits should be interpreted.

Examination of the photographs reveals that many of the etch pits are disposed along straight lines, and that these lines are roughly parallel to one another, with only a few isolated pits between them. This may imply that the crystal possesses a high degree of perfection, for the following reasons.

For many crystalline substances a point-bottomed etch pit indicates the terminal point on the surface of a dislocation. A series of equidistant parallel dislocations running through the crystal lattice represents a "mosaic" boundary (sub-grain boundary) between two somewhat differently oriented crystalline blocks [2]). Assuming that this is the case here, the fact that the boundaries observed are few in number, and run more or less parallel, is in itself an indication that the structure of the crystal as a whole is good. Moreover, we can make an estimate of the difference in orientation between two adjacent mosaic blocks. We consider here only the case of a symmetrical tilt boundary, made up of a row of parallel edge dislocations. (Similar considerations apply to other types of dislocation boundaries.) If $b$ is the magnitude of the Burgers vector (the distance over which one part of the lattice is sheared with respect to the adjacent part), and $d$ is the separation between successive dislocations (etch pits, assuming a 1 : 1 correspondence) along the grain boundary, then the angle between the two grain orientations is given by $\alpha \approx b/d$. We deduce from the photographs that $d$ has the relatively large value of approximately 25 microns. Further, the grain boundaries are seen to lie roughly in the $\langle 10\bar{1}0\rangle$ direction; the Burgers vector of the (edge) dislocations therein is then perpendicular to that direction, in other words in the $\langle 1\bar{2}10\rangle$ direction, i.e. along the $a$-axis of the hexagonal structure. Assuming the magnitude of the Burgers vector to be equal to the lattice spacing in the $\langle 1\bar{2}10\rangle$ direction, we calculate that $\alpha = b/d = 15\times10^{-6}$ radians $= 3''$. It would seem, therefore, that adjacent mosaic blocks in this crystal show extremely small differences of orientation, of the order of a few seconds of arc.

A. J. ELAND.

[1]) The technique used, which can make very small differences in height visible — often better than Zernike's phase-contrast method — was developed by Nomarski. (See G. Nomarski and A. R. Weill, Rev. Métall. **52**, 121, 1955.)

[2]) See e.g. Philips tech. Rev. **15**, 246 and 286, 1953/54, for an introduction to the subject of dislocations, in which such concepts as the Burgers vector are explained; for polygonization to form mosaic boundaries, see p. 291.

# SOLID-STATE RESEARCH AT LOW TEMPERATURES

## III. THERMAL CONDUCTION IN INSULATORS; PARAMAGNETISM; DIELECTRIC LOSSES RELATED TO CHEMICAL LATTICE IMPERFECTIONS

by J. VOLGER.

536.48

*In this third and last article in the series on solid-state research at low temperatures [1]) further examples are given of recent investigations. Those discussed here, however, are based on methods differing from those dealt with in the previous article, and are mainly concerned with research on insulators. Some of the low-temperature effects referred to here reveal the consequences of the fact that many processes proceed more slowly at low temperature.*

### Thermal conduction in insulators

The mechanism of thermal conduction in insulators closely resembles that of electrical conduction in metals. In the previous article [1]) we saw that the electrical resistance of a metal, where the temperature is not too low, is primarily determined by the thermal vibrations of the crystal lattice, and at low temperature primarily by the scattering of electrons from lattice imperfections. In the transport of heat, analogous considerations apply to the lattice vibrations themselves, which both *constitute* the thermal energy and are the carriers or "transporters" of it (in insulators they are the *only* carriers). Whereas, in the case of metals, information on lattice imperfections can be derived from the behaviour of the electrical conductivity at low temperature, in the case of insulators — where the electrical method is obviously not applicable — similar information can be obtained by studying the thermal conductivity.

Before going into detail, we shall first say a few words about the Debye theory of the specific heat of solids, in which the underlying considerations on lattice vibrations are also fundamental to the problem of thermal conduction. This theory is based on the vibrational modes of the entire lattice. Thus where a lattice consists of $N$ monatomic molecules, the number of vibrational modes (standing waves) $\mathrm{f}(\nu)\mathrm{d}\nu$ having a frequency between $\nu$ and $\nu + \mathrm{d}\nu$ is given by:

$$\mathrm{f}(\nu)\mathrm{d}\nu = 9N\nu^2\mathrm{d}\nu/\nu_{\max}{}^3. \quad \text{(III, 1)}$$

Here, $\nu_{\max}$ is the maximum frequency of the thermal vibrations of the lattice; vibrations whose frequency is so high that the wavelength is smaller than the interatomic distance are irrelevant to a consideration of the crystal as a whole [2]). The values of $\nu_{\max}$ are found on the basis of the premise of mechanics that the total number of independent modes of vibration of a system of $N$ particles cannot be greater than $3N$, i.e.

$$\int_0^{\nu_{\max}} \mathrm{f}(\nu)\mathrm{d}\nu = 3N. \quad . \ . \ . \quad \text{(III, 2)}$$

The value of $\nu_{\max}$ that follows from this expression is found to correspond to a wavelength approximately equal to twice the interatomic spacing.

If, as in the first article for the case of molecules or atoms, we consider each vibrational mode as a quantized harmonic oscillator of zero-point energy $\frac{1}{2}h\nu$, then according to formula (I, 5) the average number of quanta of energy $h\nu$ that can be present per oscillator in thermal equilibrium is equal to $\{\exp(h\nu/kT) - 1\}^{-1}$. By analogy with the term photon for a light quantum — whose energy is also $h\nu$ — a quantum of elastic vibrational energy, or acoustic energy, is called a *phonon*. The number of phonons $\mathrm{g}(\nu)\mathrm{d}\nu$ found at a temperature $T$ in the frequency range from $\nu$ to $\nu + \mathrm{d}\nu$ is thus given by:

$$\mathrm{g}(\nu)\mathrm{d}\nu = \frac{\mathrm{f}(\nu)\mathrm{d}\nu}{\exp(h\nu/kT) - 1} = \frac{9N\nu^2/\nu_{\max}{}^3}{\exp(h\nu/kT) - 1}\mathrm{d}\nu. \quad . \quad \text{(III, 3)}$$

It follows from this (cf. I, 6) that $\mathrm{g}(\nu)$ is proportional to $T$ at high temperatures — the total vibrational energy of the lattice is then proportional to $T$ and the specific heat is constant — but not at low temperatures. This is represented graphically in *fig. 1*, where, omitting the factor $9N$, $\mathrm{g}(\nu)$ is plotted in

[1]) J. Volger, Solid-state research at low temperatures, I. Introduction, Philips tech. Rev. **22**, 190-195, 1960/61 (No. 6), and II. Electron conduction in metals and semiconductors, Philips tech. Rev. **22**, 226-231, 1960/61 (No. 7).

[2]) For the derivation of these formulae, see R. Kronig, Textbook of Physics, 2nd edn., Pergamon, London 1960. See also, for example, C. Kittel, Introduction to solid state physics, 2nd edn., Wiley, New York 1957, Chapter 6, and M. Born, Atomic physics, 6th edn., Blackie, London 1957, Chapter VIII, where the formulae are derived in somewhat different form.

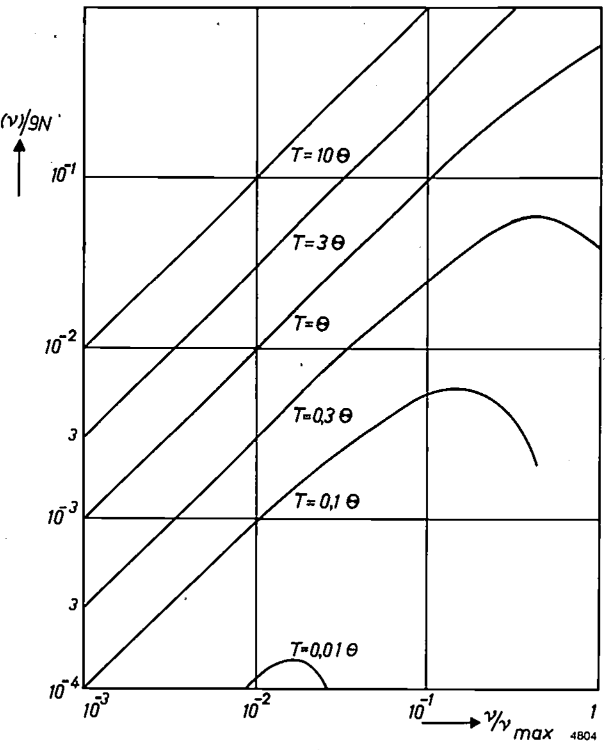


Fig. 1. The phonon spectrum in a solid. The number of phonons g($\nu$) of frequency $\nu$ (disregarding a factor $9N$) is plotted versus frequency for various temperatures. The temperature $T$ is expressed in the figure as a fraction of the Debye temperature $\Theta$, the frequency as a fraction of the maximum frequency $\nu_{\max}$

log-log coordinates against frequency (as a fraction of $\nu_{\max}$), for various values of the ratio $T/\Theta$ ($\Theta$ is the Debye temperature, given by $h\nu_{\max}/k$). It can be seen that at high temperature the curves are straight lines which displace vertically by one decade for every increase in temperature by a factor of 10. For $T \ll \Theta$, this remains valid only at low frequencies; quanta with frequencies in the region of $\nu_{\max}$ become fewer in number as the temperature decreases and finally are virtually non-existent. The maximum of the curve lies approximately at the frequency for which $\nu/\nu_{\max} = 1.6\ T/\Theta$. This behaviour affects the variation at low temperatures of both the specific heat and the thermal conductivity.

According to Debye's theory the total energy $E$ is given by:

$$E = \int_0^{\nu_{\max}} g(\nu) h\nu \, d\nu, \quad \ldots\ldots \quad \text{(III, 4)}$$

from which we find $E = 3NkT$ for the high-temperature region, where the specific heat thus has the required constant value $3Nk$, and at low temperatures:

$$E = \frac{\pi^4}{5}\left(\frac{kT}{h\nu_{\max}}\right)^3 3NkT. \quad \ldots\ldots \quad \text{(III, 5)}$$

It follows from this theory, then, that the specific heat at low temperature is proportional to $T^3$, which agrees very well with the variation found by experiment.

For a solid, in which the heat transport can be represented by the movement of phonons, the formula for the coefficient of thermal conductivity $\varkappa$, that is the ratio between the quantity of heat flowing per unit time through unit area and the temperature gradient, is arrived at by a method similar to that applied in the case of a gas. According to kinetic theory, the thermal conductivity of a gas is given by:

$$\varkappa = \gamma C v l, \quad \ldots\ldots \quad \text{(III, 6)}$$

where $C$ is the heat capacity per unit volume, $v$ the arithmetic mean velocity of the molecules, $l$ their mean free path and $\gamma$ a constant. The thermal conductivity of solids is expressed by exactly the same formula, except that phonons should be read instead of molecules in the definition of $v$ and $l$. The velocity $v$, then, is now the velocity of sound in the material concerned. The value of $\gamma$ is of the order of magnitude of unity.

Before considering how $\varkappa$ varies with $T$ in the low-temperature region, it should be recalled that the phonons fulfil two functions. In the first place they *constitute* the thermal energy, and in the second place they are responsible for the transport of the thermal energy. In a crystal where the phonons were able to move quite freely, the heat transport would thus be comparable with the transmission of energy in free space by electromagnetic radiation. In a real crystal this is not the case. Due to the scattering of the phonons for one reason or another, the heat transport takes on the character of diffusion [3]).

We shall now discuss the behaviour of $\varkappa$ with decreasing temperature by reference to a curve [4]) measured on bismuth telluride ($Bi_2Te_3$), represented in *fig. 2*. In this figure $1/\varkappa$ is plotted versus $T$. The

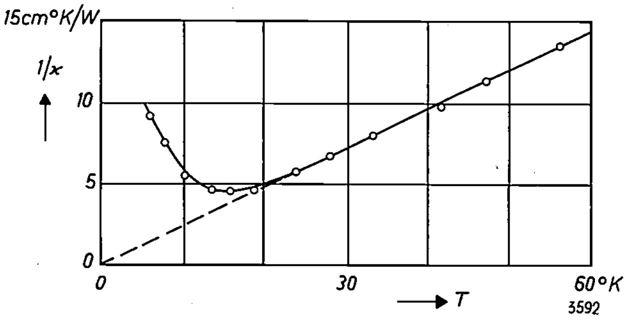


Fig. 2. The reciprocal of the coefficient of thermal conductivity $\varkappa$ of bismuth telluride as a function of temperature.

[3]) Macroscopic evidence of this is that the heat flow is proportional to the temperature gradient. If there were no scattering, the heat flow through a bar would be proportional to the temperature difference between its ends, irrespective of its length.

[4]) Measurements by G. Bosch of this laboratory.

graph shows clearly that above 20 °K the curve is virtually a straight line passing through the origin, indicating that in this region $\varkappa$ is proportional to $1/T$. Thc form of the curve may be explained as follows. If the temperature is appreciably higher than the Debye temperature $\Theta$, then, as we have seen (eq. III, 3 *et seq.* and fig. 1), the value of the phonon distribution function $g(\nu)$ is proportional to $T$. Now, in sufficiently pure crystals the scattering of phonons in this temperature range is primarily due to their mutual "collisions". The probability of such a collision is proportional to the number of phonons, and hence proportional to $T$. The mean free path $l$ is therefore inversely proportional to $T$, and the same holds for $\varkappa$, since the other quantities in formula (III, 6) are constant in this temperature range.

When $T$ approaches $\Theta$ (at 150 °K in the case of $Bi_2Te_3$), the number of phonons decreases more than proportionally with $T$, and the same applies to the probability of their being scattered, but here the specific heat also begins to decrease. At first, these effects roughly compensate one another. This partly explains why the proportionality between $\varkappa$ and $1/T$ continues to exist quite a long way below the Debye temperature.

If the collisions between the phonons werc the only cause of the scattering, then at very low temperatures ($T \ll \Theta$) the thermal conductivity would finally increase very steeply with falling temperature. It can be seen from fig. 2 that the reverse is the case. Since $g(\nu)$ is very small in this temperature range, and so is the probability of collisions between the phonons, the scattering from lattice imperfections begins to play a relatively important part here. A calculation shows that, as a consequence of this, the mean free path $l$ of the phonons varies only to a slight extent with temperature for $T < \Theta$. The specific heat however, at $T < \Theta$, varies with $T^3$, so that with decreasing temperature $\varkappa$ must indeed undergo a sharp decrease.

If the lattice contained no defects whatever, the mean free path with falling temperature would finally be determined solely by the dimensions of the material, and would thus be independent of temperature. The thermal conduction in that case would be comparable with a Knudsen flow, i.e. the flow of a gas which is so rarefied that the number of molecular collisions is negligible compared with the number of collisions with the wall of the tube through which the gas is flowing, or it might be compared with the movement of light quanta in an internally reflecting tube (see [3]). An effect of this kind has in fact been found experimentally, for example on diamond and on some alkali halides [5]).

Summarizing, it can be said that the variation of thermal conductivity as a function of temperature provides important information on the scattering of phonons by lattice imperfections; indeed, the study of the thermal conduction of insulators is no less important as a research tool than the study of electron conduction in metals. The way in which the phonon spectrum varies with temperature, has, of course, an equally important bearing on the variation of other processes which take place under the influence of lattice vibrations. A notable example of this will be discussed in the next section.

## Paramagnetism

From formula (I, 8) for the magnetic moment of paramagnetic substances we were able to deduce that from the point of view of the magnetization, we may speak of a "low" temperature where $kT \ll \mu_B H$. In this temperature range the magnetization is no longer proportional to $H$, as it is in the case of normal field strengths at room temperature, but saturation occurs (fig. 2 in I).

We have already mentioned that the rate at which some processes take place in the solid state is slowed down when the temperature falls. We shall now examine this effect more closely for the case of paramagnetism. The application of an external magnetic field does not instantaneously produce the magnetization $\overline{\mu}$ given by (I, 8); some time elapses before the new state of equilibrium is reached. If this orientation process, called paramagnetic relaxation, is an exponential function of time, it can be described by a characteristic time $\tau$, the relaxation time. The cause of the relaxation phenomenon is of a very general nature. A paramagnetic ion can orient itself in a magnetic field only if it is able to exchange energy with its surroundings, just as, for example, a pendulum pulled out of equilibrium returns finally to its vertical position undcr the action of gravity only because its movement is damped. If the ion is unable to give up energy, it will continue to describe a precessional motion without any decrease in the angle between the directions of $\overline{\mu}$ and $H$, so that the component of $\overline{\mu}$ in the $H$ direction does not change. In such a case the pendulum in our example would go on swinging with undiminished amplitude. It may thus be concluded that the magnitude of the relaxation time is determined by the extent to which energy can be exchanged.

Since $\tau$ is generally small, the inertia of the orientation process is not directly evident when a constant magnetic field is applied, but it *is* apparent when the

[5]) W. J. de Haas and T. Biermasz, Physica 5, 47, 320 and 619, 1938. A theoretical treatment is given by H. B. G. Casimir in Physica 5, 495, 1938. See also R. Peierls, Ann. Physik 2, 5, 1055, 1929.

magnetizing force is periodically varied. As long as the frequency is small compared with the reciprocal of the relaxation time, no effect is observed. When the frequency is raised, however, the susceptibility in the frequency range around $1/\tau$ begins to decrease. A phase shift also occurs between $H$ and $\bar{\mu}$. Both effects indicate that the orientation of the paramagnetic ions can no longer properly follow the variations of the magnetic field.

In order to give a qualitative explanation for the marked increase which $\tau$ must undergo with falling temperature, we shall return for a moment to the case touched on in the first article, where the spin can assume only two values, namely parallel and anti-parallel to the magnetic field. It was mentioned there that the energy difference between the two states is $2\mu_B H$. In a transition, then, the ion must either absorb or give up this amount of energy, depending on the direction of the transition. This it can do, via the lattice vibrations, in two ways. In the first place, a phonon of energy $2\mu_B H$ may be absorbed or emitted, and in the second place a phonon may be non-elastically scattered by the ion, and thereby give up or absorb the energy $2\mu_B H$. The latter process is directly comparable with the Raman effect, familiar in optics. Both processes have a greater probability of occurring the more phonons are available. As regards the Raman process, this is immediately understandable; as regards the spin-phonon process it is also understandable in so far as it concerns the *absorption* of a phonon. Quantum mechanics shows, however, that the chance of the *emission* of a phonon is also greater the greater the number of phonons. As in the case of absorption, they must likewise have the energy $2\mu_B H$ in order to be effective.

In the foregoing section (see fig. 1) we have already seen that the phonons do indeed decrease sharply in number as the temperature falls. It is therefore clear that the relaxation process will be slower, and the relaxation time $\tau$ longer, the lower is the temperature. *Fig. 3* shows the variation with temperature of the relaxation time for gadolinium sulphate and for a chromium-alum.

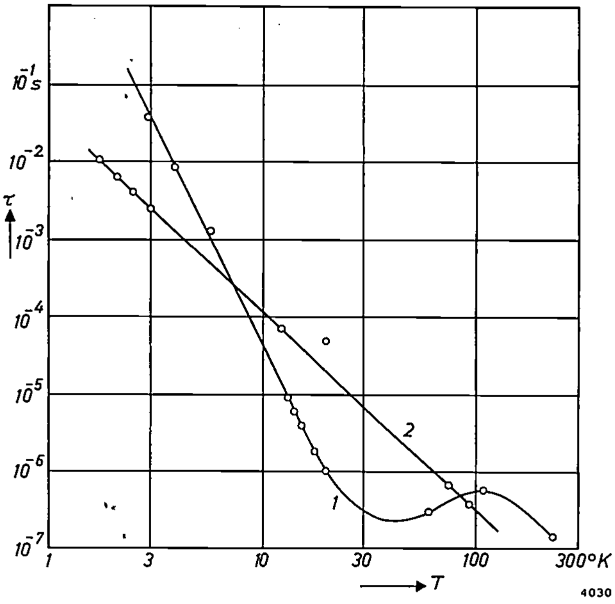


Fig. 3. The characteristic time $\tau$ of a paramagnetic relaxation increases steeply as the temperature falls. The figure gives a log-log plot of $\tau$ (in seconds) versus temperature, as measured on two substances:
*1* : $Gd_2(SO_4)_3.8H_2O$; measured at $2.6 \times 10^5$ A/m;
*2* : $CrK(SO_4)_2.12H_2O$; measured at $3.2 \times 10^5$ A/m.

Apart from the spin-lattice relaxation just discussed, a direct exchange of energy is possible between the ions mutually. The extent to which this process can take place is determined by the strength of the ionic interaction and therefore depends on the concentration of the paramagnetic ions. This interaction may thus be attenuated by "diluting" a salt with a certain quantity of non-magnetic ions. In normal, undiluted paramagnetic substances the relaxation time $\tau_{s\text{-}s}$ corresponding to this spin-spin relaxation is of the order of magnitude of only $10^{-10}$ sec. In diluted salts, $\tau_{s\text{-}s}$ is accordingly somewhat longer. The speed of this relaxation phenomenon is not affected by temperature [6]).

### *Paramagnetic resonance*

Temperature is a factor of importance in various respects in solid-state research using the method of paramagnetic resonance, which has already been described at some length in this journal [7]). In this method the degeneracy of the ground state of the paramagnetic ions is removed by the application of a magnetic field, the energy level thereby being split into a number of sub-levels the distance between which, $\Delta E$, is proportional as a first approximation to $H$. Transitions between neighbouring sub-levels can be brought about by irradiating the substance with electromagnetic waves whose frequency $\nu_r$ is such that $h\nu_r = \Delta E$. Where most of the ions occupy the lower of two such levels — which is the normal situation — the result is a net absorption; in the contrary case a net emission of radiation having the frequency $\nu_r$ occurs.

It is necessary to speak of *net* absorption and *net* emission because in each case both emission and absorption take place;

[6]) The existence of a temperature-independent relaxation phenomenon side by side with the spin-lattice relaxation, which does vary with temperature, was predicted on theoretical grounds by J. Waller in 1932 (Z. Phys. **79**, 370, 1932). The existence of paramagnetic relaxation effects was first demonstrated experimentally by C. J. Gorter (Physica **3**, 503, 1936).

[7]) J. S. van Wieringen, Paramagnetic resonance, Philips tech. Rev. **19**, 301-313, 1957/58.

in the first case the number of absorbed quanta exceeds the number of emitted quanta, in the other case the opposite applies. Both numbers are proportional to the population of the two levels and to the number of incident quanta. The fact that the emission is proportional here to the intensity of the incident radiation field is an indication that it is *stimulated* emission. Unlike the situation in optical spectroscopy, spontaneous emission is negligible here (cf. reference [7]).

The temperature has a threefold influence on the observed phenomena. In the first place it determines the degree to which one can discriminate between the details of the absorption spectrum; secondly it determines the ratio between the intensity of the incident and the absorbed radiation — i.e. the sensitivity —, and thirdly it governs the maximum permissible intensity of the incident radiation.

As regards the discrimination between details of the resonance spectra, it should be recalled that the energy levels into which the ground state is split are often themselves split into a number of very closely spaced sub-levels. It is precisely the latter splitting that provides the most important information on the substance under investigation. These fine details obviously cannot be observed if the width of the lines in the resonance spectrum is greater than the distance between adjacent sub-levels. Since this width is related to the rate of the relaxation process (the slower the relaxation process the smaller the line width), it is possible in appropriate cases, e.g. in investigations on substances containing $Ti^{3+}$, $Fe^{2+}$ or $Co^{2+}$, to achieve a sufficiently small line width by reducing thc temperature. A good example is shown in *fig. 4*.

With regard to the temperature-dependence of the sensitivity it should be recalled that the net absorption is greater the greater the excess population of the lower of the two levels between which the transitions take place. The ratio between the populations of the two levels is proportional to the Boltzmann factor exp $(-\Delta E/kT)$, and therefore the absorption can be increased by lowering $T$. Since $\Delta E$ is of the order of magnitude of $k \times 1$ °K, this would produce in all cases an appreciable gain if it were not for the fact that the temperature reduction also gives rise to an unfavourable effect — the third temperature effect mentioned above.

In a resonance experiment the population of the levels is not exactly that given by the Boltzmann distribution, the reason being that the radiation field causes transitions to the upper level, which it thus tries to "pump full". At normal temperatures, where the relaxation time is short (e.g. $10^{-7}$ sec) and where consequently the upper level is soon "emptied" again, the disparity is slight and the sensitivity is not seriously affected. When the sample is cooled, however, the relaxation time increases, as we have seen, and in some cases a situation may be reached where the population of the uppcr level is equal to that of the lower level (saturation), so that there is virtually no net absorption. In that case, to avoid saturation, one would have to work with a considerably weaker incident radiation. The choice of the temperature and the intensity of the incident radiation is therefore always a compromise, to be decided from one case to another.

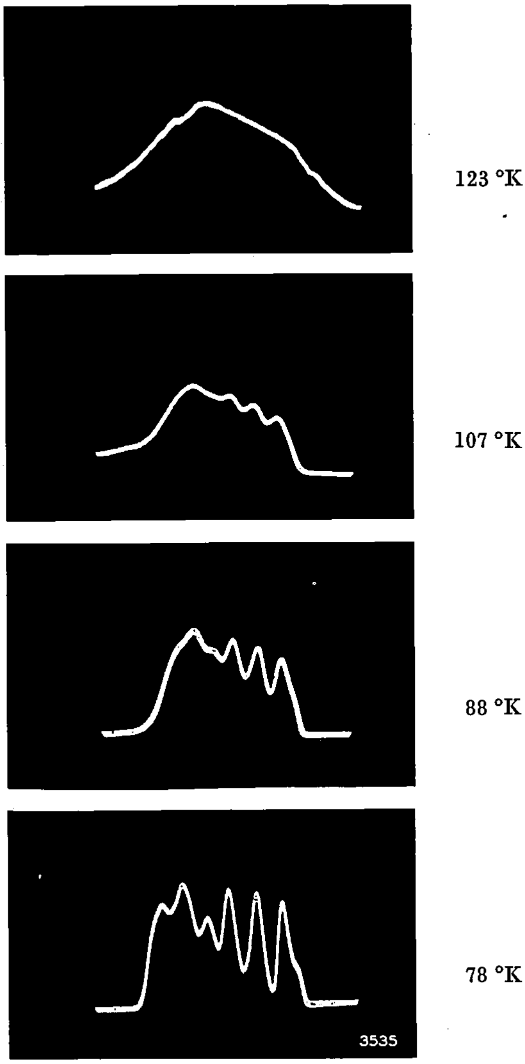


Fig. 4. Spectral line with hyperfine structure, from the paramagnetic resonance spectrum of $\alpha$-quartz irradiated with X-rays. This irradiation produces colour centres in the quartz near aluminium impurity atoms. The recordings were made at four diffcrent temperatures. It can be seen that each of the fine-structure lines becomes narrower — and therefore higher, since their integrated area remains unchanged — the lower is the temperature. At the highest temperature their width is so large relative to their separation that they are marged into each other. The frequency of the electromagnetic field was 9.7 Gc/s ($\lambda$ = 3.1 cm). The orientation of the permanent magnetic field was parallel to that of the $c$ axis of the quartz crystal. (Measurements by J. S. van Wieringen of the Philips Research Laboratories, Eindhoven.)

### *Solid-state amplifiers for microwaves*

We shall now consider, again from the point of view of temperature, the microwave amplifiers which have been given the name of maser (*M*icrowave *A*mplification by *S*timulated *E*mission of *R*adiation). We shall be concerned only with the so-called three-level solid-state maser [8]). It will be shown that the physical principles underlying the temperature effects here encountered are largely the same as those just discussed under the heading of paramagnetic resonance.

The operation of the maser depends on the fact, already touched upon, that when a paramagnetic substance is exposed to radiation of a frequency corresponding to the energy difference between two levels, the result is a net emission — and hence amplification of the incident signal — if the population of the upper level is greater than that of the lower level. In that case the transitions to the lower level exceed the transitions to the upper one.

The reason for this is evident when it is realized that, like the probability of absorption — i.e. of an upward transition — the probability of emission is also proportional to the relative populations of the energy levels and to the intensity of the incident radiation; the respective proportionality factors are the same for absorption and emission. The emission is almost entirely stimulated emission; the probability of spontaneous emission, which of course depends on the population of the upper level but is not affected by the incident radiation, is negligible or at least of minor significance.

The *inversion* of the energy-level population, on which the operation of the maser depends, can of course only be brought about by some external intervention; in a substance in thermal equilibrium the lower level always has the greater population.

In the Bloembergen solid-state maser a paramagnetic salt whose ground state, in a magnetic field, splits into at least three sub-levels, is made emissive by irradiating it with microwave energy corresponding to the transition between the lowest and highest of the three levels (*1* and *3*, respectively, in *fig. 5*). The intensity of the microwave radiation is chosen high enough to produce saturation, i.e. such that the population of *3* is equal to that of *1*. Evidently, this is only feasible in practice provided the relaxation time $\tau_{31}$ for the return to the lowest level is not unduly short. This situation can be improved by reducing the temperature of the substance, thereby reducing the required "pumping power" to an attainable value. Here, then, the effect which was a disturbing effect in the resonance measurements is turned to good use.

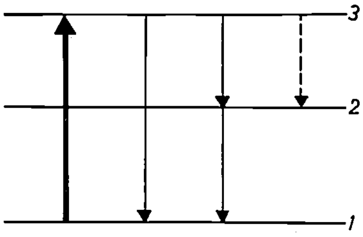


Fig. 5. Illustrating the operation of a solid-state maser, after Bloembergen. A paramagnetic salt is used whose ground state in a magnetic field splits into at least three levels.

Now an ion, in addition to returning directly from level *3* to level *1*, may also return by first moving to level *2*, thereby emitting one or more phonons of intermediate energy (thin lines in fig. 5). It can be calculated that the population $N_2$ of level *2* in the state of equilibrium is determined not only by the temperature but also by the energy differences $E_{12}$ and $E_{23}$ and the relaxation times $\tau_{32}$ and $\tau_{21}$ of these transitions. It further appears [8]) that level *2* will be less populated than level *3*, and that consequently inversion will occur where:

$$\frac{E_{12}}{\tau_{21}} > \frac{E_{23}}{\tau_{32}}. \quad . \; . \; . \quad \text{(III, 7)}$$

If $E_{12} \approx E_{23}$ (the two energies should not be identical!) this roughly amounts to the condition $\tau_{32} > \tau_{21}$. In other words, level *2* must be "emptied" faster than level *3*. The maximum positive difference $(N_3 - N_2)$ between the populations is given by:

$$(N_3 - N_2)_{\max} = N \times \frac{1 - \exp(-E_{12}/kT)}{2 + \exp(-E_{12}/kT)} \approx \frac{N}{3}\frac{E_{12}}{kT}, \quad . \; . \; . \quad \text{(III, 8)}$$

where $N = N_1 + N_2 + N_3$. It can be seen that this difference is greater the lower the temperature of the salt. For this reason, the masers we are now discussing have hitherto been designed to operate at the temperature of liquid helium.

When electromagnetic radiation of frequency $\nu_{23}$ $(= E_{23}/h)$ is incident on a paramagnetic salt whose energy levels and relaxation times are such as to satisfy (III, 7) and which is supplied with a sufficient pumping power, this radiation — owing to the inversion of the populations of *2* and *3* — induces the emission of stronger radiation of the same frequency (dashed line in fig. 5). The salt then functions as a maser. (If $\tau_{32}$ is smaller than $\tau_{21}$, the population of *1* and *2* is inverted, and the result is a maser which amplifies radiation of the frequency $\nu_{12}$.)

Of the three categories of amplifiers now known

[8]) The idea underlying this type of maser is due to N. Bloembergen (Phys. Rev. **104**, 324, 1956).

— namely radio-valve (or transistor) amplifiers, parametric amplifiers, and masers — the masers have the lowest noise figures. For this reason, in spite of their obvious drawbacks — cooling with liquid helium and amplification only in an extremely narrow waveband — masers nevertheless have interesting technical possibilities.

The noise of a maser is mainly due to the spontaneous transitions from the upper to the lower level, which, unlike the stimulated transitions, are of course not correlated with the incident radiation. This noise contribution may be expressed theoretically as the thermal noise of a negative resistance at a negative temperature [9]. This negative temperature is that which would have to be substituted in the Boltzmann factor in order to describe the inverted population of the energy levels. Since the absolute value of this temperature is of the same order of magnitude as the true temperature, it is understandable from the point of view of noise reduction that the maser should be operated at a low temperature. This also reduces, incidentally, the thermal noise of the resonant cavity in which the crystal is irradiated.

*Cooling by adiabatic demagnetization*

Although the subject does not, strictly speaking, come within the scope of the present article, it will be useful at the end of this section on paramagnetism to touch briefly on the celebrated method of cooling based on the adiabatic demagnetization of a paramagnetic salt [10]. A "pellet" of a suitable paramagnetic material, subjected to a strong magnetic field, is in thermal contact on one side with the sample to be cooled, and on the other with a bath of boiling helium. Once thermal equilibrium is established, the thermal contact with the helium bath is broken, after which the magnetic field is switched off. The temperature of the pellet then falls. Ingenious application of this principle has made it possible to reach a temperature of 0.001 °K.

The physical background of the method may be briefly explained with reference to *fig. 6*, which shows schematically the variation of the entropy $S$ as a function of temperature, both in the presence and the absence of a magnetic field. Since the magnetized state of the salt represents a state of greater *order* than the non-magnetized state, the entropy at any given temperature is greater in the absence of a magnetic field. Upon adiabatic (isentropic) demagnetization — stage *2* in fig. 6 — the temperature of the paramagnetic material therefore drops.

A very pronounced effect can be produced with salts whose ground state, when no external magnetic field is applied, may be regarded to a first approximation as completely degenerate. This means that the splitting of the ground state caused by the internal crystalline field, which therefore remains after the external field is removed, gives rise to only very small energy differences between the sub-levels. The greater this so-called zero-field splitting, the greater is the order prevailing in the non-magnetized crystal, and hence the smaller is the entropy difference between the magnetized and the normal state; the temperature reduction that can be achieved will therefore likewise be smaller.

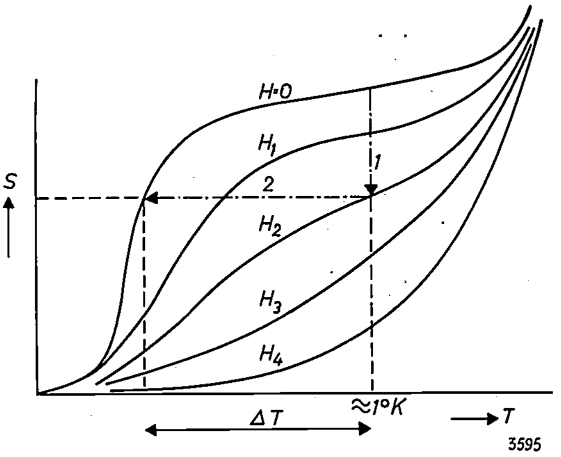


Fig. 6. The variation of the entropy $S$ of a paramagnetic salt as a function of temperature $T$ in the region between 0 and 1 °K, in the presence of magnetic fields of varying strengths. A temperature drop $\Delta T$ can be brought about by cooling the salt to about 1 °K, applying the magnetic field $H_2$ (stage 1), breaking the thermal contact with the environment, and finally switching off the magnetic field (stage *2*).

## Dielectric losses in relation to chemical lattice imperfections

In the theory of dielectrics the magnitude of the dielectric constant can be related to various mechanisms of polarization. For example, there is electronic polarizability, which arises from the displacement of the electrons relative to the nucleus; there is ionic polarizability, due to the relative displacement of ions of opposite sign in the crystal lattice; and there is orientational polarizability in substances containing molecules which have a permanent dipole moment. Since solid-state research in recent decades has evolved primarily into an investigation of lattice imperfections, we shall confine ourselves here to a few examples of dielectric phenomena related to lattice imperfections [11]. In particular we shall be concerned with cases where an imperfection contains an electric dipole of moment $p$, the orientation of which in an external electric field is subject to a certain inertia; the manner of orientation can be described in our examples by a single characteristic time, the relaxation time $\tau$.

As in the case of paramagnetism, discussed in the previous section, where the inertia of the relaxation process leads to a lower susceptibility when the frequency of the field variations is greater than $1/\tau$,

[9] R. V. Pound, Ann. of Physics 1, 24, 1957.

[10] This subject is dealt with in detail by C. G. B. Garret, Magnetic cooling, Harvard, Cambridge, 1954, and by D. de Klerk in an article in Handbuch der Physik, Part XV, p. 38-209, Springer, Berlin 1956.

[11] A survey of this field is given by J. Volger in Dielectric properties of solids in relation to imperfections, Progress in Semiconductors 4, 205-236, 1960.

here too we find that the dynamic value of the dielectric constant decreases when the frequency of the applied electric field increases to values higher than $1/\tau$ [12]). Let $\varepsilon_m$ be the relative dielectric constant of the unperturbed lattice — which, for convenience, we shall regard as constant — then the value $\varepsilon$ found at a certain angular frequency $\omega$ is given by:

$$\varepsilon = \varepsilon_m + \frac{\Delta\varepsilon}{1+\omega^2\tau^2}. \quad \text{(III, 9)}$$

The value of the quantity $\Delta\varepsilon$ follows from:

$$\Delta\varepsilon = Nf\frac{4\pi p^2}{3kT}. \quad \text{(III, 10)}$$

Here $N$ is the number of dipoles per unit volume, and $f$ is a factor defining the effect of the internal (electrical) crystalline field (the dipoles are not affected solely by the applied external field, but by the resultant of this field and the internal field). Provided the concentration $N$ of the dipoles is not too high, $f$ is given by:

$$f = (\varepsilon_m + 2)^2/9. \quad \text{(III, 11)}$$

Both the relaxation time $\tau$ and the quantity $Np^2$ can be determined directly. The latter quantity is found by measuring $\varepsilon$ at high and low frequency; with the aid of (III, 9) the value of $\Delta\varepsilon$ can then be found, and substituted in (III, 10). The relaxation time can be measured in principle by ascertaining the frequency range in which $\varepsilon$ changes in value; this is not very accurate, however, because the change in $\varepsilon$ is only a small fraction of its value.

A better method of determining $\tau$ is to ascertain the variation with frequency of the loss angle $\delta$, i.e. the phase angle between the applied alternating field and the dielectric displacement. This angle can be found by measuring the dielectric losses, which are proportional to $\tan\delta$. The variation of $\tan\delta$ with frequency is expressed by the formula:

$$\tan\delta \approx \frac{\Delta\varepsilon}{\varepsilon_m}\frac{\omega\tau}{1+\omega^2\tau^2}. \quad \text{(III, 12)}$$

It can be inferred from this expression that $\tan\delta$ depends strongly on frequency and shows a peak at $\omega = 1/\tau$.

In many cases the variation of $\tau$ with frequency is given by an expression of the form:

$$\frac{1}{\tau} = \frac{1}{\tau_0}\exp(-Q/kT). \quad \text{(III, 13)}$$

The energy $Q$ in this expression is an "activation energy", representing the height of the energy barrier which an ion or an electron must exceed to arrive at the new state. It is plausible that the formula expressing the probability of this barrier being exceeded will contain a Boltzmann factor, and so therefore will the formula for $\tau$. The nature of this energy barrier will be discussed in more detail presently. In some cases an ion can reach a new state as a result of a wave-mechanical "tunnelling" effect, or because the energy barrier which keeps it entrapped disappears for a moment owing to a favourable combination of the movement of other ions. In these cases too, an expression of the form (III, 13) is often applicable. This formula brings out clearly the important bearing which the temperature has on the study of dielectric phenomena. Especially if $Q$ is small, the relaxation time at high temperature is so short that its reciprocal lies outside the accessible frequency range. Only by working at low temperatures is it possible to observe losses corresponding to such a mechanism.

It should be added that the analogy with paramagnetic relaxation — where formulae (III, 9), (III, 10) and (III, 12) also apply, mutatis mutandis — ceases to be valid as regards the nature of the relaxation mechanism. The manner in which $\tau$ varies with $T$ is often quite different, even though $\tau$ may be large at low temperatures and small at high temperatures.

### *Dielectric losses in quartz due to local deformation of the crystal lattice*

Dielectric losses attributable to a polarization mechanism of low activation energy, which are therefore observable only at low temperature, are found (for example) in quartz. *Fig. 7* shows the result of measurements of $\tan\delta$ as a function of $T$ — and thus indirectly as a function of $\tau$ — on three different samples at a frequency of 32 kc/s. The curves are seen to exhibit a number of peaks. The temperatures at which these peaks occur — in so far as they are present in a particular curve — are the same for all three curves shown. Spectrochemical analysis of the crystals revealed the impurities which they contained, and on this basis an attempt was made to correlate each peak with the presence of a particular kind of foreign atom. On the assumption that formulae (III, 10), (III, 12) and (III, 13) are applicable here, one can calculate for the temperature $T$ at which a peak occurs that:

$$kT_m = -\frac{Q}{\ln\omega\tau_0}\left(1 - \frac{1}{(\ln\omega\tau_0)^2} + \dots\right). \quad \text{(III, 14)}$$

[12]) An extensive treatment of the theory of dielectric phenomena, giving the derivation of the formulae mentioned in this section, will be found in: H. Fröhlich, Theory of dielectrics, Oxford University Press, London 1950.

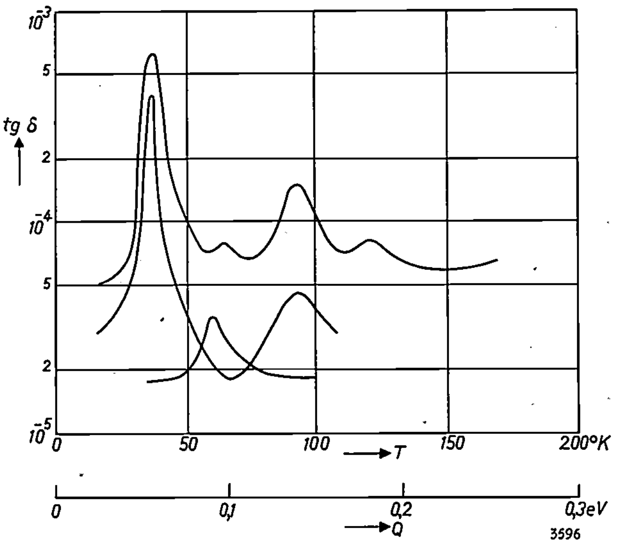


Fig. 7. The variation of tan $\delta$ with temperature $T$ for three quartz samples in an alternating electrical field of 32 kc/s. The peaks observed are attributed to polarization mechanisms connected with the presence of impurities. The values of the activation energy $Q$ corresponding to each peak can be read from the lower abscissa scale.

Since $\tau_0$ proved to have the value $10^{-13}$ sec for all peaks in all three samples, it was possible to draw under the temperature scale in fig. 7 a linear energy scale, applicable to all three samples. It can be seen that the $Q$ values are of the order of 0.1 eV. This is indeed very small; the $Q$ value for the displacement of an ion in the crystal lattice — that is to say a displacement from one lattice site to another — corresponds to a value roughly 10 times as high.

The mechanism of the polarization processes responsible for these losses is assumed to be roughly as follows. It is assumed that in "wide-mesh" crystal lattices, as found in quartz and glass, the potential wells which correspond to the lattice sites, have a "sub-structure", each consisting of a small group of separate minima. These sub-minima are thus separated from one another by tiny potential barriers whose height is much less than that of the maxima between the lattice sites (*fig. 8*). Even at temperatures so low that an ion cannot possibly move from one lattice site to another, the ions are still capable of a slight displacement, i.e. from one sub-minimum to another. The small value of $Q$ for these processes corresponds to the low height of the potential barriers separating the sub-minima.

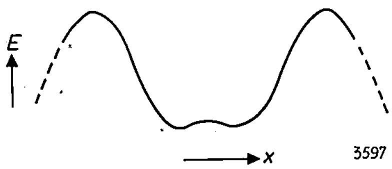


Fig. 8. The occurrence of dielectric losses at such a low temperature that an ion cannot possibly move from one lattice site to another is explained by assuming that, in "wide-mesh" lattices (quartz, glass, etc.), the lattice sites do not always correspond to a simple minimum of the potential energy $E$, but to a group of sub-minima. The drawing illustrates a cross-section of the potential variation in such a lattice site in which two sub-minima occur.

The multiple minima are assumed to be present mainly in the immediate proximity of foreign atoms. The details of the structure of chemical lattice imperfections are unfortunately not yet sufficiently known to be able to say with certainty in exactly what way the above-mentioned polarization processes take place in an actual crystal.

*Dielectric losses in oxidic semiconductors*

Another remarkable polarization mechanism, which is likewise only observable at low temperature, is found in oxidic semiconductors. We shall discuss its nature with reference to the case of nickel oxide containing trace quantities of lithium.

In the nickel-oxide lattice the Ni ions are divalent and positive. If one of the Ni ions is replaced by an Li ion, the local electric neutrality is disturbed, since Li ions are monovalent, and there is no evidence of the existence of divalent Li ions. As the ions of the iron group do possess different valencies, an obvious supposition is that the neutrality might be restored by the presence of a trivalent Ni ion in the immediate vicinity of the $Li^{1+}$ ion (*fig. 9*).

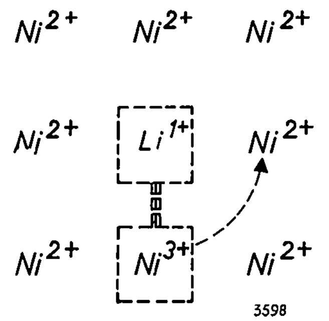


Fig. 9. The electric neutrality of a nickel-oxide crystal containing $Li^{1+}$ ions is maintained by one of the surrounding Ni ions becoming trivalent. Due to the transference of an electron, the trivalence may move to any of the Ni ions that are immediate neighbours of the Li ion. This mechanism is proposed to explain the dielectric losses found at low temperature in impure or non-stoichiometric oxidic semiconductors.

At a temperature so low that all electrons are bound to the donors, the material thus being an insulator, dielectric losses are found to occur.

It is assumed that the polarization, i.e. the alignment of the dipole formed by the $Li^{1+}$ ion (which is negatively charged relative to the surrounding $Ni^{2+}$ ions) and the $Ni^{3+}$ ion (which has a relative positive charge) is brought about by an electron transferring from one of the $Ni^{2+}$ ions to the $Ni^{3+}$ ion, apparently causing the latter to change place.

The Ni ions likely to be involved are only those surrounding the $Li^{1+}$ ion. This is somewhat analogous to the loss mechanism discovered by Breckenridge which occurs in such types of point defect. An example, discussed in this journal some time ago [13], is the case of a substitutional $Ca^{2+}$ atom in the NaCl lattice, where the electric neutrality is restored by the presence of a vacancy on an Na site.This vacancy can occupy any of the Na sites around the $Ca^{2+}$ atoms, but cannot get away at moderate temperatures.

The same phenomenon observed on nickel oxide has been found in $\alpha$-$Fe_2O_3$, where $Ti^{4+}$ ions were substituted for a number of $Fe^{3+}$ ions, and also in non-stoichiometric $\alpha$-$Fe_2O_3$. In the first case the electric neutrality is assumed to be restored as a result of one of the Fe ions around the $Ti^{4+}$ becoming divalent, and thus possessing an extra electron. *Fig. 10* shows the result of measurements on non-stoichiometric $\alpha$-$Fe_2O_3$. The relaxation time $\tau$ was governed here by the extremely small activation energy of $5\times10^{-3}$ eV and a relaxation time $\tau_0$ of $2\times10^{-7}$ sec.

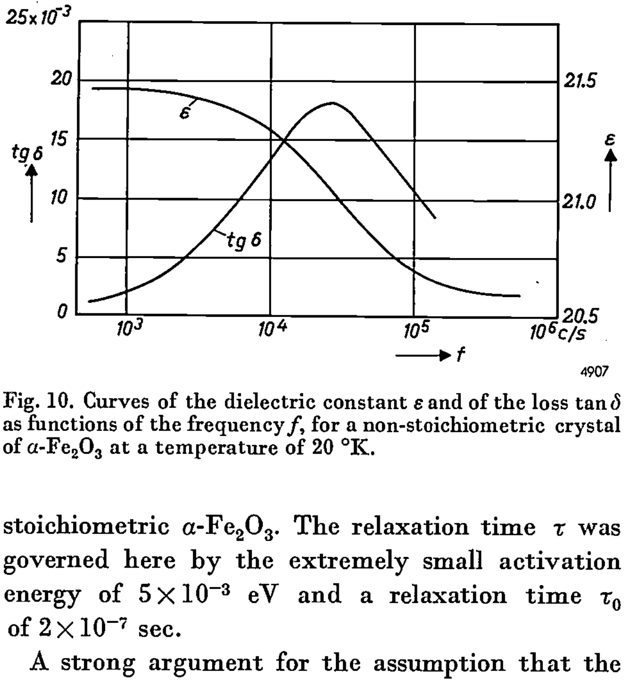


Fig. 10. Curves of the dielectric constant $\varepsilon$ and of the loss $\tan\delta$ as functions of the frequency $f$, for a non-stoichiometric crystal of $\alpha$-$Fe_2O_3$ at a temperature of 20 °K.

A strong argument for the assumption that the polarization mechanism is bound up with the presence of foreign, or at least extra, atoms, is the fact that the losses rise and fall with their concentration. In a pure state the substances do not show the effect at all.

To conclude this series of articles, a few comments will not be out of place on the technological significance of cryogenic solid-state research. Its importance, as in the case of all more or less fundamental research, may be said to be both direct and indirect. As regards the latter, it should be noted that the study of each of the effects discussed — including those that have no direct technical importance — forms part of a wider programme of research, any part of which might yield the essential factor for obtaining new insight that can be turned to practical ends. To give only one instance, the study of paramagnetic resonance and dielectric losses has increased our knowledge of lattice imperfections, and the greater insight thereby gained into the discolouration shown by certain types of glass, subjected to intense irradiation, has proved to be extremely useful in the production of glass for X-ray tubes and in X-ray dosimetry.

As regards the direct technical applications of the phenomena discussed here, it may be mentioned that some of them can be used for the purposes of analytical chemistry, and others underlie the design of electrical or electronic devices or switching elements. Methods of chemical analysis can be based on measurements of the residual resistance of metals, on paramagnetic resonance (for tracing paramagnetic ions present as impurities in a non-paramagnetic substance) and on thermal-conductivity measurements. We have encountered applications of the second kind in the maser and the cryotron. There is every reason to believe that more and more practical use will be made in future of those physical phenomena which occur, or are observable, only at low temperatures.

**Summary.** Just as the residual resistance of a metal depends on the scattering of electrons by lattice imperfections, the thermal conductivity of insulators at low temperature is determined by the scattering of phonons by the same imperfections. The behaviour of a substance is also determined, of course, by the way in which the phonon spectrum varies with temperature. As an example, the behaviour of bismuth telluride is discussed. After dealing with the subject of paramagnetic relaxation, the author discusses the importance of the temperature on the application of paramagnetic resonance and its significance in connection with a solid-state microwave amplifier (maser). Dielectric losses due to polarization mechanisms of very low activation energy (< 0.1 eV) and which can therefore only be studied at low temperatures, are found in such substances as impure quartz and non-stoichiometric or impure oxidic semiconductors.

[13] Y. Haven, Lattice imperfections in crystals, studied on alkali halides, Philips tech. Rev. 20, 69-79, 1958/59.

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

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution.

**2793:** J. G. Bonenkamp and W. Hondius Boldingh: Quality and choice of Potter-Bucky grids (Acta radiol. **51**, 479-489, 1959, No. 6; **52**, 149-157, 1959, No. 2; **52**, 241-253, 1959, No. 3).

In the first section of this paper a new and less ambiguous criterion for the quality of Potter-Bucky grids is suggested. A graphical method is given whereby the best contrast improvement can be obtained. In section II of the paper, measurements are reported on 13 grids, revealing that their contrast-improving capacity is determined mainly by their lead content. The relation between contrast, free-air dose and filters is discussed, especially at high voltages. The final section discusses the absorption of primary and scattered radiation and the relation between the number of lines per cm, ratio and lead content of grids under optimum conditions. Rules are thereby arrived at for a suitable choice of grid for a given field of application.

**2794:** J. D. Fast and H. A. C. M. Bruning: Entkohlung und Entstickung von Eisen-Silicium-Legierungen (Z. Elektrochemie **63**, 765-772, 1959, No. 7). (Decarburization and denitriding of iron-silicon alloys; in German.)

The denitriding and decarburization of Fe-Si-alloys with Si contents up to 3% by weight were investigated by heating samples for two hours in pure and in wet hydrogen at 900 °C. The rate of *denitriding* of Si-free iron is independent of the water content of the hydrogen. Denitriding of iron containing Si takes place in very dry hydrogen but is retarded considerably even by traces of water vapour, probably by formation of a $SiO_2$ skin on the surface. The rate of *decarburization* in very dry hydrogen is practically zero for Si-free iron but increases with increasing Si content.

It is concluded that the activation energy for the formation of $CH_4$ molecules on the surface is much reduced by the presence of Si. Damp hydrogen has a decarburizing effect on Si-free iron because the formation of CO on the surface requires a smaller activation energy than the formation of $CH_4$. Both CO and $CH_4$ are formed on Si-iron in damp hydrogen; decarburization would therefore be very rapid, were it not that the same retarding reaction takes place as with denitriding, viz. the formation of $SiO_2$ on the surface. The $SiO_2$ skin practically stops the formation of $CH_4$ and slows down the formation of CO. The sealing-off of the surface is, however, less complete than in the case of denitriding, owing to penetration of C into the skin with the formation of SiO and CO.

**2795:** J. L. Meijering: Störungslinien und Störungsbänder in innerlich oxydierten Kupfer- und Silberlegierungen (Z. Elektrochemie **63**, 824-829, 1959, No. 7). (Perturbation lines and bands in internally oxidized copper and silver alloys; in German.)

Perturbation of the reaction balance, e.g. of Be + O → BeO, at the subscale boundary gives rise to local variations in the BeO concentration. The perturbations may be brought about by changes in gas atmosphere or in temperature, and also by a special geometry of the specimen.

**2796:** P. Massini: Synthesis of 3-amino-1,2,4-triazolyl alanine from 3-amino-1,2,4-triazole in plants (Biochim. biophys. Acta **36**, 548-549, 1959, No. 2).

The herbicide 3-amino-1,2,4-triazole is metabolized by plants. One of the transformation products has been isolated from bean plants treated with aminotriazole, and its structure has been partially elucidated.

**2797:** M. Avinor: Effect of aluminium on the green emission of cadmium sulphide (Physica **25**, 1095-1096, 1959, No. 11).

Note on the so-called Ewles-Kröger emission in CdS, excited by about 2.4 eV at low temperatures. It is found that activation by aluminium actually enhances the intensity of the green emission while at the same time destroying the fine structure of the emission band. There is also an enhanced emission on the verge of the infrared.

**2798:** K. van Duuren, W. K. Hofker and J. Hermsen: Compact low-level counting arrangement (Proc. 2nd United Nations int. Conf. on the peaceful uses of atomic energy, Geneva 1-13 Sept. 1958, Vol. 14, pp. 339-344; Pergamon Press, London 1959).

See Philips tech. Rev. **20**, 170-172, 1958/59.

**2799:** W. L. Wanmaker, W. P. de Graaf and H. L. Spier: Luminescence of Pb- and Pb-Mn-activated lanthanum silicates (Physica **25**, 1125-1130, 1959, No. 11).

Lanthanum silicates of the composition $1La_2O_3.1SiO_2$ and activated with Pb give under 2537 Å excitation an U.V. emission (peak wavelength 3150 Å) and activated with Pb and Mn an orange emission (peak wavelength 5950 Å). The preparation of some lanthanum silicates is described and X-ray diagrams are given (viz. of $2La_2O_3.SiO_2$, $La_2O_3.SiO_2$ and $La_2O_3.2SiO_2$).

**2800:** G. A. Ovezall-Klaasen and J. Halberstadt: The preparation of $S^{35}$- and/or $Cl^{36}$-labelled $SO_2Cl_2$ (Int. J. appl. Radiation and Isotopes **7**, 145-147, 1959, No. 2).

Radioactive sulphuryl chloride labelled with $^{35}S$ and/or $^{36}Cl$ can be prepared in every quantity and with every possible specific activity by mixing calculated quantities of radioactive $^{35}SO_2$ with a small excess of $Cl_2$ gas, or $^{36}Cl_2$ gas with a small excess of $SO_2$. Mixing is done in a reaction vessel in which a small quantity of charcoal acts as a catalyst. The yields are practically 100 per cent.

**2801:** F. L. H. M. Stumpers and R. Schutte: Stereophonische Übertragung von Rundfunksendungen mit FM-modulierten Signalen und AM-moduliertem Hilfsträger (Elektron. Rdsch. **13**, 445-446, 1959, No. 12). (Stereophonic transmission of radio broadcasts with frequency-modulated signals and an amplitude-modulated auxiliary carrier; in German.)

Description of a system for transmitting stereophonic signals in the FM band. With $A(t)$ and $B(t)$ as the microphone signals, the frequency modulation of the main carrier is given by $A + B + a(1 + A - B) \cos \omega t$, where $a$ is the amplitude of the auxiliary carrier and $\omega$ its angular frequency. The equipment of the transmitter is discussed, and it is also shown that only slight modifications are required to make FM receivers, provided with two sound channels, suitable for the system discussed (this is an attractive feature of the system). A discussion is also devoted to the FM spectrum, and the expected signal-to-noise ratio is calculated.

**2802:** J. Meltzer: Unspecific resistance mechanisms in the house-fly, Musca domestica L. (Indian J. Malariology **12**, 579-588, 1958, No. 4).

By selecting insects with a certain insecticide (i.e. treating one generation with the insecticide, breeding the survivors, and so on for each generation), strains can be obtained that are resistant to the selecting agent. In selecting house-flies with DDT and lindane (both chlorinated hydrocarbons) the author has produced strains that show a high level of resistance to these agents. Resistance to "Diazinon" was of a lower degree, and no or only slight resistance was developed to S 17 (N,N-dimethylphenylcarbaminate). It was surprising, however, that all selected strains also showed resistance to the compounds with which they had not previously been in contact. All four were highly resistant to DDT and lindane (and also to various other chlorinated hydrocarbons used in the investigation); all four showed the same fairly low degree of resistance to "Diazinon" and scarcely any resistance to S 17. The relative toxicities of the various insecticides for the selected strains differed entirely from those for the original strain. These results suggest that this multiresistance is not so much due to specific decomposition mechanisms but rather to a generally-active mechanism. Such a mechanism might consist in a decrease of the permeability of cell membranes.

**2803:** W. J. Oosterkamp: The concept of absorbed dose and its measurement (Symposium on quantities, units and measuring methods of ionizing radiation, Rome, April 1958, pp. 86-99; publisher Hoepli, Milan 1959).

The "International Commission on Radiological Units and Measurements" introduced in 1953 the concept "absorbed dose", which is the ratio of the energy imparted to the tissue by ionizing particles to the mass in a small volume of irradiated tissue around the point $P$ where the dose is to be ascertained. The "exposure dose", measured in röntgens, is given by the ionization produced in air by the emission of electrons associated with the X-rays in an imaginary volume of air around the point $P$. The relation between these quantities is discussed and illustrated graphically for the case of a water phantom irradiated with X-rays of 0.2, 2 and 30 MeV.

**2804:** B. Combée and K. Reinsma: Methods of measuring the effectiveness of protection against ionizing radiation (as **2803**; pp. 320-337).

Review of the requirements to be met by instruments for measuring ionizing radiation from X-ray equipment and nuclear plant, and for determining the effectiveness of installations designed to provide protection against ionizing radiation. The advantages and disadvantages of existing instruments are considered, and some examples are dealt with in detail.

**2805:** N. V. Franssen: Enkele onderzoekingen omtrent richtingswaarneming (T. Ned. Radiogenootschap **24**, 321-335, 1959, No. 6). (Some investigations into directional hearing; in Dutch.)

A survey is given of the theory of binaural and stereophonic hearing. On the basis of an electrical model of the binaural hearing mechanism, the phenomena occurring in stereophonic reproduction are explained. Finally some methods of compressing the stereo-information are discussed.

**R 399:** H. J. Heijn: Representations of switching functions and their application to computers (Philips Res. Repts. **15**, 305-341, 1960, No. 4).

This thesis (Delft, 1960), continued in **R 407**, deals with switching problems in electronic computers. Of the three main elements of such computers, viz. the memory, the arithmetical unit and the control unit, the memory often consists of a matrix of square-loop ferrite rings (see e.g. Philips tech. Rev. **20**, 193, 1958/59). It is known that the arithmetical and control units could also be designed using ferrite rings in place of the usual valve or transistor flip-flops. This would involve some loss of speed, but the computer would gain in reliability. Boolean algebra, used in the design of computer circuits to determine the optimum circuit for a given operation, has to be specially adapted for application to systems using square-loop ferrite elements. The first half of this thesis is devoted to the development of an algebra based on that of Boole but adapted to systems using magnetic-ring elements. In the second half of the thesis, this algebra is used to design various magnetic-ring elements such as an adder, a counter and a decoder. Special attention is paid to the calculation of the time necessary for the transfer of carries in a binary adder. It is found that the adding time can be approximately halved when the adder is sub-divided into segments of suitable length.

**R 400:** S. Duinker: Durable high-resolution ferrite transducer heads employing bonding glass spacers (Philips Res. Repts. **15**, 342-367, 1960, No. 4).

The magnetic and mechanical properties of a very dense, homogeneous and fine-grained ferrite are discussed with reference to its application as a material for making transducer heads for magnetic recording and reproduction, for which it was specially developed as a substitute for laminated metallic alloys. The technique of producing transducer gaps by preparing ultra-thin (i.e., from less than one micron upwards), well-bonding and stress-free glass layers between polished ferrite surfaces is described, and the excellent qualities of such gaps as regards their optically and magnetically measured lengths and their wear-resistance are discussed. Several novel single- and multiple-track head constructions are outlined which, compared to existing assembly methods, are characterized by reduced production costs. This is because the expensive mechanical operations (polishing and gap preparation) can be confined to the most vital parts of the head (i.e., the frontal part containing the gap) and can be done for a large number of heads simultaneously in a prefabrication stage.

**R 401:** G. Diemer and J. G. van Santen: Power amplifiers based on electro-optical effects; a survey (Philips Res. Repts. **15**, 368-389, 1960, No. 4).

Electro-optical effects (such as electroluminescence, photoconduction and combinations thereof) and devices based on these effects are discussed from the point of view of power amplification. Special features arise owing to the transformation of the signal (from electric to radiative form, and vice versa) and owing to the quantized nature of radiative energy transfer. A survey of applications along these lines is given.

**R 402:** A. J. W. Duijvestijn and B. P. A. Boonstra: Numerical evaluation of functions occurring in a study of domain configuration in thin layers of $BaFe_{12}O_{19}$ (Philips Res. Repts. **15**, 390-393, 1960, No. 4).

Two functions occurring in the theory of domain configuration in thin layers of $BaFe_{12}O_{19}$ are evaluated numerically and presented graphically.

**R 403:** P. B. Braun and W. Kwestroo: On some calcium-iron-oxygen compounds (Philips Res. Repts. **15**, 394-397, 1960, No. 4).

Three new calcium oxide-iron oxide compounds are stabilized by addition of small amounts of a third component:

1) $Ca_4Fe_{14}O_{25}$, hexagonal ($R\bar{3}c$) with $a = 6.0$ Å and $c = 95.0$ Å. Stabilized by e.g. $Y^{3+}$. Ferrimagnetic, preferential plane of magnetization.

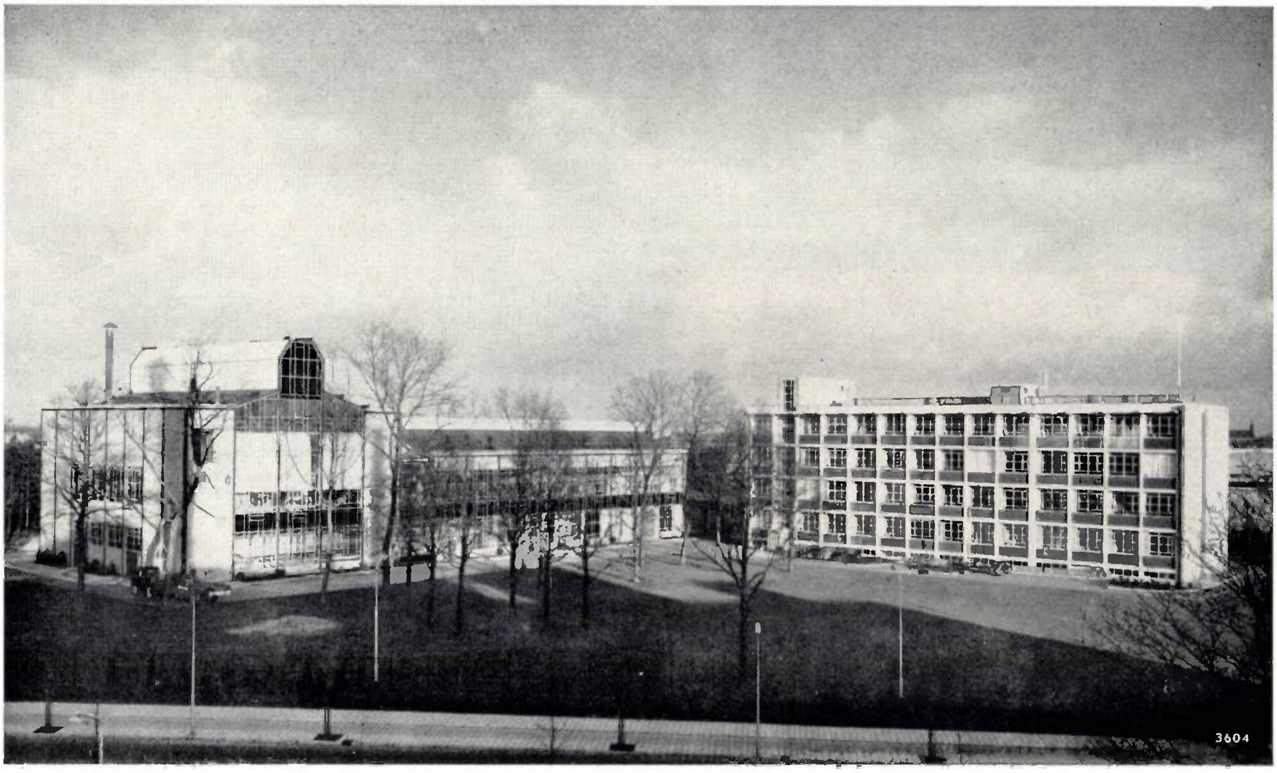
2) $Ca_4Fe_{14}O_{25}$, hexagonal ($P\bar{3}c$) with $a = 6.0$ Å and $c = 31.6$ Å. Stabilized by e.g. $Mg^{2+}$.

3) $Ca_4Fe^{2+}_2Fe^{3+}_{18}O_{33}$, hexagonal ($R\bar{3}c$) with $a = 6.0$ Å and $c = 62.3$ Å, also stabilized by e.g. $Mg^{2+}$ ions. Ferrimagnetic, preferential plane of magnetization. Magnetic properties and X-ray data are given.

VOLUME 22, 1960/61, No. 9/10 pp. 281-344 Published 27th July 1961

# Philips Technical Review

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



*This issue of our journal is entirely devoted to the material "glass" — one of the most important of the materials used in the manufacture of Philips products. Glass, in multifarious kinds and shapes, has here traditionally served the purpose of "vacuum-packaging material". The preparation and processing of glass — the modern melting furnaces with their ancillary equipment, and bulb-blowing machines, remarkable examples of mechanization — will be discussed in this issue along with the still continuously evolving theoretical views concerning the structure of glass, including such surprising ideas as the invert glasses and the "amorphous substance with lattice imperfections". A short, concluding article deals with glass-metal seals and with manufacturer-user relations in this connection.*

*This series of articles is prefaced by a concise historical survey of the uses which mankind has made of glass in the last 7000 years. We believe that articles of this kind are useful, not only to the historian but also to the engineer and to the man of science, helping them to see their own work and the work of others in a broader perspective. The survey presented here was written at our request by Professor Forbes of the University of Amsterdam, a distinguished authority on the history of technology.*

*To lend relief to the idea underlying this survey — the evolution from empiricism to science — our title photograph shows the Glass Development Centre, Eindhoven, opened in 1957, where glass research and the development of new manufacturing methods for the Philips glass factories are largely concentrated.*

# GLASS THROUGHOUT THE AGES

by R. J. FORBES *).

666.1.004.14

## Glass at the dawn of history

From the many modern types of glass and their applications, numerous threads run back through the course of history to a single point, the invention of glass-blowing in the first century before the Christian era. Only from that time onwards can glass rightfully be regarded as an independent material of common utility. Nevertheless, glass processed by a variety of more primitive methods had already been in use from the most ancient times, principally for decorative purposes.

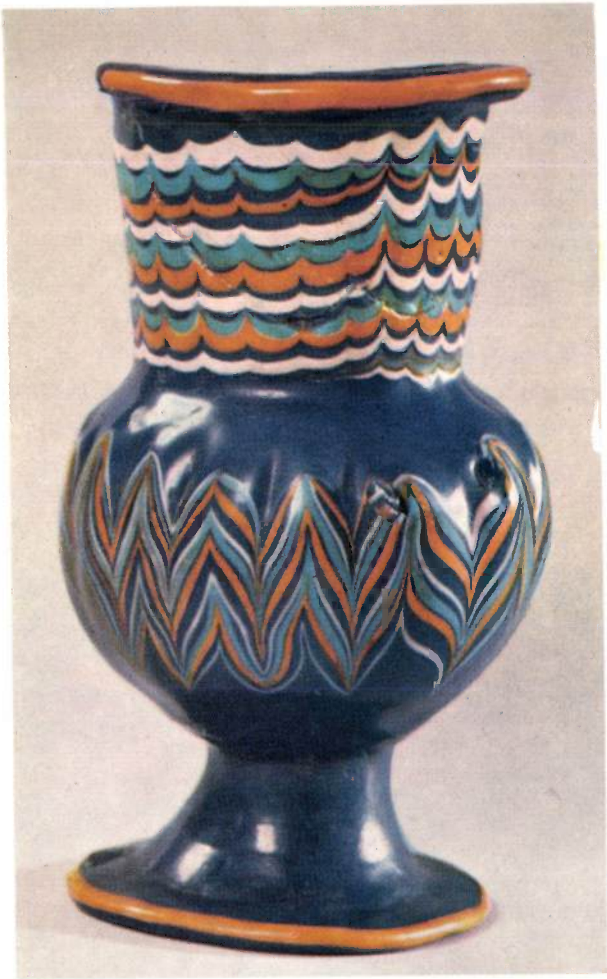


Fig. 1. Egyptian unguent jar from the eighteenth dynasty (c. 1500-1350 B.C., probably made by modelling a softened layer of glass in a mould. Reproduced actual size. (Courtesy of Rijksmuseum van Oudheden, Leyden.)

It is certain that glazed pottery was being made in the Near East as early as 5000 B.C. The glaze applied to the pottery was prepared by the fusion of silica (quartz) sand, shell lime or magnesian limestone and "alkali". The alkali used in Egypt was usually the natural soda extracted from the oases of the Western Desert; in Phoenicia and Mesopotamia it was soda ash gained from seaweed and alkaline plants. In this way even prehistoric man applied glassy coatings to brightly coloured semiprecious stones and pebbles. Variegated colouring was obtained by the addition of small quantities of copper, iron and other minerals.

Round about 1500 B.C. more and more small objects made entirely of glass began to appear in Egypt and Mesopotamia. They were made by four techniques [1]).

1) Beads or small pieces of glass were cast in moulds, and used for jewellery, mosaics or cloisonné work. Beads of this kind were exported throughout prehistoric Europe, and their shape and style now often enable the archeologist to date his finds. The technique of making these beads and imitation stones went from Mesopotamia to China, where they remained in vogue for centuries; long after, in the 18th century A.D., the technique was still used in China for making ornamental figures and small perfume flasks, which were exported to Europe.

2) A layer of hot, softened glass was pressed into shape in a mould. This technique was related to the ancient method of making faience earthenware by modelling a siliceous powder mixed with milk of lime (the heating in that case, however, being applied after the shaping). A spectacular example is the famous blue-glass neck-support found in 1922 in Tutenkhamon's tomb (14th century B.C.). In the residence of this pharaoh at El Amarna the remains of a complete glass workshop have been excavated. An Egyptian moulded vial of the period made in this way is shown in *fig. 1.*

*) Professor in the History of Science and Technology, University of Amsterdam.

[1]) R. J. Forbes, Studies in ancient technology, Vol. V, E. J. Brill, Leyden 1957.

3) Quite large arbitrarily shaped and sometimes coloured lumps of glass were cast (e.g. in clay crucibles) and then cut in the manner of rock crystal to form amulets or ornaments. Examples of this technique are known only from Mesopotamia, the oldest being a small vase inscribed with the name of King Sargon II (700 B.C.), shown in *fig. 2*.

4) Rods of softened glass were modelled around a core of sand and then reheated to fuse them together, after which the sand was removed. Sometimes disks were cut from the still warm, coloured glass rods and arranged around a core; they were then fired together to form a vase or bowl ("millefiori" technique, see *fig. 3*). The vials, unguent jars and perfume flasks made by this process were very popular with the ancient Greeks and Romans, and the history of the millefiori technique is closely bound up with that of the cosmetics industry.

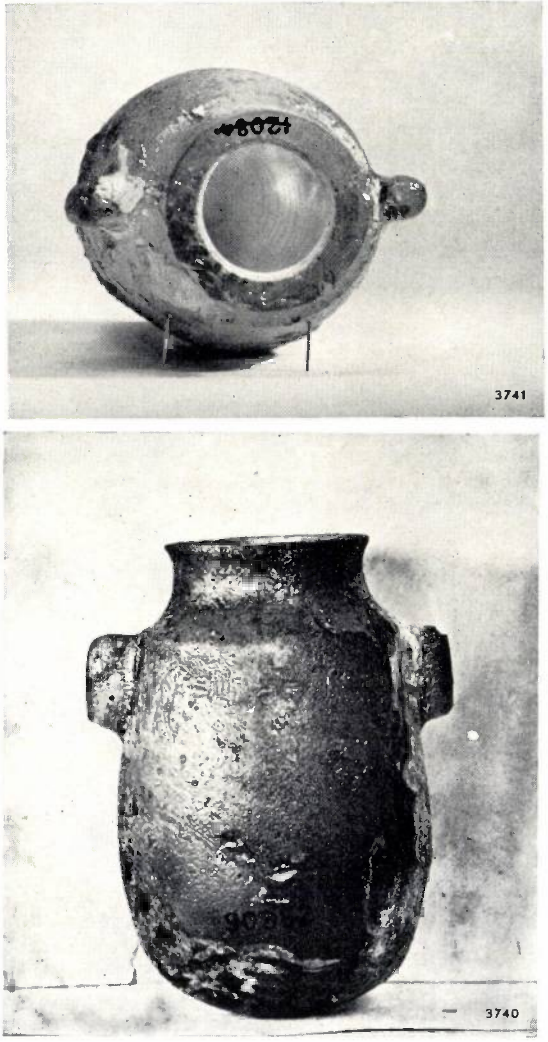


Fig. 2. Glass vase (height about 6″) bearing in cuneiform script the name of King Sargon II of Assyria (700 B.C.); found at Nimrud. The interior grooves demonstrate that the vase was ground from a solid block of glass. (Courtesy of British Museum, London.)

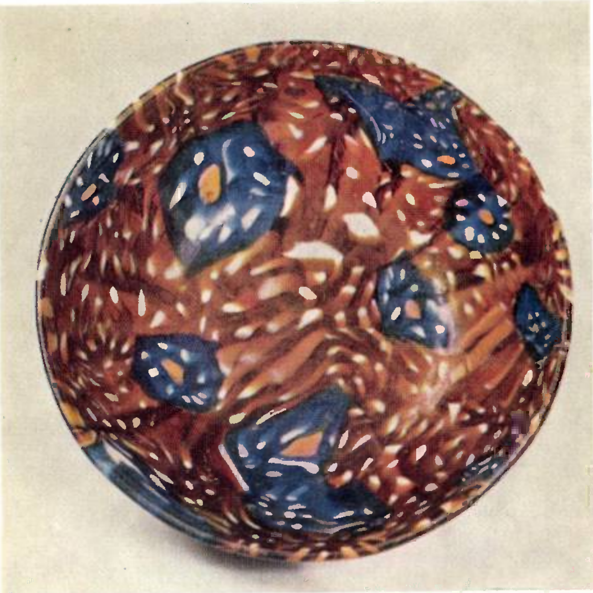


Fig. 3. Millefiori bowl from the first century A.D.; diameter 5.7 ″. Found at Nijmegen. (Courtesy of Rijksmuseum van Oudheden, Leyden.)

Early texts on glass-making, dating back to 1700-700 B.C. (*fig. 4*) and originating not from Egypt but from Assyria, give detailed information on furnaces and smelting procedures, with recipes for a crude glass which was fused with colouring minerals to produce glazes or coloured glass. The basis is still sand-soda-lime glass, which was being made in Phoenicia at least six centuries before Christ. And it was there, in about 50 B.C., that the decisive discovery was made — the blowing of glass.

### From ornament to utility material

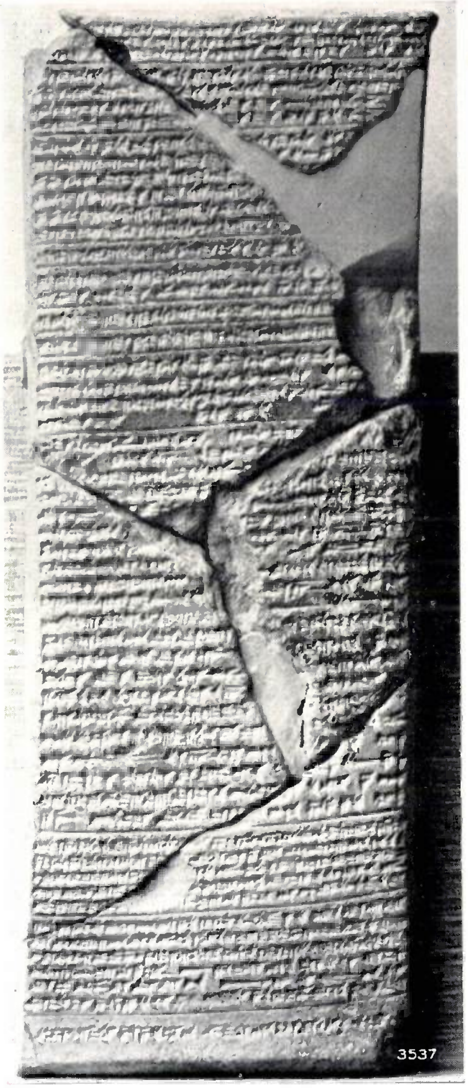
The invention of glass-blowing, which was probably first done in a mould but was very soon followed by free or "off-hand" blowing, revealed new potentialities and properties of glass as a material. This led to a much greater diversity of shapes and designs (*fig. 5*) and also to the use of glass in homes and workshops, together with or instead of metal and earthenware. It was especially the possibility of making glass which was *transparent* and not merely translucent that opened up new prospects for science and daily life.

Whether the decisive factor was the ease of working and shaping the material, or its useful and decorative properties, is difficult to say; at all events the use of glass now widened rapidly.

Within a century of the invention of soda-lime glass-blowing, the art had spread via Persia to the Orient. In the West, Syrian, Jewish and Alexandrian glass-blowers were to be found in Rome and the districts south of the capital, in the valleys of the Rhône and the Saône, and in the Rhine province.

From there the art spread to Spain, to the border districts of the Low Countries and Gaul — where timber for the furnaces was plentiful — and to Britain. The primitive methods of pressing the glass in a mould or shaping it around a core still persisted, but off-hand blowing was the most important technique. Glass-cutting established itself as a separate art, practised not by the "vitrearii" (glass-blowers) but by the "diatretarii" (glass-cutters), whose techniques sprang from the much older skills of working rock crystal and precious stones [2]).

In the numerous objects that have come down to us from that period we see the application of a variety of decorative techniques. The old millefiori process continued to flourish in ancient Alexandria, the centre of the cosmetics industry. A novel decorative technique was based directly on blowing: the *hot* lump of glass on the blowpipe was blown out into a mould, and after repeated immersion in various molten glasses and re-blowing, layered glass objects were produced. Ribs or grooves could be tooled into the shaped objects whilst still *warm*, or glass bosses or threads applied as decoration. The glass could also be *cold*-worked by polishing, painting or enamelling, and gold-leaf might be applied, often sandwiched between two glass layers. Finally, cold glass could be hollow-ground by the glass-cutters and provided with facets with a rotary grindstone and emery; the surface could be given a relief finish or it could be engraved. The latter was done with obsidian or similar hard stone, diamond



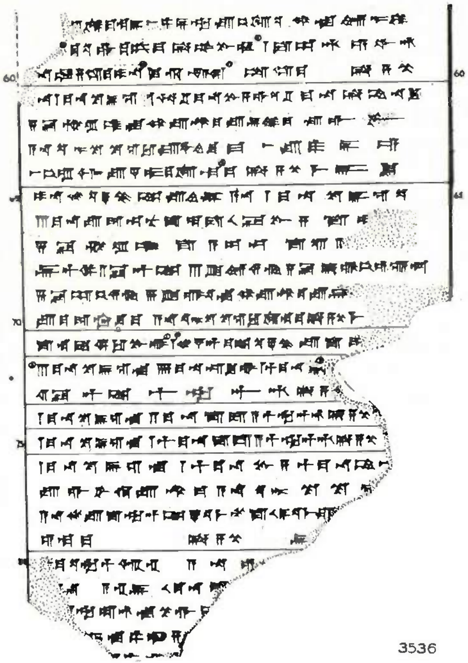


Fig. 4. The earliest document on glass-making: Mesopotamian cuneiform tablet from the seventeenth century B.C., giving recipes for making various kinds of glass and glazes. Left, photo with transcription beside it; right, translation of the first twenty lines. Investigators have followed the recipes and proved their validity (and the correctness of the translation); see p. 131 et seq. of book cited in reference [1]). (Photograph by courtesy of British Museum, London.)

1. To a mina of zukû-glass (thou shalt add) 10 shekels of lead, *)
2. 15 shekels of copper, ½ a shekel of saltpetre, ½ a shekel of lime,
3. thou shalt put (it) down into the kiln (and) shalt take out santu (red) glass of lead.
4. To a mina of zukû-glass (thou shalt add) 1/6th (mina) of lead,
5. 14 shekels of copper, 2 shekels of lime, a shekel of saltpetre:
6. Thou shalt put (it) down into the kiln (and) shalt take out Accadian santu (red) glass.
7. (Thou shalt) green the clay and in vinegar and copper shalt thou keep it.
8. At the third (day) of thy keeping
9. it will deposit a bloom and thou shalt take (it) out.
10. Thou shalt continuously pour it off and it will dry and
11. thou shalt shape it. If it is (like) marble, be not troubled.
12. Accadian santu-glass and (that of) lead
13. thou shalt take in equal parts, and
14. triturate them together.
15. After thou hast melted them together,
16. into 1 mina of the melt a shekel and a half of zukû-glass,
17. 7½ grains of saltpetre, 7½ grains of copper, 7½ grains of lead
18. shalt thou triturate together and
19. thou shalt melt and keep it (so for) one (day?)
20. and shalt take it out and cool it . . . .

*) 1 mina (505 grammes) = 60 shekels = 10 800 grains (of barley). The Assyrian scales were certainly correct to within 5 milligrammes!

[2]) Much of our knowledge of glass in those times comes from Pliny, particularly from the 26th book of his "Natural History".

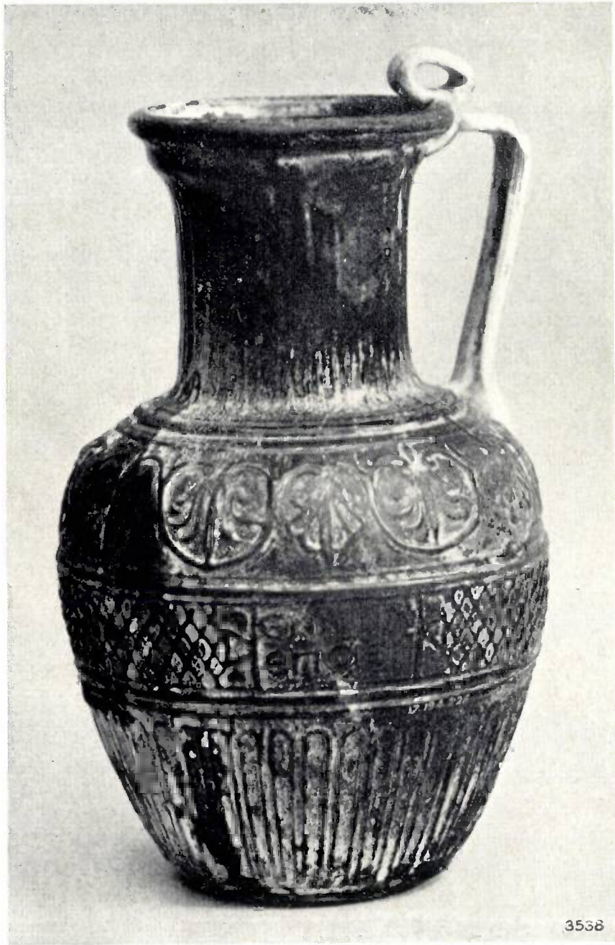


Fig. 5. Mould-blown yellow glass jug, dating from early in the first century A.D. Shortly before this time, off-hand (free) blowing must also have been invented. The jug bears the signature "Ennion", the name of a Syrian glass-blower whose products have been unearthed at numerous places in the Roman Empire of that time. (Courtesy of Metropolitan Museum of Art, New York.)

being unknown in Antiquity. An example of what these craftsmen could achieve with relatively simple means can be seen in *fig. 6*.

Decorated blown glassware first made its appearance in the homes of rich Romans as a rival to gold and silver ware. At the same time, millefiori flasks, bowls and trinkets were being imported into Rome from Alexandria. The earliest glassware was almost as dear as cut rock crystal, but round about 75 A.D. prices began to drop as a result of the quantity production of simple tableware in addition to fine glass. The glass workshops became so numerous that in Rome, from 200 A.D., the glass-blowers were concentrated in the suburbs of Mons Caelius because the smoke from their furnaces had become a nuisance. An organized trade even sprang up in broken glass, which thus found its way back to the glass works. Around this time glassware on a Roman table was in fact a sign of poverty! A similar trend took place in about 200 A.D. in the Near East, and finds in the excavations at Karanis (Fayum, Egypt) bear witness to the versatility of Alexandrian glass-blowers, even in the production of simple glassware for daily use. During the persecution of the Christians under Diocletian (300 A.D.) the headman of an Egyptian village declared that his community was poor and could therefore only afford altar vessels of glass. In the West a similar situation arose much later: in about 850 A.D. Pope Leo V forbade the use of glassware for celebrating Mass since it had become too cheap and debased a material.

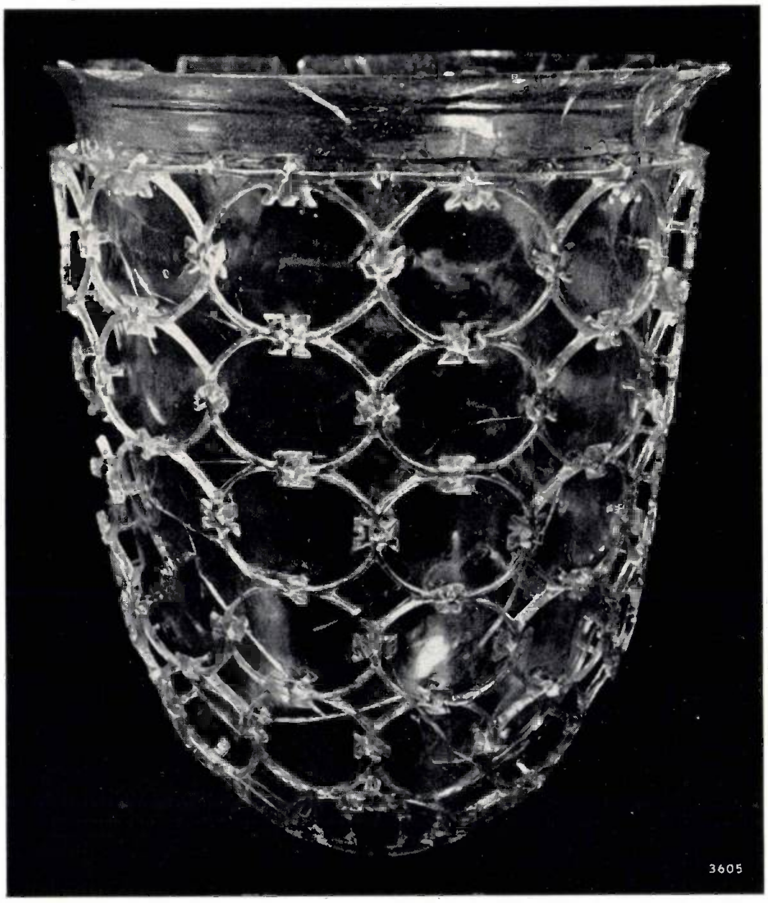


Fig. 6. Roman "cage cup" (height 7") found in 1950 in a stone sarcophagus at Niederemmel near Trier, probably made at the Cologne glassworks towards the end of the third or the beginning of the fourth century A.D. Cage cups, cut from a single, originally thick-walled, glass beaker, probably represent the pinnacle of the glass-cutters' art. The technique adopted is discussed in detail by F. Fremersdorf, Schuhmacher-Festschrift, Mainz 1930. (Courtesy of Rheinisches Landesmuseum, Trier.)

### Europe discovers new kinds of glass

The glass industry which had grown up in Gaul and in the Rhineland flourished for several centuries. Thriving centres of production were established notably in Cologne and Trier (cf. fig. 6) which, in about 300 A.D., ranked third among the cities of the Roman Empire. It was here, probably from a Germanic word for amber (or generally for a transparent, lustrous substance) that the late-Latin term *glesum* originated, to which we owe our own word "glass".

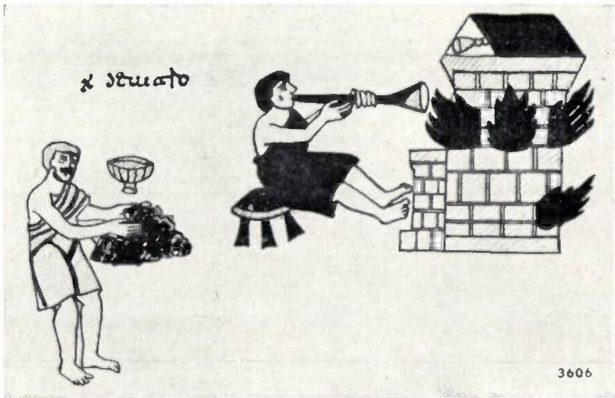


Fig. 7. Picture of a glass furnace, given by Hrabanus Maurus [3]).

The barbarian invasions and the fall of the Roman Empire in the West temporarily put an end to the development of the glass industry in Central Europe. Glass-blowers were still to be found in the cities along the Rhine and the Rhône, but many of them retreated to the comparative safety of Italy, in particular to the Po valley and to l'Altare (near Genoa), from where they later spread out again all over Europe. The gradual revival of glass craftsmanship in the Frankish Empire was partly due to the churches and monasteries, which not only acted as the wardens of ancient knowledge but were also the patrons of new and extensive applications of glass, for example glass mosaics and stained-glass windows — though its use was not approved

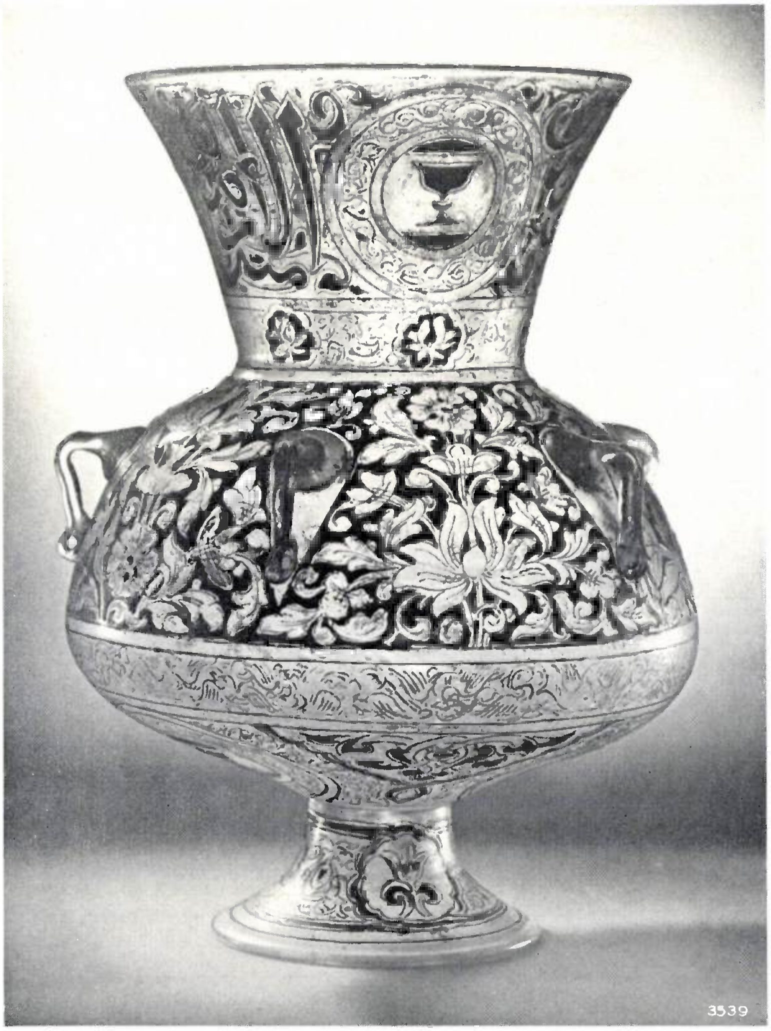


Fig. 8. Painted mosque lamp from Damascus, 1350. (Courtesy of Corning Museum of Glass, Corning, N.Y.)

for altar vessels. To Hrabanus Maurus (died 856 A.D.), Bishop of Mainz and Councillor to Charlemagne, we owe the description of glass production in those times[3]) and the oldest picture of a glass furnace (*fig.* 7). The Rhenish glass industry was hampered by the difficulty of importing soda from the South, and had to resort to plant ash, which contained potassium. Later too, after the tenth century, this potash glass remained characteristic of Central Europe when the glass-blower's art spread to Thüringia, Bohemia, Saxony and Silesia, whilst soda glass continued to be made in the coastal districts. Techniques, however, underwent very little change, partly owing to the lack of knowledge of glass chemistry. The most that can be said is that other forms appeared, such as the "tumbler" type of drinking glass. The earliest handbooks on glass making, by Heraclius and Theophilus from the ninth and tenth century[4]), disclose an unbroken tradition going back over at least 1500 years.

In the East, too, the tradition remained unbroken, even after the emergence of Islam. The same decorative technique as used for rock crystal remained in vogue, and beautifully painted, gilded and enamelled Islamic glassware was produced (*fig.* 8), some of which found its way into the treasury repositories of mediaeval cathedrals. The arrival of the Mongols saw the application of Chinese decorative elements to this glassware, but at the same time it provoked the flight of many glass-blowers from Damascus and Aleppo to the West. Here, Venice had become an important glass centre since the commissioning of glass-blowers from Constantinople (about 1050) to make the mosaics for San Marco. In about 1200 they formed a powerful guild which, in 1291, settled on the island of Murano, with the idea that such isolation would better preserve their secrets. The secrets of their fabrication process were so jealously guarded that the emigration of guild members was forbidden on pain of death. On Murano the privileged glass-blowers made decorative tableware from soda glass, which had to compete with the products of gold and silver smiths; from 1500 they followed old Roman and Eastern styles.

The death penalty on the emigration of Venetian glass-blowers did not prevent their art from filtering out into other countries in the sixteenth and seventeenth centuries. The Venetian Gridolphi, who came to work in the Netherlands, complained that the local glassware was hardly to be distinguished from genuine Venetian ware. In 1575 Verzelini acquired a 21-year monopoly in England for making "Venetian glass", on condition that he taught the English the art. In 1592 he sold his patent to English merchants, who were able to extend it for seventy years.

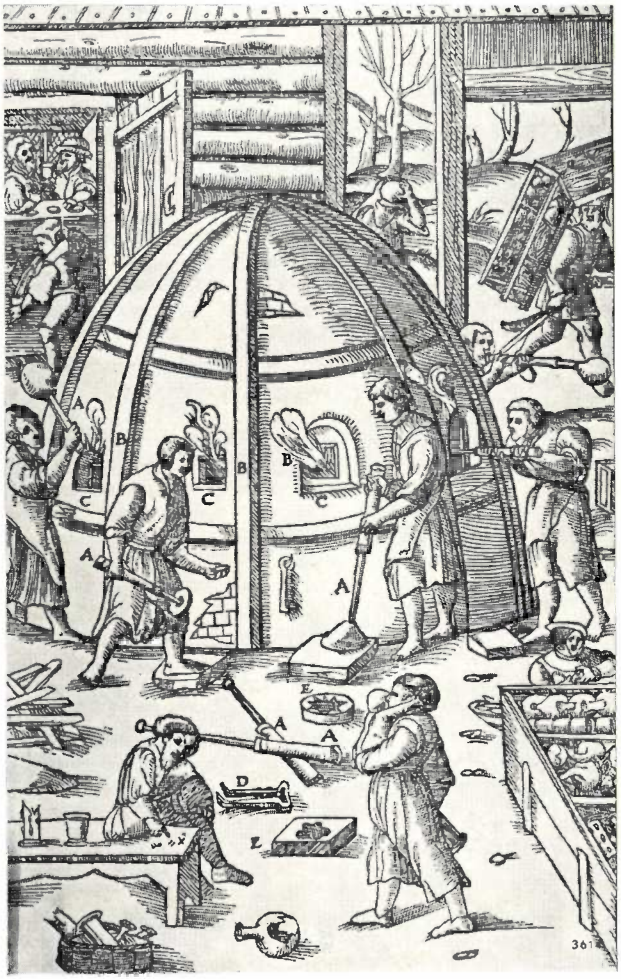


Fig. 9. Glass workshop from the year 1550, a contemporary print by Georg Agricola, De re metallica, Basle 1556.

Meanwhile glass had come into use on a large scale for household ware, for chemical equipment and other appliances, and for window panes. The large glass furnaces illustrated by Agricola in his "De re metallica" (1556) provide evidence of the volume of this production (*fig.* 9). The Italian scientific achievements of 1550-1650, to which we shall return presently, owed a debt to the Venetian art of glass-blowing. The fact that the alchemists were also keenly interested in glass is not surprising. They had a special symbol for glass (*fig.* 10), as they had for other important substances. As early as about 550 A.D. Aeneas of Gaza remarked that the emergence of the brilliant substance glass from the fusion of humble, almost worthless raw materials was an example of the "metamorphosis of matter into a

[3]) Hrabanus Maurus, De originibus rerum, Monte Cassino mss. 1023.

[4]) Compositiones Variae, Codex 490, Bibliotheca Capitulare, Lucca; Heraclius, über die Farben und Künste der Römer, Edit. Ilg, Vienna 1873; W. Theobald, Des Theophilus Presbyter Diversarum artium schedula, V.D.I. Verlag, Berlin 1933.

superior state", which was the avowed object of the alchemists. In the technology of glass, however, their efforts brought little progress, owing to the lack of essential knowledge of its composition.

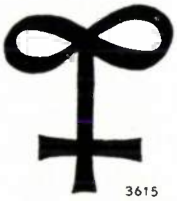


Fig. 10. The alchemists' sign for glass, composed of the symbol X for breaking, the sign ☿ for Mercury and the sign ☌ indicating a blowpipe plus bulb. (After K. F. Bauer, Glastechn. Berichte **24**, 191, 1951.)

The 17th century ushered in important changes. The year 1612 saw the publication of Neri's handbook which, soon translated into many languages, disclosed the Italian art of glass-making to the glass-blowers of other countries [5]). In 1615 the shortage of timber for the English Navy became so acute that glass-makers and others in England were forbidden to use timber as fuel. This prohibition had long been threatening, and patents had already been granted for the smelting of glass in crucibles in a coal furnace, using "sea-coal" from Newcastle. After many experiments, Ravenscroft succeeded in making a softer and more brilliant glass than the Venetian soda-lime product (1675). His potash lead-oxide glass could be made entirely from native English materials. It was marketed by the Glass Sellers' Company, and also enjoyed considerable success on the continent of Europe. In Holland, Anna Roemer Visscher, in the first half of the seventeenth century, had created a vogue for dot-engraved drinking glasses ("rummers", see *fig. 11*) which now, in about 1700, were made from the new "crystal glass".

Meanwhile in Bohemia and Germany a potash-lime "crystal glass" had come into use. Kunckel describes how much better this could be worked on a rotary grindstone than soda-lime glass, a method introduced by Caspar Lehman, jeweller at the court of Rudolph II. At the same period opaque glass was made with the aid of stannic oxide and calcinated bone or horn, and in about 1675 gold-ruby glass was made from "purple of Cassius", a compound of stannic chloride and gold chloride. The Bohemian crystal caused a decline in the Venetian glass industry, but was overshadowed in its turn by the English crystal towards the end of the eighteenth century.

In France, Jean Baptiste Colbert, Minister of the Crown under Louis XIV, set out to make the native glass industry independent of Venice by protective legislation. The factories at Baccarat, Clichy and St. Louis were soon turning out millefiori work and mirror glass of excellent quality, and in the middle of the nineteenth century French crystal glass gained the ascendancy, since heavy taxes, which were not lifted until 1845, now encumbered the English glass industry, which indeed had partly crossed over to Ireland. Nevertheless, it remained inventive. Discoveries made at that time were that good opaque glass could be made with arsenic compounds and that the addition of blast-furnace slag represented a saving in alkali; in 1755, a transparent ruby-lead glass was invented.

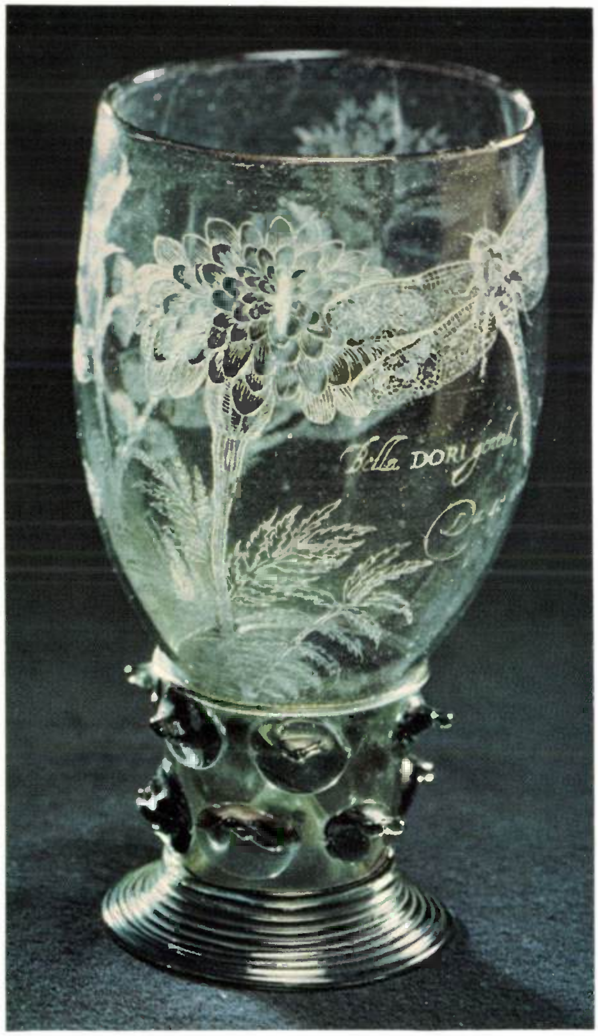


Fig. 11. Dutch drinking glass from 1621, ornamented with diamond engraving by Anna Roemer Visscher. (Courtesy of Rijksmuseum, Amsterdam, and the Stichting Openbaar Kunstbezit.)

The early development of the American glass industry was beset with difficulties. Attempts made

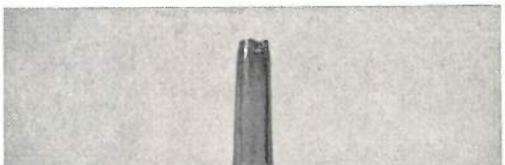
[5]) Antonio Neri, L'arte vetreria, Florence 1612.
English: C. Merret, The art of glass, London 1662.
The art of glass by Mr. H. Blancourt, London 1699.
Latin: De arte vitraria Libri VII, Amsterdam 1686.
German: J. Kunckel, Ars vitraria experimentalis, Amsterdam and Danzig 1679.
French: M. D., Art de la verrerie, Paris 1752 (annotated by Merret and Kunckel).
J. Haudiquer de Blancourt, De l'art de la verrerie, Paris 1697.

by the London Glass Company in 1608 and 1621 to start manufacturing glass in America ended in failure. Brief successes in the eighteenth century were strangled by the competition from the mother country. It was not until after 1820 that a series of improvements were introduced into the manufacture of glass in the New World which were gradually to promote the American glass industry to a position of pre-eminence. The first of these improvements was the mechanical pressing of glassware in iron moulds, a process invented in 1827 and employed by the Boston and Sandwich Glass Company; further developments will be discussed in another context.

After this more general survey of the history of the glass industry, we shall now consider in somewhat more detail the evolution of various applications.

### Glass crucibles and vessels

In the Middle Ages, scholars like Robert Grosseteste (1175-1253) were already using glass in their laboratories for reaction vessels, crucibles and retorts, glass having the obvious advantage of enabling the experimenter to see what was taking place inside the vessel. Generally, however, the early glass was not well able to withstand the protracted heating which was then an important aspect of chemical research. To make the glass vessels heat-resistant they had to be given a clay jacket, which of course destroyed the advantage just mentioned. Moreover, the glass was not well able to withstand leaching. Boerhaave (1668-1738) devoted a special treatise to the proposition that the protracted boiling of water in a glass retort would produce "earth"; this proved to be leached constituents of the glass wall [6]). *Fig. 12* shows a 17th century distilling vessel of the "pelican" type, much used by Boerhaave. Better glass for chemical equipment first appeared after the foundation of the Schott factory at Jena (1882). The borosilicate glass "Pyrex", now very



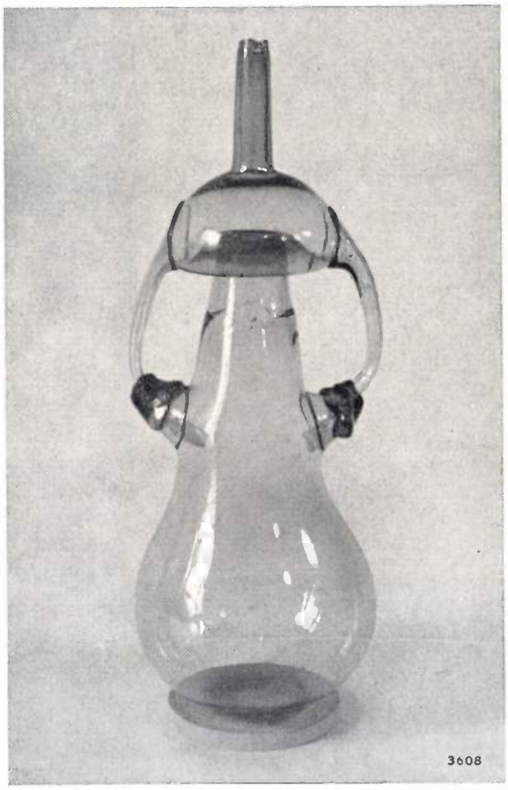


Fig. 12. Seventeenth-century "pelican" retort. An account of such retorts is given e.g. by E. J. Holmyard, Alchemy, Penguin, London 1957. The fine example shown here has become known as "Rebecca with the pitcher". (Courtesy of Rijksmuseum voor Geschiedenis der Natuurwetenschappen, Leyden.)

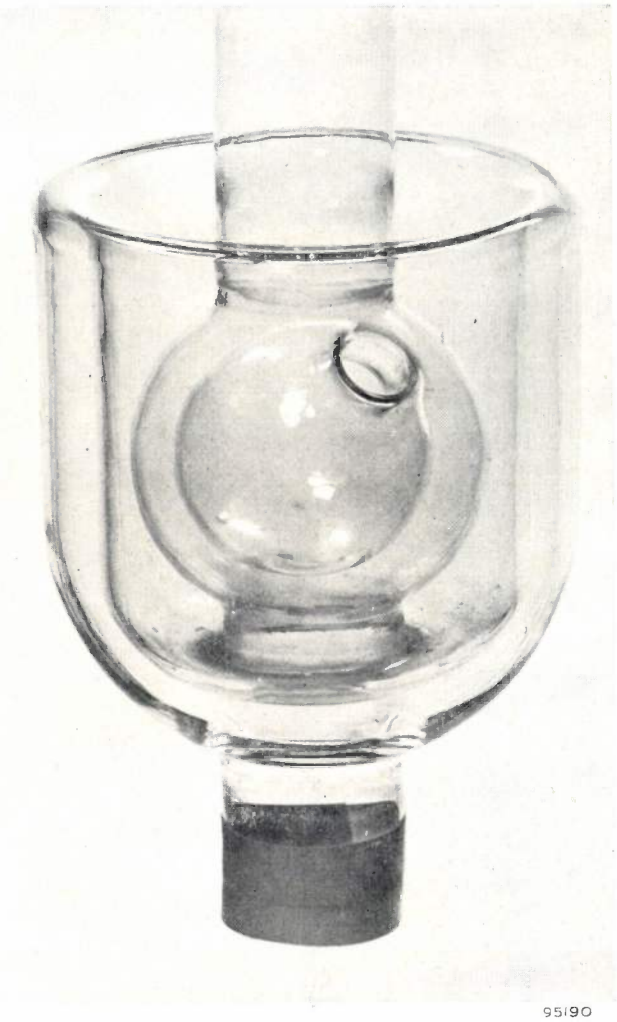


Fig. 13. Liquid-air-cooled vapour trap of very high conductance and high condensation efficiency, for an ultra-high vacuum pump built at Eindhoven (A. Venema and M. Bandringa, Philips tech. Rev. 20, 151, 1958/59). The vessel is filled by pouring liquid nitrogen (e.g.) into the "beaker"; the liquid then also enters and fills up the innermost bulb.

[6]) H. Boerhaave, Elementa Chemiae, Leyden 1731/32; see also A. S. Margraf, Mém. Berl. Akad. Wiss. 1766, pp. 20-31, where Boerhaave's hypothesis is disputed.

widely used, was developed in about 1910 by the Corning Glass Works in New York state, and was first marketed in 1915. The refinement of glass-blowing techniques and the Moissan electric furnace (1893) also made it possible to process quartz for making laboratory appliances (1904). Interchangeable ground glass joints (1929) and ground glass stoppers (1918) greatly facilitated the construction of chemical equipment, and introduced glass where metal had previously been used. Modern laboratory glassware may sometimes assume intricate shapes, as illustrated in *fig. 13*. Particularly impressive are the large apparatus for the chemical processing of liquids which must not come into contact with metals: even the pumps in such equipment may be made entirely of a special type of glass ("Duran", Schott).

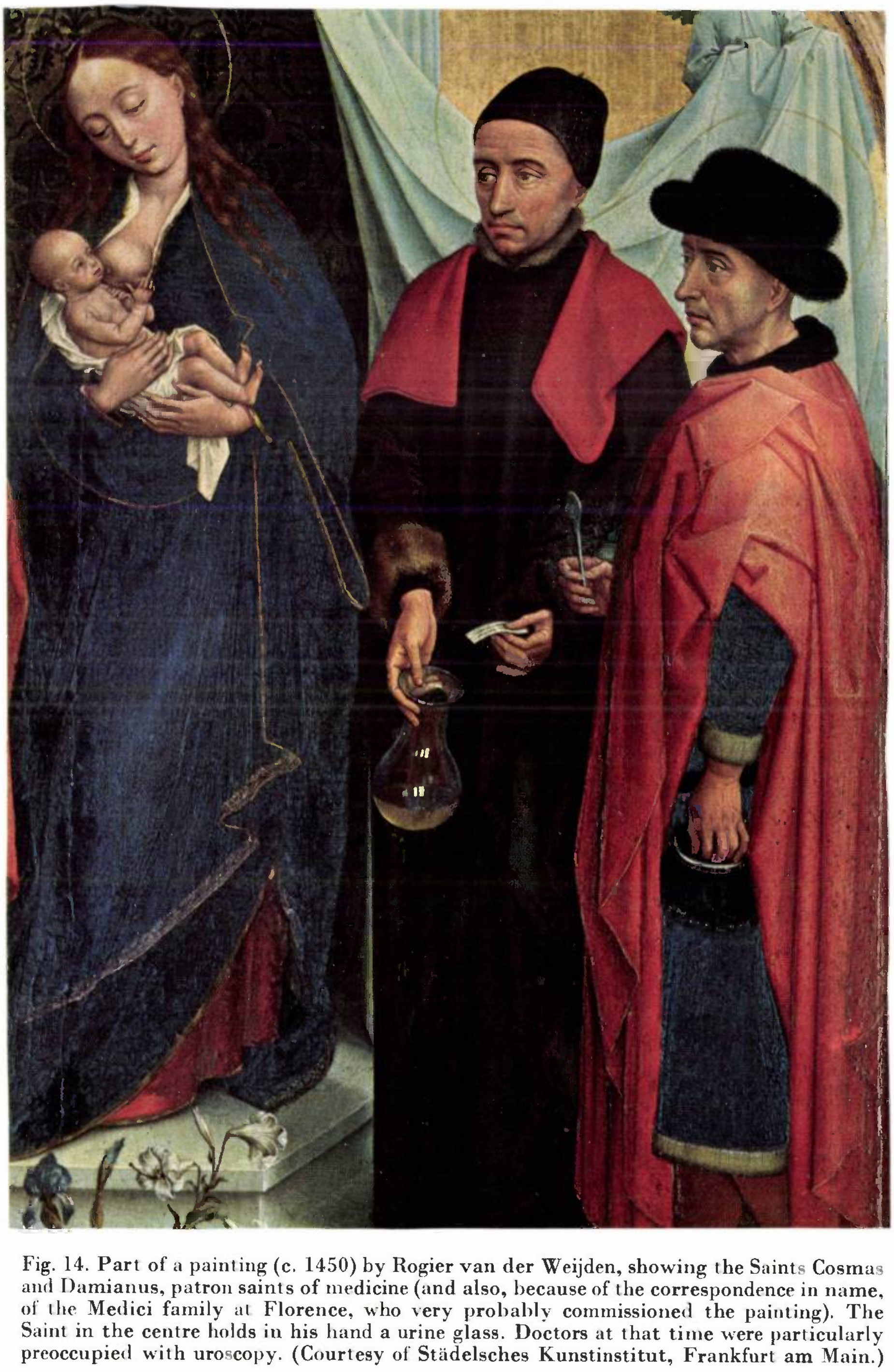


Fig. 14. Part of a painting (c. 1450) by Rogier van der Weijden, showing the Saints Cosmas and Damianus, patron saints of medicine (and also, because of the correspondence in name, of the Medici family at Florence, who very probably commissioned the painting). The Saint in the centre holds in his hand a urine glass. Doctors at that time were particularly preoccupied with uroscopy. (Courtesy of Städelsches Kunstinstitut, Frankfurt am Main.)

Glass was more easily turned to use for medical purposes than for chemical processes. The inspection of urine for colour and sediment in the transparent, appropriately shaped urine glass (*fig. 14*) played an important part in mediaeval medicine, even though it was later officially banned as a diagnostic method owing to its abuse by quacksalvers. Again, transparency and easy shaping (shapes were at first rather complicated) were the principal requirements for the thermometer, since c. 1610 an important medical instrument, which was introduced into general clinical practice by Boerhaave 100 years later (*fig. 15*). In 1584 Jeremias Martius wrote: "Das Glas braucht der Mensch auf mancherlei Weg, aber der Nutz, so es in der Artznei hat, übertrifft das ander alles" ("Man uses glass in many ways, but nowhere is it more useful than in medicine").

### Windows

The dwelling house of Antiquity, with its inner courtyard and few windows, had little need of glass window panes. The coverings used were lattices of wood or earthenware, slabs of translucent selenite or alabaster, or oil-impregnated bladders and sheets of parchment. The first glass window panes were found in excavations at Pompeii, in particular in a building dating from about 60 B.C. These were cast on a flat stone or in a mould and drawn on all sides with pincers (as described by Theophilus) to an average size of 1′×2′. There were also round glass disks measuring 2 to 2½′ in diameter, and mounted in a marble frame. In the Eastern Roman Empire blown disks 6 to 8″ in diameter first appear in the fourth century; these seem to be the forerunners of the Norman "crown glass", which was made in the Middle Ages in the West. The process applied was to take a blob of glass from the blowpipe on to a "pontil" (an iron rod) and to rotate it rapidly until it flattened into a disk thicker in the middle than at the edges (*fig. 16*). Panes made in this way were still in use up to the nineteenth century.

Another method, also described by Theophilus [4], produced what was known as "broad glass". By this method, developed in Germany, a cylinder was blown to a length of 4 to 6′ and a diameter of 1 to 2′; both ends were removed and the cylinder was cut open while hot and rolled flat on a stone table with the aid of a piece of wood, after which the pane was cooled in a kiln. Excellent results were obtained with this process. In 1832 it was imported from France to England, where it was further developed. The Chance Brothers Glass Company succeeded in making cylinders 6′ in length and 1½′ in diameter, and a better method of abrasion and polishing was discovered. It was in this way that the panes were made for the Crystal Palace in 1851, thus marking the advent of glass as a *structural material.* The same year saw the similar use of glass for building a large hothouse, in which the exotic Victoria Regia was made to bloom for the first time in Europe. (The hothouse, although in more modest dimensions, already existed in Roman times, pumpkins being grown for Tiberius Caesar in a box closed with "transparent stone" to protect them from inclement weather.) The cylinder method reached its highest stage of development when Lubbers in 1903 succeeded in drawing (whilst blowing) cylinders 40′ long and 3′ in diameter from a crucible in the furnace.

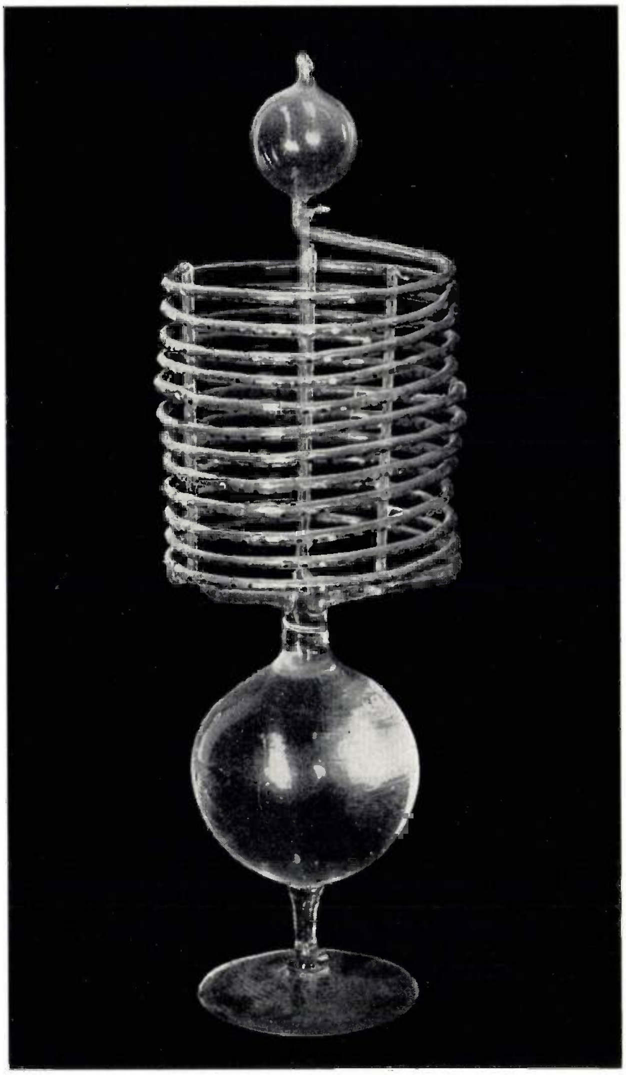


Fig. 15. One of the earliest sealed liquid-thermometers, built by Santorio at Florence in about 1650 for meteorological and medical work by the Accademia del Cimento. These thermometers were filled with alcohol ("acqua arzente bianca") and the scale graduation — usually in 100 or 50, but sometimes in 300 or even 520 "degrees" — was marked with enamel beads: black for the digits, white for the tens and blue for the hundreds. A recent calibration has shown that the inside diameter of the metres-long stem (here wound in a spiral, though straight types also exist) is remarkably constant. (Courtesy of Museo di Storia della Scienza, Florence. See also: M. L. Bonelli, Gli strumenti superstiti dell'Accademia del Cimento, Domus Galil., Pisa 1958.)

Glass panes of greater mechanical strength and also, after somewhat cumbersome working, of superior quality were produced by a casting process.

This process received considerable impetus, particularly in France, from the demand for mirrors and for carriage-door windows. In about 1690 the cast-glass process underwent considerable improvement at the Royal glass factories at St. Gobain. The molten glass was poured on to a metal table and flattened with rollers, after which the glass plate was ground and polished. Blessed with abundant supplies of timber for fuel, the industry flourished in France.

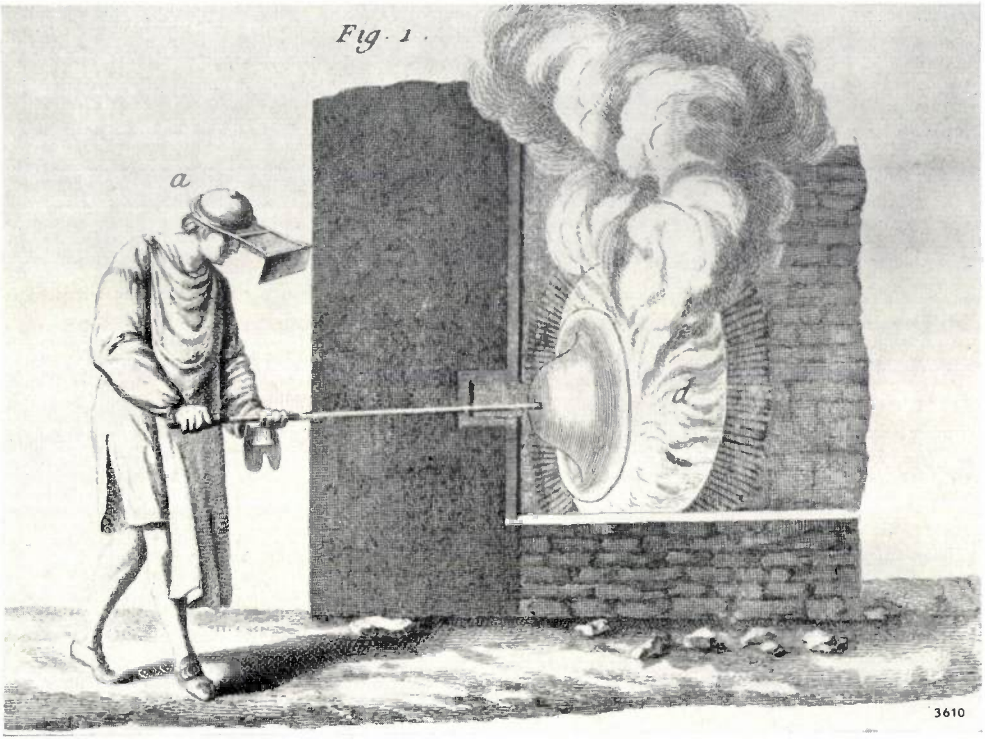


Fig. 16. Crown glass being made with the "pontil" (a solid iron rod). A gather of glass from the furnace is first blown into a sphere. This is taken over with a pontil, diametrically opposite the blowpipe. By renewed heating of the now opened sphere in the furnace and rapid rotation of the pontil, the glass is flattened into a disk having a thicker part in the centre (the "crown"). The glass workers wore protective masks during this very hot and rather hazardous process. (From "Encyclopédie ou dictionnaire raisonné des sciences, des arts et des métiers" by Diderot and d'Alembert, Paris 1752-1777; in part re-issued by Dover, New York 1959.)

Not so in England. Although the casting process had been imported there in 1670 for the manufacture of window panes, the economy of production was hampered by a tax on window space. In 1773, French skill in this field was brought to England by the Huguenots, and in 1776 the Ravenhead glassworks managed to cast panes measuring 12′×6′. Through the use of a cast-iron base plate the dimensions were increased in 1843 to 15′×9′. The growth of building activity in England around 1820 stimulated efforts to find better and preferably continuous methods of producing window glass. Bessemer attempted in 1846 to cast glass from the furnace through two rollers, but it was not yet possible to make this into a continuous process. In 1884 the Chance Brothers evolved a method of casting glass on to a sloping plate and subsequently rolling it. In 1857 a patent was granted on a process of drawing plate glass from a furnace, but it was not until 1901 that Fourcault was able to apply this process by drawing glass from the furnace under hydrostatic pressure through a slit, the "débiteuse" and then rolling it. This had become a commercial process by 1913.

### Mirrors

The earliest mirrors were of metal (bronze or silver), although the Egyptians also used mirrors of opaque black glass. Glass mirrors, with a backing of tin foil, first appeared in about 200 A.D. Mediaeval mirrors were usually octagonal, slightly convex pieces of glass, backed with lead. In 1503 the Del Gallo brothers invented the method of treating the back of mirror glass with a tin-amalgam, for which they were granted a 20 years' patent in 1507. The method of applying a backing of silver to glass by chemical means was not discovered until about 1840. (Silver has now been almost entirely supplanted for this purpose by aluminium.) The size of mirrors, like the subdivision of windows, was governed by the attainable size of the glass plate. To make such mirrors really flat was very difficult, as anyone can see who walks around a hall of mirrors of the Rococo period: up to the end of the eighteenth century, large mirrors were still built up from small sections.

### Stained-glass windows

The use of glass for church windows is mentioned in documents dating from the fifth century; coloured church glass was first reported by Anastasius in about 800. This application of coloured glass seems to have been evolved in Constantinople from the ancient Egyptian art of glass mosaics. Instructions for making leaded glass windows are given by Theophilus [4]) and other writers. Like the ancients, mediaeval craftsmen made blue glass with copper salts, green with iron and copper, dark-blue to purple with manganese, red opaque glass with cuprous oxide, white opaque glass with stannic oxide, yellow opaque glass with antimonic oxide, and black glass with a great deal of iron, copper or manganese. The fourteenth century discovered the art of producing lemon-yellow and orange coloured glass with silver chloride. Neri [5]) gives recipes for dark-blue cobalt glass.

The coloured glass was cut into pieces following the outlines of drawings; up to about 1500 this was done with a hot iron, after which time diamond began to be used. The pieces were joined together with strips of lead to form a window. The earliest stained-glass windows were built up in this way from simple pieces of coloured glass. In the 13th century the practice started of painting the coloured glass with "grisaille" (glass frit, powdered metal and gum), which was baked in to produce shading effects (*fig. 17*). In the 16th century, grisaille and enamel paints were applied directly to colourless panes, thereby degenerating the stained-glass window to an imitation of the art of panel painting.

The great advances in chemistry between 1750 and 1850 made it possible to produce types of glass in many new colours [7]), which were now not only used for artistic purposes, in leaded windows or for ornamental glass, but also for more everyday uses in all kinds of household glassware. In 1774 it was found that nickel salts (discovered in 1751) produced glass ranging in colour from brown or grey to purple, and in 1779 that ferric oxides gave a blue glass and uranium salts a yellow glass. Green chrome-glass was discovered round about the same time. Thénard (1777-1857) made a systematic investigation of the influence of added salts. The purer substances of the new chemical technology made it possible to obtain better reproducible colours.

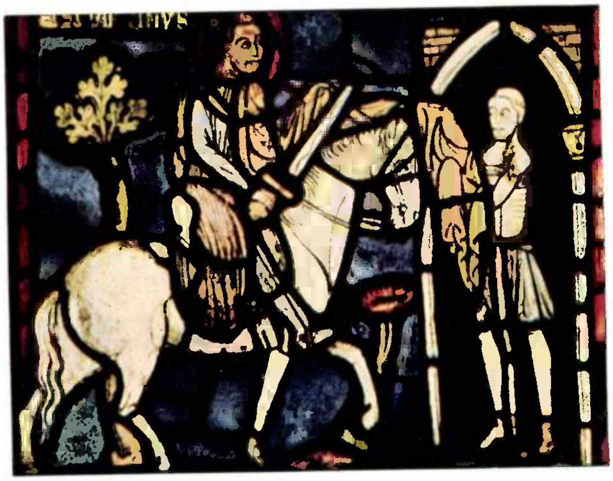


Fig. 17. Part of a stained-glass window dating from 1220, originally in Gerey Abbey, demolished at the end of the 18th century, and now in the Musée de Cluny, Paris. The panel shows St. Martin arriving at the gate of Amiens, where he gives a beggar half his mantle, which he cuts through with his sword. The lead strips holding the pieces of coloured glass are seen to be an integral part of the composition, e.g. in the contours of horse and sword. (Reproduced with the cooperation of N.V. Filmfabriek Polygoon, Hilversum.)

In the foregoing we have outlined the evolution which glass has undergone under the impact of practical requirements and the possibilities opened up by the advance of chemistry, an evolution which clearly demonstrates the cooperation that has grown up between glass-maker and natural scientist since the Renaissance. The finest example of this cooperation, however, is to be found in the history of spectacles, the telescope and the microscope.

### "Roundels for the eyes"

Lenses for use as burning-glasses were already known in Antiquity. Strepiades, a character in Aristophanes' "Clouds", says that a burning glass, for lighting a fire, might also be used for setting a Court Order alight from a distance without touching it. It is probable that lenses were used as primitive magnifying glasses by the cutters of gems and cameos. Science in Ancient Greece was acquainted with katoptrics, the theory of reflection from mirrors, etc., and also with dioptrics, the theory of the

[7]) W. Ganzemüller, Beiträge zur Geschichte der Technologie und der Alchemie, Chemie-Verlag, Weinheim 1956.

refraction of light. The astronomer Ptolemy even gives a table of angles of incidence with their associated angles of refraction, without attempting, however, to give laws for the relation between them. Hellenic scientists also studied the rainbow and the functioning of the eye, both of them problems concerning the refraction of light.

The Greek heritage was preserved and further developed by Arabian scholars, notably by Ibn-al-Haitham (965-1039), known in the Middle Ages as Alhacen, whose "Opticae Thesaurus" served as the basis of all theories on the operation of the eye, mirrors, etc. right up to the seventeenth century (*fig. 18*). To make it more accessible to European scholars, his work on the rainbow was translated into Latin in 1170, and his "Opticae" in 1269. At about this time various scholars in the West, amongst them Robert Grosseteste and Roger Bacon (1214-1294), began to take an interest in optics. Grosseteste proposed the use of lenses for magnifying small objects and bringing distant objects closer. Bacon suggested that the focus of the eye in long-sightedness due to old age might be corrected by using a segment of a sphere of glass. Soon afterwards the first spectacles appeared, although they were not yet designed and constructed on scientific principles (*fig. 19*). They were probably invented round about 1280; in his book on wine, brought out in that year, Arnald de Villanova mentions a "euphrasia", a wine that makes the eyes brighter and helps people "to read small print again without eye glasses". The Venetian regulations for "cristalerii" (glass-blowers) of 1300 also make reference to "roidi da ogli" ("roundels for the eyes"). In the fourteenth century they were being made in Flanders and Zeeland (Holland), and later in Bavaria in the towns of Nuremberg, Augsburg and Regensburg. *Concave* lenses for short-sightedness were first mentioned by Nicolaus of Cusa, in about 1450, and first illustrated in 1518. Until Kepler explained the action of eye and spectacles in his textbook on optics (Ad Vitellonem paralipomena, 1604), spectacles were made on a purely empirical basis, although the demand for them had greatly increased after the invention of printing. In the seventeenth century, spectacles were already in common use, but it was not until a century later that physicians began to study optics seriously. Bifocals, i.e. spectacle lenses of two powers, one for distant vision and the other for reading, were made by Franklin in 1760. Cylindrical lenses for correcting astigmatism were designed by Young in 1800, and introduced into Holland by Donders in 1864. By the nineteenth century, empiricism had been so far routed that people now began to consult an eye specialist before buying spectacles. In 1890 Von Rohr computed the famous "Punktal" lens for Zeiss. Subsequent developments with regard to spectacles are more a question of fashion than of science.

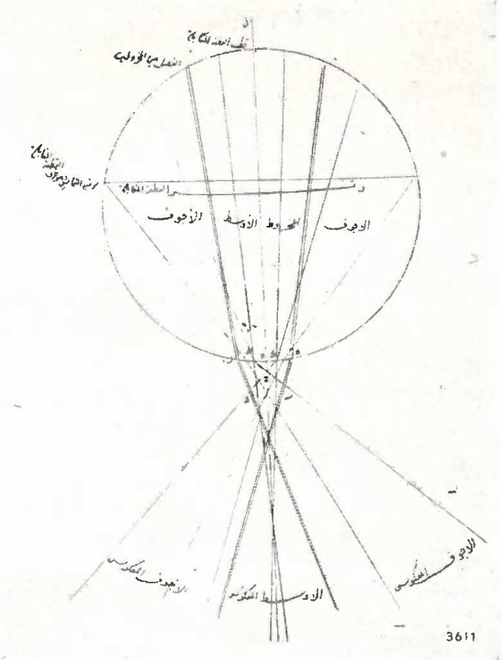


Fig. 18. Drawing by Ibn-al-Haitham (Alhacen, early 11th century) of the way in which a crystal sphere projects an image. The drawing is reproduced from a copy of the commentary written by Kamâl al Dîn in about 1310 on Alhacen's "Opticae". From the book it is evident that Alhacen knew that only the rays close to the optical axis must be used to produce a sharp focus. (Courtesy of the University Library, Leyden.)

### Glass reveals new worlds

Before the appearance of Kepler's treatise on optics Zacharias Jansen of Middelburg, Netherlands, had already made a telescope, in about 1590, consisting of a double-convex objective lens and a double-concave lens as eye-piece. In 1620 he made a binocular telescope for Prince Maurits of Orange. At about the same time telescopes were also being made in Italy. Galilei, inspired by an instrument imported from Holland, made a telescope himself in 1609, with which he observed details of the surface of the moon, the satellites of Jupiter, the phases of Venus and the spots on the sun — discoveries which he described in his "Nuncius Sidereus" of 1610.

The earliest microscopes appeared in about 1620. In 1667 Robert Hooke published his "Micrographia", but his lens systems were weaker than the single lenses ground from small glass spheres by Anthonie van Leeuwenhoek, with some of which magnifications of more than 200 times were achieved! Unfortunately, Van Leeuwenhoek never disclosed the methods he used to grind his excellent lenses.

The mathematical theory underlying the design of optical systems and lenses, and the theories of light on which the optical laws could be based, were created in the seventeenth century. Descartes, who laid the foundations of analytical geometry, of great importance to optics, provided in his "Dioptrique" (published as a supplement to his "Discours de la Méthode", Leyden 1638) a mathematical derivation of Snell's law of refraction (1621). In 1665 Grimaldi published a treatise on the reflection, refraction and diffraction of light. Christiaan Huygens, who himself made lenses together with his brother Constantijn Huygens (*figs. 20* and *21*), expounded his wave theory of light in "Dioptrica", a work otherwise concerned entirely with geometrical optics, on which he worked from 1652 to 1692. Newton's emission theory was put forward in his "Opticks" of 1704.

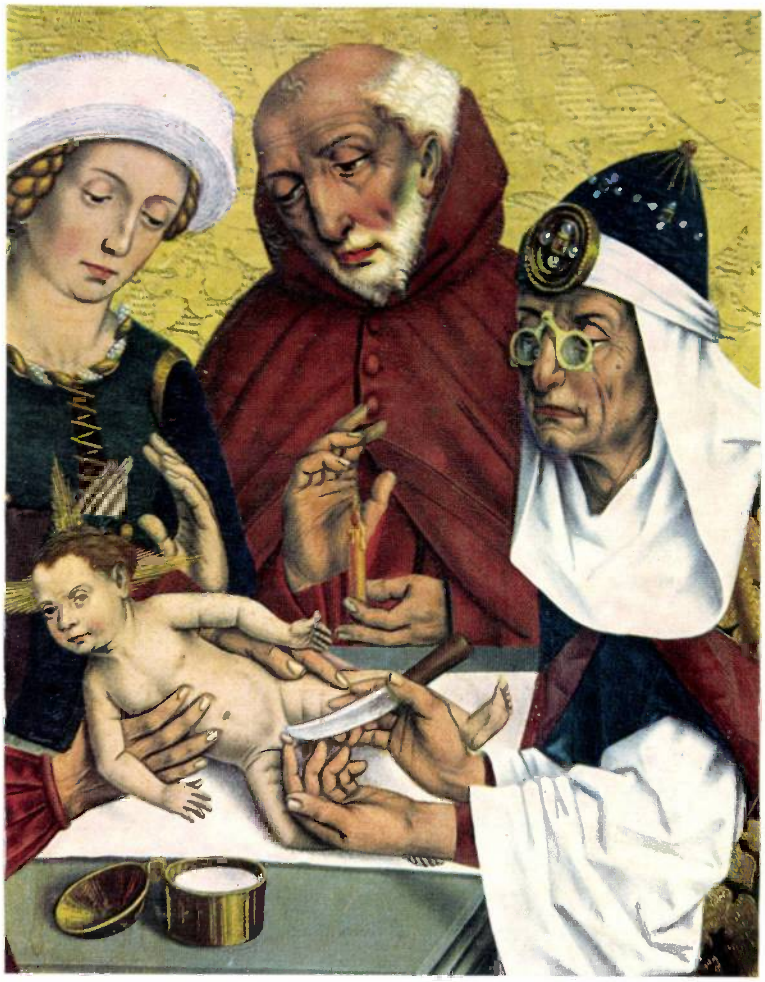


Photo Ohmeye

Fig. 19. Part of a painting by Friedrich Herlin (Circumcision of Christ) dating from 1466, on an altar in St. Jakobs Kirche at Rothenburg ob der Tauber, showing a man wearing spectacles. On the Predella of the same altar the painter also represents a bespectacled Peter the Apostle. The spectacles in both cases are of the "hinged" type, having exactly the same form as depicted more than a hundred years earlier (1352) by Tommaso Barisino in a fresco at Treviso. (Courtesy of the custodian of St. Jakobs Kirche, Rothenburg, Germany.)

The lenses used by these workers had largely to be made by themselves. Good instrument makers, like Eustachio Divini (1620-1695, in Italy) and Christopher Cock (1660-1696, in London) were few and far between. Soon, however, determined efforts were made to overcome the defects of lenses arising from shortcomings in the glass as well as from the intrinsic errors such as spherical aberration, distortion, astigmatism and chromatic aberration. In his "Géométrie" Descartes had already shown that spherical aberration might be corrected by using lenses with elliptical and hyperbolic surfaces, but the grinding machines at that time were not capable of producing such surfaces [8]). Use had therefore to be made of lenses of considerable focal length, in which the error was not so noticeable. As regards chromatic aberration, Newton believed that this was not to be avoided, but Klingenstierna at Uppsala calculated in 1760 that the solution could be found in a combination of convex and concave lenses of different types of glass, i.e. of different refractive index. Chester Moor Hall in 1733 had meanwhile arrived at the same conclusion empirically, and John Dollond, an instrument maker, was granted a patent for achromatic lens combinations of this sort in 1758. He put them on the market, but it was not until about seventy years later that they found general application in microscopes.

The shortcomings in the glass itself were mainly poor surfaces and inhomogeneities, and we shall now consider the efforts made to overcome them. Like the application of types of glass with specific optical properties, they provide another striking illustration of the fruitful exchange between the natural sciences and the glass industry.

[8]) C. A. Crommelin, Het lenzen slijpen in de 17de eeuw, H. J. Paris, Amsterdam 1929. (The grinding of lenses in the 17th century; in Dutch.)

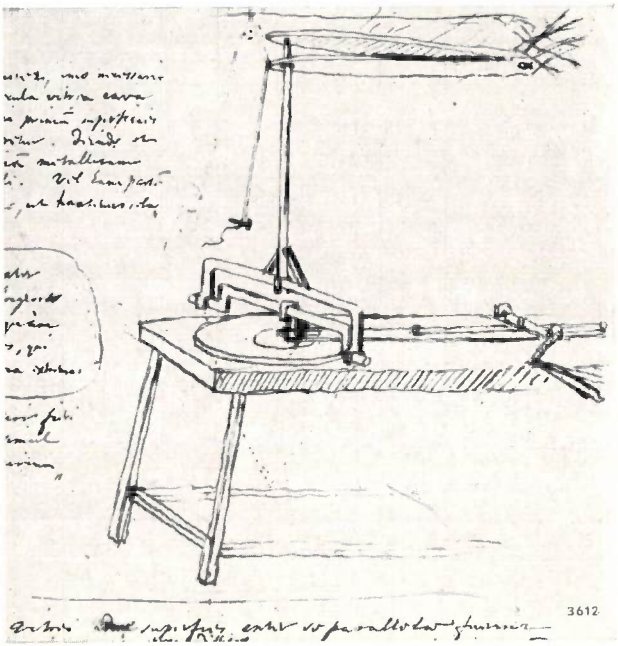


Fig. 20. Lens-grinding machine, designed by Christiaan Huygens. The grindstone on the table is set in rotation, and the lens rotating in the opposite direction is pressed onto the stone by the cross-frame and a wooden spring suspended from the ceiling. Drawing by Huygens in May 1692. (Courtesy of the University Library, Leyden.)

### Better glass and new instruments

Improvements in the finishing and polishing of lenses were introduced in the eighteenth century in France and England, when a number of leading instrument makers turned their attention to lens systems. In the nineteenth century they were joined by German instrument makers. A major contribution to improving the quality of glass itself was made by Pierre Louis Guinand of Switzerland (1748-1824). Inspired by the stirring of "fondu" (hot cheese with wine), he sought for a means of homogenizing the glass batch in the crucible, and finally devised a stirrer, a cylinder of refractory stone (1805). On the invitation of the manufacturer Utzschneider, he went to Benediktbeuren to join forces with Von Fraunhofer (1787-1826), who was investigating various types of glass on his spectrometer and designing glass-grinding machines and testing instruments. In 1832 Guinand's son Henri founded a factory at Paris for the manufacture of optical glass. He sold his "know-how" for 3000 frs. to Bontemps, the leading glass technologist of the middle of the last century. Bontemps later associated with the Chance Brothers Glass Company at Birmingham, where the new processes were patented in 1838.

Scientists in several countries now began to show increasing interest in the glass industry. On the instigation of the Royal Society, Faraday studied the processes of the Falcon Glass Works, whilst Harcourt used the new chemical methods to investigate various types of glass, and made titanate and borate glass. In 1830, Dumas published an important treatise on the composition and processing of soda-lime glass, and gradual progress was made in the difficult analysis of silicates. The chemical industry was now supplying pure raw materials — pure soda,

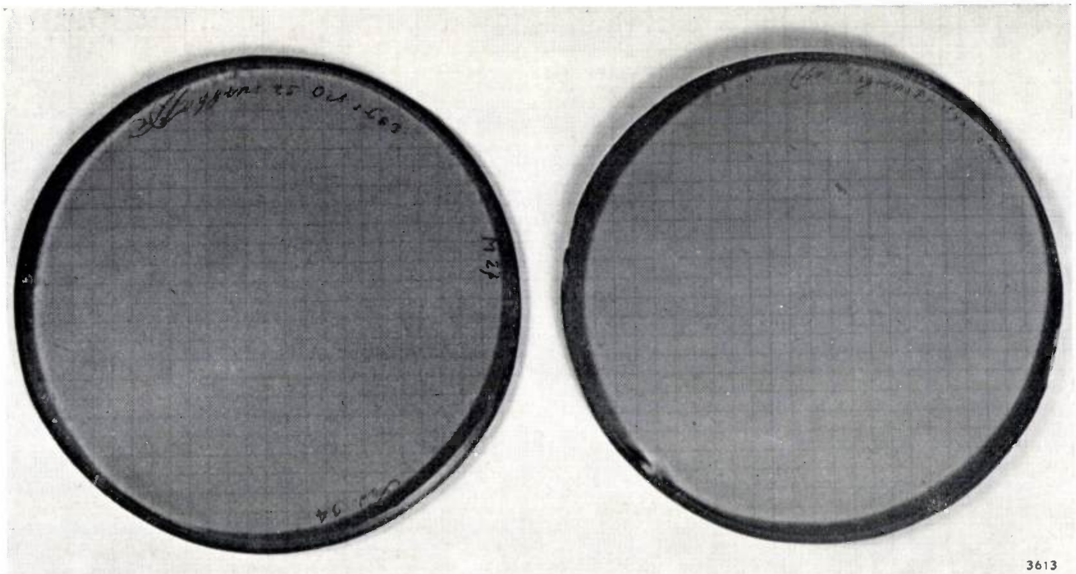


Fig. 21. Telescope lenses, ground and signed by Constantijn and Christiaan Huygens, 1683. Each square of the paper underneath the lens measures 5 mm. Owing to the considerable focal length no distortion is visible in the pattern of squares. The lenses are made of greenish glass: chromatic aberration was reduced by cutting off the red and blue ends of the spectrum. (Courtesy of Rijksmuseum voor Geschiedenis der Natuurwetenschappen, Leyden.)

for example, had been marketed since 1790 — and these gradually supplanted the impure natural products.

Only with the improved quality of lenses did it become possible to derive full benefit from achromatic objective lenses for microscopes, and the years between 1830 and 1880 saw a rich harvest of biological and medical discoveries. In the same period, too, there was a spate of new optical instruments. The following summary of the more important of these instruments will help to show their significance to science and industry:

1) The polarimeter, created by Nicol (1833, 1840), and improved by Savart and others.
2) Associated herewith, the Biot saccharimeter (1842), which evolved through a succession of improvements to its final form in 1874.
3) The primitive refractometer invented by the Duc de Chaulnes (1767) was made into a precision instrument by Abbe (1872).
4) The interference equipment of Fresnel (1822), Brewster (1831) and others evolved into the modern interferometer — the Fabry and Pérot interferometer (1899) and that of Lummer and Gehrcke (1901).
5) The spectroscope, introduced as a laboratory instrumentby Meyerstein (1856), was used by Bunsen and Kirchhoff for chemical analysis (1861).

Once again, we see the repercussions of this development on the glass industry. In 1876, Abbe and Von Helmholtz pointed out that the usefulness of optical instruments depended not only on the quality of the lenses and prisms, but also on the optical properties of the glass used. Schott, who had been looking for new types of glass with better properties, entered into cooperation in 1884 with Abbe and with Carl and Roderich Zeiss at the glassworks at Jena. At that time the optical properties of only some ten types of glass were known; within a few years they had tested thirty new elements as supplements or substitutes for the old range of seven elements used in the manufacture of glass. Schott's catalogue of 1902 offered no less than 80 types of optical glass.

### Bottles and lamp glasses

Whilst the advances in chemistry were steadily placing the manufacture of glass on a firm scientific basis, and Abbe and Zeiss, working on a foundation of theory from preceding generations, were creating optical glass as we know it today, the glass industry also began to undergo radical changes in its methods of production. The nineteenth century, with the beginnings of mass production, saw the introduction of *mechanization* into glass manufacture, as in so many other fields. Glass tubing and bottles, needed in such large quantities and for so many purposes, were obvious candidates for mechanization. Between 1859 and 1893, numerous semi-automatic machines for bottle-blowing were patented, the most successful of which was the machine invented by Ashley. The first completely automatic bottle-making machine was the work of Owens in America, who in 1898 designed his "hand-gun" for gathering from the furnace the right amount of glass to be blown into a bottle. Together with Libby, Owens built between 1899 and 1904 a completely automatic bottle-making machine with six to fifteen arms, capable of turning out hundreds of bottles an hour [9]).

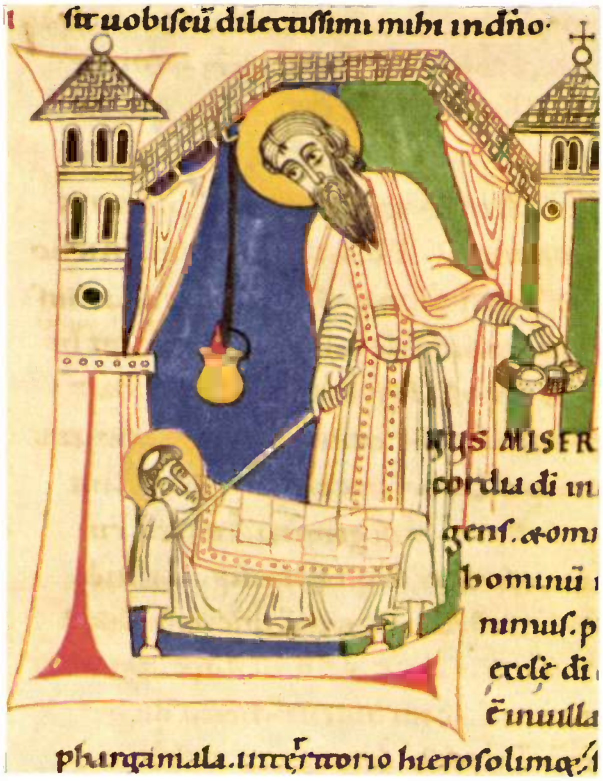

sit uobiscū dilectissimi mihi indño·
IUS MISER
cordia di in
gens. et omi
hominū
nimus. p
eccle di
ē inulla
plurima mala. inc̄ritorio hierosolimæ

Fig. 22. 12th century illumination of the initial letter L, from a Passionale (book on the sufferings of the Saints and Martyrs) originating from Hirsau Abbey. The priest Lucian, asleep in the baptistery, receives instructions from Gamaliel on finding the remains of the martyred St. Steven (415 A.D.). Above the sleeper is depicted a lamp enclosed in glass. (Courtesy of Landesbibliothek Stuttgart; see also: K. Löffler, Schwäbische Buchmalerei in romanischer Zeit, Augsburg 1928.)

The use of glass for lamps was already known in Antiquity; the word "candela", now part of the lighting engineer's vocabulary in another sense,

[9]) W. C. Scoville, Revolution in glass-making 1880-1920, Harvard Univ. Press, Cambridge 1948.

originally meant a glass oil lamp. At first, glass simply replaced the ceramic material formerly used for holding the oil. In about the seventh century A.D., however, it is recorded that oil lamps or candles in portable "lanterns" began to be enclosed in transparent glass to shield them from the wind. This use of glass as a "container" for a light is to be seen in many mediaeval drawings (*fig. 22*). Early in the nineteenth century, lamp glass began to be mass-produced for the chimneys of the new lamps burning oil, kerosene or colza oil. In 1785 Argand had to go to England to find glass-blowers able to make lamp glasses, but by about 1840 they were being produced all over Europe, and the market was further expanded after the invention of the incandescent gas mantle in 1883. This production, too, was mechanized in about 1900. The Schott Works at Jena, which employed their new refractory borosilicate glass for this product and turned out some 40 million lamp glasses a year, owed much of their prosperity and growth to this production.

In 1879 the Corning Glass Works made glass bulbs for Edison, who brought out his first incandescent electric lamps two years later (*fig. 23*). From that modest beginning arose the electric-lamp industry. Its elaborate bulb-blowing machines now produce thousands or tens of thousands of bulbs an hour [10]). The process somewhat resembles the automatic manufacture of bottles — for which reason bottles and lamp glass are here lumped together under the same heading.

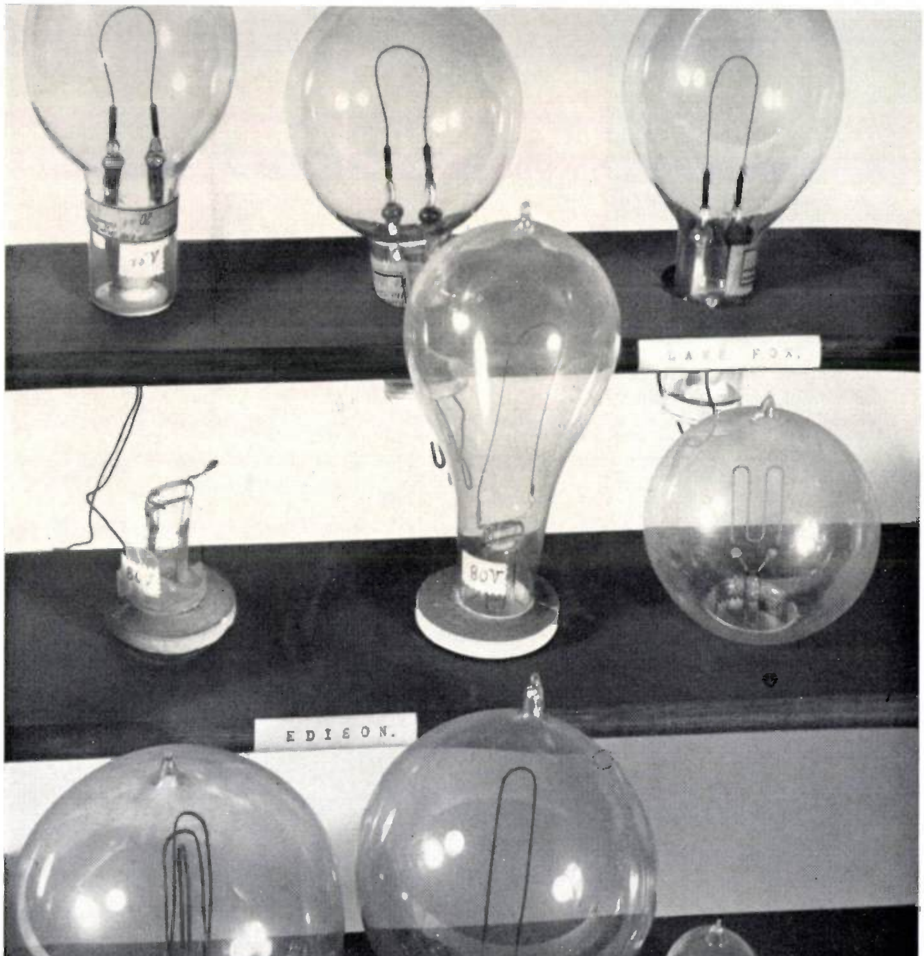


Fig. 23. Some incandescent electric lamps from a collection of the Teyler Foundation. The lamps were bought in 1881 at the Paris exhibition of electro-technology. The whole collection comprises 20 lamps, made by Edison, Swan, Lane Fox and Siemens. (Courtesy of Teyler's Museum, Haarlem, Holland.)

The last fifty years have witnessed the birth of so many new applications of glass that our relatively detailed account must end at this point. The manufacture and use of glass in the present century are dominated by the influence of scientific investigation. Chemical research led to the comparative analysis of silicates in general and to a deeper insight into the vitreous state. Physical research was no longer confined to the optical properties of glass. Interest began to be taken in its viscosity and thermal properties, such as its expansion, which are important in the glass-to-metal seals found in electric lamps and in so many instruments and apparatus. The investigation of its electrical properties led to the use of glass insulators in electrical equipment. Knowledge of its mechanical properties made glass into a modern structural material, and the study of its transmission of heat and light

[10]) See the article by P. van Zonneveld in this issue.

opened up the possibility of using glass as a heat insulator in buildings. Other developments were porous glass (frits), produced by sintering powdered glass, and glass "wool" and "textiles", made by the extrusion and spinning of fibres. These in turn led to fresh applications as a structural material (glass fibres for reinforcing polyesters, etc.) and, remarkably enough, to novel uses in optics (fibre optics [11]). It is a fascinating study to follow in this way the evolution of glass from a crude substance used for glazing the surface of pottery, or for making imitation jewellery, to one of man's most versatile materials with an ever-widening range of applications.

[11] See, for example, B. O'Brien, Physics today **13**, No. 1, 52, 1960.

**Bibliography** (some books devoted to glass, which are generally available):

R. Schmidt, Das Glas, Reimer, Berlin 1922.
B. Kollak, Glas, Technik und Kunst, Winkler, Darmstadt 1937.
O. Völckers, Glas und Fenster, Bauwelt-Verlag, Berlin 1939.
W. B. Honey, Glass, Victoria and Albert Museum, London 1946.
C. J. Phillips, Glass the miracle maker, Pitman, London 1948.
W. Schnauk, Glaslexikon, Callwey, Munich 1959.

**Summary.** The history of glass is surveyed up to about 1900, with special emphasis on applications. Glazing and glass-making techniques were used in Ancient Egypt and Mesopotamia, primarily for decorative purposes. The discovery of glass-blowing, in Phoenicia about 50 B.C., was soon followed by the quantity production of glass objects, including tableware, and by the spread of the glass industry into many parts of Europe. After briefly sketching the evolution of glass technology, touching on Arabian, Venetian and other products, the author devotes a series of short sections to the historical uses of glass for chemical and medical purposes, for windows, mirrors and stained-glass windows, and for spectacles, telescopes and microscopes, leading up to the development of optical glass and the invention of many new optical instruments in the last century. An attempt is made throughout the article to show how the manufacture and uses of glass have gradually been placed on a more scientific foundation. Finally, the mechanization of bottle and lamp-bulb production is discussed. Only very brief mention is made of the many and varied new applications of glass in the last fifty years.

Physician with spectacles and urine glass. Part of a woodcut from the Heidelberg Dance of Death, late 15th century.

# NEW LIGHT ON THE STRUCTURE OF GLASS

by J. M. STEVELS.

666.11.01:539.213.1

## Introduction: Zachariasen's theory

The present-day conception of the structure of glass differs in several respects from that presented in Zachariasen's now classical work [1]. In order to explain certain physical properties of glass, the theory of the vitreous state has been modified and refined in various points, and in this article we shall deal with these changes under three main headings.

In each part we shall examine a separate aspect of glass structure. The first will be concerned with the coherence of the atoms in glass, and one of our conclusions is that the vitreous state can occur under other conditions than were formerly held to be possible. In the second part it is shown that glass is not the homogeneous and purely amorphous substance it was long thought to be. The last part deals with some new aspects of the structure of glass which stem from a concept of major importance in the physics of the crystalline state but which seems strange in relation to glass, namely the concept of lattice imperfections.

The better insight gained into the structure of glass has made many new applications possible, and has also led to he development of a category of glasses differing quite essentially from conventional types. Before considering the results of glass research in recent years, we shall briefly recapitulate Zachariasen's theory.

According to Zachariasen a substance in the vitreous state is built up of a random three-dimensional network which, in normal glasses composed of inorganic oxides, is constructed from polyhedra (tetrahedra or triangles) of oxygen ions. As a rule, the centres of the polyhedra are occupied by multiply-charged ions such as $Si^{4+}$, $B^{3+}$ or $P^{5+}$ ions. Since these form the network together with the oxygen ions, they are called *network formers*. The oxygen ions are of two kinds, known as *bridging oxygen ions*, each of which belongs to *two* polyhedra, and *non-bridging oxygen ions*, each of which belongs to only *one* polyhedron. The degree of cohesion between the polyhedra, and hence of the network as a whole, evidently depends on the percentage of bridging oxygen ions. The excess negative charge of the network is compensated by *network-modifying* ions, located in the interstices of the network. These are generally large metal ions of low positive charge, such as $Na^+$, $K^+$ or $Ca^{++}$ ions. In conventional glasses they have little influence on the network compared with the network formers. Nevertheless, as their name suggests, they are not entirely to be disregarded.

A two-dimensional representation of the structure of an inorganic oxide glass in accordance with Zachariasen's theory is given in *fig. 1*.

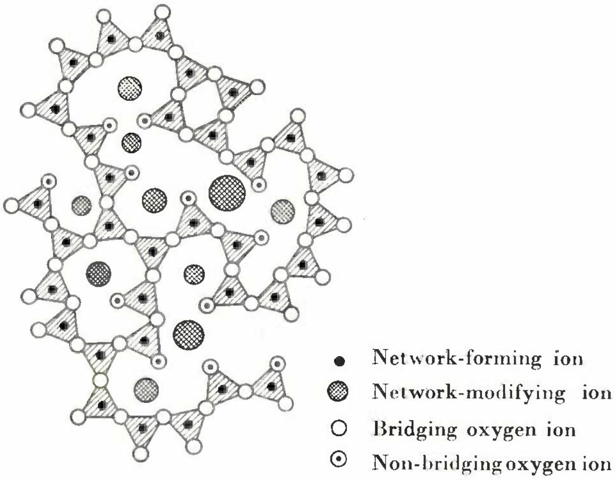


Fig. 1. Schematic two-dimensional representation of glass, according to Zachariasen's theory [1].

## The structure parameter *Y* and its influence on the properties of glass

To understand various physical properties of glass it is useful to introduce the following four quantities:

$X$ = average number of non-bridging oxygen ions per polyhedron,

$Y$ = average number of bridging oxygen ions per polyhedron,

$Z$ = average total number of oxygen ions per polyhedron,

$R$ = ratio of total number of oxygen ions to total number of network formers.

[1] Zachariasen's theory and various hypotheses from the years 1940 to 1950 are discussed by J. M. Stevels in: The vitreous state, Philips tech. Rev. **8**, 231-237, 1946, and in: The structure of glass, Philips tech. Rev. **13**, 293-300, 1951/52.

Between these quantities two simple relations exist, which are easily found by counting the ions:

$$X + Y = Z, \quad \ldots\ldots \quad (1)$$

$$X + \tfrac{1}{2}Y = R, \quad \ldots\ldots \quad (2)$$

or

$$X = 2R - Z, \quad \ldots\ldots \quad (3)$$

$$Y = 2Z - 2R. \quad \ldots\ldots \quad (4)$$

As a rule the total number of oxygen ions per polyhedron $Z$ is known ($Z = 4$ in phosphate and silicate glasses), and $R$ can usually be calculated from the composition; it is therefore generally a simple matter to determine $X$ and $Y$.

Given $Z = 4$ and $R = 2$ for quartz glass (i.e. fused silica, $SiO_2$), it follows that $X = 0$ and $Y = 4$. With $R = 3$ and $Z = 4$ for glass of formula $Na_2O.SiO_2$, we find $X = 2$ and $Y = 2$. Likewise, for glass having the composition $15\% Na_2O . 85\% SiO_2$ ($R = 1.85/0.85 = 2.175$ and $Z = 4$) we calculate $X = 0.35$ and $Y = 3.65$.

The situation is not always as simple as this. Some glasses contain ions which do not appear as typical network formers or network modifiers, but which occur both *in and between* the oxygen tetrahedra (or triangles). Where the "equilibrium" lies in such a case depends on the composition of the glass and on the conditions under which it was formed. It will be evident that the terms network former and network modifier are only appropriate in such a situation if the "equilibrium" lies in practically all circumstances completely to one side. When this is not so, the ions concerned, e.g. $Co^{++}$, $Ni^{++}$, $Pb^{++}$ and in some cases $Ca^{++}$ and $Ba^{++}$ are referred to as "intermediates".

Where intermediates are present it is not generally possible to determine $R$ exactly, since the situation of the "equilibrium" is not usualy known. The practice is then to count the intermediates together with the network-modifying ions in the above formulae. Of course this means that the values of $Y$ thus calculated ($Y_c$) will differ from the true values ($Y_r$), and it is not difficult to see that $Y_c$ will turn out smaller than $Y_r$. An example of this will be given below.

Knowledge of the value of $Y$, the average number of bridging oxygen ions, is of considerable importance in view of the surprising fact that numerous properties of glass depend primarily on $Y$. Generally speaking, the spatial coherence of the network is less the smaller the value of $Y$: the structure becomes looser and this is accompanied by the appearance of larger interstices. As a result the network modifiers will be able to move more readily, both when oscillating in their own site and when jumping from one site to another through the meshes of the network. Thus, as $Y$ decreases, we find an increasing coefficient of thermal expansion, an increasing electrical conductivity and a decreasing viscosity, and analogous changes in various relaxation phenomena [2]).

The influence of $Y$ on various properties of glass appears from *Table I*. For each pair, the compositions of each of the vitreous systems mentioned are entirely different chemically, yet they have the same values of $Y$, and consequently almost the same physical properties.

Table I. Illustrating the influence of $Y$, the average number of bridging oxygen ions per polyhedron, on the "melting temperature" (temperature at which the viscosity reaches the value $10^2$ poise) and on the expansion coefficient $\alpha$ of various glasses.

| Composition | $Y$ | "Melting temperature" °C | $\alpha \times 10^7$ |
|---|---|---|---|
| $Na_2O.2SiO_2$ | 3 | 1250 | 146 |
| $P_2O_5$ | 3 | 1300 | 140 |
| $Na_2O.SiO_2$ | 2 | 1050 | 220 |
| $Na_2O.P_2O_5$ | 2 | 1100 | 220 |

Increasing the metal-oxide content of the glasses generally results in a lower $Y$, since the fusion of metal oxides and quartz glass gives rise to an Si-O network which differs from the quartz glass network in that at various places a bridging oxygen ion is replaced by two non-brigding oxygen ions. The extra oxygen needed for this substitution is supplied by the metal oxide. Plainly, then, more non-bridging oxygen ions will appear, and $Y$ will become smaller the more metal oxide is used in forming the glass. This knowledge makes it possible — within certain limits — to give a glass the properties required for a particular purpose.

### *The boron anomaly*

A higher metal-oxide content in glass does not *always* result in a lower $Y$. Pure borate glasses are a case in point. When not too large amounts of metal oxide and $B_2O_3$ are fused together, the oxygen of the metal oxide is not taken up as a non-bridging ion. This oxygen is taken into the network by the conversion of oxygen triangles into oxygen tetrahedra, which consist entirely of bridging oxygen ions.

This implies that raising the metal-oxide content of these glasses will entail an increase in $Y$, and thus strengthen the network. Correspondingly, various physical properties change in a direction exactly opposite to that in comparable silicate glasses under the same conditions. This is known as the "boron anomaly". It is customary to refer to the range of

[2]) The latter are discussed by J. M. Stevels in: Dielectric losses in glass, Philips tech. Rev. **13**, 360-370, 1951/52.

compositions in which this effect occurs in borate glasses as the *accumulation region.*

The conversion, just mentioned, of oxygen triangles into oxygen tetrahedra is found to obey certain rules, which set a limit to the "accumulation" [3]. These rules are:

a) Each triangle must be associated with no more than one tetrahedron;

b) A tetrahedron must not be associated with another tetrahedron.

This means that there must be at least four triangles present per tetrahedron.

If a borate glass contains more metal oxide than is consonant with this condition, non-bridging oxygen ions will be formed, just as in the case of silicate glasses: as a result, the value of $Y$ will *decrease* as the metal oxide content increases. The range of concentrations where this occurs in borate glasses is known as the *destruction region.*

The effects described are well demonstrated by *fig. 2*: as the metal-oxide content increases, the coefficient of expansion $\alpha$ initially falls and then starts to rise again at a composition which exactly coincides with the point where the $Y$ curve turns over.

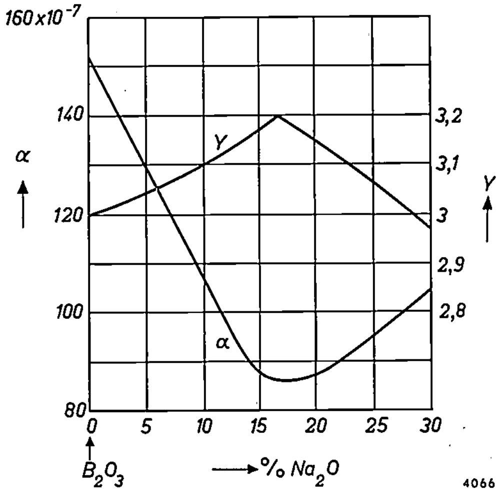


Fig. 2. Relation between $Y$, the average number of bridging oxygen ions per polyhedron, and the coefficient of expansion $\alpha$ of pure sodium borate glasses. Both quantities are plotted versus composition in mole % $Na_2O$. The structure with the largest value of $Y$ has the smallest expansion coefficient.

The agreement between theory and practice in this connection will be made clearer by a few simple calculations. The theory leads us to expect the transition from the accumulation to the destruction region to take place at $Y = Z = 3.2$ and $X = 0$, since the network in the boundary case consists of polyhedra with three and four oxygen ions in the ratio of 4 : 1, all of them bridging ions. For borate glasses the value $Y = 3.2$, according to the theory, indicates the maximum average number of bridging oxygen ions, and at this value the structure will thus be strongest. From eq. (2) - (4) it follows that the corresponding value of $R$ is equal to 1.60 which, expressed in terms of chemical composition, corresponds to $Na_2O.5B_2O_3$, or 16.7%$Na_2O$ . 83.3% $B_2O_3$ (mole percentages). Fig. 2 shows that the minimum of the expansion coefficient does in fact occur at exactly this value.

## *Changes where* $Y = 3$

It is interesting to examine what happens when $Y$ is equal to 3. It is understandable that when this value is reached — here, for *all* inorganic oxide glasses — marked changes may be expected, since polyhedra will now occur which are bound to their surroundings by only two points of contact. When this value is exceeded there will thus be a rather abrupt change in the stiffness of the network, and hence in the mobility of the network modifiers. The effects are apparent in various physical properties, as can be seen in *figs. 3* and *4*. These figures show the electrical conductivity of sodium silicate glasses and the Vickers hardness [4]) of sodium borate glasses as a function of composition (and of $Y$). In the silicate glasses the value $Y = 3$ corresponds to the chemical formula $Na_2O.2SiO_2$ (33.3%$Na_2O$, 66.7% $SiO_2$), in the borate glasses to $2Na_2O.5B_2O_3$ (28.6% $Na_2O$, 71.4%$B_2O_3$). In both figures the curves show kinks at these compositions. Fig. 4 also shows the effect of the transition from accumulation to destruction region at 16.7% $Na_2O$.

## *Invert glasses* [5])

The changes that occur when $Y$ is *smaller than 2* are particularly important. The feature of the structures then produced is that the tetrahedrons have at the most two points of contact with their environment. In other words, the spatial coherence is lost, and the structure is built up from *chains.*

Disregarding the possibility of ring formation and branching, the chains are infinitely long in the case of $Y = 2$. A smaller $Y$ implies a shorter chain. The average "chain length" $\bar{n}$, that is to say the average number of tetrahedra per chain, is given by:

$$\bar{n} = \frac{2}{2-Y}, \quad \text{or} \quad Y = 2 - \frac{2}{\bar{n}} . \quad . \quad . \quad (5)$$

[3]) T. Abe, J. Amer. Ceram. Soc. **35**, 284, 1952.

[4]) On the measurement of the Vickers hardness of materials, see e.g. E. M. H. Lips, Philips tech. Rev. **2**, 179, 1937.

[5]) For a detailed study of invert glasses, see H. J. L. Trap and J. M. Stevels, Glastechn. Ber. **32 K**, VI/31, 1959.

$Y$ and $\bar{n}$ can also be expressed in terms of chemical composition. For example, given a silicate glass of composition $(100-p)\mathrm{M_2O}.p\mathrm{SiO_2}$, where $p$ represents the mole percentage and M is a general symbol for a monovalent network-modifying ion, then:

$$Y = 6 - \frac{200}{p}, \quad \ldots\ldots \quad (6)$$

or

$$\bar{n} = \frac{p}{100-2p}. \quad \ldots\ldots \quad (7)$$

It follows from these formulae that $Y$ becomes smaller than 2 at less then 50 mole % $\mathrm{SiO_2}$ and moreover that $\bar{n}$ decreases very rapidly at percentages lower than this. At $p = 48$, for example, $\bar{n} = 12$; at $p = 45$, $\bar{n} = 4.5$; and at $p = 40$, $\bar{n}$ is as low as 2. It is seen, then, that the properties of these glasses depend largely on composition — much more so than in the case of the conventional glasses earlier discussed.

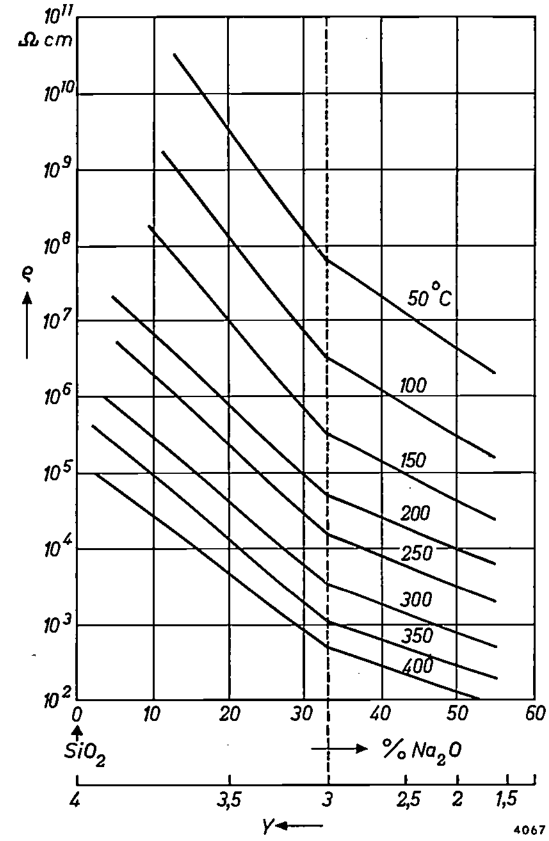


Fig. 3. Electrical conductivity $\varrho$ of sodium silicate glasses at various temperatures as a function of composition (in mole % $\mathrm{Na_2O}$) and of $Y$. A value $Y < 3$ corresponds to the occurrence of tetrahedra bound to the rest of the network by only two bridging oxygen ions. At the value $Y = 3$ there occurs a discontinuous change in the "stiffness" of the network and hence also of the mobility of the network-modifying ions. This explains the kink in the curves. (After E. Seddon, E. J. Tippett and W. E. S. Turner, J. Soc. Glass Technol. 16, 450, 1932.)

In practice it is very difficult to produce glasses in which $Y < 2$. In a system like $\mathrm{Na_2O}$-$\mathrm{SiO_2}$, for instance, vitrification ceases entirely if the molar content in $\mathrm{Na_2O}$ exceeds 50%, the value that corresponds to $Y = 2$. This is bound up with the disintegration of the structure into chains and with the fact that the system contains only one kind of metal ion. Both these circumstances enable the metal ions together with the tetrahedron chains to assume an ordered arrangement when the melt solidifies, in other words, crystallization occurs.

This crystallization can be prevented by resorting to an artifice: if *two* kinds of metal ions are used, the vitreous region can be shifted to smaller values of $Y$, the idea being that crystallization is made more difficult when metal ions of *different size and charge* are present. By taking this principle far enough, e.g. by introducing four or five different metal ions into the same system, very low values of $Y$ can be achieved [6]).

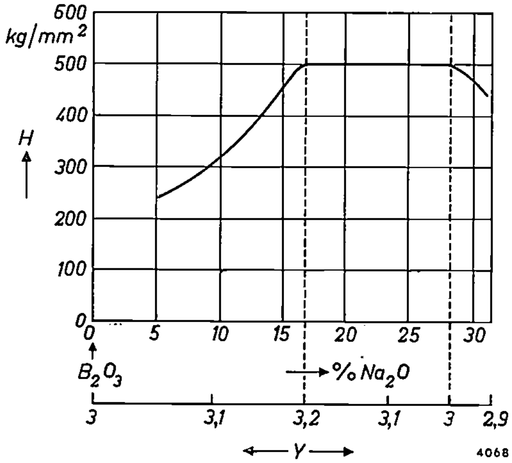


Fig. 4. The Vickers hardness $H$ of sodium borate glasses as a function of composition in mole % $\mathrm{Na_2O}$. Both the reversal point at $Y = 3.2$ (16.7% $\mathrm{Na_2O}$) and the effect at $Y = 3$ (28.6% $\mathrm{Na_2O}$) are perceptible in the curve. Compare also figs. 2 and 3. (After F. C. Eversteijn, J. M. Stevels and H. I. Waterman, Phys. and Chem. Glasses 1, 134, 1960, No. 4.)

The schematic representation of a glass in the region of $Y < 2$ is given in *fig. 5*. The average chain length is here 3.5, corresponding to $Y = 1\frac{3}{7}$.

Glasses in which $Y < 2$ have been called *invert glasses*, for two reasons. The first is that, compared with conventional glasses, their structure is in fact "inverted". The structural cohesion of conventional glasses is mainly due to the Si-O network, and the network modifiers play a subordinate part. In invert glasses exactly the opposite is the case. They con-

[6]) This has not been investigated in the case of borate glasses.

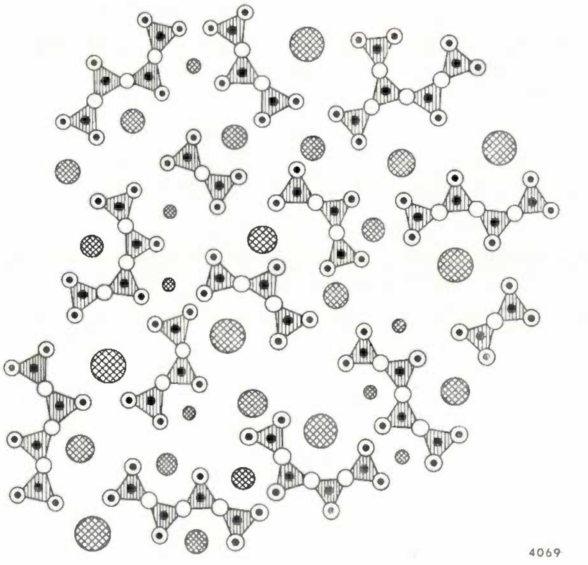


Fig. 5. Schematic two-dimensional representation of the invert glass structure. Compare with fig. 1, which explains the various circles. Instead of a coherent conglomerate of oxygen polyhedra with metal ions in the interstices, as in fig. 1, the oxygen polyhedra here form "islands" in a "sea" of metal ions. Such glasses can only be made by introducing metal ions of different size and charge to counteract crystallization.

tain only short chains of polyhedra which are embedded, as it were, in a large quantity of metal ions, and it is the forces between these metal ions and the oxygen ions of the chains that primarily determine the cohesion of the substance.

The second reason for the name "invert glasses" is that certain physical properties also exhibit an "inversion". We have seen that various properties of glasses change with decreasing $Y$ in a way that corresponds to a decrease in the coherence of the structure. In the invert glass region the influence of the (short) chains of polyhedra is slight compared with that of the metal ions. The more the latter gain the ascendency as $Y$ decreases, the stronger the structure becomes, and certain physical properties undergo a corresponding change, i.e. in the opposite direction to that in conventional glasses. Of course, the properties in question are those which are primarily determined by the coherence in the structure, as for exampl: the expansion coefficient, the viscosity and the dielectric losses.

As an example of this inversion, *figs. 6* and *7* show the viscosity (at various temperatures) and the dielectric loss angle (for various mole fractions of the metal oxides) as functions of $Y$.

In a glass containing numerous intermediates the inversion occurs at a calculated value $Y_c$ which is *smaller than 2* (*fig. 8*), in agreement with the discussion in small print on page 301. In practice, this difference between the observed and theoretical values at which inversion occurs is often used for estimating the distribution of the intermediates into network formers and network modifiers.

Invert glasses have their practical as well as their theoretical importance. For example, in glass capacitors it is sometimes required that the capacitance suffers little change during switching; for this purpose glasses with a small temperature coefficient of the dielectric constant are required. Certain invert glasses are able to meet this requirement vęry satisfactorily in that their high dielectric constant and low dielectric losses are properties that favour the low temperature coefficient required.

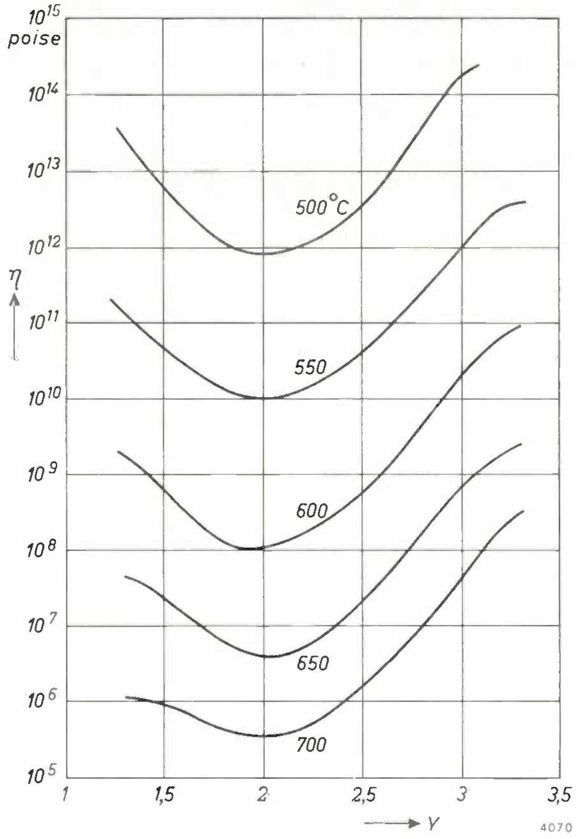


Fig. 6. Viscosity of $Na_2O$-$K_2O$-CaO-SrO-BaO silicate glasses, at various temperatures, as a function of $Y$ (equimolar quantities of the various metal oxides). The change in sign of the slope of the curves at $Y = 2$ is related to the fact that at $Y > 2$ the coherence of the structure is primarily determined by the Si-O network and at $Y < 2$ by the interaction between the metal ions and the oxygen ions of the chains.

For an explanation of this we refer the reader to an earlier article in this journal [7]). The effect of the structure inversion on the two properties mentioned can be seen in fig. 7 and *fig. 9*: the dielectric loss angle becomes smaller, whereas the dielectric constant increases monotonically. The latter is evidently not affected by the structure inversion, as this property does not depend on the coherence

[7]) See e.g. M. Gevers and F. K. du Pré, Philips tech. Rev. **9**, 91, 1947/48.

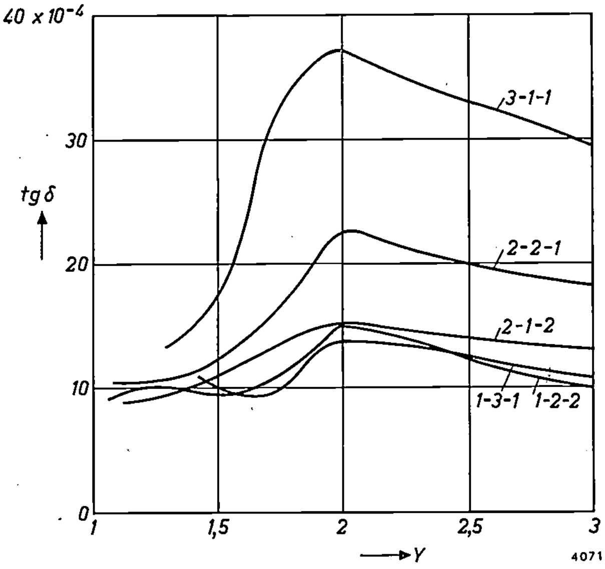


Fig. 7. The dielectric loss angle tan $\delta$ of $K_2O$-CaO-SrO silicate glasses as a function of $Y$, measured at 1.5 Mc/s. The three figures on each curve give the ratio of the metal oxides in the sequence $K_2O$-CaO-SrO. Inversion occurs at $Y = 2$.

of the structure but is solely the sum of atomic properties (this also applies, for example, to the density and the refractive index, which likewise show no inversion).

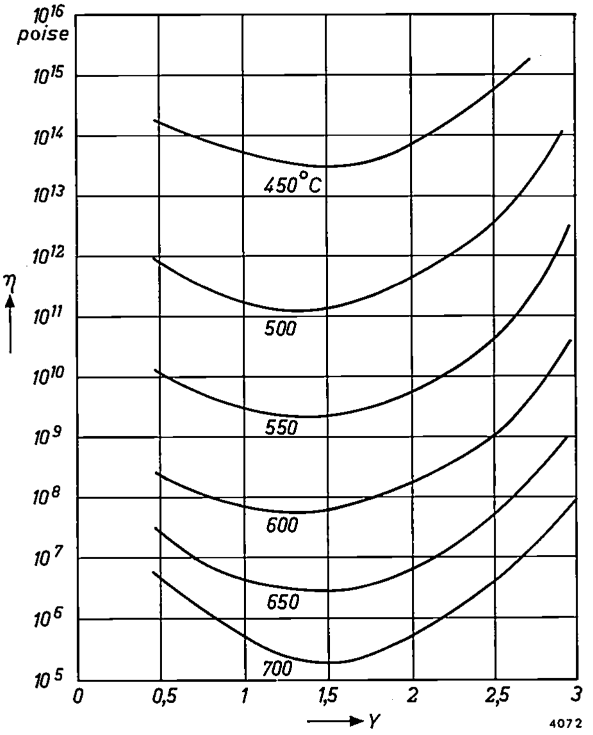


Fig. 8. Viscosity of $Na_2$-$K_2O$-MgO-CdO-ZnO silicate glasses, at various temperatures, as a function of $Y$ (equimolar quantities of the various metal oxides). Owing to the presence of intermediates ($Mg^{++}$, $Cd^{++}$, and $Zn^{++}$ ions) the calculated value $Y_c$, plotted here, is not the same as the actual value $Y_r$ (which is unknown) and will be smaller. Hence the fact that the slopes of curves reverse their sign at a lower value than $Y = 2$.

To conclude this section it may be remarked that the development of invert glasses constitutes in a certain sense a refutation of the classical theories concerning the structure of glass. The view formerly held was that a random three-dimensional network was essential to the vitreous state. We have seen, however, that such a network need not be present at all, *provided the metal ions show a sufficient variation in size and charge*. It appears, then, that the condition for realizing a vitreous state must be formulated somewhat more widely: whereas in conventional glasses it is the disorder of the network that promotes the vitreous state, in the invert glasses the vitreous properties are due to the disorder of the metal ions.

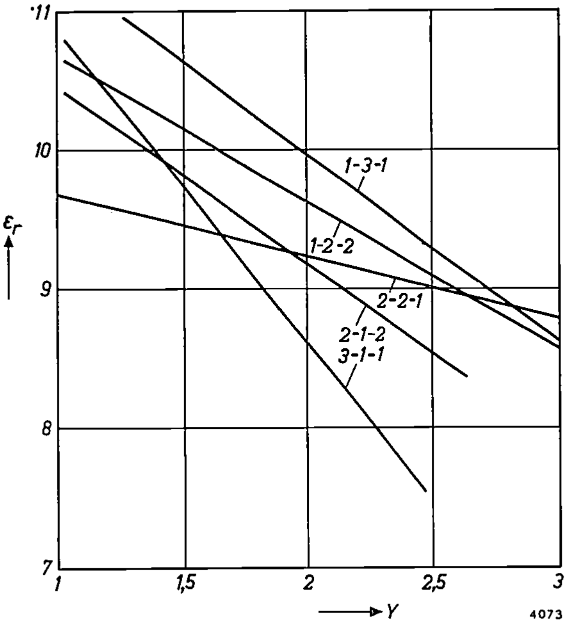


Fig. 9. Relative dielectric constant $\varepsilon_r$ of $K_2O$-CaO-SrO silicate glasses as a function of $Y$, measured at 1.5 Mc/s. The mole fractions of the metal oxides are indicated in the same way as in fig. 7. No inversion occurs. This is due to the fact that the dielectric constant is not governed by the coherence of the structure, but solely by atomic properties.

### Glass built up from microscopic domains of different structure and composition

In recent years experimental indications have been forthcoming to show that glass, which appears homogeneous to the eye, is in reality built-up from microscopic domains, differing one from the other and varying in size from 0.01 to 0.1 μ. How these domains come about is not known with certainty. It is thought that they are formed in the liquid state, when certain phases are separated. In certain glasses these domains can be observed with an electron microscope; in others heat treatment is necessary to separate the phases more distinctly (see *fig. 10a, b*).

It is probable — at least in many cases — that these domains differ in their chemical composition, but whatever view is held of their nature, their existence means at all events that the original concept of a completely random network extended throughout the entire mass of the glass cannot possibly be correct. In so far as such a network exists it can only extend over short distances [8]).

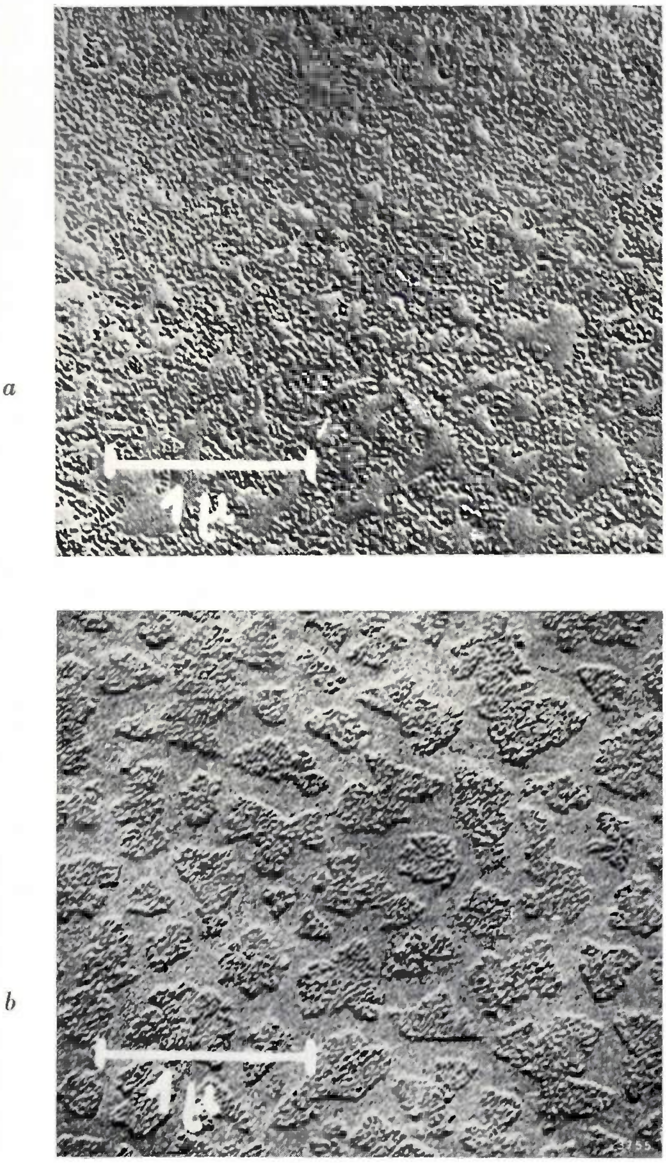


Fig. 10. Electron micrographs of sodium borate glass which is optically perfectly clear. *a*) glass that has received no after-treatment, *b*) glass heat-treated for $2\frac{1}{2}$ hours at 500 °C. (Taken from: W. Skatulla, W. Vogel and H. Wessel, Silikattechnik **9**, 51, 1958.)

In this connection reference may also be made to refinements of the theory proposed by Russian investigators [9]). In their view too, the structure does not show the degree of randomness suggested by Zachariasen's theory. Special grounds for this conclusion were provided by attempts to interpret the results of experiments with infrared absorption and small-angle X-ray scattering. They showed that it is necessary to assume the existence of domains with a more ordered structure side by side with domains having a "vitreous" structure. These "microcrystalline" domains, which are supposed to have the character of strongly deformed crystal lattices, are referred to as "crystallites". They are thought to merge one into the other *via* domains of gradually decreasing order. Variations in chemical composition might correspond to these transitions. The situation described is represented schematically in *fig. 11a, b*, after Porai-Koshits [9]).

As a commentary on fig. 11*b* we give as our opinion that the various domains will not have such divergent compositions as is suggested. It is very unlikely that a glass will contain domains side by side, the one built up of tetrahedra with four bridging oxygen ions ($SiO_2$) and the other of tetrahedrons with two bridging oxygen ions and two two non-bridging oxygen ions ($Na_2O.SiO_2$). If this were so, one could certainly not expect discontinuous changes in particular properties (examples of which have been given in the previous section) to occur at very specific values of $Y$ corresponding to a simple physical model. On the other hand it is true that $Y$ represents an average value, the material being after all a glass. In a glass with $Y = 3$, for example, there will not only be tetrahedra with three bridging oxygen ions, but also tetrahedra with four or two bridging ions. For $Y > 3$ there will undoubtedly be some tetrahedra with two bridging oxygen ions, and, conversely, for $Y < 3$ there will be some tetrahedra with four bridging oxygen ions. Their number, however, will be relatively small. It may be said then the glass will be found to contain domains of different character, as sketched in the above figures and manifested for example, in the phase-separation phenomena discussed. These domains, however, will differ only slightly in composition.

## Network imperfections

After what has been said it will be clear that various concepts concerned with the crystalline state are evidently also applicable to the vitreous state.

[8]) A rough calculation shows that a spherical domain of 0.1 μ diameter and of average vitreous composition contains only $10^6$ to $10^7$ tetrahedra.

[9]) E. A. Porai-Koshits, Glastechn. Ber. **32**, 450, 1959.

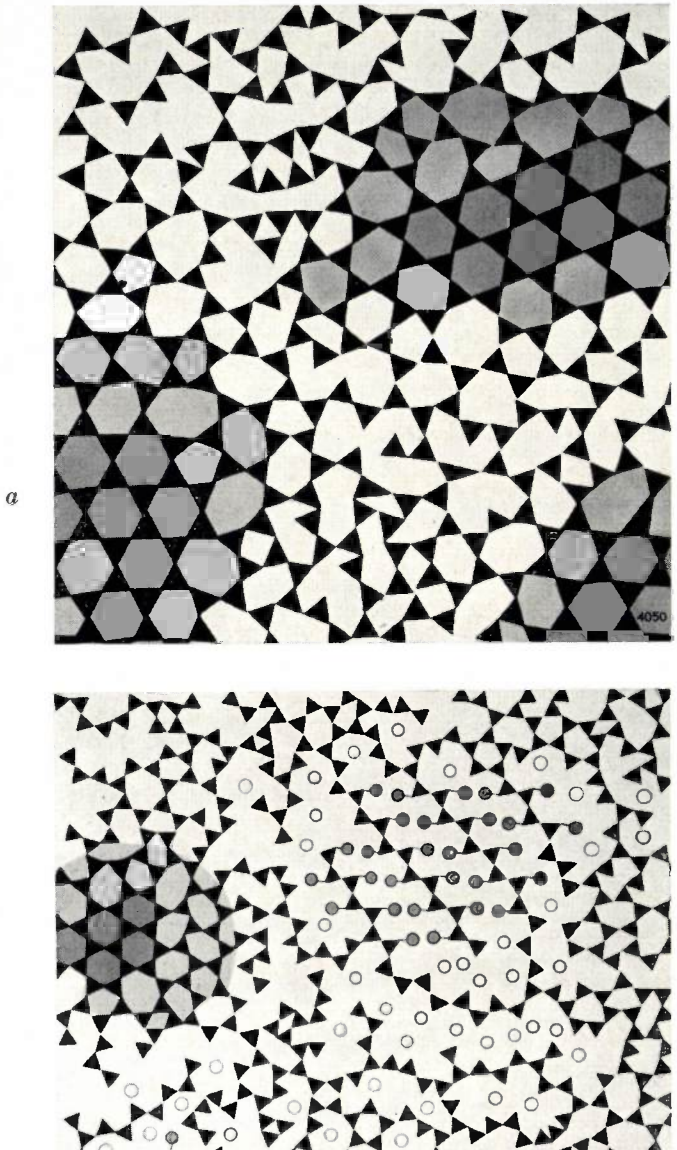


Fig. 11. Two-dimensional representation of the structure of *a*) vitreous $SiO_2$, *b*) glass of composition $Na_2O.2SiO_2$. (After E. A. Porai-Koshits, Glastechn. Ber. **32**, 450, 1959.) The crystalline domains (crystallites) are clearly recognizable. The degree of crystalline ordering is indicated by light and dark shading. In the view of the present writer, the interstices in (*a*) are in reality not so irregular as represented here. Oberlies and Dietzel have shown by X-ray diffraction measurements that the network of vitreous $SiO_2$ is largely built-up of six-ring conglomerates of tetrahedra [10]). In (*b*) we see domains of crystalline $SiO_2$ which, *via* vitreous $SiO_2$ and a vitreous region which also contains $Na^+$ ions, gradually merge into two regions whose structure closely resembles that of crystalline $Na_2O.SiO_2$ (regions with dark circles).

In the last twenty years, it has been realized that all crystalline substances show lattice imperfections. Apart from two-dimensional imperfections, such as grain boundaries, there are also linear discontinuities (dislocations) and "point defects". We shall deal at greater length with the latter, in view of their particular importance in glass. In any given lattice, atoms may be missing here and there (*vacancies*). Atoms may be found at sites where, according to the arrangement of the lattice, they should not properly be situated (*interstitial atoms*). Lattice sites may be occupied by atoms alien to the lattice (*foreign atoms*). Finally, these three types of imperfection may occur with a positive or negative charge (*F centres*, *V centres*, etc.).

The above concepts have become a commonplace to investigators of the crystalline state, and have made it possible to gain a better insight into the mechanism of numerous physical phenomena. In the case of solids of simple structure, such as halides and chalkenides, quantitative descriptions of such phenomena have in fact been given with the aid of these imperfections [11]).

In spite of the complicated and irregular nature of vitreous systems, fruitful use has been made of several of these concepts in connection with the vitreous state.

### *Point defects in the vitreous network*

As stated, point defects in particular are of importance in the vitreous network [12]). Although more or less the same classification has been maintained as for crystals, the absence of regularity in the network has made it necessary to introduce a modified convention with somewhat less sharp definitions.

*Fig. 12*, top left, represents a part of the Si-O network of quartz glass. The following three kinds of modifications may be introduced in this network.

An oxygen ion may be missing: we then speak of a vacancy. An oxygen ion may be added, which is then called an interstitial ion. (Strictly speaking, the term interstitial is out of place in relation to glass; follow-

[10]) F. Oberlies and A. Dietzel, Glastechn. Ber. **30**, 37, 1957.

[11]) See e.g. G. W. Rathenau, Imperfections in matter, Philips tech. Rev. **15**, 105-113, 1953/54; H. G. van Bueren, Lattice imperfections and plastic deformation in metals I, Philips tech. Rev. **15**, 246-257, 1953/54; Y. Haven, Lattice imperfections in crystals, studied on alkali halides, Philips tech. Rev. **20**, 69-79, 1958/59.

[12]) Linear discontinuities are hardly to be expected in glasses in view of their disordered structure. As regards superficial discontinuities, a glance at fig. 10 shows that these do exist in glass. Little is known about them as yet, but it is certain that their closer investigation will be of considerable importance to the further development of the technology and applications of glass.

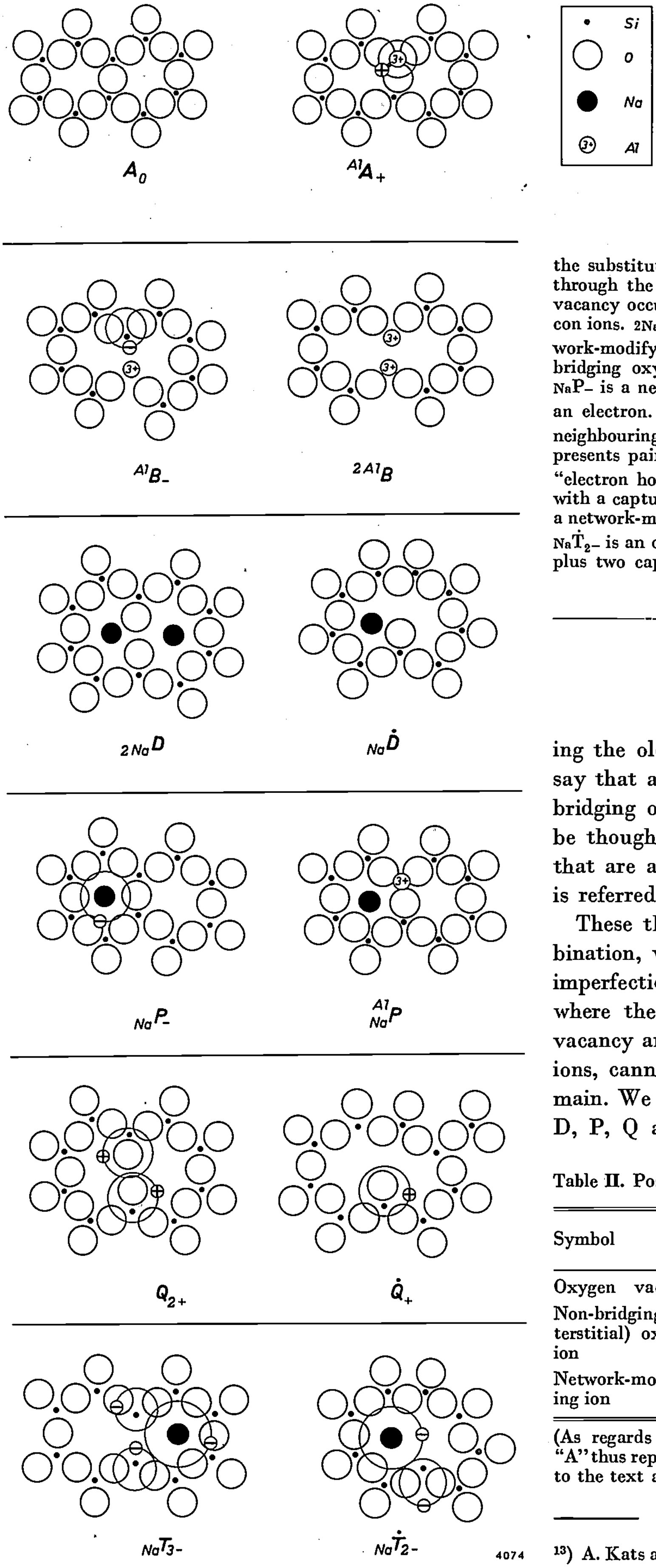


Fig. 12. Schematic representation of a "perfect" Si-O network and of various possible network imperfections. The symbol $A_0$ is assigned to the perfect network. All network imperfections given here are electrically neutral. This need not necessarily be so, of course, but it is usually the case. $^{Al}A_+$ is an imperfection in which an $Si^{4+}$ ion is replaced by an $Al^{3+}$ ion plus a captured hole. $^{Al}B_-$ is an oxygen vacancy with a neighbouring aluminium ion substituted for a silicon ion, and a captured electron. (The captured electron and the charge gained due to the substitution of $Al^{3+}$ for $Si^{4+}$ compensate the charge lost through the absence of the oxygen ion.) $^{2Al}B$ is an oxygen vacancy occupied by two aluminium ions substituted for silicon ions. $_{2Na}D$ are two paired non-bridging ions and two network-modifying sodium ions. $_{Na}\dot{D}$ represents an unpaired non-bridging oxygen ion and a network-modifying sodium ion. $_{Na}P_-$ is a network-modifying sodium ion which has captured an electron. $^{Al}_{Na}P$ is a network-modifying sodium ion with a neighbouring aluminium ion substituted for silicon. $Q_{2+}$ represents paired non-bridging oxygen ions with two captured "electron holes". $\dot{Q}_+$ is an unpaired non-bridging oxygen ion with a captured "electron hole". $_{Na}T_{3-}$ is an oxygen vacancy, a network-modifying sodium ion plus three captured electrons. $_{Na}\dot{T}_{2-}$ is an oxygen vacancy, a network-modifying sodium ion plus two captured electrons.

ing the old terminology, it is also the practice to say that a bridging oxygen ion is replaced two non-bridging oxygen ions.) Finally, a metal ion may be thought to occupy one of the large interstices that are always present in the Si-O network: this is referred to as a foreign ion (network modifier).

These three possibilities may also occur in combination, which, in principle, yields eight types of imperfection. The two combinations, however, where the bridging oxygen ion is replaced by a vacancy and at the same time by two non-bridging ions, cannot of course occur, so that only six remain. We shall call these, rather arbitrarily, A, B, D, P, Q and T; see *Table II* [13]).

Table II. Possible imperfections in Si-O networks.

| Symbol | A | B | D | P | Q | T |
|---|---|---|---|---|---|---|
| Oxygen vacancy | — | × | — | — | — | × |
| Non-bridging (interstitial) oxygen ion | — | — | × | — | × | — |
| Network-modifying ion | — | — | × | × | — | × |

(As regards the three imperfections specified in the table, "A" thus represents a perfect network; but the reader is referred to the text and to fig. 12.)

[13]) A. Kats and J. M. Stevels, Philips Res. Repts. **11**, 115, 1956.

The imperfections designated by these group symbols are more closely specified by superscripts and suffices to the letters. Typical examples of network imperfections are shown in fig. 12, illustrating this closer specification of possible imperfections. The place top left of the symbol is reserved for a superscript which indicates the ion replacing the silicon ion (e.g. ${}^{\mathrm{Al}}\mathrm{A}_{+}$ and ${}^{2\mathrm{Al}}\mathrm{B}$). The subscript in front of the symbol denotes the nature of the network modifier (e.g. ${}_{2\mathrm{Na}}\mathrm{D}$ and ${}^{\mathrm{Al}}_{\mathrm{Na}}\mathrm{P}$). A superscript behind the symbol denotes the replacement of oxygen by another ion (e.g. substitution by F, not indicated in fig. 12). The subscript behind the symbol denotes the number of captured electrons or "electron holes" (e.g. ${}^{\mathrm{Al}}\mathrm{A}_{+}$, ${}^{\mathrm{Al}}\mathrm{B}_{-}$, ${}_{\mathrm{Na}}\mathrm{P}_{-}$, $\mathrm{Q}_{2+}$ and ${}_{\mathrm{Na}}\mathrm{T}_{3-}$).

Some network imperfections in fig. 12 require further explanation. In the case of vitreous silica in which a small amount of $\mathrm{Na_2O}$ is built in, the imperfections will be of the type ${}_{2\mathrm{Na}}\mathrm{D}$. Similarly, paired non-bridging oxygen ions will also be found in silicate glasses having a relatively low metal-oxide content. However, in glass networks with a relatively high concentration of metal oxides, which are much more loosely built, the non-bridging oxygen ions will be found *unpaired*. These cases are denoted by a point above the symbol. It need hardly be said that, in addition to $\dot{\mathrm{D}}$ centres (e.g. ${}_{\mathrm{Na}}\dot{\mathrm{D}}$), the corresponding $\dot{\mathrm{Q}}$ centres (e.g. $\dot{\mathrm{Q}}_{+}$) and $\dot{\mathrm{T}}$ centres (e.g. ${}_{\mathrm{Na}}\dot{\mathrm{T}}_{2-}$) may also occur.

### *Network imperfections and physical phenomena in glasses*

We shall now briefly outline the grounds on which the presence of certain centres may be inferred, after which we shall consider various examples of phenomena which are closely related to the existence of these centres.

Some idea concerning the centres likely to be encountered may be given by the results of a chemical analysis. In many cases, however, this is not enough, for the analysis provides no information on the manner in which the components or impurities occur, and the quantities involved in practice are often too minute to be detected by chemical analysis.

Various physical methods of measurement have therefore been used, and with their aid — and the suitable combination of results — it has proved possible to detect quite a number of centres. Measurements of optical absorption and of paramagnetic resonance have been of particular value in investigations of the vitreous state [14]).

Silicate glasses can often be described as an Si-O network containing $\dot{\mathrm{D}}$ and/or $\dot{\mathrm{D}}$ centres. As a rule, these cannot be demonstrated by the physical methods mentioned. The situation is quite different, however, if the glass is exposed to ultraviolet, X- or gamma radiation. In glasses containing numerous $\dot{\mathrm{D}}$ centres, these are the points where the radiation takes hold, as it were: an electron is released from the oxygen ion, the associted metal ion usually moves a certain distance and again captures an electron, so that ultimately a reaction takes place which can be described as follows:

$$ {}_{\mathrm{X}}\dot{\mathrm{D}} \rightarrow {}_{\mathrm{X}}\mathrm{P}_{-} + \dot{\mathrm{Q}}_{+}, $$

where X denotes, say, an alkali atom.

The resultant ${}_{\mathrm{X}}\mathrm{P}_{-}$ centres show optical absorption at a wavelength that depends on the nature of the alkali ion. For example, the ${}_{\mathrm{Li}}\mathrm{P}_{-}$ centre exhibits absorption at 4200 Å, the ${}_{\mathrm{Na}}\mathrm{P}_{-}$ centre at 4600 Å, and the ${}_{\mathrm{K}}\mathrm{P}_{-}$ centre at 4750 Å. In the cases considered, absorption has also been observed at about 3000 Å which is independent of the nature of the alkali ion and must be attributed to the $\dot{\mathrm{Q}}_{+}$ centre.

A similar picture has been obtained from paramagnetic resonance measurements. These are capable of demonstrating two kinds of centres, one of which must contain an unpaired electron and the other an unpaired electron-hole. The so-called *g*-factors thereby found for the first kind of centre were shown to be dependent on the nature of the alkali ion (1.960, 1.964, 1.966, 1.976 and 1.974 for ${}_{\mathrm{Li}}\mathrm{P}_{-}$, ${}_{\mathrm{Na}}\mathrm{P}_{-}$, ${}_{\mathrm{K}}\mathrm{P}_{-}$, ${}_{\mathrm{Rb}}\mathrm{P}_{-}$ and ${}_{\mathrm{Cs}}\mathrm{P}_{-}$, respectively). The *g*-factor of the centre with the electron hole, on the other hand, was found to be independent of the alkali ion (2.011) [15]).

For completeness it should be noted that, apart from releasing an electron, irradiation may also, through a secondary reaction, cause the release of an oxygen ion, which gives rise to a B centre. Where a P centre has been formed near this site, T centres may occur, detectable by absorption peaks in the region of 6200 Å.

Effects as here described are responsible for the discolouration of glass sometimes observed during and after irradiation. A case in point is the discolouration of the glass in X-ray tubes after a certain period of operation. Since the reactions that produce the centres responsible for such discolouration

[14]) On paramagnetic resonance, see e.g. J. S. van Wieringen, Philips tech. Rev. **19**, 301, 1957/58.

[15]) J. S. van Wieringen and A. Kats, Philips Res. Repts. **12**, 432, 1957.

are reversible, moderate heating is sufficient to remove the effect. At room temperature and below, however, the colour centres can be preserved for a considerable time.

Discolouration of this kind can often be very troublesome, and therefore means have been sought to counteract it. One successful method of dealing with the difficulty is to introduce small quantities of cerium oxide. The electrons liberated by irradiation in such a glass are captured by complexes of cerium ions and P centres. As a result, the absorption of light takes place at a lower wavelength than in the original glass, namely at <3500 Å, where the abosrption is no longer perceptible to the eye.

Examples of such "controlled compositions", which lend special properties to glass and thus make special applications possible, have become too numerous to mention since the introduction of the concept of network imperfections. Glasses can be "composed", for instance, which discolour strongly when exposed to certain radiation. Glasses of this kind can be used for measuring extremely small doses of X-rays or gamma rays, the discolouration produced in the glass being a measure of the intensity of the radiation investigated.

Special importance also attaches to glasses which are capable of almost completely absorbing shortwave radiation without discolouring. Such material is suitable, for example, for observation-windows exposed to nuclear radiation.

Another effect studied is the presence of OH groups in glass, as a result of traces of water originating from the raw materials or from the flames in the melting furnace. Scholze has investigated the network imperfections involved (in our terminology $_{\rm H}\dot{\rm D}$) by infrared absorption measurements [16]).

The OH groups in glass can occur in two different states. They may be "free", or they may be associated with a non-bridging oxygen ion and thus form a hydrogen bridge. Scholze showed that the absorption bands in the infra-red spectrum at 2.75-2.95 μ must be attributable to the first kind of OH group and that at 3.35-3.85 μ to the second kind [17]). The knowledge of this relationship affords a better understanding of various effects encountered in glass, as the following examples illustrate.

In the neighbourhood of a non-bridging oxygen ion there is always a metal ion to be found (to compensate the charge). The greater the electrical field of the ions the stronger will this metal ion be bound to the oxygen. It thereby enters into competition with the OH groups that form the hydrogen bridge. As the field-strength of the metal ion decreases (e.g. $Li^+ \rightarrow Na^+ \rightarrow K^+$) more OH groups of the second kind may thus be expected, the alkali-ion content remaining constant. It has in fact been found that the absorption at 3.35-3.85 μ does indeed increase under these conditions.

Another example. The aluminium ion, like the silicon ion, is a network former. When $SiO_2$ mixed with $Al_2O_3$ is used as a glass-forming oxide, a network is produced which is poorer in oxygen than if only $SiO_2$ had been used, there being fewer non-bridging oxygen ions (smaller X). If the molar fraction of $Al_2O_3$ is made equal to that of the alkali oxide it is even possible to produce a network *without* non-bridging ions. This is observable for example, from the complete absence of infra-red absorption of the second kind of OH groups (which are bound to non-bridging oxygen).

Finally, when the temperature is raised it may be inferred from the change in the infra-red absorption that the first kind of OH group increase in number at the expense of the second kind; this, again, is fully in accordance with the general picture outlined.

The knowledge of these phenomena has been turned to use in various ways. It is owing to the presence of OH groups that glass transmits infrared radiation well only at wavelengths shorter than 2.8 μ. For certain purposes it is desirable to widen the spectral range to be transmitted, for example where prisms are to be made for spectrometers that can also be used for the infrared part of the spectrum. Quartz glass is employed in such cases, and in order to obtain transmission to as far as possible in the infra-red efforts are made to expel all hydrogen from the glass. This can be done by heating the quartz glass for some time at 1000 °C in a CO atmosphere.

Another method of increasing the infrared transmission is to substitute deuterium for the hydrogen. This substitution is accompanied by a shift of the infrared absorption edge, which then occurs at wavelengths of roughly 3.7 μ instead of 2.8 μ.

From the foregoing it will be clear that, by making use of the many and various imperfections that can occur in the vitreous network, it is possible to give glass widely diverse properties. The ability to control these imperfections, in conjunction with other modern developments discussed earlier, enables the manufacturer to produce a versatile range

[16]) H. Scholze, Glastechn. Ber. **32**, 81, 142, 278, 314, 381, 421, 1959.

[17]) The significance of a band at 4.25 μ is not yet quite clear. Cf. R. V. Adams and R. W. Douglas, J. Soc. Glass Technol. **43**, 147, 1959.

of glasses to meet the multifarious requirements of science and technology.

Summary. In recent years a better insight has been gained into the structure of glass, through improvements and refinements of Zachariasen's theory. The relation between $Y$ (average number of bridging oxygen ions per polyhedron) and various physical properties helps to explain the "boron anomaly" and other peculiarities in the behaviour of the physical properties of glass with changing composition. If $Y$ is smaller than 2, i.e. if the network is no longer coherent, the system tends towards crystallization. This tendency is counteracted by incorporating metal ions of differing size and charge. This has led to an entirely new category of glasses, called invert glasses. A new development is the discovery that glass is built-up from small domains of varying composition and structure (domains in which various phases have segregated and "crystallites"). The theory of the so-called network imperfections provides an explanation of the discolouring of glass under short-wave irradiation, and of various other vitreous properties. Some practical applications of this theory are discussed.

# AUTOMATIC CONTROL IN GLASS MANUFACTURE

by P. M. CUPIDO [*]).

621-53-79:666.1.031.2

As in so many mass-producing industries, automatic control techniques have made their entry into the glass factory. In this article, after a brief survey of the manufacturing processes, we shall deal with the parameters that can be subjected to automatic control, and discuss the requirements to be met by the instrumentation.

The following considerations are primarily concerned with glass manufacture as it is known at Philips, where glass is produced in the form of bulbs, tubes and rods. In many respects, however, they also apply to the mass production of glass in other forms.

## Principles of glass manufacture

In glass manufacture a mixture of the raw materials (the batch) is heated in such a way that the temperature varies with time according to a specified programme. This time-temperature programme may be either discontinuous or continuous in nature.

An example of a *discontinuous* method is the melting of glass in pots. This is the oldest method of glass melting, and is still in use for glasses of special quality or for making relatively small quantities. The empty pots are placed in a furnace, which may be designed for only one pot at a time or for as many as 20. After the pot has been raised to a high temperature, the batch mixture is introduced. This consists of powdered oxides and carbonates together with waste glass (cullet). Because of the amount of air entrapped in the mixture, the thermal conductivity is low. It therefore takes a considerable time to heat. As the temperature rises, the melting reaction sets in, and this is accompanied by the generation of gases which partly disappear into the furnace atmosphere and are partly trapped as gas bubbles in the foaming, viscous mass. The higher the temperature is raised, the lower becomes the viscosity, enabling the gas bubbles to rise more readily to the surface and escape.

The displacement of the gas bubbles sets up a movement in the glass which effectively promotes the desired chemical reactions. When the latter have ceased, no further gas bubbles are generated. The gases still entrapped now gradually escape, and the glass becomes more homogeneous in composition. This part of the process is known as "fining". The glass is then allowed to cool until it has reached the viscosity required for further working.

The mechanical production of large quantities of glass imposes demands on the manufacturing process that cannot be met by pot melting. A process better adapted to the purpose is that whereby the melting is done in a *tank furnace.* Here the charge and the molten glass are in direct contact with the walls and floor of the furnace. The floor, together with the part of the walls, forms the melting tank (or "end"). It is an oblong trough built up from blocks of refractory material. The batch is fed in at one end (the "doghouse") and the molten glass withdrawn at the other. At the melting end the charge is subjected to the necessary time-temperature programme, which consists of creating an appropriate temperature gradient along the trough and carefully controlling the rate at which the glass traverses that gradient. As opposed to pot melting, this is a *continuous* process.

It would be ideal if all the glass issuing from the melting tank had been subjected to exactly identical conditions. This is not the case, however. Owing to

[*]) Glass Division, Eindhoven.

of glasses to meet the multifarious requirements of science and technology.

Summary. In recent years a better insight has been gained into the structure of glass, through improvements and refinements of Zachariasen's theory. The relation between $Y$ (average number of bridging oxygen ions per polyhedron) and various physical properties helps to explain the "boron anomaly" and other peculiarities in the behaviour of the physical properties of glass with changing composition. If $Y$ is smaller than 2, i.e. if the network is no longer coherent, the system tends towards crystallization. This tendency is counteracted by incorporating metal ions of differing size and charge. This has led to an entirely new category of glasses, called invert glasses. A new development is the discovery that glass is built-up from small domains of varying composition and structure (domains in which various phases have segregated and "crystallites"). The theory of the so-called network imperfections provides an explanation of the discolouring of glass under short-wave irradiation, and of various other vitreous properties. Some practical applications of this theory are discussed.

# AUTOMATIC CONTROL IN GLASS MANUFACTURE

by P. M. CUPIDO [*]).

621-53-79:666.1.031.2

As in so many mass-producing industries, automatic control techniques have made their entry into the glass factory. In this article, after a brief survey of the manufacturing processes, we shall deal with the parameters that can be subjected to automatic control, and discuss the requirements to be met by the instrumentation.

The following considerations are primarily concerned with glass manufacture as it is known at Philips, where glass is produced in the form of bulbs, tubes and rods. In many respects, however, they also apply to the mass production of glass in other forms.

## Principles of glass manufacture

In glass manufacture a mixture of the raw materials (the batch) is heated in such a way that the temperature varies with time according to a specified programme. This time-temperature programme may be either discontinuous or continuous in nature.

An example of a *discontinuous* method is the melting of glass in pots. This is the oldest method of glass melting, and is still in use for glasses of special quality or for making relatively small quantities. The empty pots are placed in a furnace, which may be designed for only one pot at a time or for as many as 20. After the pot has been raised to a high temperature, the batch mixture is introduced. This consists of powdered oxides and carbonates together with waste glass (cullet). Because of the amount of air entrapped in the mixture, the thermal conductivity is low. It therefore takes a considerable time to heat. As the temperature rises, the melting reaction sets in, and this is accompanied by the generation of gases which partly disappear into the furnace atmosphere and are partly trapped as gas bubbles in the foaming, viscous mass. The higher the temperature is raised, the lower becomes the viscosity, enabling the gas bubbles to rise more readily to the surface and escape.

The displacement of the gas bubbles sets up a movement in the glass which effectively promotes the desired chemical reactions. When the latter have ceased, no further gas bubbles are generated. The gases still entrapped now gradually escape, and the glass becomes more homogeneous in composition. This part of the process is known as "fining". The glass is then allowed to cool until it has reached the viscosity required for further working.

The mechanical production of large quantities of glass imposes demands on the manufacturing process that cannot be met by pot melting. A process better adapted to the purpose is that whereby the melting is done in a *tank furnace*. Here the charge and the molten glass are in direct contact with the walls and floor of the furnace. The floor, together with the part of the walls, forms the melting tank (or "end"). It is an oblong trough built up from blocks of refractory material. The batch is fed in at one end (the "doghouse") and the molten glass withdrawn at the other. At the melting end the charge is subjected to the necessary time-temperature programme, which consists of creating an appropriate temperature gradient along the trough and carefully controlling the rate at which the glass traverses that gradient. As opposed to pot melting, this is a *continuous* process.

It would be ideal if all the glass issuing from the melting tank had been subjected to exactly identical conditions. This is not the case, however. Owing to

[*]) Glass Division, Eindhoven.

the complex currents that arise, and because of unavoidable "blind spots", the duration of flow through the melting tank differs for the glass in different regions of the tank. Good glass can only be obtained if no part of the mass is heated for too short a time, i.e. the duration of flow must be above a certain minimum. A long duration of flow is equally undesirable, however, for economic reasons and also for reasons of quality: if the glass stays for a lengthy period in the melting tank the composition of the molten mass changes, due to the disparate evaporation of certain constituents and to the solution of wall material. The duration of flow, then, must lie within critical limits.

An unchanging flow pattern in the melting tank is therefore of particular importance, and the conditions governing that pattern must be kept carefully constant. These conditions are: the method of feeding-in the batch mixture, the withdrawal of the molten glass, and, above all, the heat transfer in the melting tank and the temperature at various places above and below the glass surface.

To give some idea of how these and other conditions can be controlled in a continuous process, we shall examine a tank furnace in somewhat more detail.

## *A regenerative tank furnace*

The heart of the furnace is the melting tank. This is built, without mortar, from flush-fitting blocks of refractory material. In the case we shall consider, the furnace walls contain a number of ports as inlets for the flames from oil burners. The distribution of the fuel over several burners allows accurate adjustment of the axial temperature gradient in the furnace. The oxygen required for combustion is drawn from the air entering the furnace through the burner ports. For reasons of fuel economy, the air is preheated in a heat exchanger, which derives its heat from the combustion gases passing to the chimney stack.

There are two preheating systems — recuperative and regenerative. The *recuperative* system is generally used only for relatively small furnaces, with a capacity of no more than twenty or thirty tons a day. In this system the air to be heated and the combustion gases flow through separate channels (the recuperator) divided by ceramic or metal partitions through which heat is continuously transferred.

For larger furnaces the *regenerative* system is more economical. The furnace is flanked on both sides by a box-like brick structure (*fig. 1*) with a refractory lining and containing an open-stacked structure of fire-bricks ("checkers"). This structure (the regenerator) has a high heat capacity and offers little resistance to the alternating flow of the hot combustion gases and the air for preheating. As explained in the caption to fig. 1, the burners at each side of the furnace are operated alternately. The air needed is preheated in the one regenerator, whilst the combustion gases, on their way to the chimney stack, flow through the other regenerator, which thereby accumulates heat. After, say, half an hour the burner flames are extinguished, the gas flows are reversed, and fuel is fed to the burners at the other side, and so on.

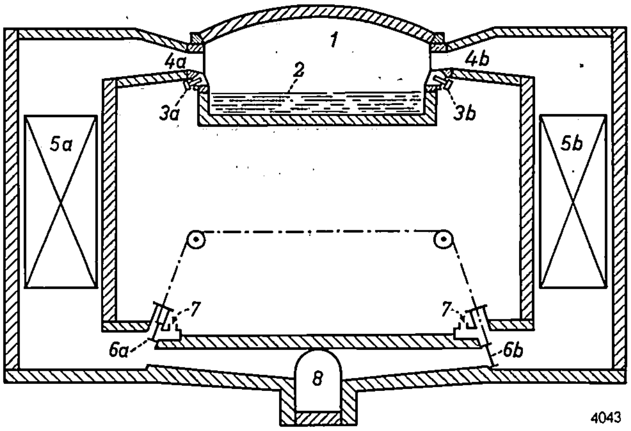


Fig. 1. Cross-section of a regenerative tank furnace. *1* furnace. *2* glass level. *3a*, *3b* oil burners. *4a*, *4b* burner ports. *5a*, *5b* regenerators. *6a*, *6b* reversal dampers. *7* combustion-air feed. *8* chimney flue.

With the dampers *6a* and *6b* as drawn, the burners *3b* are in operation, the combustion air being preheated in regenerator *5b* and the combustion gases delivering heat to regenerator *5a*; burners *3a* are extinguished. The situation is reversed every 20 to 30 minutes.

In the recuperative system the burners work continuously, there is no reversal and a temperature equilibrium is established in the heat exchanger; as a result, constant processing conditions are easy to achieve. By contrast, in the regenerative system the temperature of the heated air is continually changing, and each reversal severely disturbs the conditions prevailing in the furnace. In spite of this drawback — which can be reduced by shortening the reversal periods — the regenerative system is preferred for quantity production because, as we have said, it is more economical. The reason is the better heat transfer in the regenerative heat exchanger, in which the air is preheated to a temperature of 1000 or 1100 °C, i.e. 300 to 400 °C higher than in a recuperator.

*Further processing*

After leaving the melting tank the glass enters the working tank (or "end"), likewise a refractory structure but of smaller dimensions. (The terms "melting end" and "working end" date from the time when glass manufacture was not yet an automatic process; raw materials were always added at one end of the tank and the glass-blower worked at the other.) In automatic production — with which we are primarily concerned — the main function of the working tank is for letting the glass settle to a lower temperature. The glass flowing out of the melting tank has a temperature of 1400 or 1500 °C, which must be reduced to 1000 or 1100 °C for working. Part of this reduction, which must be uniform and thus calls for suitable time and space, is effected in the working tank.

From the working tank the glass flows into one or more *feeders*. These are long channels of refractory material, in which a proper combination of cooling and heating gives the glass a constant and uniform temperature. The feeders terminate at the glass-working machines. Depending on the product to be made, the glass is fed to the machine continuously or discontinuously. A machine that turns out glass tubing or rods receives a continuous supply, whereas a bulb-blowing machine (see the relevant article in this number) is fed with successive portions termed "gobs" which are cut off with special shears (*fig. 2*; see also figs. 3 and 4 on pp. 321 and 322).

## Parameters amenable to automatic control

Having outlined the production process in a regenerative tank furnace we shall now consider how the processing conditions can be controlled [1]).

*Automatic control of furnace temperature*

We have already mentioned the great importance of a constant heat transfer from the furnace atmosphere to the glass. To this end it is a first prerequisite to keep the temperature in the furnace and its distribution carefully constant. In the case of an oil burner the temperature is easy to control by placing a suitable sensing element, e.g. a thermocouple, near the flame and transmitting the signal to a valve-operated control device which regulates the rate of supply of oil. The operation of such a simple control system will now be examined, and will be shown to fall short of requirements.

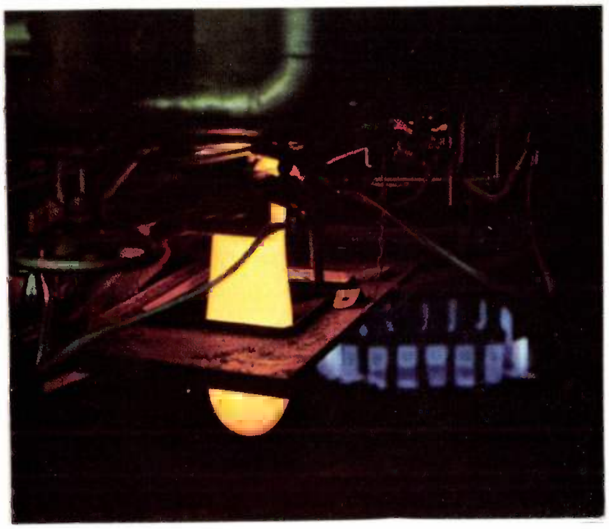


Fig. 2. The stream of glass flowing through the feeder to the glass-working machines is cut into "gobs" by automatic shears. The photograph shows a gob on the point of dropping into a machine for pressing the screens of television picture tubes.

The fuel is atomized by air or steam under pressure. As the pressure rises, the droplets get smaller and mingle more rapidly with the air feed, thus accelerating the combustion and making the flame shorter. The flame transfers its heat to the ambient in two ways: by *radiation*, mainly from the luminous cone, and by *convection* via the combustion gases. A shorter flame thus means that less heat is given up by radiation. Since the same amount of energy per second is still being supplied to the burner, the result is that the combustion gases get hotter. If the thermocouple is located at a position where the heat transfer is mainly governed by convection, the consequence of increased atomizing pressure is that the control system reduces the oil flow, which is of course not the action required. The controller should return the atomizing pressure to the correct value. The simple control system described would therefore be unsatisfactory.

A rise in oil temperature would have a similar effect. The fuel used is a heavy oil whose viscosity only permits satisfactory atomization at a temperature between about 80 and 90 °C. Its viscosity is so temperature-dependent that a variation of a few °C is sufficient to cause a marked change in the atomization. A higher oil temperature thus results in a finer oil mist, a shorter flame and, in the case described, a higher thermocouple temperature, again reducing the oil feed.

Improvement can be sought by mounting the thermocouple at a place where it is least directly influenced by the flame, for instance in the crown of the furnace. The disturbances mentioned will then be less noticeable. A disadvantage, however, is that

[1]) See also P. M. Cupido, Some views on automatic control in glass factories, Glastechn. Ber., 5th Internat. Congress on Glass, Sonderband **32 K**, Heft I, pp. I/1-I/5, 1959.

the thermocouple is now more or less part of the crown, whose heat capacity is very high. Changes in flame temperature will thus be measured with a considerable lag, making rapid and stable automatic control impossible.

The right answer to this problem is to provide every condition governing the properties of the flame with its own control circuit. The first thing to do is to insert a *thermostat* in the oil line to keep the fuel temperature constant, and a *pressure regulator* in the atomizing system.

The next step is to keep the *oil flow rate* constant. For this purpose a constriction is included in the line (orifice plate or venturi), and the pressure drop across it acts, via a regulator, on a control valve in the oil line. This system quickly corrects disturbances without any risk of instability. A fall in oil pressure — due, for example, to partial clogging of a filter — can be corrected in a matter of a few seconds.

Also of importance are the *flow rate of combustion-air* and its *temperature.* Let us consider the flow rate first. This directly determines whether the flame is surrounded by an oxidizing or a reducing atmosphere, and also influences the length and radiation of the flame. It is therefore desirable to measure and regulate the flow of the combustion air. The construction of most furnaces does not allow this, however. The resistance which the burner ports and the regenerator offer to the air flow causes the air to be distributed over the ports in a way that can hardly be controlled. Attempts to control the air distribution by means of dampers in the burner ports have been made, but have all come up against severe practical difficulties.

A system that is both feasible and highly effective relies on a separate regenerator for each burner port. The air required by each burner is fed-in at the bottom of the appertaining regenerator. The air feed can be controlled by a system similar to that used for the oil feed, that is to say the air flow is measured with an orifice plate or venturi tube and kept constant by means of a regulator and control valve. The oil and air regulators are coupled via a *ratio controller*, which maintains the desired ratio of oil to air when one or the other is being varied.

The temperature of the combustion air remains an intractable problem. This temperature is highest immediately after the burners have been reversed, and then gradually falls. The temperature drop can be reduced by shortening the period between successive reversals. The period must not be too short, however, since every reversal — as mentioned above — severely disturbs the prevailing conditions: all flames are extinguished during that operation and the composition of the furnace atmosphere alters radically. The most favourable compromise is usually to fix the period between 20 and 30 minutes. To minimize the disturbance, the reversal itself must be made to take place as rapidly as possible.

### *Automatic reversal*

The reversal of the system is obviously an operation that should be done automatically, the more so since a furnace with separated regenerators in any case involves various operations which must be carried out in rapid succession. We shall now see what these are and the sequence in which they are required to take place.

Suppose that the burners on the left are in operation. When the moment for reversal arrives the first operation is to shut off the oil feed to the left-hand side, and the second to lower the pressure of the atomizing air.

It is not permissible to shut off the supply of atomizing air completely, air being required for cooling the idle burners in order to prevent residual oil carbonizing and causing a blockage. It is equally impermissible to leave the full atomizing pressure on the burners, in view of the quantity of cold air that would then enter the furnace (about 5% of the total combustion air). Hence the *reduced* atomizing pressure on the burners when not in operation.

The next operation is to supply combustion air via the regenerator chamber on the right, its first function being to dispel the combustion gases still present in the chamber. Not until this has been properly done may the atomizing air for the right-hand burners be raised to the requisite pressure and the oil feed turned on. If the latter were done too soon, lack of oxygen would prevent immediate combustion of the atomized oil, resulting in a badly smoking flame and possibly after-burning in the regenerator.

To make these simple operations take place automatically it is necessary to introduce a number of reliable safety measures. Defects in the reversing mechanism, for example, must never give rise to a situation where oil continues to flow to the burners but no combustion or atomizing air. The consequence could obviously be a serious explosion.

*Fig. 3* shows a safeguarded reversing system which has been developed for the automatic furnaces in the Philips glass factories. The initial reversing command is given by a clockwork control timer, and each successive command cannot be issued until the preceding one has been properly executed. Reversal is effected in a time of 10 to 15 seconds. The system is so designed as to shut off the oil feed automatically in the event of a fault or failure, e.g. of the mains voltage or the compressed air. The operation of the system is described in the caption to fig. 3.

*Automatic control of the gas pressure in the furnace*

A parameter not yet mentioned, but whose constancy has an important bearing on the process, is the *pressure inside the furnace*. The combustion process gives rise in the furnace to a certain pressure distribution and the gases present acquire certain velocities. By measuring and controlling the pressure at a suitable point it is possible to stabilize the distribution pattern as a whole. The pressure in the furnace must be somewhat higher than outside, otherwise cold air would be sucked in through various apertures and adversely affect the tempe-

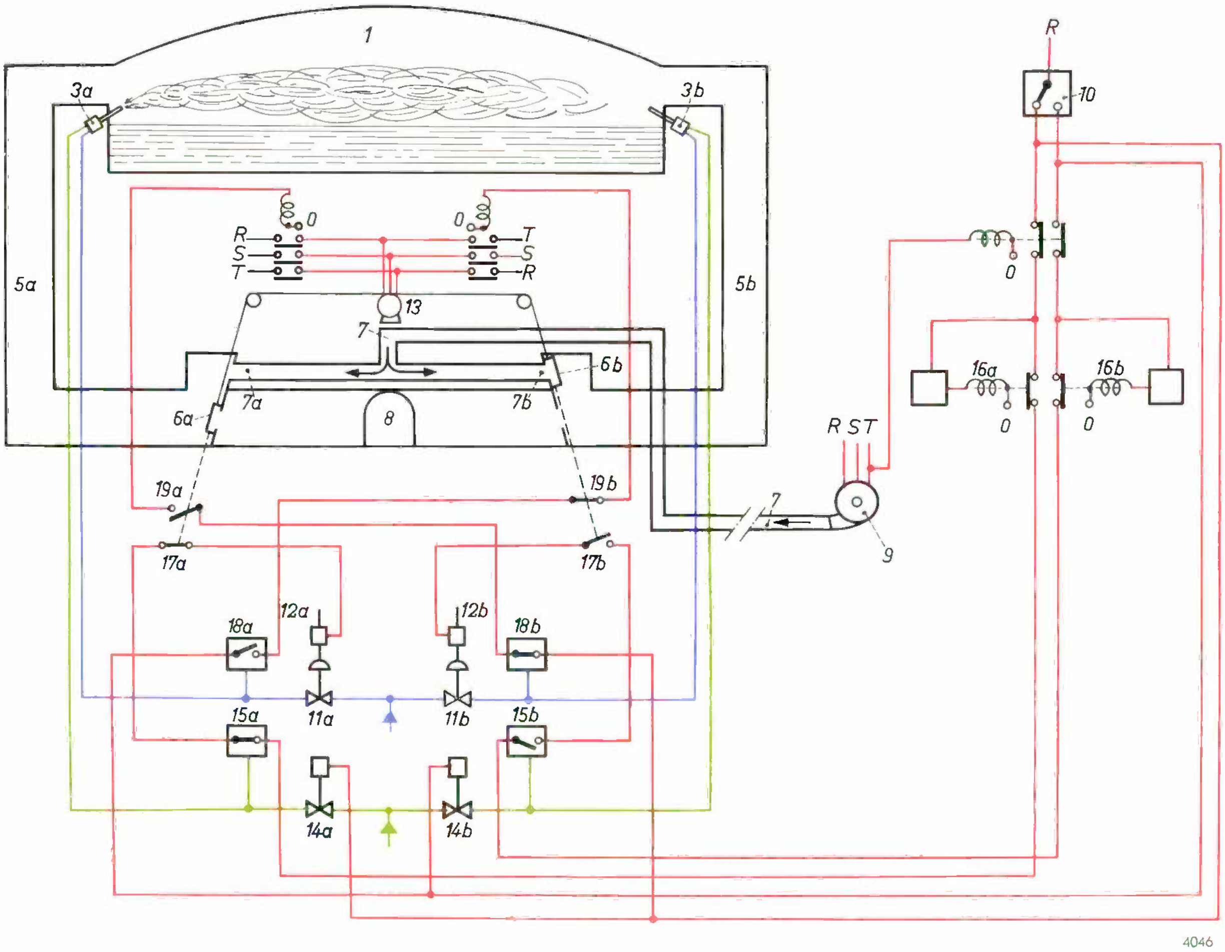


Fig. 3. Simplified layout of automatic reversal system in use in Philips glass factories. *1* tank furnace with burners *3a* and *3b*, regenerators *5a* and *5b*, reversal dampers *6a* and *6b*, duct *7* for the supply of combustion air from blower *9*, and chimney flue *8*. The clockwork control timer *10* delivers at preset intervals the signal initiating the reversal.

*Blue:* oil lines. *Green:* atomizing-air lines. *Red:* electric wiring. *RST0* three-phase mains.

In the situation as drawn, burners *3a* are working. Oil is fed to them via valve *11a*, atomizing air via valve *14a*, and combustion air via channel *7a* and the hot regenerator *5a*. The combustion gases are removed via the cooled regenerator *5b* and the chimney flue *8*. The oil valve *11a* is held open against the action of a strong spring by a pneumatic motor, which receives air only so long as the electromagnetic cut-out *12a* is energized.

The reversal is initiated by the switching arm in timer *10* changing its position. The cut-out *12a* ceases to be energized, oil valve *11a* closes, and the flames on the left are extinguished. Switch *18a*, which was *off*, changes to the *on* position as the oil pressure drops, and switches on the electric motor *13*. The latter now effects the reversal: damper *6a* rises (connecting regenerator *5a* to the chimney and closing the air feed *7a*), and damper *6b* descends. On reaching their new positions, the dampers switch off the motor *13* by actuating switch *19b*.

By altering the position of control valve *14a* the initial signal from timer *10* causes the atomizing air on the left to drop to the required low pressure. As a result the pressure-sensitive switch *15a* moves to the *off* position. This switch, in series with contacts *17a*, is incorporated in the line for energizing cut-out *12a*; the oil feed to the burners *3a* cannot therefore be restored until the pressure of the atomizing air has been raised and the dampers are in their correct positions.

To supply oil to burners *3b*, valve *11b* must be opened. This involves energizing cut-out *12b*, which is not possible until switches *15b* and *17b* are on. Switch *17b* comes on when damper *6b* descends. For *15b* to close, the atomizing pressure of burners *3b* must be sufficiently high, i.e. valve *14b* must be wide open. The timer *10* further ensures that *14b* is opened wide, not immediately, however, but only after a delay determined by relay *16b*. This delay is necessary to allow the combustion air to dispel the combustion gases from regenerator *5b*. This latter operation is promoted by a special "flushing" signal which fully opens the control valve for the combustion air and the valve in the chimney flue, thus enabling the oil feed to be switched on earlier. In the event of a failure in the electric supply or the compressed-air supply, the oil feed is automatically shut off.

rature distribution, and perhaps also the combustion. The difference in pressure should only be slight, however, since the escape of hot gases entails a loss of heat and moreover accelerates furnace wear.

In manually-regulated tank furnaces the positive pressure is adjusted with stack dampers, which are generally broad and heavy cast-steel plates in the flue to the chimney stack. The control range is covered by moving the damper up or down a few centimetres. For automatic control such a damper is scarcely suitable: its large mass (often increased by counter-weights) does not lend itself to rapid control, and the available maximum displacement of only a few centimetres precludes the precision and stability required. It would be an improvement to mount a narrower and much lighter damper on the existing damper; the reduced width then gives a greater stroke and the smaller mass allows faster and more accurate control.

An effective and inexpensive method of automatic control depends on the use of an adjustable "artificial draught" (*fig. 4*). In the chimney flue *1*, in front of the damper *3*, a round aperture is made which is fitted with a short side-pipe *5*. This contains a butterfly valve *6*, which can be moved by a pneumatic or electric servo-motor *7* much faster than is possible with the stack damper in the systems just discussed. The butterfly valve controls the pressure prevailing in front of the damper by admitting a greater or smaller quantity of air.

Even with this system, however, the control action is not fast enough to deal with the consequences of gusts of wind at the mouth of the chimney. Gusts of wind give rise to often steep-fronted pressure waves which travel at the speed of sound into the furnace itself, where they upset the prevailing conditions. In this respect a high brick stack is inferior to a short, possibly metal chimney in which an appreciable negative pressure can be created with the aid of an air-injector (*fig. 5*), thereby minimizing the effect of pressure variations caused by strong gusts. The correct pressure in the furnace is obtained by suitably adjusting the flow resistance in the flue by means of a damper.

The system using a short chimney, however, is not without its drawbacks. The short chimney does not conform to the regulation in many countries that large quantities of combustion gases containing a high percentage of sulphur (as in the case of fuel oil) may only be discharged into the atmosphere above a certain height. Furthermore the energy required for the injection pushes up the running costs to a multiple of those incurred with a high stack.

Some of Philips glass factories use high chimneys, others low. *Fig. 6* shows an outside view of the pressed-glass works at Eindhoven, with its two high chimney stacks.

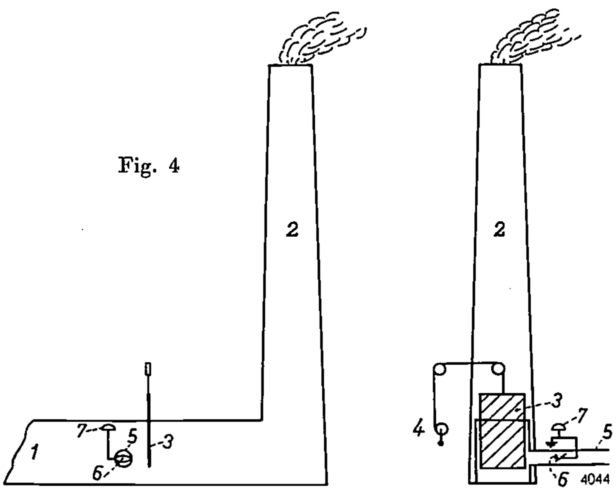


Fig. 4

Fig. 4. Automatic control of the pressure inside the furnace by an artificial draught. *1* flue through which the combustion gases from the furnace enter the chimney stack *2*. *3* damper, adjustable with hand winch *4*. *5* side-pipe with butterfly valve *6*, rotated by servomotor *7*.

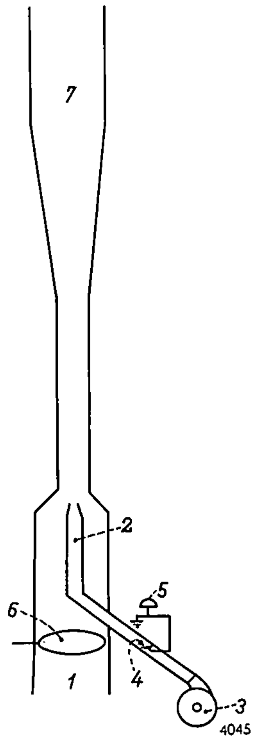


Fig. 5

Fig. 5. A short chimney using air injection avoids the difficulties caused by gusts of wind with high chimney stacks. *1* flue for exhausting the combustion gases. *2* air-injector with blower *3*, control valve *4* and servomotor *5*. *6* adjustable damper producing the correct under-pressure upstream of the damper. *7* short chimney.

### *Automatic control of glass temperature and level in the feeders*

We now come to the feeders, which represent the last stage of the molten glass on its way to the machines where it is worked. The feeders are required to deliver per unit time a *constant quantity* of glass of *constant viscosity*. Since there is no practical means of directly measuring and regulating the viscosity, an indirect means of controlling it is adopted, that is by controlling the *temperature* of the glass. However, the viscosity is not solely dependent on the temperature but also on the composition of the glass. Changes in composition sometimes occur in the form of irregularly distributed inhomogeneities (cords), which may be due to non-uniformities in the batch composition. The less such deviations occur, the better is the temperature control system able to maintain the desired viscosity.

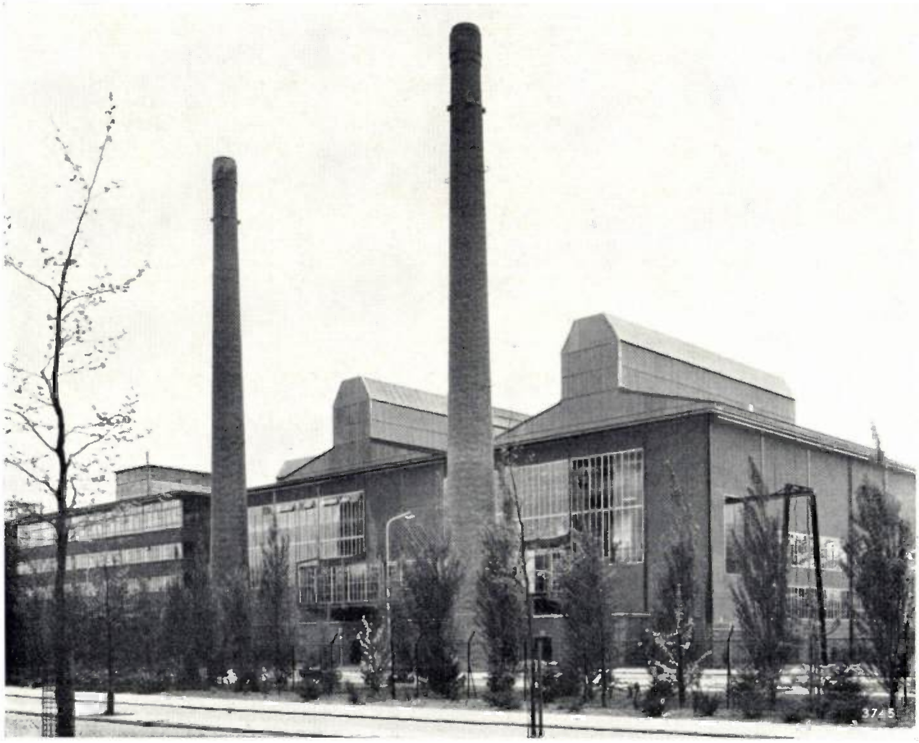


Fig. 6. Philips pressed-glass factory at Eindhoven.

The *quantity* of glass which the feeder delivers per unit time depends on the viscosity and level at the effluent end. It is therefore necessary to keep the level of the glass constant. This level is therefore continuously measured, and deviations from the desired height are fed-back to act on the rate at which the batch mixture enters the furnace. Controlling the automatic batch feed in this way also helps to keep the melting conditions constant, as well as the flow pattern in the tank furnace.

Experience has shown that accurate level control prolongs the life of the melting tank appreciably.

The molten glass attacks the walls of the tank, especially at the level of the glass surface. This ultimately causes a groove to appear, which is narrower the less the level of the glass varies. That this is attended by a longer tank life is presumably to be explained by the fact that, in the narrow groove, the glass is virtually stationary, so that after some time it becomes saturated with material dissolved from the wall, and further corrosion proceeds much more slowly.

**Performance criteria required of control systems**

A principal requirement to be met by control equipment is that, if the value of the controlled quantity changes, the correct value should be rapidly restored. A limit is set to the speed of the control action by the properties of the controlled process and of the control instruments used. If this speed is exceeded the system becomes unstable, that is to say, after a disturbance the value of the controlled quantity continues to oscillate at a certain amplitude around the desired value. This phenomenon has been discussed in this journal in an article concerned with process control systems in general [2]). We shall briefly recapitulate the salient points.

The operation of a control system is based on comparing the measured value of the controlled quantity with the desired value. Two main categories may be distinguished: discontinuous and continuous control. The first category, for example simple two-step or on-off control, has the advantage of considerable simplicity and reliability. A process controlled by a discontinuous controller, however, is by nature unstable: the controlled value oscillates continually around the desired value. In cases where the amplitude of the oscillation is excessive continuous control is indicated. The controller may have

[2]) H. J. Roosdorp, On the regulation of industrial processes, Philips tech. Rev. 12, 221-227, 1950/51.

proportional or integral action, whereby the position of the final control element is proportional to the deviation $\varepsilon$ from the desired value, or to $\int \varepsilon \mathrm{d}t$ respectively. Purely proportional action has the disadvantage that the desired value is never entirely reached, which is not the case with integral action. The best result is obtained as a rule by combining proportional and integral action. A derivative action, proportional to $\mathrm{d}\varepsilon/\mathrm{d}t$, may be added. It has a damping function and is only employed to improve the stability of control systems where particularly fast corrective action is required.

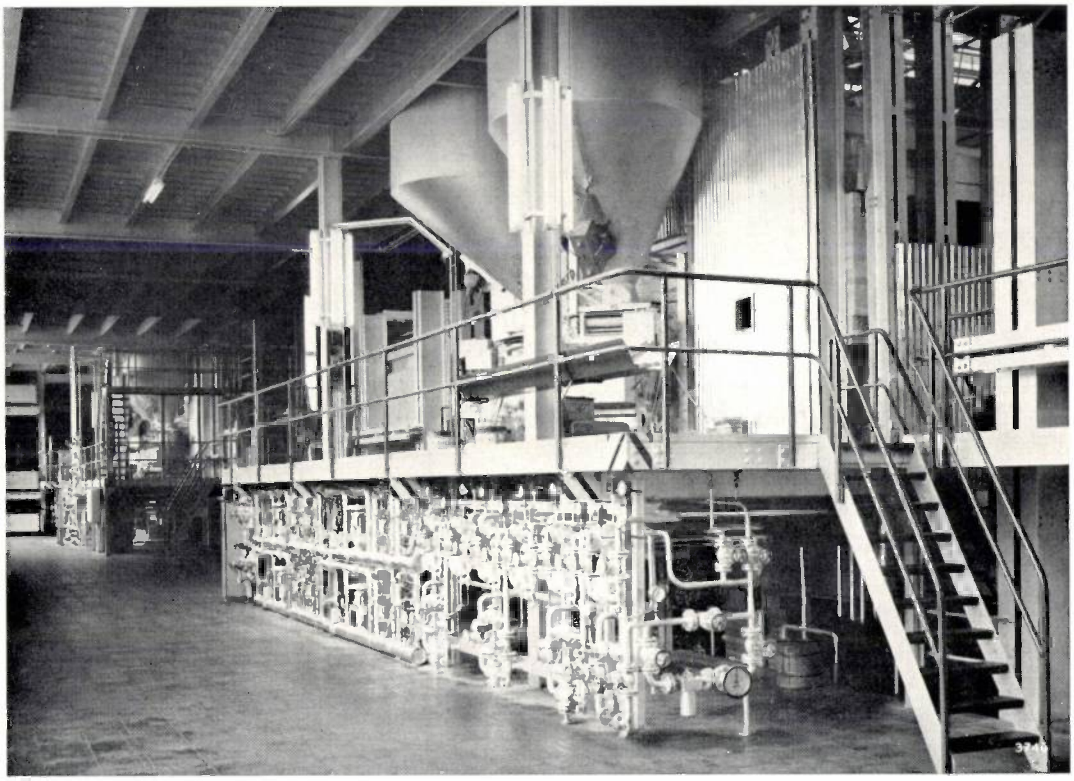


Fig. 7. The two furnaces in the pressed-glass factory in Eindhoven. Above the gallery can be seen the raw-material hoppers that open onto the feed machines. The corrugated plates shield off radiation from the furnace. Under the gallery are some of the valves and control elements of the control system.

Sufficient stability is almost invariably to be ensured by not raising the sensitivity of the control equipment too far. There are cases, however, where the response of the system would be too slow, because of the transfer lag or dead time of the process itself, of the measuring unit or of the final control element. Reduction of the lag of a final control element was discussed above in connection with controlling the pressure in the furnace (where a heavy damper was replaced by a lighter one). We shall discuss two other practical examples to illustrate how it was possible in certain cases to reduce the lag in a part of the process and in the measuring unit.

We have seen that, to control the level of glass, it is necessary to measure the level continuously, deviations from the desired height being made to act on the rate at which the batch mixture is fed in. Between the melting tank and the working tank — i.e. in front of the point where the level is measured — there is a communicating passage, the "throat", which serves for skimming-off impurities and inhomogeneities on the surface of the glass and for preventing undesired currents along the bottom of the melting tank. For the latter reason the throat must not be too wide. In some existing plants, however, it was so narrow that a disturbance of the level in the melting tank reached the detecting element only after a considerable delay. This delay was reduced by widening the throat.

Our last example relates to the equipment for measuring the temperature at the top of the furnace. Following the old and familiar practice, a block was placed in the crown of the furnace with a hole drilled into it for the thermocouple. The block protected the thermocouple and also acted as a radiation shield, and had a wall thickness of a few centimetres.

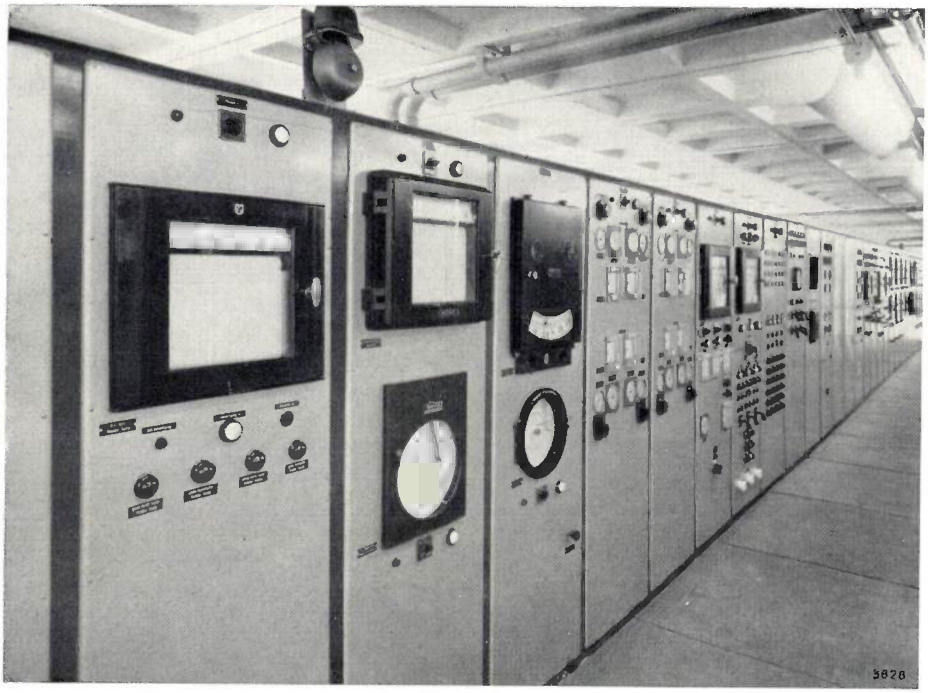


Fig. 8. Instrument racks of the (largely pneumatic) control system for the plant shown in fig. 7. Recording and other instruments, pilot lamps and control buttons can be seen.

The result was that the measured temperature was appreciably lower than that of the furnace, and that the response to changes in the furnace temperature was subject to a considerable dead time. The temperature difference and the response lag were substantially reduced by placing the thermocouple in a thin-walled refractory sheath passing through the crown block and projecting about 10 cm into the furnace. A marked improvement was thus obtained in the automatic temperature control.

A property which all automatic control equipment must possess, especially for continuous processes, is a high degree of *reliability*. This is a matter of both plant design and choice of control equipment. In the plant it is necessary to ensure that any component which is at all vulnerable is readily accessible and can be quickly replaced. This also applies to control valves, orifice plates, etc., incorporated in the pipelines. It should be possible to isolate all these components from the rest of the equipment, and they should be provided with individual bypass lines that can be shut off, the latter to allow replacement without having to close down the plant. To facilitate fault-finding, readily accessible and well-planned points should be provided where the necessary check measurements can be made. It is desirable that defects should be automatically localized and signalled by pilot lamps. In the choice of control equipment, too, reliability is the dominant factor, the cost being a secondary consideration.

*Fig. 7* shows a view of the furnaces in the pressed-glass factory at Eindhoven. Under the platform can be seen the valves of the largely pneumatic control system. The associated control panels, with meters, pilot lamps, etc., are to be seen in *fig. 8*.

Hitherto, the control devices used in glass manufacture have been almost exclusively pneumatic. Recently, however, the use of electronic controllers has started to gain ground. The fact that these devices possess the high degree of reliability required for industrial processes is partly attributable to the transistor. Electronic systems are particularly suitable wherever fast automatic control is called for.

**Summary.** A discussion of the parameters amenable to automatic control in glass production is introduced by a short description of the glass-manufacturing process. Recuperative and regenerative tank furnaces are touched upon. Stabilization of the process by individual automatic control of all quantities influencing the process, is shown to be desirable. These quantities are: the temperature of the fuel oil, the oil flow rate, the pressure of the atomizing air, the combustion air flow rate, the oil to air ratio, the pressure inside the furnace, and the temperature and level of the molten glass in the feeders to the glassworking machines. A discussion is devoted to the automatic reversal of the burners in a regenerative furnace, which must take place every 20 to 30 minutes. In the choice and installation of control equipment — hitherto largely pneumatic — emphasis is placed on reliability. The use of electronic control equipment in glass production is gradually gaining ground.

# MECHANICAL PRODUCTION OF BULBS FOR ELECTRIC LAMPS AND RADIO VALVES

by P. van ZONNEVELD *).

666.1.032:621.326.61

It is now many years since hollow glass objects, such as bottles, jars and similar containers, have been manufactured mechanically, largely replacing manual methods. Manual as opposed to mechanical production is preferred only where small quantities are involved and possibly also where it is desired to make a variety of products at the same time or to change quickly from one type of glass to another.

Machines for producing lamp or valve bulbs must be capable of turning out a fairly thin-walled product without marked variations in thickness and possessing — for lamps, at least — reasonable optical properties. Bulbs cannot therefore be made on a bottle machine.

The machines that fabricate articles direct from *molten glass* fall into two main categories, namely those which pick up the glass by suction from the surface of a glass tank, and those which receive their supply from an orifice in the bottom of the reservoir containing the molten glass. The latter may again be sub-divided into machines that receive their charge in portions or "gobs", and machines to which glass is fed continuously.

As we shall see, it is not economical to make very small objects with these machines. For this purpose special machines have been developed whose starting material is not molten glass but *glass tubing*.

In this article we shall discuss both main types of machine. Three machines designed for the working of molten glass will be dealt with. Particular attention will be devoted to a machine developed at Eindhoven which has the merit of working without loss of glass: the gobs fed to the machine need be no heavier than the bulbs to be fabricated.

As regards machines that work from glass tubing, we shall deal first with an 18-head machine designed, again at Eindhoven, for working short sections of tubing, one per bulb. This will be followed by a brief discussion of the way in which new machines have been and are being evolved on the same principle.

Before proceeding to mechanical glass-working, it will be useful to recall the method of blowing glass by hand. The glass-blower starts by dipping and rotating the end of his blowpipe into the molten glass and thus making a "gather" (*fig. 1a*), care being taken to gather no more and no less than is roughly required for the product. He then rolls the mass of

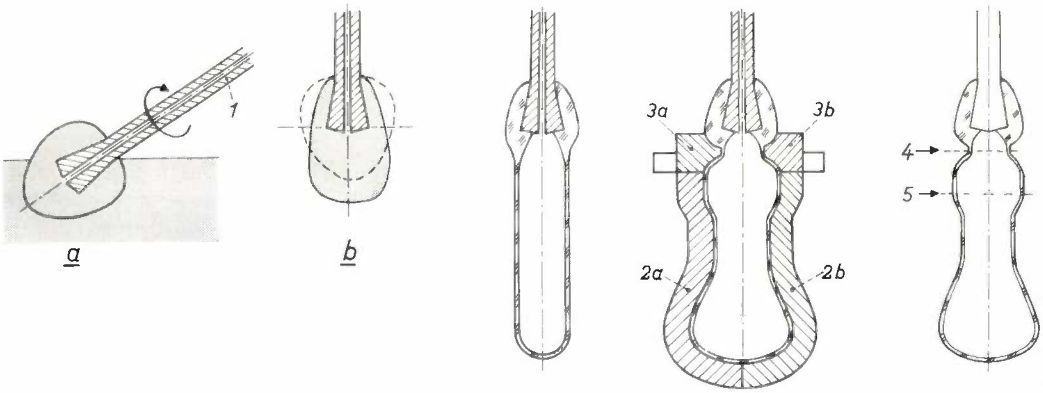


Fig. 1. The principal phases in off-hand glass-blowing.
*a*) "Gathering" the glass: the end of the blowpipe *1* is dipped into the molten glass (shaded) and rotated to pick up a lump of glass of the required size.
*b*) The gather is rolled on a plate to reduce the amount of glass that cannot be blown (above the horizontal dashed line).
*c*) The hollow mass of glass — the "parison" — to be enclosed in a blow-mould. The glass-blower swings, inverts and otherwise manipulates the parison to distribute the glass in such a way that the blown product will not show excessive variations in wall thickness.
*d*) The blown bulb inside the two halves of the mould (*2a* and *2b*), which mate against the halves of the cap (*3a* and *3b*). As can be seen, the blowpipe remains outside the mould.
*e*) After blowing the bulb is severed from the blowpipe (at *4*) and the excess glass is cut off from the bulb (at *5*). The glass still adhering to the blowpipe is removed before the glass-blower makes a fresh gather.

*) Glass Division, Eindhoven.

hot glass on a plate to reduce the portion which, adhering around the pipe, cannot be blown (fig. 1*b*). This being done, he blows for a moment into the mouthpiece and then closes it with his thumb. Thermal expansion of the air causes the hollow mass thus formed — called the "parison" — to swell out. By swinging the parison to and fro the glass-blower helps it to reach the required elongation faster than it would under the force of gravity alone. While the air is expanding he may also hold it up vertically or obliquely for a moment, to ensure the required distribution of the glass. Next, he encloses the parison between the two halves of a split mould ("open and shut" mould), and blows until the glass fits against the sides (fig. 1*c* and 1*d*). During this process the glass cools down to rigidity. Finally, it is severed from the blowpipe, and the blower is able to make a fresh gather. His product is not, however, complete at this stage. As a rule, excess glass has to be sheared away from the top (fig. 1*e*). In *fig. 2* a number of glass-blowers can be seen at work. Some of the steps in the production process described will be encountered in the machines used for working molten glass, which we shall now discuss.

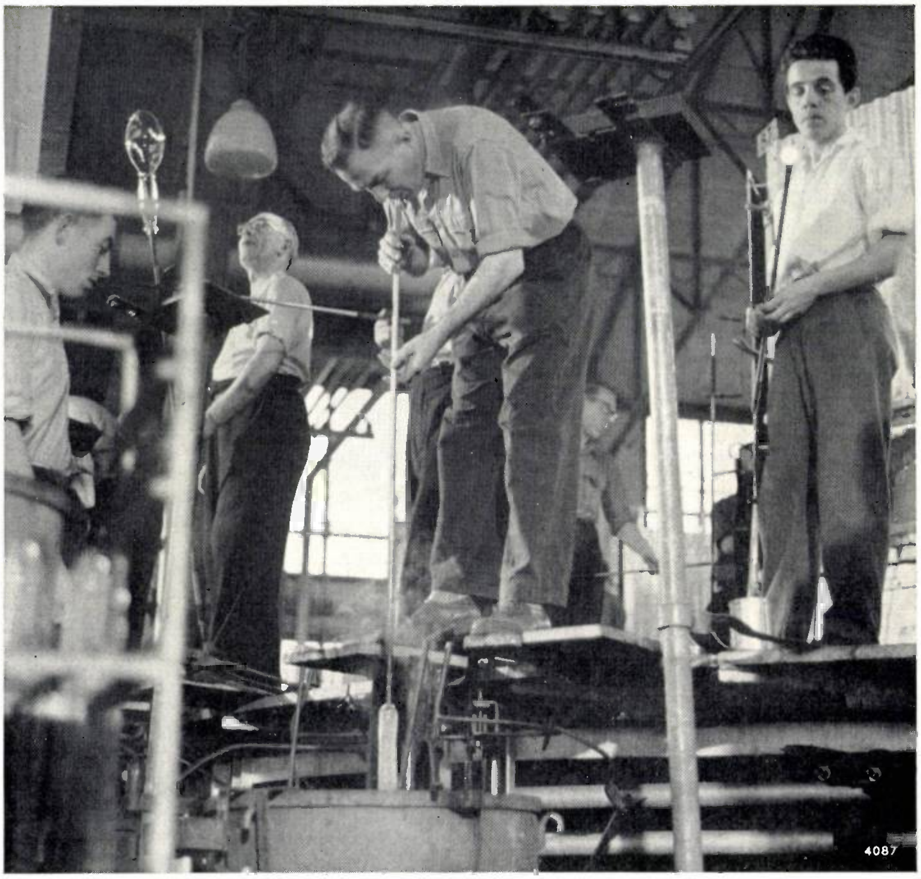


Fig. 2. Glass-blowers at work. The man on the right has just made a "gather" and has started to form the parison. The man in the centre (foreground) has completed the parison and is about to close the blow-mould (with a pedal). The third is inspecting a product fresh from the mould. The workman on the left removes the bulbs from the blowpipes. Between the first two glass-blowers (foreground) is a table for the rolling operation.

## Machines for making bulbs from molten glass

### *The Eindhoven 16-head machine*

The gobs of molten glass worked by the 16-head machine developed at Eindhoven are automatically cut at regular intervals from the viscid stream of glass issuing in a constant delivery from an orifice in the furnace floor. The cutting shears (see *figs. 3* and *4*) are two knives with roughly V-shaped cutting edges, which slide one over the other.

The machine itself consists of a continuously rotating turntable on which are mounted 16 gob-receiving stations and 16 complete glass-blowing units; the former are operated mechanically, the latter for the most part pneumatically. Each glass-blowing unit has its own set of control valves, which ensure that the unit turns out one bulb on every

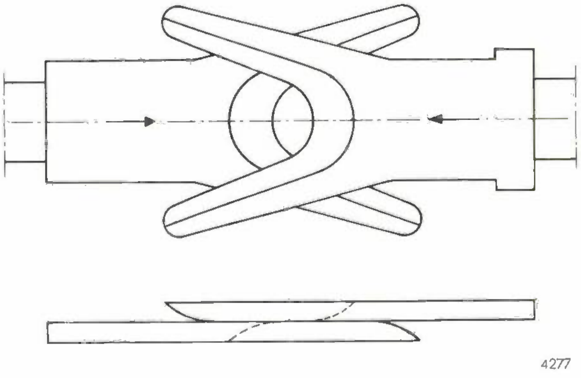
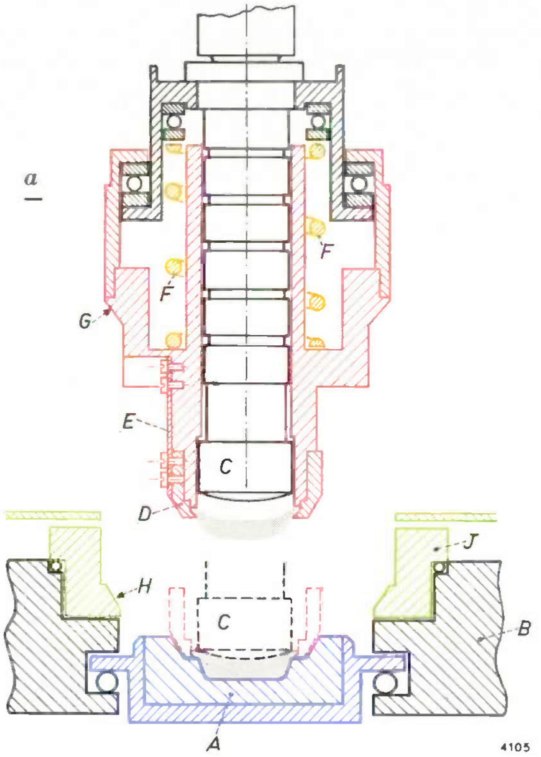


Fig. 3. Shears for cutting the viscid stream of glass issuing from underneath the furnace into "gobs" (schematic). The two V-shaped knives slide one over the other in the direction of the arrows.

revolution of the turntable. The machine thus produces 16 bulbs per revolution.

As soon as a gob is severed it drops into a small tray of a given receiving station. The tray, which at that moment is situated close to the periphery of the turntable, is then slid back under the blow-head of the corresponding glass-blowing unit. The manner in which this takes over the glass from the tray and turns it into a bulb will be described with reference to *fig. 5a*.

In this figure, *A* is the gob tray (blue). It is mounted in a holder (also shown in blue) which can move radially in a groove in the turntable *B*. When the tray is centred under the blow-head, the latter descends (see dashed lines) and the plunger *C* presses the glass (shaded) into the claws *D*. These are flexibly mounted to leaf-springs *E*, and close together at the sides so that the glass cannot flow out between them. In this blanking operation the red part of the blow-head rests on the edge of the tray, and the plunger slides downwards with respect to this part, thereby slightly compressing the spring *F* (yellow). The plunger and the remaining part of the blow-head then return to their initial position (this is the position represented in fig. 5*a*; the red part is now suspended from the part outlined in black) and the empty gob receiver returns to the edge of the turn-



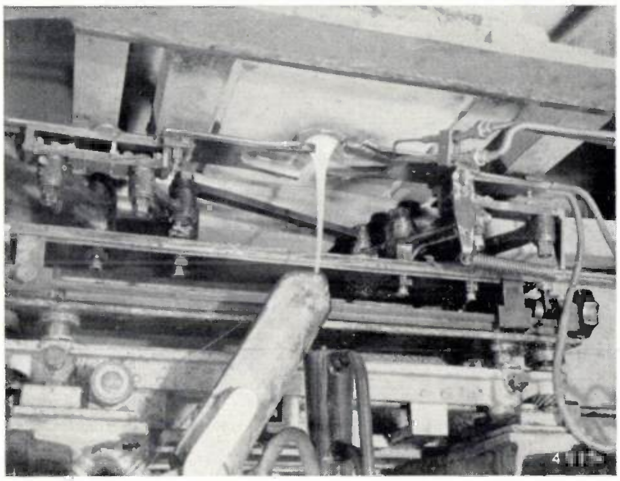


Fig. 4. Gob shears mounted under the furnace floor. Since the machine had to be stopped for taking the photograph, the glass stream is led off through a gutter at the side.

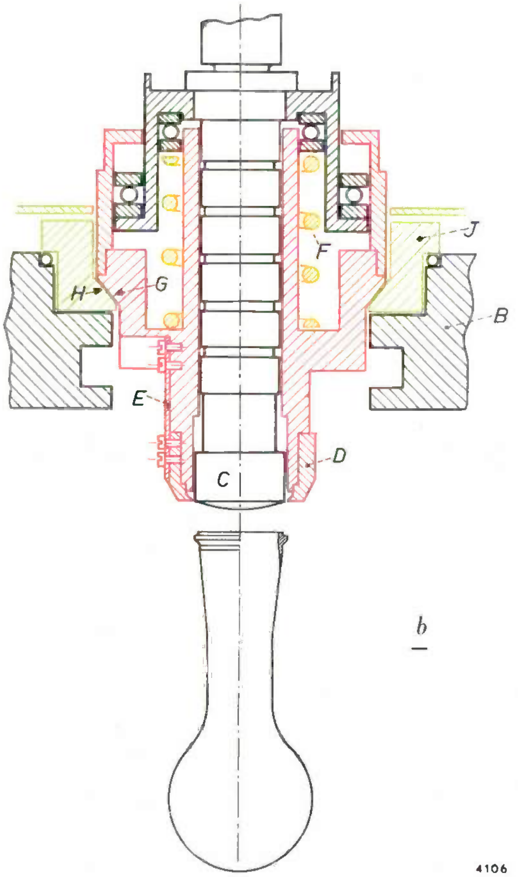


Fig. 5. *a*) Schematic cross-section of a blow-head (red-black-yellow), a gob receiver (blue) and part of the turntable. *A* is the gob tray, mounted on a slide which moves in a radial groove in the turntable *B*. *C* blank plunger. *D* one of the claws which hold the glass (shaded) forced into them by the plunger. (The blanking position is represented by dashed lines.) *E* leaf-spring to which *D* is attached. *F* spring (yellow). *G* tapered edge of red part which, during the elongation of the parison, the blowing of the bulb and the ejection from the head, rests on the tapered edge *H* of the permanently rotating ring *J* (green).
*b*) The blown bulb is ejected by the plunger *C*, the red part of the head still resting on the ring *J*. In this operation the spring *F* is compressed somewhat more than in the blanking operation.

table. The blow-head is now able to carry out the further operations. It again descends and abuts with its tapered edge $G$ against the mating edge $H$ of a ring $J$ (green), which is in continuous rotation on the turntable. The red part of the blow-head is thus set in rotation too. Its lower portion, carrying the transferred gob or blank, projects under the turntable and is immediately above the open blow-mould. At this point the glass, which began to elongate the moment it was picked up, is reheated (in the case of certain shapes of bulb) by a burner mounted underneath the turntable.

To accelerate the elongation process and ensure that a bulb of roughly constant wall-thickness is produced, a few short puffs of air are blown into the parison. The air for this "puffing" operation is supplied through a duct inside the plunger stem, and enters the parison by flowing around the plunger when the latter is in the position shown in fig. 5*a*. It will be noted that, compared with the glass-blower, the machine is limited in its manipulations for properly distributing the glass in the parison; it is therefore particularly important for the puffs to be of the correct strength and duration. When the elongation is far enough advanced, the two halves of the split blow-mould close and air is blown into the parison, which thus acquires the contour of the mould. The mould then opens again and the bulb is released from the claws. This is done by the plunger, which again descends and presses on the mouth of the bulb, thereby forcing open the claws. (The part shown red cannot descend since it rests on the turntable, and spring $F$ is fairly strongly compressed — fig. 5*b*.) The ejected bulb drops on to a conveyor belt. Finally, the blow-head returns to its starting position, and the gob receiver and blowing unit are ready to accept a fresh charge. The movement of the turntable is synchronized with that of the shears by driving them both with synchronous motors.

It should be added that so-called "paste" moulds are used, i.e. moulds lined with adherent carbon — in the present case a layer of graphite. While the mould is open the lining is sprayed with cold water. During the blowing operation, when the bulb continues to rotate inside the mould, a small quantity of water vapour forms which precludes actual contact of the glass with the mould, giving the product a very smooth surface.

The vertical movement of blow-head and plunger is brought about pneumatically by means of a vertical piston above the blow-head. In *fig. 6*, which shows the arrangement schematically, the positions are marked as assumed by the piston during the various stages of the fabrication process. It can be seen that only the lowest (after ejection of the bulb) is governed by the dimensions of the cylinder; the blanking and blowing positions are determined by the blow-head resting respectively on the gob receiver and the rotating ring.

The valves in the lines that supply and release the compressed air are mechanically operated. They, and the valves for the gas and air feed to the burner, are operated as follows. Mounted above each blowing unit, about 1 metre above the turntable, is an off-radial beam attached to the table and carrying the valves pertaining to the unit. The valves are operated by adjustable cams fitted underneath a fixed plate (*fig. 7*). As the machine rotates, the rocker arms of the valves pertaining to the sixteen blowing units all pass successively the same cams. Since it must be possible to vary the starting time

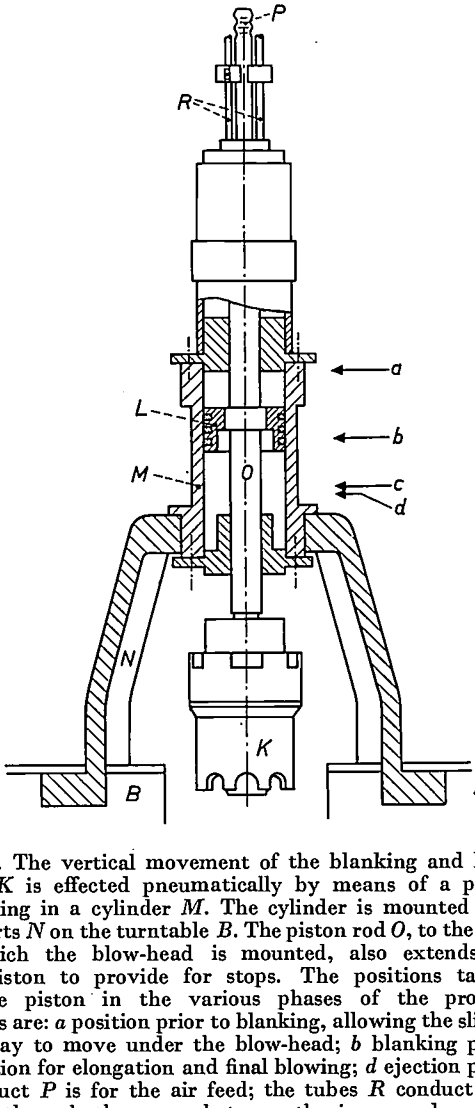


Fig. 6. The vertical movement of the blanking and blowing head $K$ is effected pneumatically by means of a piston $L$ travelling in a cylinder $M$. The cylinder is mounted by two supports $N$ on the turntable $B$. The piston rod $O$, to the bottom of which the blow-head is mounted, also extends above the piston to provide for stops. The positions taken up by the piston in the various phases of the production process are: $a$ position prior to blanking, allowing the slide with gob tray to move under the blow-head; $b$ blanking position; $c$ position for elongation and final blowing; $d$ ejection position. The duct $P$ is for the air feed; the tubes $R$ conduct cooling water through the space between the inner and outer tube forming the piston rod.

and duration of every operation in order to establish the optimum conditions necessary for blowing good bulbs, each cam consists of two parallel adjacent plates which can be displaced with respect to each other in a direction tangential to the circular cam mounting plate.

A considerable advantage of this system, in which the machine keeps turning continuously, is that a bulb never has to "wait" for the beginning of the next operation, as it must do in machines not equipped with complete blowing units, where each operation has to be performed at a particular position whilst the machine is stationary.

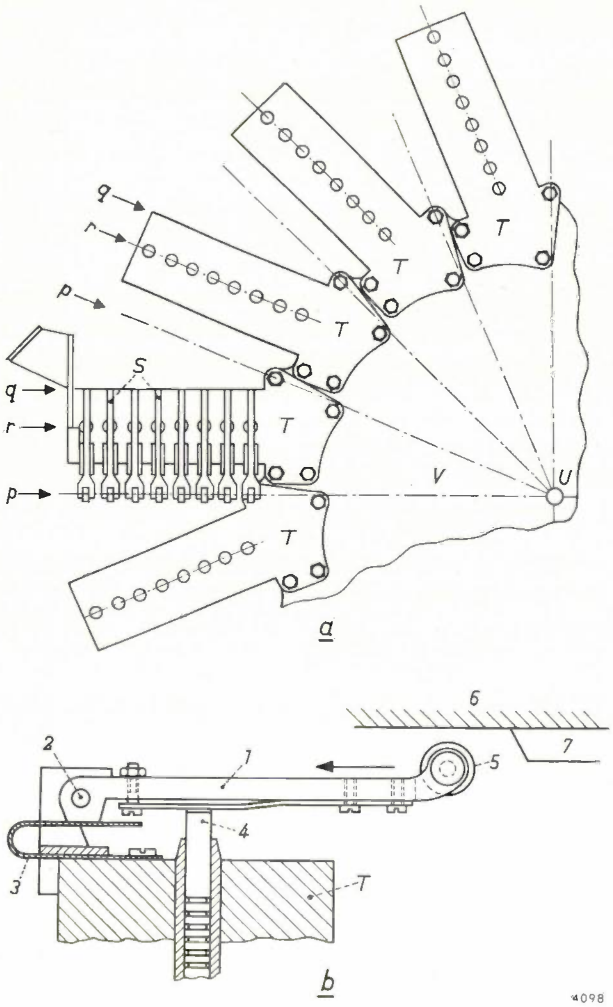


Fig. 7. *a*) The rocker arms *S*, which operate the valves, are mounted tangential to the turntable on the cross-beams *T* in such a way that the cam-rollers lie on the lines *p* intersecting the turntable axis (*U*). The rocker-arm pivots lie on the lines *q*, the valve plungers on the lines *r*. The 16 valve cross-beams are bolted to the central plate *V*, which is fixed to the turntable. The entire assembly thus rotates when the machine is in operation.

*b*) Side view of one of the rocker-arm assemblies. *1* rocker arm. *2* pivot, so designed that the lever can be removed by simply pressing-in the leaf-spring *3* and the valve plunger *4*. *5* roller which travels along under the cam plate *6*. *7* cam. The valve cross-beam moves in relation to the cam plate in the direction of the arrow.

In such machines the rate of production is governed by the longest operation at a given position, whereas in our case the speed at which the turntable revolves is the decisive factor, and this is governed solely by the total time needed to produce one bulb. The latter is about 15 seconds, and therefore the rate of production is just over one bulb per second, or roughly 4000 an hour. A continuously rotating machine is also mechanically to be preferred to one with intermittent action, particularly where large machines are involved. A third advantage is that machines equipped with complete blowing units can produce *different kinds of bulbs at the same time.* The only restriction is that the bulbs must all be of the same weight and must all be made with the cams in the same position.

Photographs showing a machine of the type described are to be seen in *figs. 8* and *9. Fig. 10*

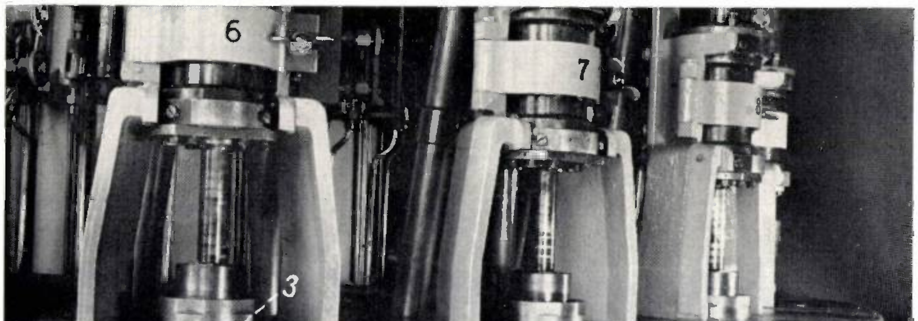


Fig. 8. Detail of the machine in fig. 9. Among the components that can be identified are the two mating halves of a blow-mould, *1a* and *1b*, and the slide *2*. As soon as a gob has fallen through the hole *3* into the tray, the slide is pushed inwards by the lever *4*, the forked end of which engages a roller. This lever is itself turned on a spindle by a cam *6* (extreme right of photograph) on the cross-beam *7*. (The relevant spindle in one of the neighbouring blow units is indicated by the figure *5*.)

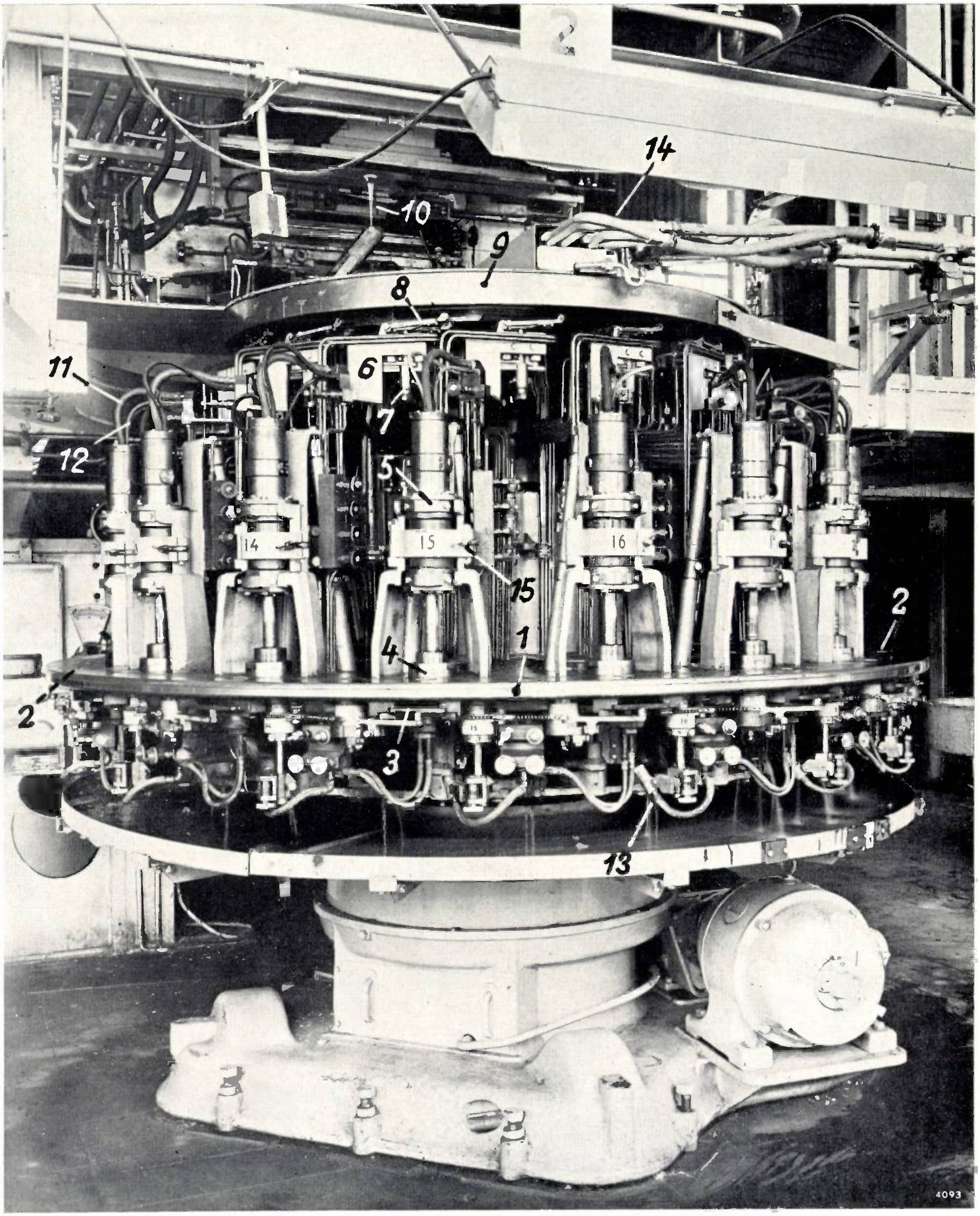


Fig. 9. Machine for making electric-lamp bulbs, equipped with 16 blowing units. *1* turntable on which the blowing units are mounted. *2* holes in the turntable through which the gobs fall into the receiving trays. These trays are mounted in slides *3*. *4* blow-head in blowing position (in which case the blow-head projects under the turntable). *5* cylinder with plunger. *6* end of a cross-beam carrying control valves. *7* outer valve of the eight carried on one cross-beam. *8* rocker arm which operates this valve. *9* cam plate. *10* stream of molten glass flowing from an orifice underneath the furnace (the glass was not fed to the machine when this photograph was taken). *11* air feed for puffing and blowing. *12* cooling-water line for piston rod. *13* burner for heating the elongating parison. *14* supply lines for gas, air (both for burners *13*), oil and compressed air (for the blow-heads). After shutting-off the feed, disconnecting the hoses and unscrewing the fly-nut *15*, a single manipulation is all that is needed to remove a blow-head together with its cylinder.

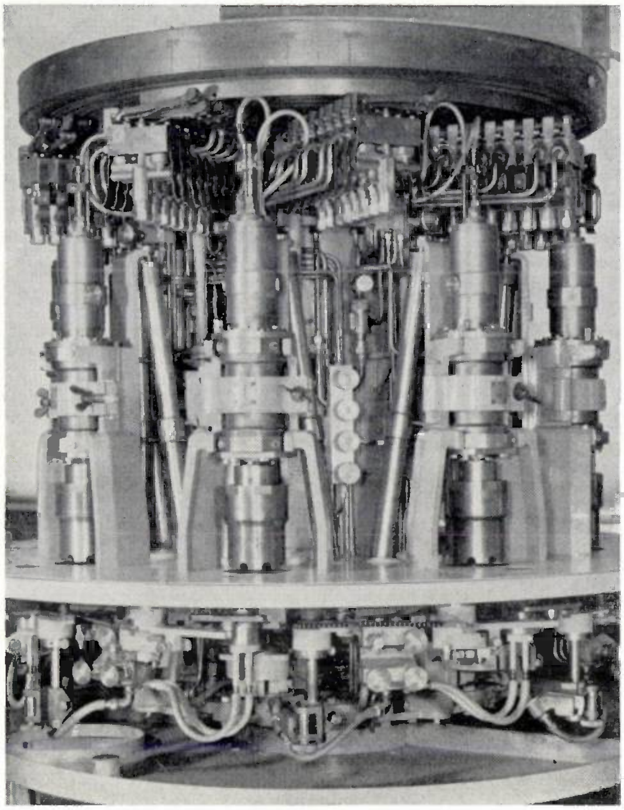


Fig. 10. Eight-head machine from which the 16-head type was developed. The construction here gives a better view of the valve-beams and the cam plate.

shows a similar 8-head machine in which various details are better visible. *Fig. 11* illustrates the layout of such a machine with respect to a glass furnace, and *fig. 12* the method of obtaining a constant delivery of glass gobs from the furnace. Since the machines work with no loss of glass, the furnace can be fairly small; a machine whose hourly output is 4000 bulbs of 32 grams each consumes in that time only 128 kg of glass. A small selection from the numerous kinds of bulbs that can be made on these machines is shown in *fig. 13*.

Although we speak of glass losses, no glass is in fact wasted. Any glass that may have to be cut away from a blown bulb to produce the shape required by the lamp factory is of course fed back to the furnace. The actual loss is in the heat needed for remelting, which amounts roughly to 75% of the heat necessary for making glass of the same temperature from the pure raw materials. In this connection it should be remembered, however, that in practice glass is never made from 100% raw materials but from a batch containing 30 to 50% of cullet (broken glass), added to obtain more readily a fluid mass possessing better thermal conductivity.

Finally, it should be mentioned that the machine is designed to allow numerous elements to be changed *during* operation. For instance, with a few manipulations an entire blow-head can be replaced, and the same applies to such components as the rockers

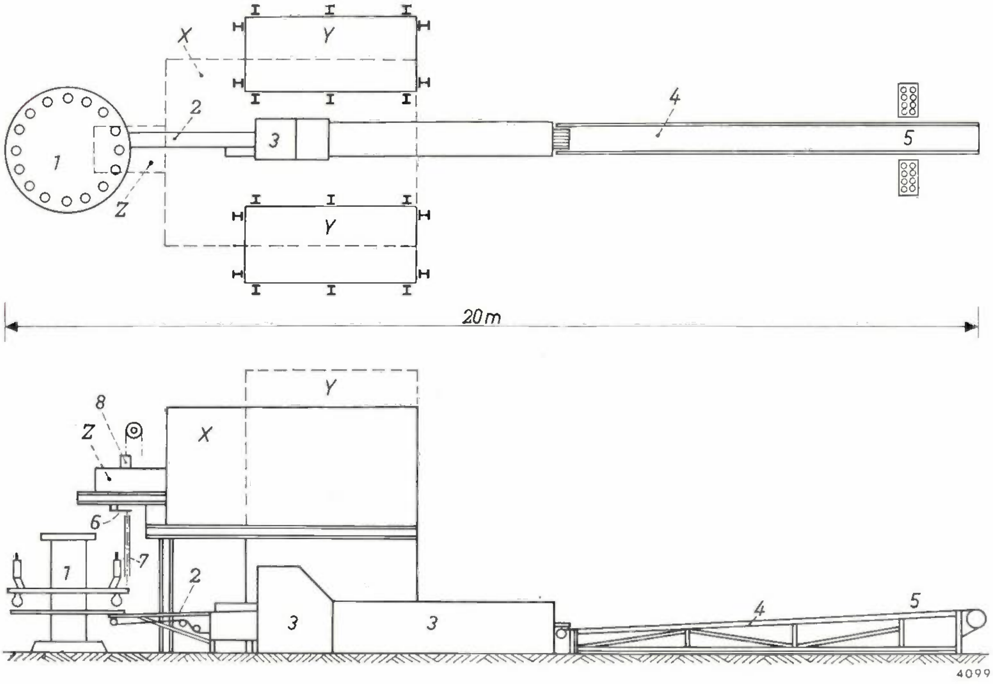


Fig. 11. Schematic layout of a manufacturing unit for the mechanical production of electric-lamp bulbs. *a*) Plan, *b*) side view. The whole plant is contained within a rectangle of 20 × 5½ metres. *X* glass furnace. *Y* furnace recuperators, in which the outflowing combustion gases heat the air flowing to the burners. *Z* part of glass furnace from which the bulb-blowing machine *1* is fed. The blown bulbs are carried by the conveyer *2* to the annealing oven (lehr) *3* (underneath the glass furnace), and by the conveyor belt *4* to the sorting room *5*, where all bulbs are inspected. *6* shears for cutting gobs. *7* chute through which gobs drop into the receiving tray on the machine. *8* mechanism ensuring uniformity of the glass stream and constancy of delivery (see fig. 12).

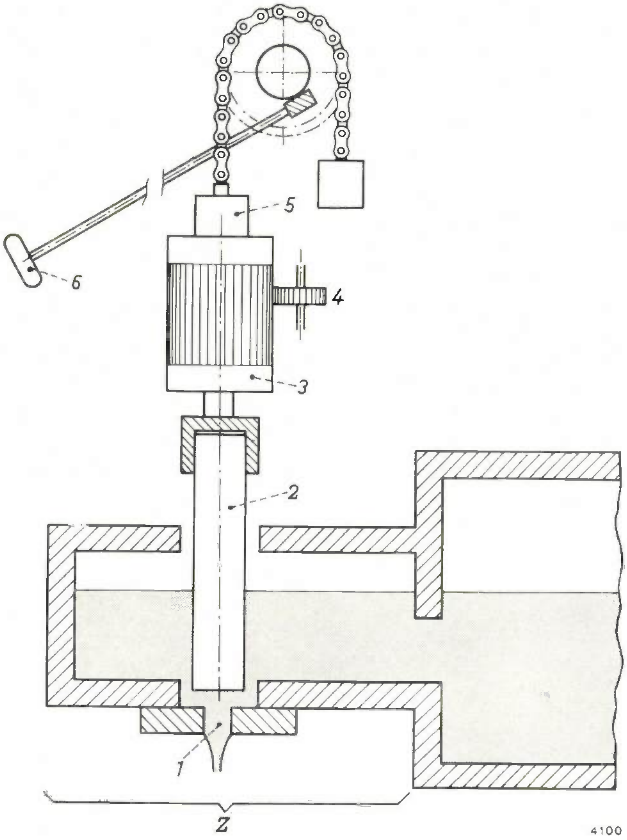


Fig. 12. Gobs of equal size and composition are obtained by introducing a rotating ceramic spindle *2* above the orifice *1* in the floor of the feeder end *Z* of the glass furnace. The spindle is raised if the gobs are too small, and lowered if they are too large. The spindle, which is rigidly mounted to the bushing *3*, is set in rotation by the pinion *4*, whose teeth engage in those around the bushing. Height variation is made possible by the long teeth cut in the bushing. The whole assembly is mounted to a non-rotating bushing *5* suspended from a chain which passes over a sprocket wheel, operated via a worm transmission by the hand wheel *6*.

on the valve cross-beams, the valves themselves, the slide carrying the gob trays, and the halves of the split blow-moulds. Where necessary the cams can also be adjusted whilst the machine is turning.

At the end of this section we shall return to the performance of the 16-head machine by way of comparison with the machines now about to be discussed.

### *The ribbon machine*

Another machine which receives the glass from an orifice in the furnace floor is the so-called "ribbon machine", developed by the Corning Glass Works [1]. The glass is not delivered in gobs, but flows from the furnace in a continuous stream and passes between two water-cooled rollers, which produce a ribbon from it. One of the two rollers is plain, the other has circular recesses or pockets. As a result the ribbon shows a series of regularly spaced protrusions, giving it rather the appearance of a strip of detonating caps for a child's pistol. Each of these protrusions comes exactly above a hole in a continuously moving metal conveyor belt, which carries along the ribbon on a horizontal plane, and the glass begins to sag down through each hole and assume a bulb shape. The ribbon now passes under a series of blow-heads, each of which is centred above the holes (the machine has scores of such heads, which are mounted on an endless belt and are thus able to travel along some way with the ribbon). As in the previous machine, air is puffed into the parisons to promote the elongation.

[1] A description of this machine is given by F. V. Tooley, Handbook of Glass Manufacture, Ogden, New York 1953, pages 356 and 386.

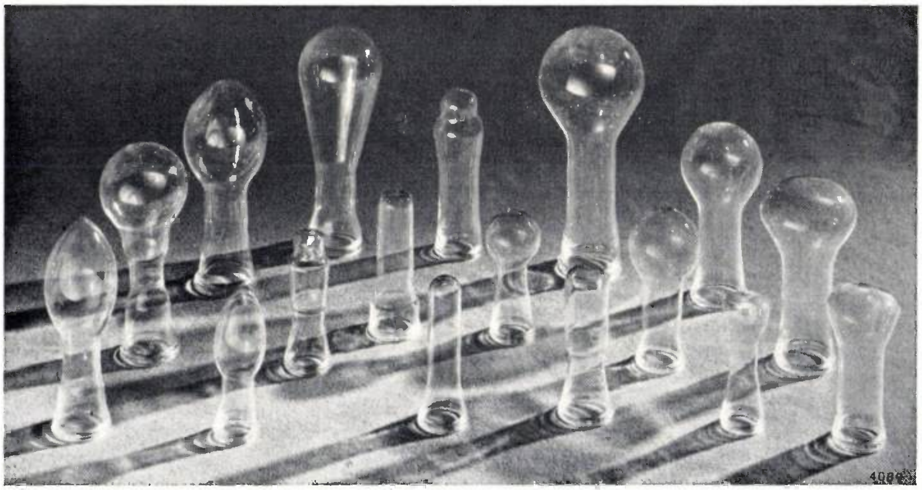


Fig. 13. Small selection from the many and various kinds of bulbs that can be made on the 16-head machine. Apart from the diameters, all bulbs have the same shape of rim, where they were held by the blanking and blowing heads.

Next, the parisons are enclosed in blow-moulds (a series of which are also mounted on an endless belt, this time under the ribbon) and the bulb is blown. After the moulds have opened again, the bulbs are allowed to cool somewhat and are finally broken off from the ribbon. Ribbon machines, then, always involve a certain loss of glass. Their production capacity, however, may be very high; the largest machines of this type produce some 60 000 bulbs an hour with a diameter of 60 mm (i.e. 1000 per minute or more than 1.4 million per 24 hours) or about 100 000 smaller bulbs (1600 per minute or 2.3 million per 24 hours).

*The vacuum-and-blowing machine*

The third kind of machines to be discussed in this section uses the so-called vacuum-and-blowing process, the portions of molten glass being gathered from the surface of the tank by suction. These machines are equipped with a swinging arm (or ram) which introduces the gathering cup into the furnace. As soon as the cup has gathered the correct volume of glass, it is swung outwards by the arm, automatically shearing off the tail of glass following it. The blank is then transferred to one of a series of blowpipes on a turntable, the blowpipe being at that moment vertical. Since these machines are also in continuous rotation, the arm must move around with the table when transferring the blank, which is of course a technical complication. After the transfer the arm turns back to its initial position, the blank cup is swung into the furnace again, and the process is repeated.

As soon as a glass gob is dropped on to a blowpipe, jaws at the end of the pipe close around it. These are so shaped and situated that, when the blowpipe is in the "blowing position", i.e. directed vertically downwards, the glass is held at the edge. As in the other machines discussed, a puffing operation now follows, and the parison is finally blown to shape in a split blow-mould mounted under each blow-head. This being done, the blow-mould opens and the bulb is then released by the jaws. The bulb now drops on to a conveyor belt and is carried to a burn-off machine where the thick glass edge held by the jaws (the moil) is removed. In these machines, too, a certain amount of glass is lost.

Bulb-making machines of the vacuum-and-blowing type exist in 6-head and 8-head versions, and may have one, two or four blowpipes per head [2]; a 6-head machine with four blowpipes per head thus has altogether 24 blowpipes and delivers 24 bulbs per revolution. According to the number of blowpipes per head, there are one-, two- or fourfold ram arms.

The production capacity of these machines is roughly between 3000 and 6000 bulbs an hour, which is thus of the same order as that of the Philips 16-head machine.

*Comparison of the various machines*

To conclude this section on machines for working molten glass, we shall comment briefly on the possibilities and limitations of the various types of machine. We shall be particularly concerned with: 1) the production capacity, 2) the glass losses, 3) the versatility of the machines as regards diversity of products.

As regards production capacity, we have seen that the gob-fed machines are roughly equivalent. Their minimum time of revolution is determined by the time needed to make one bulb. The production of a machine can therefore only be stepped up by increasing the number of blow-heads. This is possible only to a limited extent, however, otherwise the machines would be unmanageably large. A limit is also set to the speed at which portions of glass are taken over from the furnace — a speed that increases in proportion to the number of blow-heads. Since the temperature and hence the viscosity of the glass are confined to specified limits, it would be necessary to make the orifice wider. The length of the severed gobs would then soon become too small in proportion to the thickness of the glass stream. A better solution is to use, say, two orifices and two pairs of shears, but this presents serious constructional difficulties. If a very high production capacity is required, the ribbon machine is therefore the appropriate type, its capacity, as we have seen, being a factor of 10 greater.

As regards glass losses, these have been shown to be zero in the Philips machine. In the other machines the usefully employed glass amounts on an average to roughly 50% of the total processed quantity, and very much less where small bulbs are made, e.g. of 15 grammes. It should be added, however, that bulbs as small as this cannot be produced entirely without loss of glass even on the Philips machine. These machines are not suitable for handling extremely small portions of glass, which cool down too quickly to be worked.

With regard to versatility, we have seen that the Philips machine can not only make bulbs of different types at the same time (given the same gob weight) but that the blow-mould can also be changed very quickly.

The considerable versatility of the Philips machine

[2] The single-arm machines are known as Westlake machines, those with two arms as Ohio and those with four arms as Ivanhoe machines. The Westlake machine is described in the book mentioned under reference [1]), and the Ivanhoe machine by W. S. Turner, J. Soc. Glass Techn. 13, 393,1929.

described, and its low glass consumption, make it possible to equip a bulb factory with them which can economically turn out any kind of bulb in almost exactly the quantities required by a producer of electric lamps and radio valves. A glass factory thus equipped is able to modify its production programme at virtually a moment's notice, and need therefore hold only small stocks. Since the production units are relatively small (furnace-machine-lehr-conveyor belt; see fig. 11), the factory space taken up is also modest. The least versatile process is the one employing the ribbon machine. This machine is most suitable for the production of very large runs of bulbs of the same type. As a rule, their products cannot be directly assembled and finished on the same premises, and this means large stocks and considerable storage space, both in the glass works and in the lamp or valve factories.

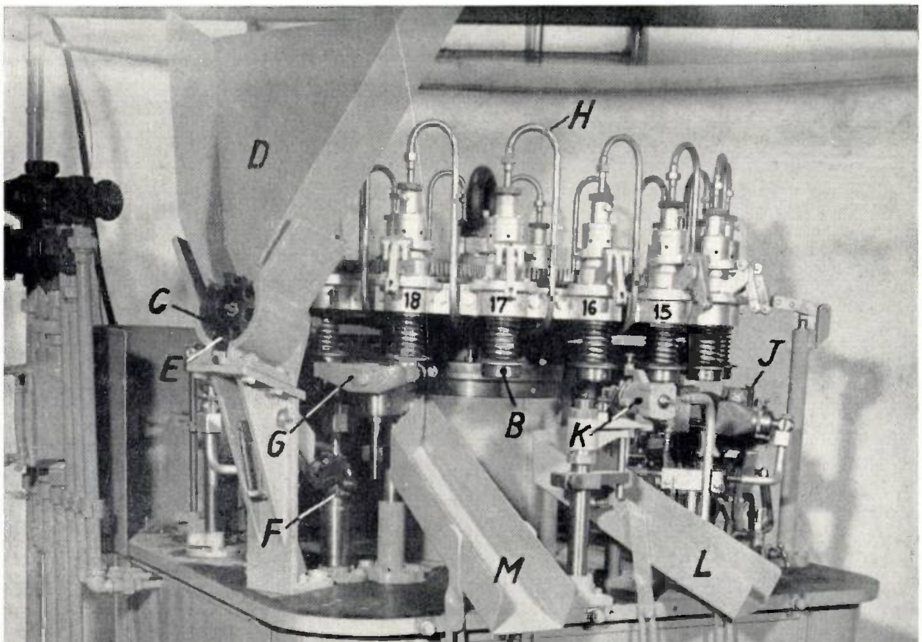


Fig. 14. Eighteen-head machine for making bulbs from short sections of glass tubing. *A* housing for drive mechanism. *B* one of the 18 chucks, mounted on a horizontal, intermittently rotating turntable. *C* sectored rotor which, turning synchronously with the turntable, delivers tube lengths, one by one, from the magazine *D* into the chute *E*. *F* glass-inserting device. *G* chuck opener. The production cycle begins for each chuck at the position *1*, and ends at position *18*. *H* one of the air lines connected to each chuck for blowing the bulbs. *J* blow-mould. *K* shearing device. *L* chute into which the sheared bulbs fall. *M* chute for removing the "moil".

**Machines for making bulbs from glass tubing**

As mentioned above, machines designed for working molten glass cannot produce very small bulbs without involving a substantial loss of glass; small bulbs are usually made mechanically from glass tubing. In principle the process is as follows. A length of tubing of, say, one metre, the diameter of which is roughly equal to the neck of the bulbs to be produced, is sealed at one end and then, while it is rotating, heated over a certain length to a temperature at which it can be worked. When this temperature is reached, a blow-mould is closed around the heated part and air is blown into the tube. After the bulb thus formed has cooled sufficiently, the blow-mould opens and the bulb is severed from the tube. It is also frequently the practice to start from small lengths of tubing, which have been cut beforehand to the length needed for making one bulb.

It is evident that the rudimentary form of blowing technique involved here can only be applied if the maximum diameter of the bulb is not much greater than that of the tube. Otherwise, bulbs would be obtained whose wall at the position of the widest diameter would be much too thin. It is not possible to get around this difficulty by the use of tubing having a thicker wall or larger diameter, since the permissible diameter and wall thickness of the neck are rather narrowly restricted by the sealing technique used in the mechanical production of electric lamps. We shall return to this at the end of the article.

Before dealing with the latest machines for working glass tubing, we shall discuss in broad lines the design and operation of such a machine in the context of an 18-head machine for working short lengths of tubing, which can only produce bulbs whose bulb diameter is not much greater than the neck diameter.

### *An 18-head machine for working short lengths of glass tubing*

In the Eindhoven 18-head machine [3]), as in all machines for working glass tubing, a horizontal turntable which rotates about a vertical axis and carries round its periphery a number of chucks (in this case 18) is situated some 30 or 40 centimetres above the housing for the drive mechanism. Each of these chucks contains one of the lengths of tubing to be worked, mounted vertically. The turntable does not rotate continuously but in 18 equal steps, so that after each step a particular chuck occupies the position of its predecessor. The chucks are in rotation, except in those positions where they are required to be stationary. The space between the top of the housing referred to and the turntable contains the burners and other equipment which together produce the bulb. A general view of the machine is to be seen in *fig. 14*, in which *A* is the housing for the drive mechanism and *B* one of the chucks.

The operations undergone by a piece of glass tubing in the machine are the following. First of all it is taken by the sectored rotor *C* from the magazine *D* and propelled into the chute *E*. It is taken from the chute by the insertion device *F* (see also fig. 18), which inserts the glass from under into the chuck in position *1* (on left of turntable in fig. 14). In this operation, the chuck opener *G* rises and opens the chuck to receive the glass. The tube is now heated just above its base for a period corresponding to several turntable positions. The unheated part is then torn off by claws on the end of a vertically reciprocating spindle (see *fig. 15*). The tube thereby collapses and its lower end is reheated, the bottom being pushed in slightly by a rising rod to promote a neat sealed end. After further reheating at various

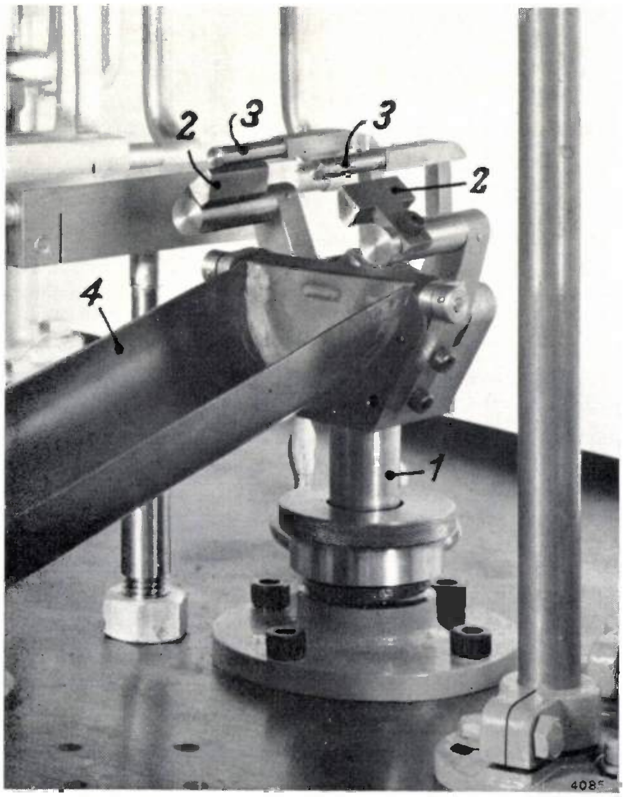


Fig. 15. The part of a glass-tubing machine where the tube is closed from underneath. *1* vertical reciprocating rod surmounted by tear-off tongs. *2* jaws of tongs. *3* stationary pins which, as the tube moves to the next station, break off the glass thread produced by the tear-off. *4* chute for discharging the torn-off piece of tubing.

[3]) This machine was developed from a similar 12-head machine of Osram GmbH.

stations the tube reaches the (open) blow-mould. The mould closes around the tube and air is blown in (*fig. 16*). In the three following stages the bulb has time to cool, and finally it is severed from the upper portion (see *K* in fig. 14). It is discharged from the machine through the chute *L*. Upon reaching station *18*, the chuck holding the remaining portion of tube is opened by *G* and the tube drops into the chute *M*. The chuck is now ready to receive a new charge in position *1*.

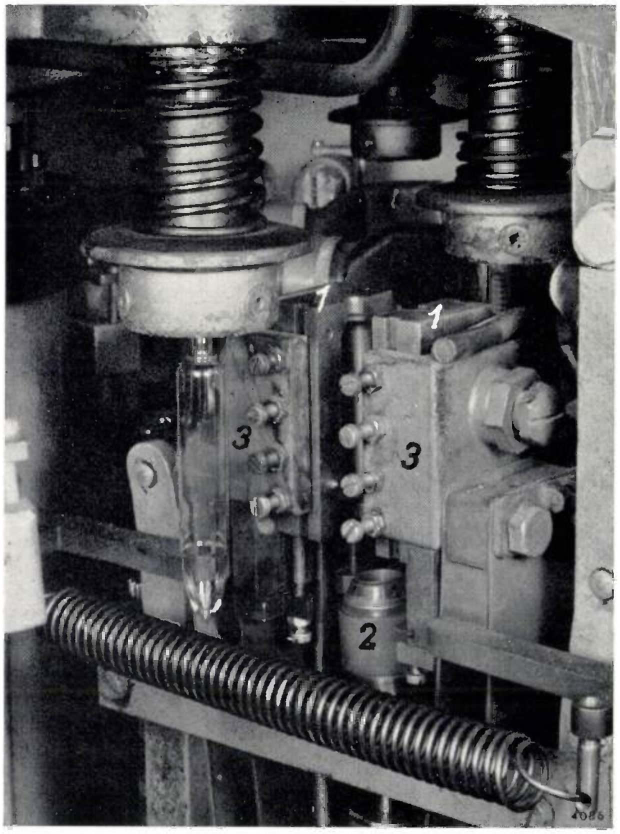


Fig. 16. Part of the 18-head machine showing a blow-mould. The mould shown here consists of three sections, i.e. two semi-cylindrical parts *1* which move towards each other horizontally to close, and a rising base *2*. By unscrewing the bolts, the mould halves can be taken from the holders *3* and replaced by others.

The bulb that has now left the machine is not yet finished. The sharpness of the edges has to be removed by a fire-polishing process, and sometimes the end of the neck must be widened, tapered or flared, by a process known as flanging. These operations, like the conveyance of the bulbs from the blowing machine to the fire-polishing and flanging machines, may also be fully mechanized.

The method of producing the puff of air for blowing the bulb is illustrated in *fig. 17a*. Here *1* is the glass tube, the top of which is pressed into the flared end *2* of a bushing *3*, located above the jaws *4* of the chuck (all of which rotates together). Fitted in the flared top end *5* of the bushing is the nozzle of the air line *6* (*H* in fig. 14). This is effected in such a way as to leave a gap through which air can leak away during the blowing process. Varying the width of the gap provides a simple means of controlling the maximum pressure of the air.

It will be noted that the number of stations of the machine in a complete cycle — and thus the number of chucks — must at least be equal to the number of operations to be performed. Since the time spent at each station is obviously the same, protracted operations (particularly most heating processes) can better be distributed over several stations. Consequently the total number of stations is in practice two or three times greater than the actual number of operations.

*Some comments on machines for working longer lengths of tubing*

By suitably shaping the end of the line *6* (fig. 17*a*), the air-feed system described can also be usefully applied in machines for making longer sections of tubing. The end of the air line (see fig. 17*b*) is then contained *inside* the glass tube (which in this case projects some way above the chuck). Provided the sections are not unduly long, so that they can still be pushed into the chuck from underneath, a rigid

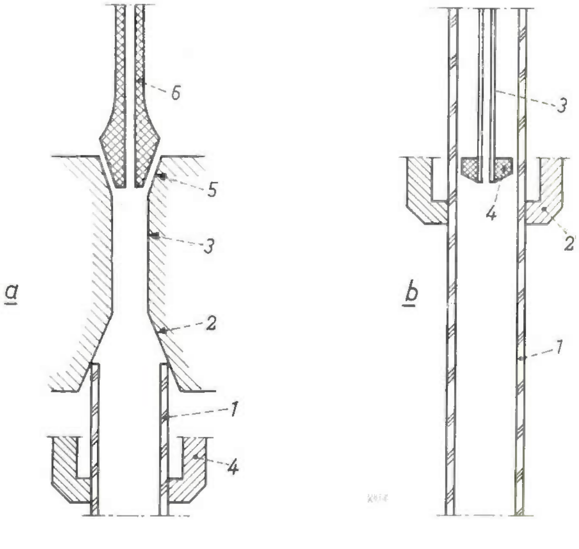
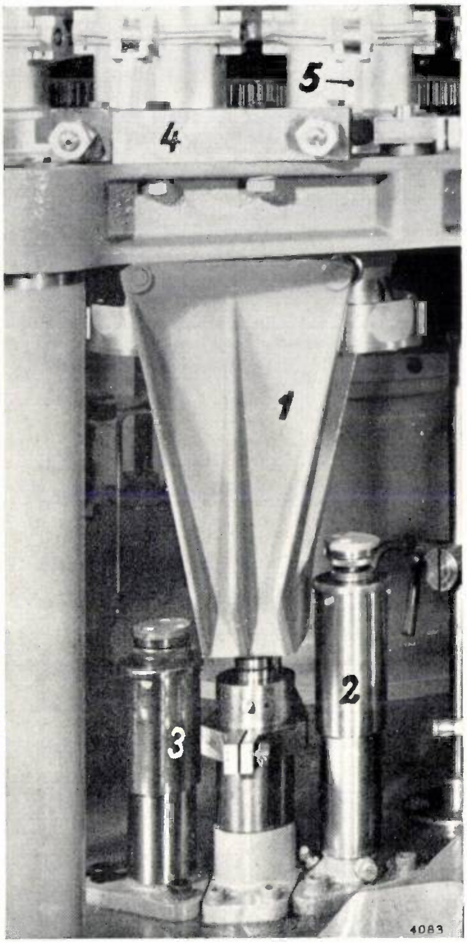


Fig. 17. Air feed for blowing the bulb in a glass-tubing machine. *a*) Method of feed in machines working short pieces of tubing. *1* top of glass tubing, pressed firmly into the flared bottom end *2* of the bushing *3* which, together with the clamp *4*, forms part of a chuck. *5* flared top end of bushing. *6* air line inserted in *3* out allowing for some air leakage. *b*) Air feed system in machines for working long sections of tubing. *1* glass tubing. *2* clamp of chuck. *3* air duct with collar *4*. Between this collar and the inside wall of the tube a gap is again provided to allow for air leakage.

air line may be used; in machines that handle tubing a metre long, a hinging point must be provided near the top of the line. The mouth of the air line is always as low as possible. For machines of this kind, this construction has considerable advantages over the system whereby the air line (in this case a hose) has to be connected via a bung fitted in the end of the tubing (the Cleveland system).



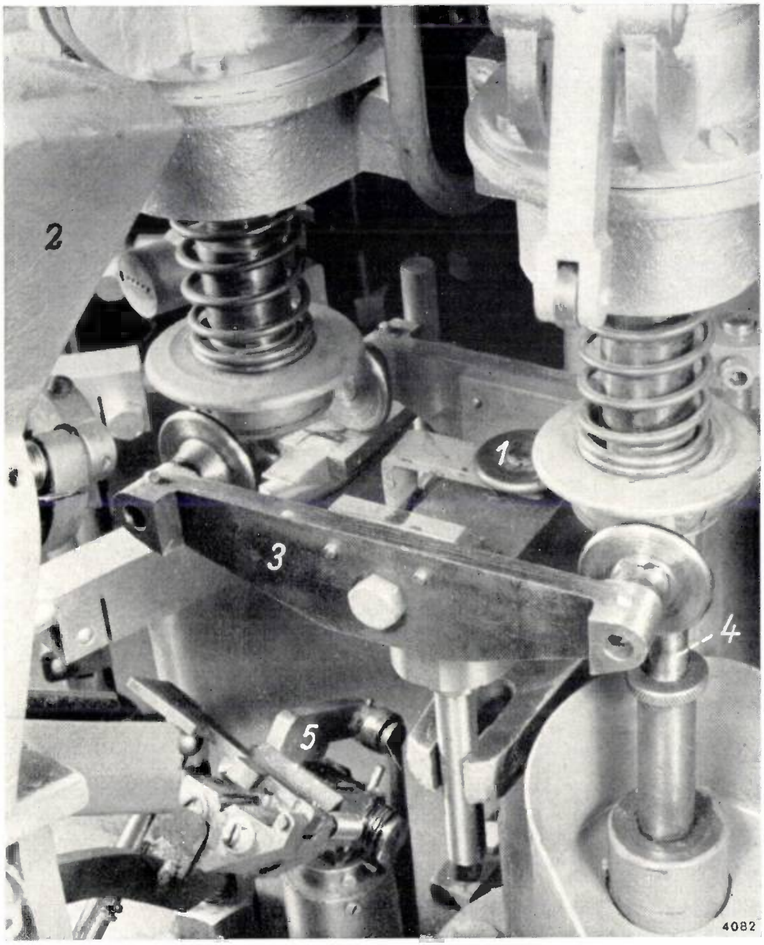


Fig. 18. *a*) Detail of 32-head machine for working long sections of glass tubing (in course of development). In the part shown, the tubes are being adjusted to the correct height. This is done here in two stages. *1* chuck opener. *2* first glass-gatherer. *3* second glass-gatherer. During the glass-gathering operations the chucks are not in rotation; this is achieved by the cam *4* actuating the lever *5*.
*b*) Detail of a machine for working lengths of tubing from which more than one bulb can be made, but which are short enough to be inserted in the chuck from underneath. These machines are equipped with a feeler device *1* which signals the moment for delivering a fresh tube from the magazine *2*. *3* chuck opener. *4* rod of glass-gatherer. *5* insertion device. (To make certain components visible, the valve heads were not in their working positions when this photograph was made.)

In the machines for working long sections, each tube has to be moved up a bit after completing one cycle in its chuck, and for this purpose the machines are equipped with a so-called gatherer. This is a small plate mounted on a vertically movable rod, which takes up the tube as the chuck is momentarily released and carries it for some distance downwards. The chuck then closes again and the machine moves on to the next position. In some cases two glass-gatherers are fitted. During this operation the tubes are not in rotation. *Fig. 18a* shows the two glass-gatherers of a 32-head machine in course of development. In machines for working moderate lengths of tubing, and where the insertion process is identical with that in the machines for handling very short lengths, the insertion mechanism is fitted with a feeler device for ascertaining whether or not a tube has to be inserted (fig. 18*b*).

### *Machines without blow-moulds*

For producing bulbs which are simple cylinders closed at one end (like many radio valves and transistor envelopes), the use of a blow-mould may be dispensed with. In this case the starting point is tubing which has the diameter and wall thickness required for the bulb. After the tube has been closed by tear-off claws as described, the closed end is given the required shape by means of a *bottom mould*. The principle of this method is illustrated in *fig. 19*. In automatic machines equipped with a bottom mould, the small amount of air needed is admitted to the tube in the same way as described in fig. 17. The presence of a gap to allow air to leak away is especially important here. A too powerful puff, or thermal expansion of the enclosed volume of air, would have disastrous consequences in the absence of a blow mould.

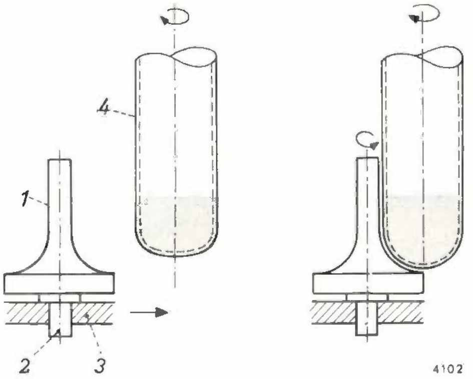


Fig. 19. Principle of making a cylindrical bulb with a "bottom mould". The roller *1*, which turns freely on a spindle *2* in the plate *3*, is slowly moved towards the rotating glass tube *4* (already closed by the tear-off tongs) until its upper end abuts against the unsoftened part of the tube (the softened part is shaded). In this position a very gentle puff of air is blown into the tube, causing the bottom to follow the shape of the now rotating roller. Finally, the roller returns to its original position (shown on the left).

### *Machines for making narrow-necked bulbs*

Narrow-necked bulbs are made from tubing whose diameter lies between that of the neck and the bulb. The required wall thickness and neck diameter are produced by the "*stretching and rolling process*" [4]). The tube is heated at the place where the neck is to be, and as soon as the glass is soft enough it is stretched a little. With the tube rotating, the heated portion is then passed between three rollers to reduce it to precisely the required diameter. The wall thickness of the rolled part which, without stretching, would be greater than that of the original tubing, can be controlled by varying the degree of elongation. The manner in which this stretching and rolling process takes place in the machine under development at Philips is represented schematically in *fig. 20*. The most characteristic features are 1) the fact that the tube is clamped only momentarily *under* the heated part for the purpose of stretching, but is otherwise freely suspended (apart from its enclosure between three centring rollers during the rolling process), and 2) the fact that the rolling process takes place in two stages, which improves the precision. Because of the fact that the tubes are not clamped from underneath, the design of the machine need not be essentially different from that of the 18-head machine described above. This leaves considerable freedom for the positioning of the burners and allows the mechanism for closing the blow-moulds to be simpler than is possible in machines equipped with two rotating turntables each carrying rotating chucks (one for the top and one for the bottom of the tubes). To enable the entire production process, including the stretching and rolling, to be carried out on a single machine, the number of stations was increased, as mentioned in fig. 20, to 32.

A view of the relevant part of the new machine is shown in *fig. 21*. Various details are explained in

[4]) This principle has long been used in machines for making ampoules, and has also been applied for making bulbs (E. Mickley and M. Thomas, Glastechn. Ber. 26, 197, 1953).

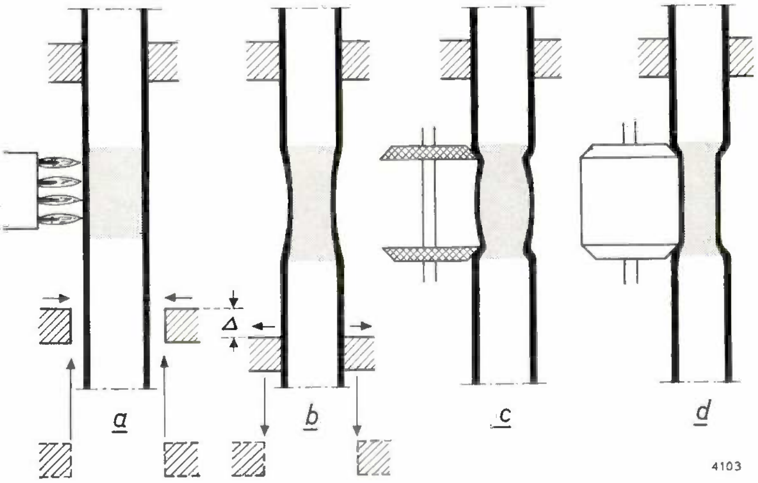


Fig. 20. Stages of the *stretching* and *rolling* process in the 32-head machine. The part to be stretched is first heated (shaded). When it is hot enough, a chuck rises, grips the tube from underneath and stretches it over the required elongation $\Delta$. The chuck then opens and drops back to its original position, enabling the tube to move to the next position (*a* and *b*). The hot part is now indented by two wheels, which accurately define the section to be rolled (*c*), and finally, possibly after reheating, it is rolled to the required diameter (*d*).

the caption. As regards the centring rollers (*8*) it should be noted that they are not driven but are set in motion by the tube itself. The tube must not be too hot at the rolling position, otherwise there is a danger of it twisting. The rollers themselves are very light and rotate very easily.

The 32-head machine differs somewhat from the other in various points of construction. The housing consists of a lower section fitted with vertical ventilation vents at the sides. The gas, air and oxygen lines, and various valve assemblies (three for each burner) are mounted here at the height of the upper section. The motors and drive mechanism are all in the lower section of the housing. All that the upper section contains, apart from various vertical shafts and rods, are hoses leading from the valves to the burners. This construction makes it possible to adjust or replace the cams of the drive mechanism without having to remove the valves (the cams are designed so that they can be replaced without taking the shaft out of its bearings). Moreover, because of the use of hoses instead of metal pipes for the connections between valves and burners, it is possible to change quickly and simply a gas-air burner, for example, into a gas-oxygen burner. The ventilation obviates the risk of explosions resulting from any slow leakages.

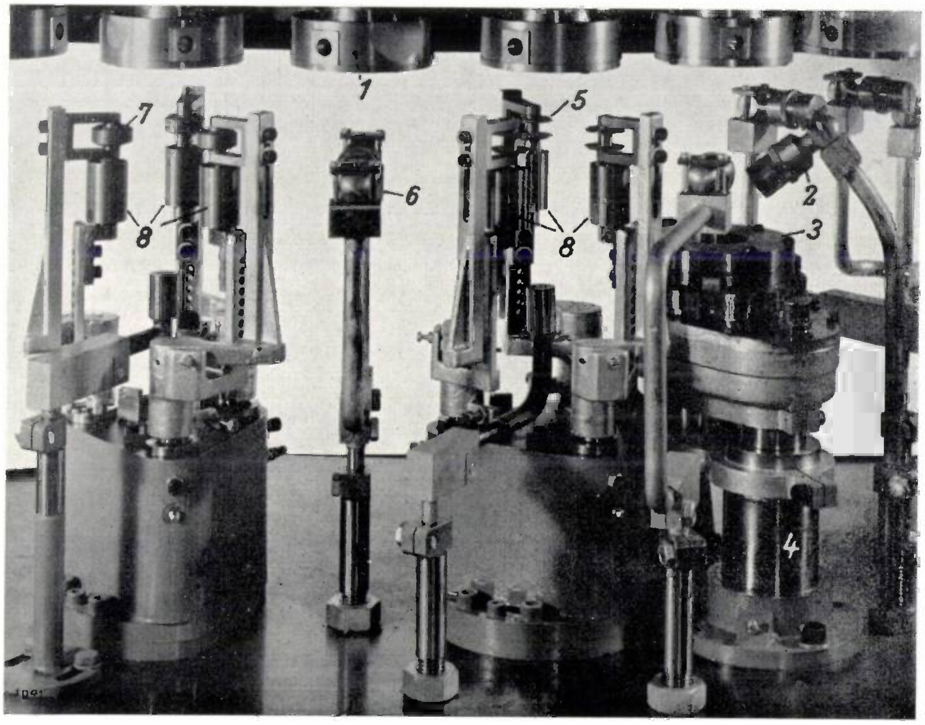


Fig. 21. Part of the 32-head glass-tubing machine where the stretching and rolling process takes place. *1* one of the six chucks visible. *2* one of the burners. *3* chuck on vertically reciprocating rod *4* for stretching. *5* indenting wheels (for the first stage; see fig. 20). *6* burners for reheating. *7* roller for second stage. During the rolling process the lower part of the tube is centred by the rollers *8*.

A selection of bulbs mechanically produced from glass tubing is to be seen in *fig. 22*.

The production capacity of the machines for working glass tubing is somewhat lower than that of the earlier discussed machines that work molten glass. Depending on the type of bulb produced, the capacity of the 18-head machine first described varies between 1000 and 1600 bulbs an hour. Roughly the same figures hold for the later-developed variants of this machine. The 32-head machine now under development has already proved itself capable of a production rate of 1600 bulbs an hour. This is expected to be improved on in the near future.

## Choice of production method

A point already mentioned in passing, and which the bulb manufacturer always has to decide, is which method of production is to be preferred. We have seen that in a few cases production by hand leads to the lowest production costs, and also that, by mechanized methods, large bulbs can on the whole be made more cheaply from molten glass and small ones more cheaply from glass tubing. In this connection, glass losses were shown to be a very important factor. Cylindrical bulbs for radio valves, for example, can be made most economically from glass tubing, the glass loss then being very small.

We shall now examine these rough indications in more detail and expand on them. Closer consideration is especially important where the best method of production is to be decided for making bulbs that are neither large nor very small, e.g. a car headlamp bulb.

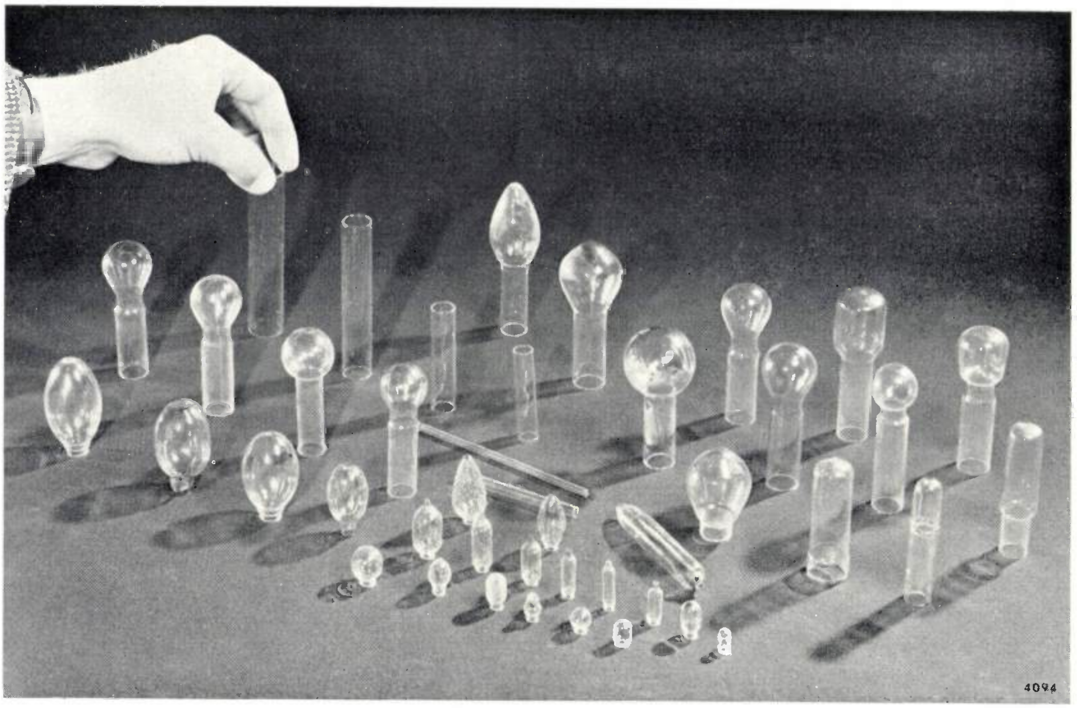


Fig. 22. A selection of bulbs mechanically produced from glass tubing.

The first factor to be considered is the size of the production run. If this is small, production from glass tubing will generally be more advantageous, provided at least that glass of the required kind and dimensions is in stock. There is then no need to start up a glass furnace for this specific purpose. If the manufacture can be combined with that of other bulbs of similar dimensions and the same type of glass, economic production is also readily possible with the Eindhoven 16-head machine for working molten glass, this machine being capable of turning out different models at the same time.

The glass loss is governed not only by the weight of the bulb but also by the shape of the neck. This is bound up with the various methods by which the mount, which carries the filament, is later to be sealed to the bulb in the lamp factory.

For the first method, called the "drop-seal" process and represented schematically in *fig. 23a*, the bulb on leaving the glassworks must have a

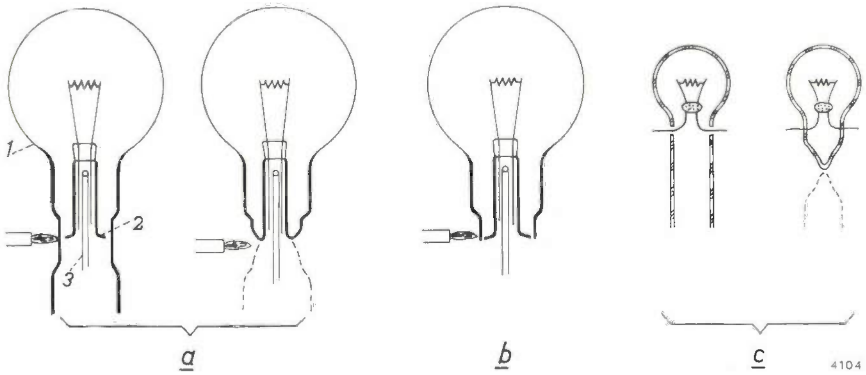


Fig. 23. The various methods by which the mount, which carries the filament, is sealed to the bulb in the lamp factory.
*a*) Drop-seal process. Here the bulb from the glassworks (left-hand sketch) has a longer neck than the lamp to be made from it. After the bulb (*1*) and the tapering bottom end (*2*) of the mount (the "flare") are sealed together, the bottom section of the bulb (dashed line in the right-hand sketch) is burnt off. The gap between flare and neck is here fairly wide. The tube *3* is the pump stem.
*b*) Rim-seal process. The bulb here is cut to the definitive length and the gap between neck and flare must be narrow.
*c*) Butt-seal process. Here too the bulb neck is short. A butt-seal is made with a tube having the same diameter as the neck; after evacuation the tube is sealed off close below the butt-seal. The lead-in wires are incorporated in the butt-seal.

considerably longer neck than the electric lamp to be made from it. The neck must also have a large enough inside diameter for it to fit with a wide clearance around the stem. After the mount, carrying the filament assembly, has been properly positioned inside the bulb, the whole is rotated and the bulb is heated around the bottom of the mount. As soon as the glass softens, the neck becomes gradually constricted and finally touches the "flare" of the mount. Further heating fuses the two together. The bottom part of the neck is then burnt off by positioning the flame somewhat lower, in which process it may be pulled with tongs or allowed to sag under its own weight.

In the "rim-seal" process the neck of the bulb leaving the glass works is just as long as required for the finished lamp, and its inside diameter is such that the flare of the mount fits it fairly exactly. The rim of the neck and of the mount are brought into alignment and sealed together (fig. 23*b*).

A favourable feature of the drop-seal method is the fact that the neck diameter may vary between fairly wide limits. The fact that the bulb wall immediately above the seal may be fairly thin — as a result of elongation — is sometimes, though not usually, a drawback. The rim-seal process produces stronger seals, but fairly strict demands are imposed on the roundness of the bulb necks, of the spread in their diameter and on the constancy of the wall thickness in the sealing zone. The same applies to the cleanliness and soundness of the edges of bulb and flare. In the rim-seal process the clearance between neck and flare is required to be small, whereas in the drop-seal process the opposite is the case, since a small clearance would result in a thick glass rim around the fusion zone and increase the risk of cracking.

The third method, called the "butt-seal" process (fig. 23*c*), also uses a short-necked bulb. A tube of the same diameter is butt-sealed to the neck and, after evacuation, burnt off a short distance below the seal. In this case the lead-in wires are fused into the joint.

Turning now to the economy of the various bulb-blowing machines, it is found that, in the lighter ranges, production from molten glass can compete longest with the glass-tubing method provided the drop-seal process is used. Apart from the difference in glass losses, production from molten glass is in the disadvantage if other methods of sealing are applied, in view of the finishing (burn-off) operation then required, which considerably increases the production costs. Moreover, the necks of bulbs produced from glass tubing better conform to the above-mentioned requirements of constant wall thickness, etc., for the rim-seal process. This is especially the case where stretching and rolling is involved.

Apart from the above considerations, which are primarily of an economic nature, the manufacturer has various other factors to take into account. For example, the choice between production from molten glass or from glass tubing is sometimes influenced by the type of glass required. In this case, the glass-tubing machines have the advantage; the mechanized production of hard-glass valves, for instance, presents difficulties if molten glass is used. The quality and finish of the bulbs are also, of course, factors that carry some weight. As regards finish, bulbs made from molten glass are definitely superior. This also applies to the smaller models, which cannot be made without some loss of glass. In so far as the strength and distribution of the wall thickness are concerned, it should be mentioned that the bulbs, both large and small, made on the Eindhoven 16-head machine are entirely satisfactory. In this respect, however, bulbs made from glass tubing are scarcely, if at all, inferior to the others if the production involves a stretching and rolling process.

**Summary.** Machines for producing lamp bulbs or valve bulbs are designed for working either molten glass or glass tubing. Of the first category the author discusses the ribbon machine, the vacuum-and-blowing machine and a machine, developed at Eindhoven, which operates without glass loss. The latter is a continuously rotating machine equipped with 16 blowing units. Each unit is capable of turning out a different model, subject only to the limitation that the amount of glass used for each is the same. The production capacity is about 4000 an hour. Since very small gobs cool too quickly to be worked with such machines, small bulbs are made from glass tubing. In this category an 18-head machine is discussed, also developed at Eindhoven, which handles short lengths of tubing. Some variants are then described, including machines that work long lengths of tubing and machines using a bottom mould instead of a blow-mould. Their production capacity ranges from 1000 to 1600 bulbs an hour. Mention is made of the production of narrow-necked bulbs from glass tubing by a process of stretching and rolling. The last section is devoted to the economic and other factors governing the choice of machine.

# TESTING OF MATERIALS FOR GLASS-TO-METAL SEALS BY MEANS OF STRESS BIREFRINGENCE

by J. de VRIES *).

536.413.082.532:666.1.037.5

One of the principal causes of stresses in glass-to-metal, glass-to-ceramic or glass-to-glass seals is the difference in thermal expansion between the sealed materials. Since the stresses must not become so high as to cause the glass to crack, it is important to have some method of checking beforehand the expansion of the materials to be used, in order to avoid difficulties during production. In many laboratories and factories it is still common practice to measure the expansion directly as a function of temperature with a dilatometer and to compare the resultant curves. A single expansion coefficient, derived from the dilatometer curve, is sometimes used to specify a material in a simpler (though less adequate) way.

Now that the marked increase in industrial production in Europe is giving rise to closer international cooperation in the field of glass processing, it is becoming more and more apparent that the dilatometric method is not accurate enough. It therefore seems desirable to draw attention once again to the stress-birefringence method of measuring the relative expansion. This method, in which the stress is deduced from the birefringence (double refraction) in test seals of standard shape, has been applied at Philips for more than twenty years, and was described in this journal in 1947 [1]). To avoid unnecessary repetition, we shall lay particular emphasis here on glass-to-*metal* seals, which were only touched on in passing in the 1947 article.

## The stress-birefringence method

The birefringencc is measured on a test seal of one of the shapes illustrated in *fig. 1*. In order to compare glass samples, or samples of glass and ceramic, small plates of these materials are fused together (fig. 1*a*). For the fusion of sheet metal and glass the so-called plate seal in fig. 1*b* is used; the cylindrical form in fig. 1*c*, the bead scal, is appropriate where metal wire is concerned. If the diameter of the wire is greater than 3 mm (in which case it can better be referred to as a rod), a flat edge is ground on the rod to which a plate seal is then made. Various examples of such test seals are shown in *fig. 2*.

If the fused materials have different coefficients of expansion, stresses will arise upon cooling. Cooling the samples in an annealing oven ensures that the residual stresses are due solely to the difference in the expansion of the two materials. It is well known that stresses in glass cause double refraction [2]). The intensity of the birefringence, expressed in optical path difference per cm thickness of the birefringent material, is a measure of the stress, and hence of the difference in expansion. The optical path difference is measured in mμ with the aid of polarized light, the intensity of the birefringence thus being given in mμ/cm.

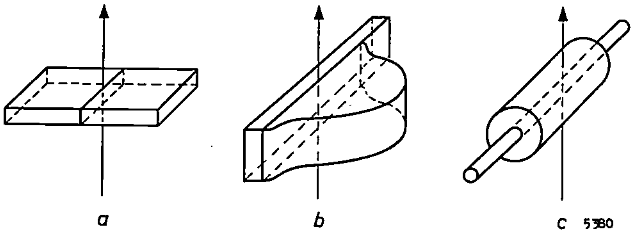


Fig. 1. Standard shapes of test seals for determining expansion differences by the stress-birefringence method. Polarized light is passed through in the direction of the arrows. Shape (*a*) serves for glass joints and glass-to-ceramic seals, sometimes also for metal and glass, (*b*) for plate-type glass-to-metal seals, and (*c*) for the fusion of metal wire and glass (bead seals).

It may be useful to recall briefly the principle of birefringence. If a plane-polarized electromagnetic wave falls perpendicularly on a birefringent plate, the wave is split into two components polarized in directions at right angles to one another. These components are distinguished by a somewhat different velocity of propagation, so that the original wave front splits into two wave fronts. After passing through the plate, the two wave fronts show an optical path difference which is proportional to the thickness of the plate, provided the birefringence in the plate is constant over the entire thickness.

In test seals using a glass plate (fig. 1*a* and *b*) the light is incident perpendicular to the plate and the optical path difference is measured on the light ray passing close to the interface, where the stress — and optical path difference — is greatest. The latter is then divided by the thickness of the plate.

When comparing the results obtained on various plate seals (fig. 1*b*), it must be remembered that the stress for a given glass-metal combination is smaller the thinner the metal. This is duc to the fact that a thin piece of metal can more easily deform and thus relax the stresses. The effect is noticeable in metal

*) Glass Division, Eindhoven.

[1]) A. A. Padmos and J. de Vries, Stresses in glass and their measurement, Philips tech. Rev. 9, 277-284, 1947/48.

[2]) See e.g. J. Partridge, Glass-to-metal seals, Society of Glass Technology, Sheffield 1949.

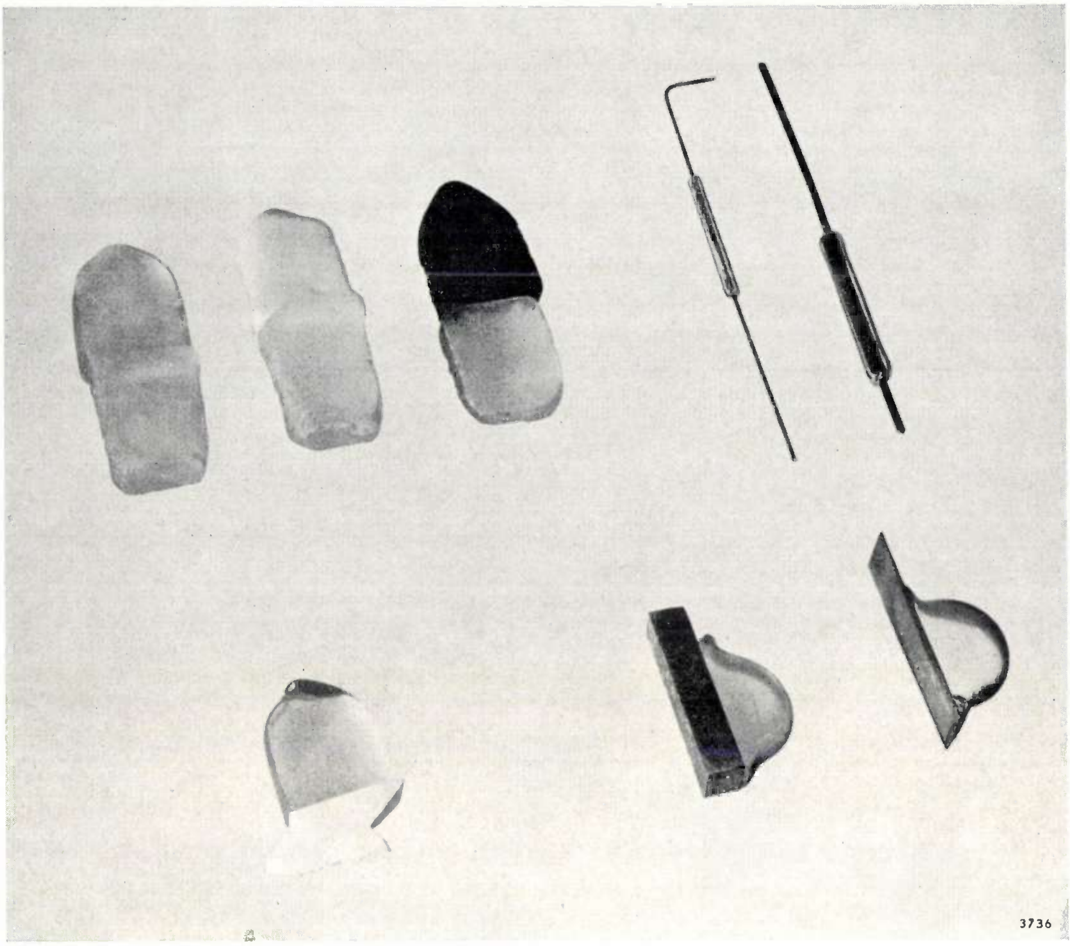


Fig. 2. Various test seals. Top left: glass-to-glass, viz. translucent, opalescent and black glass sealed to translucent glass. Top right: bead seals. Bottom left: glass-to-ceramic. Bottom right: plate seals. Roughly actual size.

plates thinner than about 0.5 mm. Results for thin plates are therefore reduced to those for thick plates by applying corrections established experimentally.

In bead seals the light is directed on to the seal perpendicular to the metal wire, and the optical path difference is measured for a light ray tangential to the metal wire. In this measurement the bead is placed in a glass cell containing a liquid which has practically the same refractive index as the glass. This avoids the complication of reflection and refraction at the surface of the cylinder. The measured optical path difference is divided by the distance covered by the light ray through the bead (i.e. by the distance $A_1A_2$ in *fig. 3*) and a result is again obtained in m$\mu$/cm.

In bead seals the result found for a given combination of glass and metal depends on the ratio between the diameters of bead and wire. To obtain comparable figures it is therefore necessary to reduce the results to one particular ratio; the ratio 3 is commonly used. For this purpose, Hull and Burger have worked out a graph on the basis of a theory put forward by Poritsky [3] [4]). For example, where the measurement concerns a bead seal in which the ratio of the diameters is 4, it follows from this graph

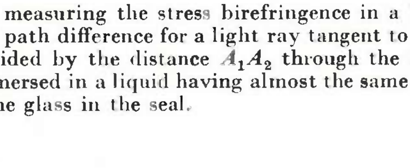


Fig. 3. For measuring the stress birefringence in a bead seal, the optical path difference for a light ray tangent to the metal wire is divided by the distance $A_1A_2$ through the bead. The bead is immersed in a liquid having almost the same refractive index as the glass in the seal.

[3]) A. W. Hull and E. E. Burger, Glass-to-metal seals, Physics **5**, 384-405, 1934.

[4]) H. Poritsky, Analysis of thermal stresses in sealed cylinders and the effect of viscous flow during anneal, Physics **5**, 406-411, 1934.

(*fig. 4*) that the result must be divided by 0.7 in order to reduce it to a ratio of 3 [5]).

It is sometimes necessary to compare results obtained on plate and bead seals. Obviously, a direct comparison is not possible, and in this case, too, correction factors must be applied that are determined by experiment.

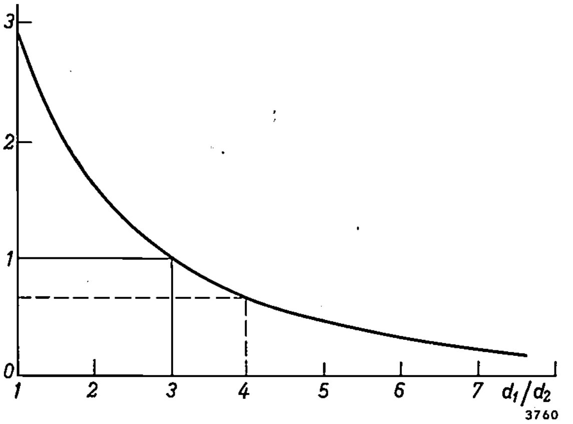


Fig. 4. Hull and Burger correction graph [3]) for comparing measurements on bead seals having different ratios of bead diameter $d_1$ and wire diameter $d_2$. A result obtained for $d_1/d_2 = 4$ is divided by 0.7 to reduce it to the result for $d_1/d_2 = 3$.

### Standard glasses

An important advance has been the introduction of standard glasses. A particular batch of glass is reserved for test purposes and thereby ranked as a standard glass with which other materials are compared by means of the test seals shown in fig. 1. In this way, expansion values are found which can be compared one with the other.

For any given type of glass, i.e. soda-lime glasses, lead glasses, borosilicate glasses, etc., a standard glass of the same type is required. The various metals appropriate for sealing to each particular type of glass are tested with respect to the standard glass of the same type. Standard glasses are given code numbers, e.g. 28/4, and the expansion of the investigated materials is expressed for brevity as, for example, T50 m$\mu$/cm in 28/4 (T means a tensile stress in the standard glass), or C100 m$\mu$/cm in 28/4 (C means compressive stress in the standard glass).

A distinction is made between *basic* standard and *secondary* standard glasses. A secondary standard is for regular use. When the supply is exhausted, glass from a fresh batch is taken as the secondary standard. The secondary standards, which are all given their own code numbers, are chosen to compare as closely as possible with the basic standard. The latter is used only for the purpose of selecting fresh secondary standards and for determining any corrections that may be necessary to the measurements made with the secondary standards. In this way, stress values are obtained which can immediately be compared one with the other, even though years may have elapsed between the tests, for all values are expressed in stresses relative to the same basic standard.

### Comparison of the stress-birefringence and dilatometric methods

The stress-birefringence method gets around one principal drawback of the dilatometric method, which is that a relatively *small* difference between two quantities has to be deduced by measuring the quantities themselves and subtracting one from the other. The stress-birefringence method is therefore much more accurate and, moreover, quicker and simpler. A theoretical advantage of the dilatometric method is that the results are in principle independent of the method of measurement. The results obtained in the different laboratories should thus be directly comparable, whereas the results of the stress-birefringence method are comparable only in so far as identically made test seals and identical standard glasses are used. In practice, however, the extent to which dilatometric results can be compared has proved to be very disappointing, so that the advantage is really illusory.

An advantage of the stress-birefringence method, not yet mentioned, is its close relevance to practice, i.e. to actual seals. For tracing the causes of rejects, and in the search for better methods of production and control, the stress-birefringence measurements provide much more assistance than dilatometric measurements. An example will be discussed at the end of this article.

The wider adoption of the stress-birefringence method calls for close cooperation between all parties concerned. Such cooperation, on an international level, promoted by the Physics Laboratory of the Philips Glass Development Centre, is gradually gaining ground.

### Expansion tolerances for glass and metal

Suppose that a glass $G$ and a metal $M$ are sealed together in a certain product. As regards their expansion, $G$ and $M$ are of course chosen to match, but there is always a certain spread in the values of

[5]) In certain cases, where the delayed elasticity (elastic aftereffect) of the glass plays a part, the Hull and Burger graph leads to false conclusions. This subject is dealt with in detail by A. L. Zijlstra and A. M. Kruithof, L'élasticité différée d'un verre borosilicate et son influence sur la formation de contraintes dans des scellements de ce verre, Verres et Réfractaires **12**, 127-141, 1958.

the actual expansion coefficients. In unfavourable cases, i.e. where the expansion of the glass is on the high side and that of the metal on the low side — or vice versa — the glass may therefore be in danger of cracking. The question to be decided is what tolerances must be specified for the expansion of glass and metal separately in order to be sure of a good joint.

To answer this question, it is necessary first of all to know what *differences* in expansion are permissible. In a few cases it is possible to discover this systematically by making a test run of the product concerned using materials with increasing differences in expansion. That is the ideal method, but in the factory it is seldom practicable. As a rule a reject analysis has to be resorted to. The glass and metal of a cracked specimen are taken and used to make a test seal. Depending on the shape of the metal (sheet or wire), this may be a plate or a bead seal. The measured difference in expansion is evidently impermissible for the product in question. By investigating a series of cracked specimens in this way, an assessment can be made of the differences in expansion (expressed in mμ/cm) that may be tolerated for the product. (At this stage, then, a standard glass is not used.)

The next step is to draw a graph in the following way. A test seal is made of a sample $G_1$ of glass $G$, to the corresponding standard glass $S$. Let the birefringence measured in $S$ be $g_1$ mμ/cm. A test seal (plate or bead) is also made of $G_1$ to a sample $M_1$ of the metal $M$. Let the birefringence measured in $G_1$ be $a_1$ mμ/cm. The points $g_1$ and $a_1$ are now plotted on the abscissa and ordinate, respectively, of a graph, defining the point $A$ (see *fig. 5*). In the same way we deal with a second glass sample $G_2$ which shows a slightly different expansion, e.g. because it comes from a different batch, and we find point $B$. (In both cases the same metal sample $M_1$ is used.) Other glass samples yield points which all lie virtually on the straight line $AB$, so that in principle we need only establish two points. The straight line $AB$ is thus characteristic of the metal sample $M_1$. Other metal samples (showing a somewhat different expansion) yield lines parallel to $AB$. The points at which these lines intersect the vertical axis obviously represent the stresses to which these metal samples give rise in the standard glass.

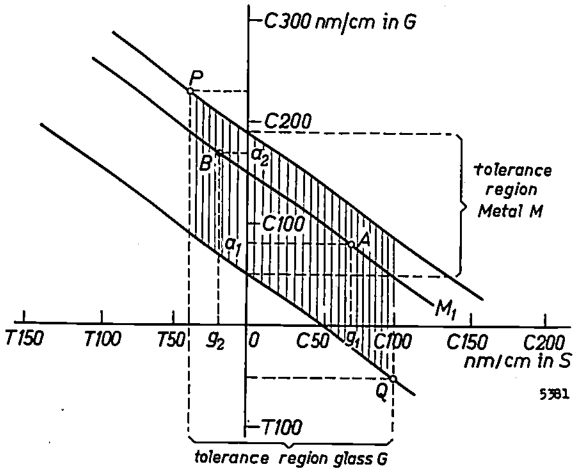


Fig. 5. Graphic determination of the expansion tolerances for glass in a glass-to-metal seal where the tolerances for the metal are given, or vice versa. (1nm = 1 nanometer = 1mμ.)

We can now proceed to specify the expansion tolerances for the glass and the metal separately. In this connection we have some freedom of choice, in that the tolerance for the glass will be wider the closer we make the tolerance for the metal, and vice versa. Assume for example that it is desirable for certain reasons to specify for the metal an expansion corresponding to between C50 and C190 in $S$. We now draw lines in fig. 5 parallel to $AB$ through the points C50 and C190 on the ordinate (where the birefringence in $S$ is zero, i.e. the expansion of the glass $G$ is identical to that of $S$). Suppose that reject analysis has shown the permissiblc differences in expansion between $G$ and $M$ to lie between C220 and T50 in a standard test seal. Horizontal lines corresponding to these values yield the points of intersection $P$ and $Q$ (see fig. 5). The abscissae of $P$ and $Q$ then represent the required tolerances for the glass $G$, namely T43 and C100, or, in round figures, T50 and C100 in $S$.

## Quality control based on stress-birefringence measurements

To conclude this article we shall consider an example of the assistance afforded by stress-birefringence measurements in tracing and removing the causes of fracture. The example in question concerns the transmitting tube shown in *fig. 6*. This contains a fernico ring $A$, to one side of which the glass envelope $B$ is sealed, and to the other the glass cap $C$. Under certain conditions during operation the ring $A$ may be heated by eddy currents to above 200 °C, as a result of which repeated cracking occurred in the neighbourhood of the seal between the envelope $B$ and the ring $A$. The plane of fracture was always found to lie roughly parallel to the plane of the seal.

Stress-birefringence measurements of the expansion difference as a function of temperature were made on plate-type standard seals between glass and metal taken from tubes which had fractured. Curves such as curve *1* in *fig. 7* were obtained. For recording "polarimeter" curves of this kind, the test

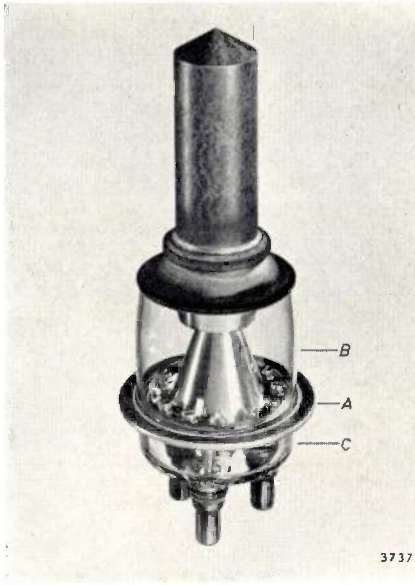


Fig. 6. A transmitting tube where the cause of fracture near the seal between the glass envelope *B* and the fernico ring *A* was traced and removed with the aid of stress-birefringence measurements on standard test seals. *C* is a glass cap sealed to the other side of ring *A*. Height of the tube approx. 20 cm.

seal is heated in a small oven specially designed for the purpose. The remarkable shape of the curves is due to the non-uniform expansion of both glass and metal [6]).

It can be seen that a maximum tensile stress prevails in the test seal at 200 °C. This does not imply that the same stress will be present in the actual tube seal; the conditions in the actual seal are never exactly the same as in the test seal, and this again has its effect on the stresses. Nevertheless, the peak at 200 °C in the polarimeter curve does give reason to suppose that excessive expansion of the metal in relation to the glass is the cause of the trouble experienced. Since the heat is generated in the metal ring, this will be hotter than the glass, and this will accentuate the higher expansion of the metal. Suppose that at room temperature there are virtually no stresses present (*fig. 8a*). At higher temperatures a deformation as sketched in fig. 8*b* will then occur. This would be accompanied by a tangential tensile stress together with an axial compressive stress on the inside, and an axial tensile stress on the outside of the fusion zone. If the latter stress is responsible for the rupture, it should be possible to remove the trouble by using only such combinations of metal and glass for which the polarimeter curve is somewhat higher, e.g. such as curve *2* in fig. 7. It was found that this measure indeed produced the expected result. At room temperature a tangential compressive stress will now prevail in the glass, together with an axial compressive stress on the outside of the fusion zone and an axial tensile stress on the inside. When the tube is switched on, this tensile stress will decrease, whilst the compressive stress on the outside will only change to a tensile stress at elevated temperatures; this stress, however, can now no longer assume a value capable of causing fracture. At room temperature the glass seal is now pre-stressed. The prestress must not, however, be unduly large, otherwise the tensile stress prevailing on the inside of the seal at room temperature and below, e.g. during transport, may prove dangerous. The limits within which the polarimeter curves must lie have been established by experiment.

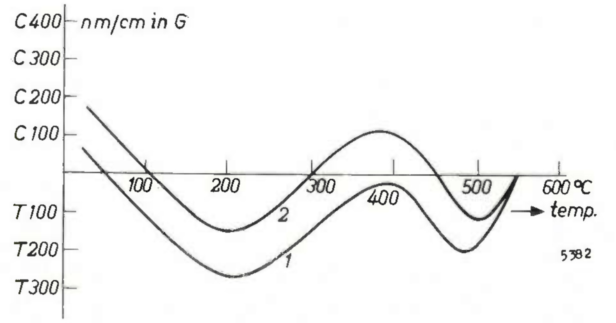


Fig. 7. Birefringence as a function of temperature (so-called "polarimeter" curves), for a standard glass-to-metal seal made with the glass and metal used in the tube type shown in fig. 6.

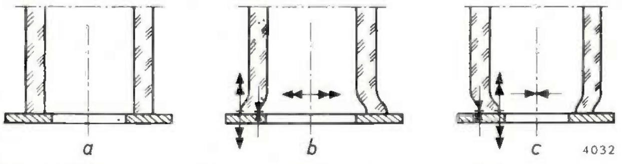


Fig. 8. Illustrating the cause of fracture in a seal between the envelope *B* and the fernico ring *A* as in fig. 6, and the corrective measure adopted. It is assumed that in a glass-to-metal seal to which curve *1* in fig. 7 applies, no or hardly any deformation, and hence no stress, is present at room temperature (*a*). The deformation that then occurs at higher temperature is shown in (*b*). Tensile stresses are denoted by double arrows, compressive stresses by single arrows. In a glass-to-metal seal to which curve *2* in fig. 7 applies, the seal is pre-stressed at room temperature (*c*).

**Summary.** A sensitive, simple and quick method of checking the relative expansion of glass-glass, glass-metal or glass-ceramic combinations is to measure the stress birefringence, i.e. the double refraction, in standard-shaped test seals of the materials concerned against a suitable standard glass. The large-scale adoption of this method instead of the still commonly used but less satisfactory dilatometer tests calls for close cooperation between the relevant industries. International cooperation is gradually increasing in this field. The article gives a description of the stress-birefringence method in comparison with the dilatometric method, and discusses the graphic determination of expansion tolerances for glass and metal in a given seal. It concludes with an example of the application of stress-birefringence measurements in tracing and eliminating the causes of fracture in a particular product.

[6]) The irregular expansion of fernico, which is due to the magnetic properties of this alloy, is turned to advantage for ensuring that, with an appropriate glass, the difference in expansion fernico-glass as a function of temperature does not become excessive.

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

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution.

**2806:** C. G. J. Jansen and A. Venema: A McLeod manometer with prescribed volumes for use as a standard instrument (Vacuum **9**, 219-230, 1959, No. 3/4).

The McLeod gauge is one of the few instruments with which it is possible to measure the actual values of low gas pressures (5-$10^{-5}$ mm Hg) without changing the composition of the gas. The McLeod gauge is therefore indispensable for calibrating the faster indicating instruments which only register relative pressure units, e.g. ionization gauges, heat-conductivity manometers, etc. Since it is very cumbersome to make and calibrate a McLeod gauge that can be used as a standard, it is worth while to improve the construction in such a way that a calibration can be omitted. In connexion with this problem a critical study was made of the formulae used to calculate the gas pressure from the readings of the McLeod gauge, which lead to expressions for the relative systematic and random errors limiting the measuring accuracy. By using a very accurate set of jigs in making the capillaries and the transition volumes at the connexion between capillaries of different diameters, and by addition of a small correction volume to the compression bulb, it was possible to construct a McLeod gauge indicating the actual gas pressures in the range of 3.5 to $10^{-5}$ mm Hg on one convenient set of scales. The volume of the compression bulb and the volume and diameters of the capillaries were calculated in such a way that full centimetres on the various scales correspond to $10^{-3}$, $10^{-2}$ and $10^{-1}$ mm Hg, respectively. The zero lines of the different sub-ranges of the instrument lie at whole centimetres from each other and at multiples of half-centimetres from the top of the compression capillary.

**2807:** W. L. Wanmaker and C. Bakker: Luminescence of copper-activated calcium and strontium orthophosphates (J. Electrochem. Soc. **106**, 1027-1032, 1959, No. 12).

Copper produces a strong luminescence in $Ca_3(PO_4)_2$. Luminescence in $Sr_3(PO_4)_2$ occurs only in the presence of small amounts of foreign ions, such as Ca, Zn, Cd, Mg, or Al. These additions give rise to a new crystal phase which is probably isomorphous with $\beta$-$Ca_3(PO_4)_2$. The emission peaks of $\beta$-$Ca_3(PO_4)_2$, $\alpha$-$Ca_3(PO_4)_2$, and of $Sr_3(PO_4)_2$ modified with Al, under excitation with 2537 Å, are found at 4800, 5700, and 4950 Å, respectively. Sensitization occurs with Mn, giving rise to a red emission peak. The fluorescence intensity remains good up to quite high temperatures (~300 °C), especially that of $\beta$-$Ca_3(PO_4)_2$-Cu and $Sr_3(PO_4)_2$ partly substituted with Mg and Ca. The application of these phosphors in lamps presents some difficulties due to the materials' sensitivity to air at binder bake-out temperatures.

**2808:** H. Bremmer: Méthodes mathématiques appliquées dans la théorie de la propagation des microondes (Conferenze del Seminario di Matematica dell'Università di Bari, March 1959, Nos. 45 and 46; publisher N. Zanichelli, Bologna 1959). (Mathematical methods in the theory of microwave propagation; in French.)

The first part of this article shows how the propagation of radio waves through the troposphere is influenced by local turbulences and by the presence of regions (of mainly horizontal extent) of differing refractive index. In the first case a role is played by the autocorrelation function of the spatial distribution of the refractive index. In the second case the horizontal dimensions of the regions in question in relation to those of the associated Fresnel zones are of importance. The second part of the article discusses the mathematical treatment of fading phenomena which, in general, involve the superposition of a constant and a fluctuating component. Joint Gaussian distributions for more than one variable are essential here. As an example, the rapidity of the fading of amplitude and phase of one and the same received signal are discussed and compared to each other.

**2809*:** B. Combée and P. J. M. Botden: Image intensification in medical X-ray technology (Tools of biological research, pp. 154-159; Blackwell, Oxford 1959).

After a brief survey of the evolution of X-ray techniques since 1895, the operation of the X-ray image intensifier is explained. Different versions of the intensifier are dis|cussed and illustrated with photographs. The X-ray image intensifier has opened the way to novel diagnostic techniques hitherto not feasible. Further developments are expected, including X-ray supervision during surgical operations.

**2810:** J. L. Melse and P. Baeyens: Protecting silver and copper against tarnishing by means of a chromate passivating process (46th Annual Technical Proceedings, pp. 293-297, American Electroplaters' Society, Newark N.J. 1959).

The paper describes the experimental work that forms the basis of improvements in methods of obtaining protection of silver or copper by immersion in solutions of hexavalent chromium compounds. To obtain good passivation consistently, without interfering appreciably with the solderability of either silver or copper, certain metal-complexing agents are added to the hexavalent chromium solution so as to maintain a defined relationship between the *p*H value of the solution and the metal-solution potential.

**2811:** P. C. van der Linden and J. de Jonge: The preparation of pure silicon (Rec. Trav. chim. Pays-Bas **78,** 962-966, 1959, No. 11).

An apparatus is described for the preparation of pure silicon by thermal decomposition of trichlorosilane on a hot tantalum wire in a hydrogen atmosphere. The polycrystalline silicon is obtained in the form of rods with a diameter of about 15 mm and a length of 20-40 cm. The yield of Si is 45-50% if calculated on the Si-content of $SiHCl_3$. From resistivity measurements, the boron content of the silicon obtained can be calculated to be $0.5 \times 10^{-6}$ % (resistivity of single crystals about 300-500 ohm.cm, p-type).

**2812:** H. C. Hamaker: Adjusting single sampling plans for finite lot size (Appl. Statistics **8,** 210-214, 1959, No. 3).

Most single sampling plans assume that the lot size is large compared with the sample size, and the calculated operating characteristic curves are strictly valid only under these conditions. This article describes a simple method for finding the sample size and acceptance number appropriate to a lot of finite size, so that the resulting operating characteristic curve closely approximates to that for a given plan with an infinite lot.

**2813:** J. Davidse: N.T.S.C. colour-television signals (Electronic and Radio Engr. **36,** 370-376 and 416-419, 1959, Nos. 10 and 11).

Investigations into the choice of parameters for an N.T.S.C. colour-television system have shown that the statistical properties of the signal are of importance. This article deals with measurement techniques and circuits used to obtain the required statistical data. The results are mentioned of a large number of measurements, using signals obtained by scanning colour slides and also camera signals. The author deals successively with the distribution of the colour information over the two colour signals, the distribution of the instantaneous level of the sub-carrier amplitude, and the distribution of the instantaneous level of the luminance signal. The bearing of the results on the transmission of colour-television signals is briefly commented on.

**2814:** J. A. Greefkes and F. de Jager: Voice radio systems for high noise paths (Electronics **32,** No. 50, 53-57, 1959).

Description of two Frena systems for speech transmission. The original signal is split into its frequency and amplitude components, the two types of information are transmitted on separate channels, and are then recombined into the original sound. This system is highly insensitive to noise, and can thus be used where the level of interference is high. Block diagrams and circuits for transmitter and receiver are given. See also Philips tech. Rev. **19,** 73-83, 1957/58.

**2815:** H. Koopman, J. H. Uhlenbroek, H. H. Haeck, J. Daams and M. J. Koopmans: Investigations on herbicides, II. 2-alkyloxy- and 2-aryloxy-4,6-dichloro-1,3,5-triazines; 2-alkylthio- and 2-arylthio-4,6-dichloro-1,3,5-triazines (Rec. Trav. chim. Pays-Bas **78,** 967-980, 1959, No. 11).

An improved synthesis of 4,6-dichloro-2-phenoxy-1,3,5-triazine from cyanuric chloride and phenol in the presence of collidine (2,4,6-trimethylpyridine) induced the authors to investigate the scope and limitation of the substitution of one chlorine atom in cyanuric chloride and the influence of the base used as an acid acceptor. Phenols, thiophenols, alcohols, thiols and oximes generally reacted with cyanuric chloride in the presence of collidine giving good yields of the corresponding substitution derivatives. The herbicidal and fungicidal properties of the compounds are dealt with and briefly discussed. The influence of the alkyl or aryl side chain on the biological activity was determined. For this reason some derivatives mentioned in the literature were included for comparison in the biological tests.

**2816:** A. Verloop, A. L. Koevoet, R. van Moorselaar and E. Havinga: Studies on vitamin D and related compounds, IX. Remarks on the iodine-catalysed isomerizations of vitamin D and related compounds (Rec. Trav. chim. Pays-Bas **78,** 1004-1014, 1959, No. 11).

As a sequel to previous publications, details are given on the iodine-catalysed reactions of previtamin D, tachysterol, cis- and trans-vitamin D (influence of solvent, concentration, wavelength of light). The possibility of a previtamin D determination based on the cis/trans isomerization is indicated. Some results and products obtained from the iodine-catalysed reactions in the vitamin $D_3$ series are reported.

**2817:** M. J. Sparnaay: Gas adsorption on germanium surfaces (Solid state physics in electronics and telecommunications, Proc. int. Conf., Brussels, June 1958, edited by M. Désirant and J. L. Michiels, Vol. I, pp. 613-618, Academic Press, London 1960).

Physical adsorption, mainly of argon gas, was used as a tool for investigations concerning germanium surfaces. It appeared in the measurements that the amount of gas adsorbed at a certain temperature and pressure depends on the oxygenated state of the surface. From the adsorption isotherms at different temperatures thermodynamic quantities can be derived and conclusions can be drawn concerning the behaviour of the adsorbed gas on the different adsorbents.

**2818:** C. Wansdronk: On the influence of the diffraction of sound waves around the human head on the characteristics of hearing aids (J. Acoust. Soc. Amer. **31,** 1609-1612, 1959, No. 12).

A small hearing aid, hanging in an anechoic room, is made to drive an AVC circuit, the output signal of which is conducted to a power amplifier and loudspeaker and can be recorded on a tape. During playback of this tape, with the output of the recorder connected to the power amplifier, the same sound field as existed around the hearing aid is reproduced. If the hearing aid is placed on a person in the position where it is to be worn and that person is situated so that the hearing aid is at the same point as during the recording, the output of the hearing aid during playback of the tape will indicate the influence of the diffraction around the human head.

Three specimens of hearing aid were measured on different people. The results showed that there exists a large difference between the hearing aids but no fundamental differences between the persons. The curves plotted for males and females showed the same trend, and no correlation was found with the hairdress. No success was achieved in an attempt to replace the human head by a simple model, such as a wooden sphere or a wooden box, the agreement of the diffraction phenomena between model and head being too poor.

**2819:** H. Koelmans and H. G. Grimmeiss: The photoconductivity of $CdIn_2S_4$ activated with Cu or Au (Physica **25,** 1287-1288, 1959, No. 12).

Polycrystalline $CdIn_2S_4$ was prepared by heating equimolecular quantities of pure CdS and $In_2S_3$ in sulphur vapour at 1150 °C. The absorption limit was found to be at 2.2 eV. The dark resistance of the samples proved to be dependent on the sulphur pressure. Unactivated samples showed photoconductivity with a spectral response peak at 2.1 eV. Activation with Cu or Au (optimum molar concentration approx. $2\times10^{-3}$) increased the photoconductivity by a factor of $10^3$. The decay time of the photocurrent was about $10^{-3}$ sec. The dependence of the photocurrent on the light intensity was found to be linear in the lower and higher intensity ranges, and superlinear in a transition range. By grinding-off part of the samples, crystals were obtained that showed the same sensitivity over the whole range as the untreated samples under strong illumination.

**2820:** A. E. Korvezee and J. L. Meijering: Validity and consequences of Schreinemakers' theorem on ternary distillation lines (J. chem. Phys. **31,** 308-313, 1959, No. 2).

In about 1900 Schreinemakers showed that the distillation lines of a ternary mixture, i.e. the loci of the points in the ternary diagram indicating the composition of the liquid phase during distillation, are tangential at their end points to one of the sides of the triangle. This was recently disputed by Redlich and Kister. In this paper a further proof is given of the correctness of Schreinemakers' theorem. Certain parameters are defined that govern the form of the distillation lines in each corner of the composition triangle. The values of these parameters can in principle be derived from accurate binary boiling-point curves. The occurrence of distillation lines with one or two points of inflection in a ternary system without complications from azeotropes or demixing is discussed at some length.

VOLUME 22, 1960/61, No. 11 pp. 345-376 Published 27th July 1961

# Philips Technical Review

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

## GENERAL CONSIDERATIONS ON DIFFERENCE AMPLIFIERS

by G. KLEIN and J. J. ZAALBERG van ZELST.

621.317.725.083.6:621.375

*For amplifying the voltage difference between two points neither at earth potential (in which case the difference to be amplified may be much smaller than the potentials to earth), increasing use is made of amplifiers specially developed for this purpose. Here, and in a subsequent article, consideration will be given to the problems involved in the design of such "difference amplifiers". The present article deals with general principles. Various fields of application are mentioned and the requirements to be met by these amplifiers are examined.*

A problem often encountered in electrical measuring techniques is to measure a voltage between two points both of which have potentials with respect to earth which are large compared with the voltage between them. This potential *difference*, as well as the potential common to both points, may be a DC or an AC voltage, or a combination of both.

In some cases it is a fairly simple matter to carry out such a measurement. In order, for example, to measure the potential difference $E_x$ of points $a$ and $b$ in *fig. 1*, it is often possible, even where the voltage $E_s$ is high, simply to use a voltmeter $V$ as shown, provided it is sufficiently insulated to prevent $E_s$ influencing the deflection; there are simple and obvious ways of verifying whether this is so.

The situation becomes more complicated when an amplifier has to be used. In many cases one of the two input terminals of such an amplifier is earthed, and it will be evident that this cannot then be used for measurements of the kind referred to. One might consider separately measuring the potentials of points $a$ and $b$ with respect to earth and then finding $E_x$ from the difference between them. If $E_x$ is small compared with $E_s$, however, the result of the measurement would be far from accurate — this is after all a typical drawback of difference measurements [1]).

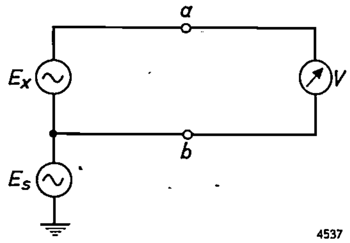


Fig. 1. With a floating voltmeter $V$ the voltage $E_x$ between points $a$ and $b$ can be measured irrespective of a voltage $E_s$ with respect to earth.

Even with an amplifier neither of whose input terminals is earthed (e.g. a balanced amplifier), difficulties are still encountered owing to unavoidable imperfections in symmetry, for only in an ideally balanced amplifier will a voltage $E_s$ on both input terminals $i$ and $i'$ (see *fig. 2*) not give rise to a voltage between the output terminals $u$ and $u'$. If a non-ideal amplifier is used to amplify and measure the potential difference $E_x$ between $i$ and $i'$

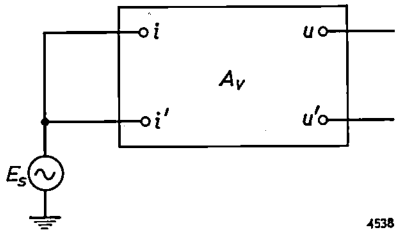


Fig. 2. Difference amplifier $A_v$ with a common voltage $E_s$ on the input terminals $i$ and $i'$. If the amplifier is perfectly balanced, no voltage then appears between the output terminals $u$ and $u'$.

[1]) In the case of alternating voltages — assuming that $E_s$ and $E_x$ have the same frequency, otherwise such a measurement would not be possible — it would also be necessary to determine any phase difference existing between the two voltages to be measured, which would be an additional complication and make the result even more inaccurate.

(see *fig. 3*), an external voltage $E_s$ may make a spurious contribution to the result.

Where the frequency of $E_x$ differs from that of the external voltage $E_s$, filters can be used to eliminate the influence of $E_s$ on the output signal. This is not always possible, however, due to partial or entire overlapping of the frequency spectra of $E_x$ and $E_s$.

The growing importance of measurements of this kind has led to the design of special amplifiers for amplifying the potential difference of two points which have a common potential to earth. The properties and applications of such "difference amplifiers" will be the subject of this article. To illustrate their importance, we shall first briefly consider various cases in which small potential differences have to be amplified and sometimes measured.

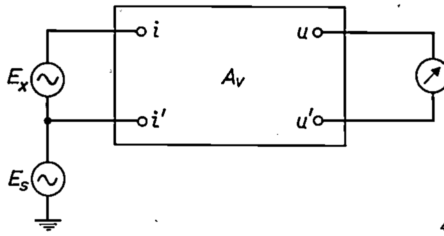


Fig. 3. Difference amplifier $A_V$ for measuring the voltage $E_x$ between the input terminals when a common voltage $E_s$ is present.

## Applications of difference amplifiers

### *Medical uses*

In electrocardiography and encephalography, electrical potentials between two points on the human body are recorded and sometimes measured. In electrocardiography two points are chosen, for example one on each arm, between which the action of the heart muscle produces a varying voltage of at the most a few millivolts. To record the voltage generated by cerebral activity (encephalography), two electrodes are applied to the skull. Provided the electrodes are properly positioned, a varying voltage is produced between them of the order of some tens of microvolts. The variation of these voltages with time can tell the physician a great deal about the presence or absence of certain cardiac or cerebral disorders.

One of the major difficulties in recording such oscillograms is the interference due to leads carrying the current for lamps, amplifiers, etc. This consists of an induced alternating voltage between the body of the patient and earth which is usually many times higher than the potential difference to be recorded between the two electrodes. To minimize this interference, the patient is earthed, usually by a conductor attached to one leg (*fig. 4*). Owing to the fairly high resistance of the human skin, the earthing is never perfect, and it is therefore not possible in this way to prevent induction from electrical leads giving rise to spurious potentials on the body [2]).

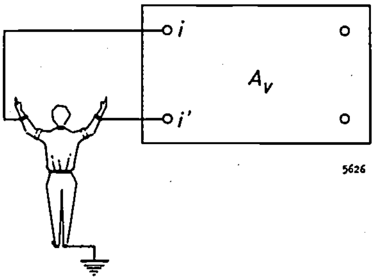


Fig. 4. Illustrating the use of a difference amplifier $A_V$ in electrocardiography. The input terminals are connected to the arms of the patient, one of whose legs is earthed.

One way of reducing these undesired voltages would of course be to keep all electrical leads out of the vicinity of the patient and the measuring equipment, but this is obviously no easy proposition. It is simpler to use a difference amplifier, which amplifies only the potential difference between the two points and does not respond to a voltage to earth common to both points [3]).

### *Technical applications*

In electronic engineering the need often arises to measure, amplify or record the voltage between two points neither of which is at earth potential. We shall briefly consider some typical cases.

Our first example is the measurement of the voltage between the grid and cathode of a valve in which neither electrode is earthed, e.g. a valve in a grounded-anode arrangement. A valve circuited in this way is often used as a cathode follower (*fig. 5*). Here the alternating voltage between cathode and

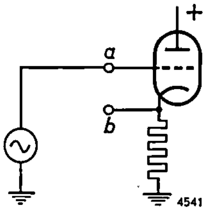


Fig. 5. Simplified circuit of a cathode follower. In the ideal case, points *a* and *b* have the same alternating potential with respect to earth. How closely this ideal is approached can be ascertained with the aid of a difference amplifier.

[2]) Spurious voltages, due to induction, can also arise *between* parts of the body, e.g. between the arms. These are much smaller, however, owing to the low resistance of the interior of the human body.

[3]) The use of difference amplifiers for a special form of electrocardiography, namely vector-cardiography, has already been described in this journal: G. C. E. Burger and G. Klein, Philips tech. Rev. **21**, 24-37, 1959/60 (No. 1).

earth is virtually identical with that between grid and earth. In the ideal case, no alternating voltage at all will be present between these two electrodes. To see how far the circuit approaches the ideal, one can try to measure the alternating voltage between points *a* and *b*. Since this voltage is always small compared with the potential of *a* and *b* with respect to earth, the use of a difference amplifier is called for.

A difference amplifier is also needed where a small voltage is to be measured at a position that can only be reached by long leads. One example is the measurement of temperature with the aid of a thermocouple at an inaccessible site. The thermocouple is then connected by a fairly long twin-cable to the measuring instrument or amplifier (*fig. 6*). These conductors almost invariably pick up interfering voltages [4]), which are again often many times larger than the voltage to be measured. If the two conductors are close enough together, the interfering voltage will be in phase at the two input terminals of the amplifier. Here too, a difference amplifier makes it possible to amplify and measure the potential difference between the wires, even though the induced spurious voltages may be relatively high.

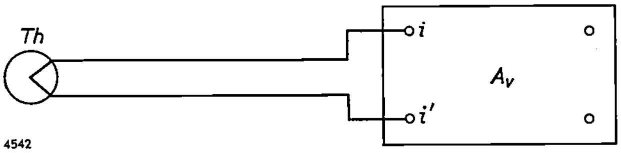


Fig. 6. Thermocouple *Th*, connected by a long twin-cable to the input terminals *i* and *i'* of a difference amplifier $A_v$.

Another very useful application of a difference amplifier is to be found in measurements using bridge circuits. A circuit of this kind is shown in *fig. 7*. The voltage source is connected between points *a* and *b*, the measuring instrument between *c* and *d*. During a measurement, then, the voltage source and the measuring instrument cannot both be earthed. If one of the instrument terminals is connected to earth, for example terminal *d*, neither of the terminals *a* and *b* must be earthed. This can adversely affect the accuracy of the measurement, in that the capacitances of the two output terminals of the "floating" voltage source with respect to earth are then in parallel with the impedances $Z_3$ and $Z_4$. Particularly at higher frequencies this may give rise to considerable errors [5]). The difficulty is avoided if one of the voltage-source terminals, for example *b*, is earthed. In that case, however, the voltage between *c* and *d* must be measured with an unearthed instrument. If the bridge is nearly balanced, the voltage between the latter two points is small compared with their voltage with respect to earth. A difference amplifier is then the ideal means of determining the voltage between *c* and *d*.

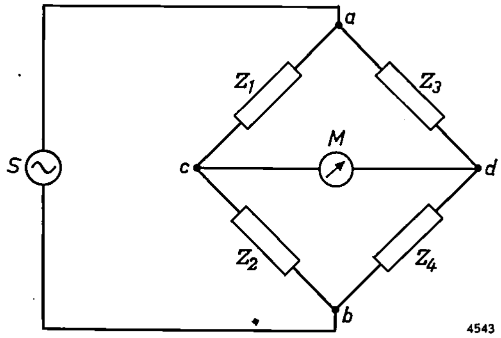


Fig. 7. Bridge circuit, consisting of impedances $Z_1$, $Z_2$, $Z_3$ and $Z_4$. *S* voltage source. *M* detecting instrument.

Difference amplifiers are also frequently used for precision measurements on electronic equipment. As an example we mention a method of measuring with a high degree of accuracy the distortion of an amplifier. A signal $E_i$ is applied to the amplifier input (see *fig. 8*) and from the output $E_u$ a voltage $E_u'$, which is made equal to $E_i$, is tapped by means of a voltage divider. The equalization is only exact provided there is no distortion in the amplifier. If distortion is present, then $E_i$ and $E_u'$ can only be equalized in respect of the fundamental frequency. (We shall not be concerned here with the phase-shift in the amplifier.) The voltage between terminals *i* and *z* then consists solely of higher harmonics due to distortion in the amplifier. Measurement

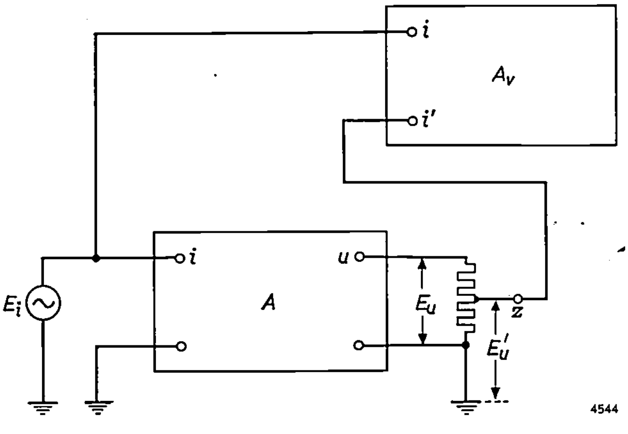


Fig. 8. Schematic representation of a method for measuring the distortion in an amplifier *A*, using a difference amplifier $A_v$.

[4]) These interfering voltages can be caused by induction, and also if the thermocouple and the amplifier are taken to earths at different places: a certain voltage often exists between "earth" at different places. Contact between the thermocouple and earth may be unavoidable if the temperature in a furnace is to be measured, since various insulating materials become conducting at high temperatures.

[5]) A method described by K. W. Wagner for avoiding these errors, the so-called Wagner earth, is in many measurements unsuitable because of the complication it involves.

of the voltage difference between $i$ and $z$ is now a means of very accurately determining the distortion, for it consists in measuring the higher harmonics without the fundamental component of the output voltage, which is always much larger. For measuring the small potential difference of points $i$ and $z$ a difference amplifier is again a useful tool.

In a similar way the frequency response of an amplifier (amplification as a function of frequency) can also be measured with considerable accuracy, and it is possible by a method based on the same principle to measure the phase shift between the input and output voltages of an amplifier.

We shall now conclude this survey of the possible applications of difference amplifiers, but a few others will be dealt with in a subsequent article, which will describe a number of circuits used for these amplifiers.

## Characteristics of a difference amplifier

The simplest form of a difference amplifier is a normal balanced amplifier, and we shall take this as our starting point. For simplicity we assume first of all that an amplifier of this kind consists of two separate and perfectly identical parts. In *fig. 9* these are denoted by $A$ and $A'$. The input and output terminals are respectively $i$-$i'$ and $u$-$u'$. The voltages between these terminals and earth are $E_{\rm i}$, $E_{\rm i}'$, $E_{\rm u}$ and $E_{\rm u}'$. If $E_{\rm i}$ and $E_{\rm i}'$ are equal, so too are $E_{\rm u}$ and $E_{\rm u}'$. There is then no voltage present between terminals $u$ and $u'$. If $E_{\rm i}$ and $E_{\rm i}'$ are equal and opposite, the same will apply to $E_{\rm u}$ and $E_{\rm u}'$. The average value of these two voltages is then zero.

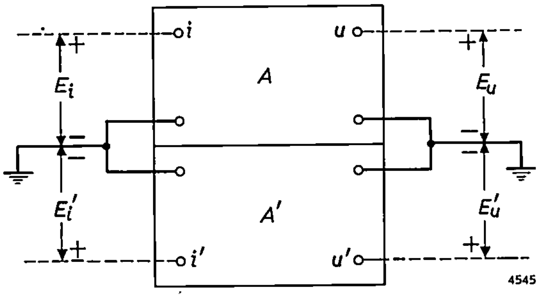


Fig. 9. Block diagram of a balanced amplifier consisting of two identical and independently operating parts, $A$ and $A'$.

If $E_{\rm i}$ and $E_{\rm i}'$ have arbitrary values, each of these voltages can be regarded as consisting of two components, the first components of both voltages being equal to one another, and the second components equal in magnitude but opposite in sign. If these components be $E_{\rm if}$ and $E_{\rm it}$, respectively, then between $E_{\rm if}$ and $E_{\rm it}$ and the input voltages $E_{\rm i}$ and $E_{\rm i}'$ there exist by definition the following relations:

$$\left.\begin{aligned} E_{\rm i} &= E_{\rm if} + E_{\rm it}, \\ E_{\rm i}' &= E_{\rm if} - E_{\rm it}, \end{aligned}\right\} \quad . \; . \; . \; . \quad (1)$$

whence

$$\left.\begin{aligned} E_{\rm if} &= \tfrac{1}{2}(E_{\rm i} + E_{\rm i}'), \\ E_{\rm it} &= \tfrac{1}{2}(E_{\rm i} - E_{\rm i}'). \end{aligned}\right\} \quad . \; . \; . \; . \quad (2)$$

In *fig. 10*, where $A_{\rm v}$ again denotes a difference amplifier, the two input voltages are shown together with their components obtained as described.

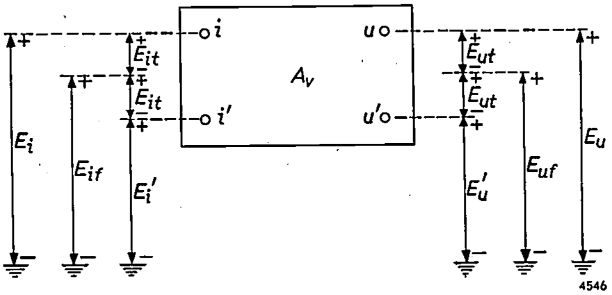


Fig. 10. Difference amplifier $A_{\rm v}$, with input voltages $E_{\rm i}$ and $E_{\rm i}'$, and output voltages $E_{\rm u}$ and $E_{\rm u}'$. Also indicated are the in-phase components $E_{\rm if}$ and $E_{\rm uf}$ and the anti-phase components $E_{\rm it}$ and $E_{\rm ut}$ of these voltages.

If $E_{\rm i}$ and $E_{\rm i}'$ are alternating voltages, it can be said that $E_{\rm if}$ is *in phase* on the two input terminals, and $E_{\rm it}$ in *anti-phase.* In the same way the output voltages, $E_{\rm u}$ and $E_{\rm u}'$, can each be resolved into a component $E_{\rm uf}$ and a component $E_{\rm ut}$ or $-E_{\rm ut}$. At the output terminals the component $E_{\rm uf}$ is in phase, and the component $E_{\rm ut}$ in anti-phase. In the following we shall therefore refer for convenience to the *in-phase component* and *anti-phase component* of the input and output voltages, irrespective of whether we are concerned with AC or DC amplifiers.

It can be useful to formulate the second equation (2) as follows: *the anti-phase component of the input voltages is equal to half the voltage difference across the input terminals that is to be amplified.*

It is easily seen that, in a difference amplifier consisting of a combination of two perfectly identical parts, the anti-phase component $E_{\rm it}$ of the input voltages will give rise only to an anti-phase component $E_{\rm ut}$ in the output voltages, whilst the in-phase component $E_{\rm if}$ will cause only an in-phase component $E_{\rm uf}$. Since, however, the two amplifier sections are never in fact perfectly identical, the ideal situation is not achieved, and a pure in-phase signal at the input gives rise on the output terminals to a combination of two voltages containing an anti-phase component as well as an in-phase component. Likewise, a pure anti-phase signal on the input terminals will give rise both to an anti-phase and an in-phase component at the output. If we apply to the input a combination of two voltages containing both an in-phase component $E_{\rm if}$ and an anti-phase component $E_{\mathrm{it}}$, the two components of the output voltages are given by the following equations:

$$\left.\begin{aligned} E_{\mathrm{ut}} &= A\, E_{\mathrm{it}} + B\, E_{\mathrm{if}}\,, \\ E_{\mathrm{uf}} &= C\, E_{\mathrm{if}} + D\, E_{\mathrm{it}}\,. \end{aligned}\right\} \quad \ldots \quad (3)$$

The extent to which a difference amplifier serves its purpose can now be expressed by three factors: the *rejection factor* [6] $H$, the *discrimination factor* $F$, and a third factor $G$, which is less important and therefore nameless. These three factors are related to $A$, $B$, $C$ and $D$ in eq. (3) in accordance with the following definitions:

$$\left.\begin{aligned} H &= \frac{A}{B}\,, \\ F &= \frac{A}{C}\,, \\ G &= \frac{A}{D}\,. \end{aligned}\right\} \quad \ldots\ldots \quad (4)$$

Confining ourselves for the time being to $H$ and $F$, we can express their significance in the following terms. The rejection factor $H$ is equal to the ratio between an in-phase voltage and an anti-phase voltage that have to be applied to the input in order to produce the same anti-phase voltage at the output. The discrimination factor $F$ is the ratio between the amplifications undergone by a pure anti-phase signal and a pure in-phase signal. As appears from the role of $A$ and $C$ in eq. (3), the amplification $A$ refers to the ratio of the anti-phase output signal to the anti-phase input signal, and the amplification $C$ to the ratio of the in-phase output signal to the in-phase input signal.

### Required values of rejection factor and discrimination factor

It will be clear from the above that $B = 0$ in the case of an ideal difference amplifier, giving an infinitely large rejection factor $H$. The wave form $E_{\mathrm{ut}}$ is then an undistorted amplified version of $E_{\mathrm{it}}$, and is not affected by the presence of an in-phase voltage $E_{\mathrm{if}}$ at the input (eq. (3)). In reality a finite rejection factor is acceptable; the requirements it must meet depend on the purpose for which the amplifier is to be used, and on the conditions under which the amplifier is to operate. Of particular importance is the magnitude of the undesired in-phase signal at the input — which is after all the reason why a normal amplifier cannot be used — or, to be more exact, the ratio of this signal to the anti-phase signal which is to be amplified.

To give some idea of the required magnitude of $H$ in a specific case, we shall again turn to applications in the medical field. The body of a patient, even when it is well-connected conductively to earth, is often found to have an alternating voltage of about 10 mV with respect to earth. The voltage produced between two points on the surface of the body by the action of the heart muscle is of the order of magnitude of 1 mV. Now if the aim is that the anti-phase signal at the output end of the difference amplifier used should not be falsified by more than one per cent by the in-phase signal at the input, the absolute value of the rejection factor will have to be at least 1000 [7].

In encephalography the voltages recorded are at least ten times smaller than in cardiography, although the spurious in-phase voltage is just as large. Higher demands are therefore made on the rejection factor of a difference amplifier in encephalography, and the minimum absolute value required is usually 10 000.

Apart from the requirement that the waveform $E_{\mathrm{ut}}$ should be identical to $E_{\mathrm{it}}$, it is in general also required of a difference amplifier that the in-phase component should be less dominant in the output signal than in the input signal. This implies that the in-phase component should undergo less amplification than the anti-phase component, in other words, the discrimination factor $F$ must be greater than unity [8]. How much greater depends on the circumstances, but it may be noted here, anticipating the last section of this article, that a reasonably large value of $F$ is particularly important in the first stage of an amplifier. If the in-phase component is only slightly amplified in this stage, the input signals of the second stage will contain only a small in-phase component, which means that relatively lower demands can be made on the rejection factor of this stage. The design of this and successive stages can thus be considerably simplified [9].

If the first stage is designed with a large enough discrimination factor $F$ — the circuitry will be

[6] This factor was used in the article cited under [3]. In the literature it is also called the common-mode rejection, transmission factor, rejection ratio or antiphase/in-phase ratio.

[7] There is little point in a higher value of $H$, because the anti-phase signal at the input terminals already contains a spurious component due to induction between the positions of measurement.

[8] $F$ might be equal to unity if the two identical parts of the amplifier in fig. 9 were entirely independent in their operation: $A = C$ and $B = 0$, $D = 0$. This shows that the condition $H = \infty$, although necessary, is not sufficient for obtaining an ideal difference amplifier.

[9] There is a certain analogy in this respect with the question of the noise level in receivers and amplifiers. Here, too, the noise of the first stage usually governs the noise level of the whole apparatus.

described in a subsequent article — the circuit of the second stage can be kept quite simple: the in-phase signal at the input of the second stage is then so attenuated with respect to the anti-phase signal that it is sufficient for this stage to have a rejection factor of $H > 100$, a requirement that can be met by fairly simple means.

After these comments on the factors $H$ and $F$, we need only say a few words about the third factor, $G = A/D$ (eq. (3)). This expresses the ratio between the anti-phase and the in-phase signals produced at the output as a result of the same anti-phase signal at the input. The fact that $D$ is not zero ($G$ is not infinite) has no influence at all on the magnitude of the required anti-phase signal at the output. It is therefore reasonable to suppose that the factor $G$ is of much less importance to the evaluation of a difference amplifier than the factors $H$ and $F$. Closer analysis shows that $G$ does have some influence on the resultant values of $H$ and $F$ in a difference amplifier consisting of several stages (see next section). In this respect too, however, the effect of $G$ is found to be of minor significance at the normal values of $H$ and $F$ for the individual stages.

It should finally be pointed out that the above-mentioned requirements for $H$ and $F$ are intended as the minimum requirements to be met by a difference amplifier if it is to serve its purpose. These requirements have to be satisfied under the most unfavourable conditions as regards the symmetry of the two sections of the amplifier. If tubes or other components are replaced, or if the temperature varies, both $H$ and $F$ will usually be affected. By means of a manually operated volume control on one or both sections of the amplifier, it is possible in most cases to adjust the symmetry with a very high degree of accuracy, and thus to achieve high values of $H$ and $F$. As a rule, however, $H$ and $F$ are required to remain above the minimum permissible value without any adjustment. In the design of a difference amplifier it is therefore necessary to calculate values for the rejection and discrimination factors that will meet the contingency where all conditions that can influence them are adversely operative, so that their effects on $H$ and $F$ are not compensatory. Only then can specific minimum values be guaranteed. The actual rejection and discrimination factors will in practice almost invariably be much greater.

## Rejection factor and discrimination factor of a multi-stage amplifier

We shall now express the characteristics of a two-stage difference amplifier in terms of the factors $H$, $F$ and $G$ of each stage separately. For the first and second stage we use, respectively, the subscripts 1 and 2, and we start from equations (3) written in the form:

$$\left.\begin{aligned} E_{\mathrm{ut}} &= A\,E_{\mathrm{it}} + \frac{A}{H}\,E_{\mathrm{if}}, \\ E_{\mathrm{uf}} &= \frac{A}{F}\,E_{\mathrm{if}} + \frac{A}{G}\,E_{\mathrm{it}}. \end{aligned}\right\} \quad \ldots \quad (3a)$$

We assume that an anti-phase signal $E_{\mathrm{it}}$ and an in-phase signal $E_{\mathrm{if}}$ are present at the input end of the amplifier. If the anti-phase gain of the first stage is $A_1$, the anti-phase signal appearing at the output of this stage is given by:

$$A_1E_{\mathrm{it}} + \frac{A_1}{H_1}\,E_{\mathrm{if}}. \quad \ldots\ldots \quad (5)$$

The in-phase signal at that point is

$$\frac{A_1}{F_1}\,E_{\mathrm{if}} + \frac{A_1}{G_1}\,E_{\mathrm{it}}. \quad \ldots\ldots \quad (6)$$

Both these signals are applied to the second stage. The anti-phase gain of this stage being $A_2$, the anti-phase signal at the output of the second stage is found by simple calculation to be:

$$E_{\mathrm{ut}} = A_1A_2\left(1 + \frac{1}{G_1H_2}\right)E_{\mathrm{it}} + \\ + A_1A_2\left(\frac{1}{H_1} + \frac{1}{F_1H_2}\right)E_{\mathrm{if}}. \quad \ldots \quad (7)$$

The in-phase signal at the output terminals is similarly found to be:

$$E_{\mathrm{uf}} = A_1A_2\left(\frac{1}{F_1F_2} + \frac{1}{H_1G_2}\right)E_{\mathrm{if}} + \\ + A_1A_2\left(\frac{1}{G_1F_2} + \frac{1}{G_2}\right)E_{\mathrm{it}}. \quad \ldots \quad (8)$$

From (7) we find the total gain for anti-phase signals:

$$A_{\mathrm{tot}} = A_1A_2\left(1 + \frac{1}{G_1H_2}\right), \quad \ldots \quad (9)$$

and again using (7) we arrive at the rejection factor $H_{\mathrm{tot}}$ of the two-stage amplifier:

$$H_{\mathrm{tot}} = H_1\,\frac{1 + \dfrac{1}{G_1H_2}}{1 + \dfrac{H_1}{F_1H_2}}. \quad \ldots \quad (10)$$

From (7) and (8) we obtain for the discrimination factor $F_{\text{tot}}$:

$$F_{\text{tot}} = H_1G_2 \frac{1 + \dfrac{1}{G_1H_2}}{1 + \dfrac{H_1G_2}{F_1F_2}}. \quad . \; . \; . \quad (11)$$

Finally, from (7) and (8) we can also derive a quantity $G_{\text{tot}}$, which determines the degree to which an in-phase output signal arises from an anti-phase input signal. We find:

$$G_{\text{tot}} = G_2 \frac{1 + \dfrac{1}{G_1H_2}}{1 + \dfrac{G_2}{G_1F_2}}. \quad . \; . \; . \quad (12)$$

It is important to note that $H$, $F$ and $G$ may be either positive or negative. Where the amplifier consists of a single stage, the sign of these quantities is generally of no importance. Where the amplifier consists of two stages, there is a possibility that, owing to different signs of $H_1$ and $H_2$, for example, the quantity $H_{\text{tot}}$ will assume a very high absolute value, although the absolute values of $H_1$ and $H_2$ are fairly small. A large rejection factor obtained in this way can never be guaranteed, however. In order to calculate the magnitudes of $H_{\text{tot}}$, $F_{\text{tot}}$ and $G_{\text{tot}}$ for the most adverse circumstances, we must assume that all factors are operative in the same (unfavourable) direction.

Formulae (9) to (12) may be written to a good approximation in a simpler form. Since, in all cases likely to be encountered in practice, $G_1H_2$ is large compared with unity (e.g. $G_1 = 10$ and $H_2 = 100$), we can neglect $1/G_1H_2$ with respect to unity, and so write the above formulae as follows:

$$A_{\text{tot}} = A_1A_2\,, \quad . \; . \; . \; . \; . \; . \quad (9a)$$

$$\frac{1}{H_{\text{tot}}} = \frac{1}{H_1} + \frac{1}{F_1H_2}\,, \quad . \; . \; . \quad (10a)$$

$$\frac{1}{F_{\text{tot}}} = \frac{1}{H_1G_2} + \frac{1}{F_1F_2}\,, \quad . \; . \; . \quad (11a)$$

$$\frac{1}{G_{\text{tot}}} = \frac{1}{G_1F_2} + \frac{1}{G_2}\,. \quad . \; . \; . \; . \quad (12a)$$

It follows from (10a) that both $H_1$ and the product $F_1H_2$ must be large in order to make the value of $H_{\text{tot}}$ that can be *guaranteed* large. (It is true that a large value of $H_{\text{tot}}$ is obtained if $H_1$ and $F_1H_2$ are opposite in sign; however, such high values cannot be *guaranteed* since $H_1$ and $F_1H_2$ are completely random in sign.) The first aim will therefore be to give the first stage the highest possible rejection factor. In the circuits to be described (in a subsequent article) $F_1H_2$ is large compared with $H_1$, even when a simple circuit is used for the second stage. As a result, $1/F_1H_2$ is usually small compared with $1/H_1$ and $H_{\text{tot}}$ is roughly equal to $H_1$. The guaranteed value of the rejection factor of a two-stage difference amplifier is therefore, as we have already shown qualitatively, in most cases determined by the guaranteed rejection factor of the first stage.

From equation (11a) another important conclusion can be drawn. This equation may be re-written in the form:

$$\frac{1}{F_{\text{tot}}} = \frac{1}{H_1}\left(\frac{1}{G_2} + \frac{H_1}{F_1F_2}\right).$$

Even if the circuit of the second stage is very simple, the terms between brackets still usually add up to less than unity. We see, then, that the discrimination factor $F_{\text{tot}}$ of the whole two-stage amplifier is greater than the rejection factor $H_1$ of the first stage, and hence greater than $H_{\text{tot}}$. This is always sufficient to make fairly small rejection factors acceptable for a third stage and any other following stages.

The quantity $G_{\text{tot}}$, where $F_2$ is sufficiently large, appears from (12a) to be roughly equal to $G_2$. For the same reasons mentioned in connection with the single-stage difference amplifier, no particular attention need in general be paid to this factor. ($G_{\text{tot}}$ is significant where an exceptionally large ratio between anti-phase and in-phase signals is required at the output terminals.) The problem of building a good difference amplifier, which amplifies the anti-phase signal at the input terminals without being appreciably influenced by an interfering in-phase signal, can therefore be reduced, as we have seen, to the appropriate design of the first stage.

Circuits that can be used to meet the high requirements for the first stage will be described in a second article to be published shortly.

Summary. After a reference to various medical and technological applications of difference amplifiers, the requirements to be met by such amplifiers are dealt with in some detail. A definition is given of the rejection factor and the discrimination factor, which describe the most important characteristics of these amplifiers. For a two-stage amplifier these factors are calculated from those of the two stages separately. It is shown that the requirements to be met by a good difference amplifier are satisfied if the *first* stage is given a high rejection factor and a high discrimination factor. This makes it possible to use fairly simple circuits for the subsequent stages. Circuits for difference amplifiers will be dealt with in a forthcoming article.

# FM RECEPTION UNDER CONDITIONS OF STRONG INTERFERENCE

by J. van SLOOTEN.

621.396.621:621.376.3:621.391.82

*The effects observed in the reception of frequency-modulated signals under conditions of strong interference are capable of exact analysis. However, the mathematical difficulties involved, although not insurmountable, make the theory difficult to grasp. In the article below, the problem is approached with the aid of simple expressions which, though not new, are seldom employed. The result is a relatively simple formula which satisfactorily describes the effects concerned.*

In the last ten years there has been a marked increase in the use of frequency modulation (FM). Most countries now have networks of FM broadcast transmitters, which are steadily being expanded. Many countries, too, use frequency modulation for the sound channel in television broadcasts. A third important application is found in the transmission of large numbers of telephone conversations by means of radio links or coaxial cables.

Frequency modulation calls for a much greater bandwidth than amplitude modulation (AM). In the applications mentioned, this requirement is not an overriding objection, and moreover the signal-to-noise ratio in reception is as a rule favourable. Because of the large bandwidth in these applications, full profit can be derived from the typical advantage of frequency modulation, which is that it minimizes the nuisance experienced from interference.

There are other instances, however, where a large bandwidth is not readily possible and where the interference at the receiving end will often be relatively heavy. Transceiver communications are a case in point. The growing number of mobile transmitters makes it desirable to limit their bandwidth and power as much as possible. In such cases, FM has a serious rival in AM, particularly in single-sideband AM transmissions.

After the publication in 1936 of Armstrong's method of frequency moduation [1]), there was at first some uncertainty regarding the extent to which FM was superior to AM in the improvement of noise conditions in reception. The first theoretical treatments yielded exact results, but were too complicated to overcome existing prejudices. In the long run the same results were arrived at by simpler theories [2]), but they still had the drawback of being valid only in the case of relatively weak interference. In the interesting region of transition to relatively strong interference, where it is doubtful whether FM is superior to AM, they yield results which are too heavily weighted in favour of FM, and which are belied by measurements. Later, exact theories were evolved for this region too [3])[4]), but their mathematical intricacy was a bar to a clear understanding of the effects. An attempt to obtain exact results by a more straightforward approach [5]) was not entirely successful.

In the following we approach the case of relatively noisy FM reception by a theoretical method which, without pretending to be entirely exact, leads to a satisfactory picture both qualitatively and quantitatively. The method is so general that its results are applicable not only to a simple *LC* circuit (to which we shall confine ourselves here), but equally well to more complicated networks, provided a numerical correction factor is introduced. In conjunction with the elementary theory for weak interference [2]), this approach leads to a general formula which, applicable to both weak and strong interference, is in reasonable agreement with known results of measurements as well as with exact calculations.

Some of the ground covered in the two articles on frequency modulation by Weijers, published in this journal a good many years ago [2])[6]), will have to be retraced here for the purpose of the present discussion.

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

[2]) T. J. Weijers, Comparison of frequency modulation and amplitude modulation, Philips tech. Rev. **8**, 89-96, 1946.

[3]) F. L. H. M. Stumpers, Theory of frequency modulation noise, Proc. Inst. Radio Engrs. **36**, 1081-1092, 1948.

[4]) D. Middleton, On theoretical signal-to-noise ratios in f-m receivers, J. appl. Phys. **20**, 334-351, 1949.

[5]) J. Cohn, A new approach to FM threshold reception, Proc. Nat. Electronic Conf. **XII**, 211-236, 1956.

[6]) T. J. Weijers, Frequency modulation, Philips tech. Rev. **8**, 42-50, 1946.

## The sum of two oscillations modulated in frequency

A frequency-modulated oscillation $v$ may be represented as a function of time $t$ by the expression:

$$v = V \cos(\omega_0 t + m \sin qt), \quad . \; . \; . \quad (1)$$

where $V$ is the amplitude, $\omega_0$ the central angular frequency of $v$, $m$ the maximum phase deviation (in radians) with respect to the unmodulated signal, and $q$ the angular frequency with which the phase or the frequency is modulated (for the sake of convenience we shall use simply "frequency" for "angular frequency"). The quantity $m$ is usually called the modulation index.

The instantaneous frequency $\omega$ is defined as the time derivative of the argument of the cosine function in (1):

$$\omega = \frac{\mathrm{d}}{\mathrm{d}t}(\omega_0 t + m \sin qt) = \omega_0 + mq \cos qt. \quad (2)$$

The maximum value that the frequency deviation can have, which we shall call the frequency excursion (half the frequency swing), is given by:

$$\Delta\omega = \omega_{\max} - \omega_0 = mq. \quad . \; . \; . \quad (3)$$

This relation — frequency excursion equal to the product of modulation index and modulation frequency — will be referred to frequently in the following pages.

As an introduction to general interference theory, we shall first consider the case where a frequency-modulated signal (1) is disturbed by a weaker signal of constant amplitude $V_s < V$. We assume that the frequency of this disturbing signal also varies, but remains within the frequency band to which the receiver is tuned for optimum reception of the desired signal.

*Fig. 1* represents the vector $\mathbf{V}$ (modulus $V$) of the desired signal. We imagine this vector to be rotating at the varying angular velocity $\omega$ given by (2). The disturbing signal is represented by a second vector, $\mathbf{V}_s$ (modulus $V_s$), which rotates in relation to $\mathbf{V}$ at an angular velocity equal to the instantaneous frequency difference $\omega_s - \omega$ between the two signals. It can be seen from the figure that the maximum phase deviation $\alpha_s$ shown by the sum of the two signals in relation to the undisturbed signal, is always smaller than $\frac{1}{2}\pi$ ($\frac{1}{4}$ period), since $\sin \alpha_s = V_s/V < 1$. The larger the phase deviation due to the modulation — this deviation being $m$ — the less significant and hence the less troublesome will be the phase deviation attributable to the disturbance. It is immediately evident, then, that the nuisance of interference in FM can be reduced by increasing the frequency excursion. For this reason the excursion at the highest modulation frequency $q_a$ is chosen many times larger than $q_a$.

It is readily inferred from (3) that at a given $\Delta\omega$, the disturbance will be minimum at the lowest modulation frequencies, where $m$ is largest. It follows from this that the disturbing voltage in the signal obtained after detection will increase in proportion with the audio frequency. This phenomenon is known as the "triangular noise spectrum" of the detected signal. (The noise *power* per unit bandwidth is then proportional to the *square* of the audio frequency.)

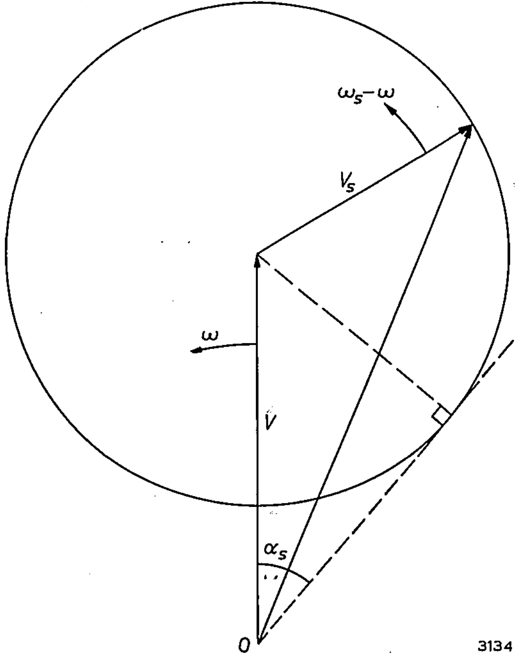


Fig. 1. The vector $\mathbf{V}$, rotating about $O$ at a varying angular velocity $\omega$, represents the frequency-modulated signal to be received; the vector $\mathbf{V}_s$ represents an interfering frequency-modulated signal ($V_s < V$). The sum of $\mathbf{V}$ and $\mathbf{V}_s$ shows in relation to $\mathbf{V}$ a maximum phase deviation $\alpha_s$, which remains smaller than $\frac{1}{4}$ period ($\frac{1}{2}\pi$).

### *FM compared with single-sideband AM, in conditions of weak interference*

By $(\Delta\omega)_{\max}$ we denote the maximum frequency excursion used in FM. The width of the radio spectrum is approximately $2(\Delta\omega)_{\max}$ in FM [6]), and $q_a$ in AM for single-sideband transmission. The ratio between these two bandwidths is called $k$:

$$k = \frac{2(\Delta\omega)_{\max}}{q_a}.$$

In order to compare these two systems with one another in regard to their freedom from interference, we write further:

$$y^2 = \frac{\text{signal power}}{\text{average noise power within frequency band } q_a}. \quad (4)$$

Both powers are taken at the input of the receiver, where the ratio of the signal power to the total noise power is thus equal to $y^2/k$. The quantity $y$ will be referred to as the "normalized signal-to-noise ratio" at the input.

From the power ratio $y^2$ at the receiver input we must now proceed to the power ratio $(S/N)^2$ at the output of the detector, $S$ being the signal voltage and $N$ the noise voltage there. In the case of FM under weak interference we find $(S/N)^2$ by means of a simple integration over the above-mentioned triangular noise spectrum [2]):

FM under weak interference:

$$(S/N)^2 = \tfrac{3}{8}\, k^2 y^2 . \quad \ldots\ldots \quad (5)$$

In the case of AM, the signal-to-noise ratio after detection is the same as before detection, and therefore, for strong as well as weak interference, we may write:

AM with single-sideband transmission:

$$(S/N)^2 = y^2 . \quad \ldots\ldots\ldots \quad (6)$$

The extent to which FM gains over AM with single-sideband transmission under conditions of weak interference is thus given by the factor $\tfrac{3}{8}k^2$. One might be inclined to conclude from this that $k$ — that is the width of the FM spectrum — should be made as large as possible. We shall see, however, that this conclusion is not correct when the interference is occasionally stronger than the signal.

*FM under strong interference*

The situation is entirely different from that given above when the interference becomes stronger than the signal ($V_s > V$). The sum of the vectors $\mathbf{V}$ and $\mathbf{V}_s$ (fig. 1) will then approximately follow the phase of the disturbance instead of the phase of the signal. What remains of the modulation of the desired signal is then totally distorted and may be regarded as part of the interference. It may therefore be stated that *when two frequency-modulated signals are applied to a frequency detector, the stronger of the two will be received; the weaker signal is entirely distorted and is perceived as "noise"*.

This statement can be verified experimentally: if we allow the weak signal to increase in strength, there comes a point when it suddenly supersedes (and "drowns") the other signal; in the narrow transitional region, only noise is heard.

To arrive at a formula for the signal-to-noise ratio after detection, which will apply in the case of strong interference and will transpose to equation (5) in the case of weak interference, we must take into consideration the amplitude statistics of the noise, and the frequency spectrum of the noise after detection.

## Statistical behaviour of amplitude and phase of noise interference

With a view to devising a readily manipulated model of a noise disturbance, it is useful — from a mathematical as well as from a technological standpoint — to consider a number of identical events, the average number of which remains constant over a long period of time, but which occur in that time in an entirely random manner. Examples of such sequences of events are the incidence of electrons on the anode of a saturated diode, and — though somewhat less accurate — the incoming calls in a telephone exchange and the occurrence of traffic accidents. The statistical distribution of such events within relatively short periods of time is known as the Poisson distribution [7]).

*Poisson distribution of noise current*

Random disturbing pulses of this kind can be made to act on a mechanical oscillator or an electrical ($LC$) oscillatory system. The pulses give rise to oscillations whose amplitude and phase are continuously changing. Given a sufficiently large average number of disturbing pulses within the period of oscillation, the amplitude and phase are found to behave in a manner that depends, for all practical purposes, only on the bandwidth of the system (width of the resonance curve). This makes it possible to interpret the sequence of random pulses as corresponding to a spectrum of periodic disturbances whose phases are not correlated. The disturbance energy is then seen to be distributed uniformly over the spectrum which, if the average number of pulses per unit time is large, extends to very high frequencies. Phenomena possessing such a spectrum are known as "white" noise. The result of letting the pulses act on the oscillatory system is thus the same as filtering-out a frequency band from a very wide spectrum of disturbances. The statistical properties of the filtered noise are of fundamental importance if we are to form a picture of the potentialities of FM in reducing interference.

We shall derive these statistical properties by using a method already described in this journal in connection with another problem of radio engineering [8]). The problem there concerned the case of interfering pulses acting on an $LC$ circuit having a resistance $r$ in series with the inductance. The magnitude of $r$ determines the bandwidth of the system.

[7]) J. van Slooten, Oscillations and noise, Philips Res. Repts. **11**, 19-26, 1956.

[8]) J. van Slooten, Mechanism of the synchronization of $LC$ oscillators, Philips tech. Rev. **14**, 292-298, 1952/53.

The oscillation produced by the system under the influence of the disturbing pulses has the tendency to persist in the instantaneous phase state. Each fresh pulse, however, changes not only the amplitude of the oscillation but also its phase. The total effect or summation of the phase disturbances is analogous to the random-walk problem (or Brownian movement in one dimension). The result obtained in this way is also valid for more complicated filter systems.

### *Simpler model of a noise current*

We shall now consider a series of noise pulses having a somewhat less general shape than that of a Poisson distribution, but still sufficiently irregular to enable us to arrive at all the properties of the noise spectrum. We consider a consecutive sequence of adjacent positive and negative current pulses, all with the same amplitude $I_0$ and the same duration, and superposed on a direct current $I_0$ (*fig. 2*).

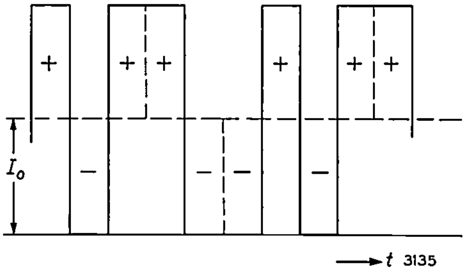


Fig. 2. Interfering pulses of amplitude $+I_0$ or $-I_0$, and superposed on a direct current $I_0$. The sign of each pulse is completely random (but with equal probability for + and —).

The sign (+ or —) of the pulses is a matter of chance, as if one were to toss a coin heads or tails. The current is $2I_0$ during the positive pulses, and zero during the negative pulses. (Of course the current thus defined may also be regarded as consisting of pulses of amplitude $2I_0$, the presence of which is decided by the same random process.)

The statistical distribution of the values of the deviation $a_n$ of the current with respect to $I_0$, summated over $n$ pulses, is readily found with the aid of the "Pascal triangle":

| $a_n =$ | —5 | —4 | —3 | —2 | —1 | 0 | +1 | +2 | +3 | +4 | +5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n = 1$ | | | | | 1 | | 1 | | | | |
| 2 | | | | 1 | | 2 | | 1 | | | |
| 3 | | | 1 | | 3 | | 3 | | 1 | | |
| 4 | | 1 | | 4 | | 6 | | 4 | | 1 | |
| 5 | 1 | | 5 | | 10 | | 10 | | 5 | | · 1 |
| etc. | . | . | | . | | . | | . | | . | . |

In this distribution each number is equal to the sum of the two neighbouring numbers one line higher. We see that after, for example, three pulses ($n = 3$) there is one chance of a deviation +3 (three successive positive pulses), three chances of a deviation +1 (two positive pulses and one negative), likewise three chances of a deviation —1 (one positive and two negative pulses), and again one chance of a deviation —3 (three negative pulses).

If, after a certain number of pulses $n$, we take the mean square of the possible deviations $a_n$ (where $a_n$'s occurring more than once must be counted more than once), the following familiar relation is applicable:

$$\overline{a_n^2} = n \, . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

After, for example, $n = 5$ pulses we have:

2 chances of a deviation $a_n = \pm 5$,
10 chances of a deviation $a_n = \pm 3$ and
20 chances of a deviation $a_n = \pm 1$.

The mean of $a_n^2$ is therefore:

$$\overline{a_n^2} = \frac{2\times(\pm5)^2 + 10\times(\pm3)^2 + 20\times(\pm1)^2}{2 + 10 + 20} = \frac{160}{32} = 5,$$

which is indeed equal to $n$.

The relation (7) can be demonstrated by a more complete inductive reasoning.

We call the number of pulses per unit time $n_1$, and the charge per pulse $\pm\frac{1}{2}e$ (if we ignore $I_0$, the current consists of pulses of charge $+ e$). We then have $I_0 = \frac{1}{2}n_1 e$. After a time $T$ the number of pulses is $n = n_1 T$. The amount $i_s$ by which the time average of the current, $\overline{I}$, deviates from $I_0$ is

$$i_s = \overline{I - I_0} = \overline{\Delta I} = a_n\,(\tfrac{1}{2}e)/T \, . \quad . \quad . \quad (8)$$

This deviation may be either positive or negative, depending on whether $a_n$ is positive or negative. The possible values of $a_n$ may be found from the Pascal triangle. We now regard $i_s$ as the AC component in the noise current, where $T$ signifies the smallest time interval of interest in the following calculations. It should again be recalled that the smallest time interval $T$ must contain at least a number of noise current pulses.

With the aid of (7) we easily find the mean square value of the AC component $i_s$ as defined by (8):

$$\overline{(\overline{I - I_0})^2} = \overline{i_s{}^2} = (e/2T)^2\overline{a_n^2} = (e/2T)^2 n_1 T = \frac{I_0 e}{2T} . \quad (9)$$

In connection with the following considerations, it is important to note that (9) is the result of double averaging: the first time over a time interval $T$, and the second time, after squaring the result, over a large number of similar time intervals.

For completeness it should be mentioned that if, instead of the current according to fig. 2, we consider noise pulses in an actual Poisson distribution (with charge $e$ and mean current $I_0$), the result is [7]):

$$\overline{i_s^2} = \frac{I_0 e}{T}.$$

Another point to be noted is that the fluctuations in the noise current given by fig. 2 obviously are more rapid than would follow from (9) or (8). The energy corresponding to these rapid fluctuations, however, falls in a frequency range which is much higher than the band of frequencies filtered out by the receiver (or by the $LC$ circuit presently to be considered).

We now let the noise current given by fig. 2 act on an $LCr$ circuit (*fig. 3*). This produces across the capacitor an alternating voltage $v$ which fluctuates in amplitude and in phase. The average frequency $\omega_0$ of $v$ is given by $\omega_0^2 LC = 1$.

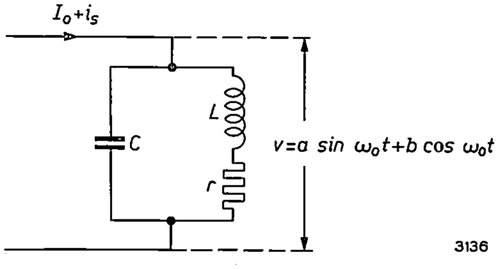


Fig. 3. The interfering current (see fig. 2) is applied to an $LCr$ circuit. The voltage $v$ over the circuit has the average frequency $\omega_0 = 1/\sqrt{LC}$, but is subject to amplitude and phase fluctuations. The voltage $v$ may be written as the sum of a sine and a cosine component of varying amplitude $a$ and $b$, respectively.

### *Statistical behaviour of the amplitude*

The statistical properties of the fluctuating amplitude of $v$ can be found most simply with the aid of an artifice due to Fresnel [9]). This consists in resolving the voltage $v$ into a sine component and a cosine component:

$$v = a \sin \omega_0 t + b \cos \omega_0 t.$$

The fluctuations both in the amplitude and the phase of $v$ can be expressed in terms of the fluctuations of the amplitudes $a$ and $b$. A considerable advantage of this method of approach is, as we shall see, that $a$ and $b$ have an equal probability of being positive or negative.

When a disturbing pulse of charge $i_s \Delta t$ is now applied to the circuit, we may write for the amplitude fluctuation $\Delta v$, because of the continuity of the current through $L$:

$$\Delta v = \frac{i_s \Delta t}{C}, \quad \dots . \quad (10)$$

provided the time $\Delta t$ is short compared with the average period $2\pi/\omega_0$, but longer than the duration of the elementary pulses in fig. 2. After simple calculation, on the same principles as described earlier [8]), it follows from (10) that:

$$\left.\begin{aligned} \Delta a &= \Delta v \sin \omega_0 t, \\ \Delta b &= \Delta v \cos \omega_0 t, \end{aligned}\right\} \quad \dots\dots \quad (11)$$

where $\Delta a$ and $\Delta b$ are the fluctuations of $a$ and $b$ [10]).

To simplify the notation we put $I_0 e = A$, so that (9) becomes

$$\overline{i_s^2} = \frac{A}{2T}. \quad \dots\dots \quad (12)$$

Using the relation $\Delta t = T$, we can now easily find from (9), (10), (11) and (12) the expression [7]):

$$\overline{(\Delta a)^2} = \overline{(\Delta b)^2} = \frac{A \Delta t}{4C^2}. \quad \dots \quad (13)$$

Taking into account the damping $r/L$ — which is proportional to the bandwidth — this result can be integrated with respect to time $t$, and we thus find the mean square values $\sigma^2$ and $V_0^2$:

$$\left.\begin{aligned} \sigma^2 = \overline{a^2} = \overline{b^2} &= \frac{A}{4C^2}\frac{L}{r}, \\ \overline{V_m^2} = V_0^2 = \overline{a^2 + b^2} &= \frac{A}{2C^2}\frac{L}{r}. \end{aligned}\right\} \quad (14)$$

Here $V_0$ is the r.m.s. value of the voltage amplitude ("instantaneous peak value") $V_m$ of the filtered noise, as illustrated in *fig. 4*. Since the bandwidth is given by $r/L$, except for a numerical factor,

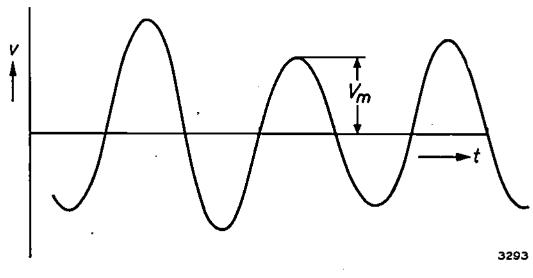


Fig. 4. Illustrating what is meant by the "instantaneous peak value" $V_m$ of the filtered noise $v$. The quantity $V_0$, which occurs in (14), is the r.m.s. value of $V_m$.

[9]) A. Blanc-Lapierre and R. Fortet, Analyse spectrale de l'énergie dans les phénomènes de fluctuations, Ann. Télécomm. 2, 222-230, 1947.

[10]) Without carrying out the calculation, one can see that (11) is correct by noting that (11) is identical with the earlier result at moments where $a$ or $b$ is zero.

it follows from (14) that $V_0$ depends only on the bandwidth, and not on the frequency. The latter agrees with the characteristic of white noise, which is that all audible frequencies are uniformly represented in its spectrum.

In order that $v$ may have any phase, $a$ and $b$ must be able to become either positive or negative with equal probability. In that context it may be assumed that the values of $a$ and $b$ are distributed at either side of the zero value in a Gaussian or normal-distribution curve as shown in *fig. 5a*. In this figure the abscissa is proportional to the amplitude $a$ of the sine component, whilst the ordinate is proportional to the probability $\mathrm{d}P$ with which a given amplitude $a$ occurs. The most probable amplitude of $a$ is seen to be zero; the same of course holds for the amplitude $b$ of the cosine component. The proportionality factors on abscissa and ordinate are so chosen that the total (integrated) probability $P$ has a value of 1. For this reason the abscissa is taken as the ratio $x = a/\sigma$, that is, the ratio of $a$ to its "standard deviation" $\sigma$ given by (14). The ordinate is then $\mathrm{d}P/\mathrm{d}x = \sigma\,\mathrm{d}P/\mathrm{d}a$. The mathematical expression for the normal-distribution curve in fig. 5$a$ is:

$$\mathrm{d}P = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-a^2}{2\sigma^2}\right)\mathrm{d}a. \qquad (15)$$

This may be read as expressing that $\mathrm{d}P$ is the probability that $a$ lies between $a$ and $a + \mathrm{d}a$. An entirely analogous formula holds for the amplitude $b$ of the cosine component. The fact that $a$ and $b$ will indeed exhibit the behaviour described follows from general considerations of probability theory. It may be noted in this connection that the binomial distribution of $a_n$ as given by Pascal's triangle approximates to a normal distribution for large values of $n$.

The statistical distribution of the total amplitude $V_m = \sqrt{a^2 + b^2}$, which is always positive, can now be found by introducing polar coordinates in the $a$-$b$ plane and then integrating [9]). This yields:

$$\mathrm{d}P = \frac{2V_m}{V_0^2} \exp\left(-\frac{V_m^2}{V_0^2}\right)\mathrm{d}V_m, \qquad (16)$$

where $\mathrm{d}P$ is the probability that the (positive) instantaneous peak value $V_m$ of the voltage $v$ lies between $V_m$ and $V_m + \mathrm{d}V_m$. Equation (16) does not represent a Gaussian but a Rayleigh distribution, illustrated by the solid curve in fig. 5$b$.

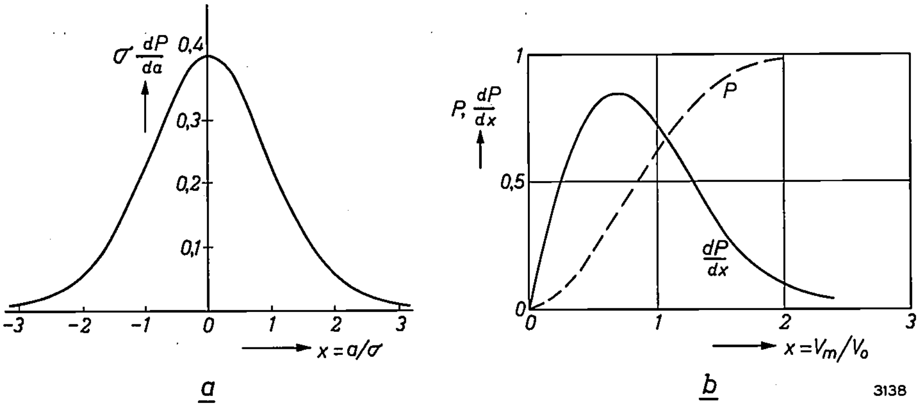


Fig. 5. $a$) Normal (Gaussian) curve representing the distribution of the varying amplitudes $a$ and $b$ of the two components of the voltage $v$ in fig. 3.
$b$) The solid curve represents a Rayleigh distribution ($\mathrm{d}P$ is the probability that $V_m$ lies between $V_m$ and $V_m + \mathrm{d}V_m$). The dashed curve represents the probability $P$ that $V_m/V_0$ is smaller than $x$.

Equation (16) can easily be found with the aid of *fig. 6*. The probability that the total amplitude lies within the elementary rectangle $\mathrm{d}a\mathrm{d}b$ is, according to (15):

$$\mathrm{d}P = \frac{1}{2\pi\sigma^2} \exp\left(\frac{-V_m^2}{2\sigma^2}\right)\mathrm{d}a\mathrm{d}b,$$

where $V_m = \varrho = \sqrt{a^2 + b^2}$. The surface element $\mathrm{d}a\mathrm{d}b$ can be directly integrated along the periphery of the circle, and then becomes $2\pi V_m \mathrm{d}V_m$. Including (14) in the calculations, we thus arrive at (16).

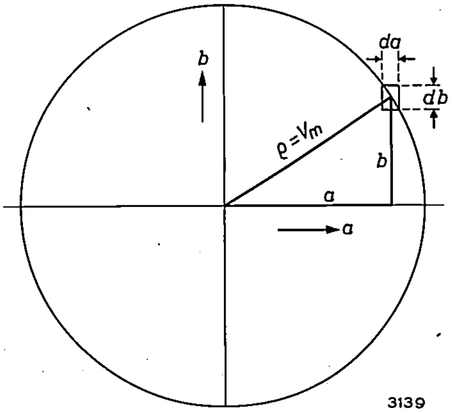


Fig. 6. Illustrating the derivation of equation (16).

By integrating (16) we can determine the probability $P(x)$ that $V_{\mathrm{m}}/V_0$ will be smaller than a given value $x$. We find:

$$P(x) = 1 - \mathrm{e}^{-x^2}. \quad . \quad . \quad . \quad . \quad (17)$$

The variation of $P(x)$ is represented by the dashed curve in fig. 5*b*. The chance that $V_{\mathrm{m}}$ will be greater than, say, $2V_0$, is then smaller than 2%.

*Statistical behaviour of the phase*

In the foregoing we have calculated, perhaps by a somewhat unconventional method, a number of well-known relations. We now come to less familiar properties, and in the first place to the behaviour shown by the *phase* of the filtered noise.

As shown in the article cited in reference [8]), the phase discontinuity produced in an oscillating *LC* circuit by a current pulse is given by:

$$\Delta\varphi = -\frac{\Delta v}{V_{\mathrm{m}}} \sin\varphi. \quad . \quad . \quad . \quad (18)$$

$\Delta v$ can be regarded as given by (10), $\varphi$ is the phase angle through which the oscillation has passed since the last positive voltage peak. The successive phase discontinuities are positive or negative and distributed in an entirely random manner. They may therefore be expected to add up in a way completely analogous to that discussed in connection with the Pascal triangle. On formal grounds alone, then, we may expect a relation of the form [7])

$$\overline{(\Delta\varphi)^2} = \frac{r}{L} T_\varphi, \quad . \quad . \quad . \quad . \quad . \quad (19)$$

where $r/L$ is proportional to the bandwidth, and $T_\varphi$ is the time difference over which the phase discontinuities may be thought to be summed. In general one should add to the right-hand side of (19) a numerical factor which depends on the network under consideration; for our simple *LC* circuit this factor is of the order of magnitude of unity.

Equation (19) is readily reducible to an expression for the average frequency deviation $\overline{\Delta\omega} = \overline{\omega - \omega_0}$ over the interval $T_\varphi$, since the identity

$$\Delta\varphi \equiv \overline{\Delta\omega}\, T_\varphi$$

and equation (19) yield directly:

$$\overline{(\overline{\Delta\omega})^2} = \frac{r}{L}\cdot\frac{1}{T_\varphi}. \quad . \quad . \quad . \quad . \quad . \quad (20)$$

An important point is that (20) has exactly the same form as (9) and (12).

We now consider a frequency detector which is "centred" on the centre frequency $\omega_0$. When the filtered noise is fed to this detector, it will deliver a voltage proportional to the instantaneous frequency deviation $\Delta\omega$. When averaged over a time interval $T_\varphi$, this rapidly varying $\Delta\omega$ gives the result expressed by (20), which varies with time in a similar way as (9) and (12). This warrants the conclusion that $\Delta\omega$ itself must have properties corresponding to those of our original noise current $\Delta I$. Equation (14) shows that this noise current contains all frequencies equally up to a certain limit. We may now assume the same in regard to $\Delta\omega$. In practical terms, this means that the frequency detector considered will deliver a voltage in which the energy is equally distributed over the relevant range of possible modulation frequencies of a frequency-modulated signal. A limit is imposed by the above-mentioned restriction concerning $T_\varphi$, and also by the bandwidth of the detector. From (20) it may be argued that the detected voltage will show the characteristic of white noise over a frequency range extending from the lowest audio frequencies up to at least $(\Delta\omega)_{\max}$.

We shall now put equations (20) and (12) into a different form, in order to bring out more clearly the uniform distribution of the power over the range of frequencies. For this purpose we write instead of (12) and (8) the formal expression:

$$\overline{(\Delta I)^2} = A\Delta f. \quad . \quad . \quad . \quad . \quad . \quad (21)$$

This expresses that our original noise current $\Delta I$ has an r.m.s. value proportional to the width of the filtered frequency band $\Delta f$, entirely in accordance with (14). The fact that (21) is indeed numerically equivalent to (12) and (8) can be demonstrated by showing that the current given by (21) would give rise to the fluctuating voltage, given by (14), over the *LC* circuit. The elementary calculation involved, which need not be given here, also shows what bandwidth $\Delta f$ is filtered out of the noise spectrum (21) by the *LC* circuit. This bandwidth is $r/4L$. Taking this to correspond to $2(\Delta\omega)_{\max}/2\pi$, we find:

$$r/L = 4\Delta f = 4(\Delta\omega)_{\max}/\pi.$$

This proportionality between $r/L$ and $(\Delta\omega)_{\max}$ will presently be used again.

In view of the equivalence of (21) and (12), and because $i_{\mathrm{s}} = \overline{\Delta I}$, we may now also replace (20) by

$$\overline{(\Delta\omega)^2} = \frac{2r}{L}\,\Delta f. \quad . \quad . \quad . \quad . \quad (22)$$

This equation again shows that the detected (audio-frequency) voltage will exhibit the character of white noise.

We now have all the data needed to describe the effects observed in the reception of a strongly disturbed FM signal.

### Extension of equation (5) to the case of a relatively high interference level

In the case of a noise disturbance which is constantly smaller than the signal, the signal-to-noise ratio at the detector output is given, as we have found, by the equation

$$(S/N)^2 = \tfrac{3}{8}\,k^2y^2 . \quad . \quad . \quad . \qquad (5)$$

The signal power $S^2$ is then given by:

$$S^2 = \tfrac{1}{2}D(\Delta\omega)_{\max}{}^2, \quad . \quad . \quad . \quad . \qquad (23)$$

where $D$ is a proportionality factor. The noise power $N^2$ after detection follows from (5) and (23):

$$N^2 = \tfrac{4}{3}\,D\,\frac{(\Delta\omega)_{\max}{}^2}{k^2y^2} . \quad . \quad . \quad . \qquad (24)$$

Where, however, the disturbance is weaker than the signal only during the part of the time given by (17), a factor $(1 - e^{-y^2/k})$ must be added to the right-hand side of (23). The detected noise now consists of two terms, $N_1{}^2$ and $N_2{}^2$. The first is the "normal" noise, given by (24), likewise with a factor $(1 - e^{-y^2/k})$ on the right-hand side. The second term, $N_2{}^2$, accounts for the "anomalous" noise, which arises when the noise disturbance is stronger than the signal and the noise disturbance itself is detected, as discussed above with reference to fig. 1. Taking into account that the detected noise is disturbing only in so far as it lies within the audio-frequency band $q_a$ to be transmitted, and that $q_a = 2(\Delta\omega)_{\max}/k$ and $r/L = 4(\Delta\omega)_{\max}/\pi$, it follows from (22) that

$$N_2{}^2 = \frac{8}{\pi^2}\,D\,\frac{(\Delta\omega)_{\max}{}^2}{k}\,e^{-y^2/k} . \quad . \quad . \quad . \qquad (25)$$

The total noise power $N^2$ is the sum of $N_1{}^2$ and $N_2{}^2$.

After including the above-mentioned factor $(1 - e^{-y^2/k})$ in the calculation, it follows from (23), (24) and (25) that the signal-to-noise power ratio at the output of an FM detector is given by:

$$(S/N)^2 = \frac{\tfrac{3}{8}\,k^2y^2}{1 + \dfrac{6}{\pi^2}\,\dfrac{ky^2}{e^{y^2/k} - 1}} . \quad . \quad . \qquad (26)$$

The numerator on the right-hand side of (26) is the square of the signal-to-noise ratio after detection for the case of weak interference (equation 5). The denominator indicates to what extent this ratio is affected by the circumstance that the disturbance is sometimes stronger than the signal. The more favourable the signal-to-noise ratio at the receiver input (i.e. the larger $y^2/k$ is with respect to unity), the closer does the denominator approach unity.

In *fig.* 7 $(S/N)^2$ calculated from (26) is plotted against the "normalized" signal-to-noise power ratio $y^2$, defined by (4), for various values of the bandwidth ratio $k$. The scale has been chosen so as to make a direct comparison possible with the results of other theoretical investigations [3][11]. For AM with single-sideband transmission, $k = 1$ and $S/N = y$ (see equation (6)); this case thus corresponds to the dashed line for $k = 1$. The solid curves relate to equation (26) for bandwidths $2(\Delta\omega)_{\max}$ which are greater than the audio-frequency band $q_a$ by a factor $k = 4$, 10 and 25. As $y$ increases, these curves approach asymptotically to the dashed straight lines, which represent equation (5) — the case of weak interference.

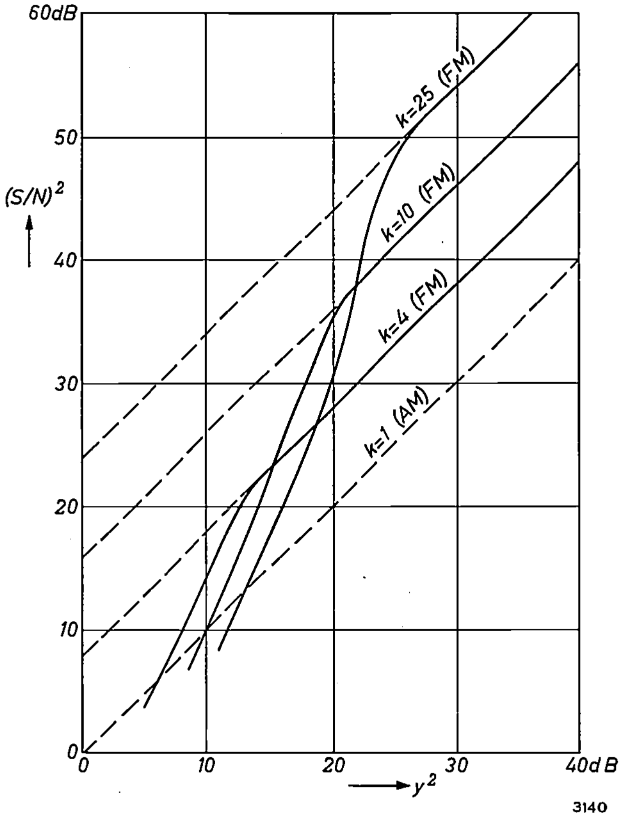


Fig. 7. The ordinate $(S/N)^2$ is the signal-to-noise ratio after detection; the abscissa $y^2$ is the ratio of the RF signal power to the average noise power within the frequency band $q_a$, at the receiver input. The parameter $k$ is the ratio of the total RF bandwidth $2(\Delta\omega)_{\max}$ to $q_a$; the value $k = 1$ applies to AM for single-sideband transmission, values $k > 1$ apply to FM. The solid curves relate to equation (26). As $y^2$ decreases, the advantage of a large bandwidth becomes smaller and smaller and eventually a larger bandwidth becomes a disadvantage.

[11] F. de Jager, Les limites théoriques de la transmission en cas de niveau de bruit élevé, Onde électr. 74, 675-682, 1954. See also Philips tech. Rev. 19, 75 (fig. 1), 1957/58.

From the shape of the solid curves it can be seen that the larger the maximum frequency excursion is made, the ratio $y^2$ decreasing, the sooner does a deviation from (5) become noticeable. This means that increasing the bandwidth — which is favourable under weak interference — has an adverse effect when the interference is *not* always weaker than the signal. For example, the case $k = 25$, compared with $k = 10$, is in the advantage as long as $y^2 >$ 22 dB, but is disadvantageous where $y^2 <$ 22 dB.

In practice, an advantage of at least 15 to 20 dB is required from FM as compared with single-sideband AM before the former is economically justifiable. Fig. 7 shows that this advantage is gained where $k \approx 10$, i.e. where $(\Delta\omega)_{\max} \approx 5q_a$, a figure which is generally accepted as standard for radio broadcasting and television purposes.

In fig. 7 the solid curves are broken off on the left at values of $y^2$ in the neighbourhood of $k$. The reason is that our formula (26) differs somewhat from the theoretical results [3][4] at values of $y^2/k$ smaller than unity, at which the power ratio of RF signal and noise thus also becomes smaller than unity. This disparity is of little practical importance, since FM at such high noise levels gives no appreciably better reception than single-sideband AM, whilst the larger bandwidth then needed for FM is an overriding objection. In the field where the use of FM is interesting, however, the method of approach described above is in good agreement both with the theoretical calculations referred to and with earlier published results of measurements [12]. It therefore provides a satisfactory picture of the effects observed.

[12] M. G. Crosby, Frequency modulation noise characteristics, Prod. Inst. Radio Engrs. **25**, 472-514, 1937.

Summary. By means of a vector diagram it is shown that the sum of two signals modulated in frequency (or phase) will approximately follow the phase of the stronger signal. This means that for reception of disturbed FM signals, during the moments when the interference is stronger than the desired signal, it is largely only the disturbance that is detected. Taking into account the "anomalous" detected interference then occurring, it is possible to extend a familiar elementary formula for the signal-to-noise ratio to the case of a relatively high interference level. To do this, it is necessary to know the statistical behaviour of both the amplitude and phase of a noise disturbance. Both are derived in an elementary way from a somewhat simplified model of such a disturbance. The method adopted (earlier described in this journal) consists in integrating the reaction of an oscillatory network to short interfering pulses. A formula is found which satisfactorily describes the effects.

# *P-N* LUMINESCENCE IN GALLIUM PHOSPHIDE

535.376

The emission of light by a solid as a result of the direct conversion of electrical energy is known as electroluminescence. The most widely studied form of this is the Destriau effect [1], where collision ionization in a sufficiently strong electric field gives rise to electrons and holes capable of causing light emission. (The production of electrons and holes is another way of describing what was referred to for simplicity in the article cited under [1] as the excitation of certain bound electrons.) Another form of electroluminescence is the Lossev effect or *P-N* luminescence [2]; here, by contrast with the first case, electrons and holes are already present in the solid in the unexcited state. The effect may be described as follows.

In a semiconductor, one part of which shows hole conduction and another part electron conduction, a potential barrier [3] occurs in the transition region, i.e. at the *P-N* junction. By electrically biasing the junction in the forward direction, the blocking action of the potential barrier is reduced and minority charge carriers are injected, i.e. electrons from the *N* region enter the *P* region, and holes from the *P* region penetrate into the *N* region. These minority charge carriers recombine with the numerous charge carriers present of opposite sign, either directly or via a level between the valence and conduction bands that acts as a trap for holes or electrons. In some solids the energy released upon this recombination is emitted in the form of light radiation. This is known as the Lossev effect.

[1] See G. Diemer, H. A. Klasens and P. Zalm, Electroluminescence and image intensification, Philips tech. Rev. **19**, 1-11, 1957/58.

[2] O. W. Lossev, Phys. Z. **34**, 397, 1933; C. R. Acad. Sci. U.R.S.S. **39**, 363, 1940. K. Lehovec, C. A. Accardo and E. Jamgochian, Phys. Rev. **89**, 20, 1953. See also: C. A. A. J. Greebe and W. F. Knippenberg, Grown *P-N* junctions in silicon carbide, Philips Res. Repts. **15**, 120-123, 1960 (No. 2).

[3] For some of the semiconductor concepts used here, in particular relating to the band scheme, energy gap, conduction mechanisms, etc., see e.g. R. E. J. King and B. E. Bartlett, Properties and applications of indium antimonide, Philips tech. Rev. **22**, 217-225, 1960/61 (No. 7).

From the shape of the solid curves it can be seen that the larger the maximum frequency excursion is made, the ratio $y^2$ decreasing, the sooner does a deviation from (5) become noticeable. This means that increasing the bandwidth — which is favourable under weak interference — has an adverse effect when the interference is *not* always weaker than the signal. For example, the case $k = 25$, compared with $k = 10$, is in the advantage as long as $y^2 >$ 22 dB, but is disadvantageous where $y^2 < 22$ dB.

In practice, an advantage of at least 15 to 20 dB is required from FM as compared with single-sideband AM before the former is economically justifiable. Fig. 7 shows that this advantage is gained where $k \approx 10$, i.e. where $(\Delta\omega)_{\max} \approx 5q_a$, a figure which is generally accepted as standard for radio broadcasting and television purposes.

In fig. 7 the solid curves are broken off on the left at values of $y^2$ in the neighbourhood of $k$. The reason is that our formula (26) differs somewhat from the theoretical results [3][4] at values of $y^2/k$ smaller than unity, at which the power ratio of RF signal and noise thus also becomes smaller than unity. This disparity is of little practical importance, since FM at such high noise levels gives no appreciably better reception than single-sideband AM, whilst the larger bandwidth then needed for FM is an overriding objection. In the field where the use of FM is interesting, however, the method of approach described above is in good agreement both with the theoretical calculations referred to and with earlier published results of measurements [12]. It therefore provides a satisfactory picture of the effects observed.

[12] M. G. Crosby, Frequency modulation noise characteristics, Prod. Inst. Radio Engrs. **25**, 472-514, 1937.

Summary. By means of a vector diagram it is shown that the sum of two signals modulated in frequency (or phase) will approximately follow the phase of the stronger signal. This means that for reception of disturbed FM signals, during the moments when the interference is stronger than the desired signal, it is largely only the disturbance that is detected. Taking into account the "anomalous" detected interference then occurring, it is possible to extend a familiar elementary formula for the signal-to-noise ratio to the case of a relatively high interference level. To do this, it is necessary to know the statistical behaviour of both the amplitude and phase of a noise disturbance. Both are derived in an elementary way from a somewhat simplified model of such a disturbance. The method adopted (earlier described in this journal) consists in integrating the reaction of an oscillatory network to short interfering pulses. A formula is found which satisfactorily describes the effects.

# *P-N* LUMINESCENCE IN GALLIUM PHOSPHIDE

535.376

The emission of light by a solid as a result of the direct conversion of electrical energy is known as electroluminescence. The most widely studied form of this is the Destriau effect [1], where collision ionization in a sufficiently strong electric field gives rise to electrons and holes capable of causing light emission. (The production of electrons and holes is another way of describing what was referred to for simplicity in the article cited under [1] as the excitation of certain bound electrons.) Another form of electroluminescence is the Lossev effect or *P-N* luminescence [2]; here, by contrast with the first case, electrons and holes are already present in the solid in the unexcited state. The effect may be described as follows.

In a semiconductor, one part of which shows hole conduction and another part electron conduction, a potential barrier [3] occurs in the transition region, i.e. at the *P-N* junction. By electrically biasing the junction in the forward direction, the blocking action of the potential barrier is reduced and minority charge carriers are injected, i.e. electrons from the *N* region enter the *P* region, and holes from the *P* region penetrate into the *N* region. These minority charge carriers recombine with the numerous charge carriers present of opposite sign, either directly or via a level between the valence and conduction bands that acts as a trap for holes or electrons. In some solids the energy released upon this recombination is emitted in the form of light radiation. This is known as the Lossev effect.

[1] See G. Diemer, H. A. Klasens and P. Zalm, Electroluminescence and image intensification, Philips tech. Rev. **19**, 1-11, 1957/58.

[2] O. W. Lossev, Phys. Z. **34**, 397, 1933; C. R. Acad. Sci. U.R.S.S. **39**, 363, 1940. K. Lehovec, C. A. Accardo and E. Jamgochian, Phys. Rev. **89**, 20, 1953. See also: C. A. A. J. Greebe and W. F. Knippenberg, Grown *P-N* junctions in silicon carbide, Philips Res. Repts. **15**, 120-123, 1960 (No. 2).

[3] For some of the semiconductor concepts used here, in particular relating to the band scheme, energy gap, conduction mechanisms, etc., see e.g. R. E. J. King and B. E. Bartlett, Properties and applications of indium antimonide, Philips tech. Rev. **22**, 217-225, 1960/61 (No. 7).

The height of the potential barrier at the *P-N* junction is roughly equal to the energy band gap of the semiconductor, i.e. the energy difference between the top of the valence band and the bottom of the conduction band [3]). At the most, therefore, it amounts to a few volts, and this is also the order of magnitude of the DC voltage to be applied in order to excite *P-N* luminescence of reasonable intensity. Light emission by the Destriau effect requires, as explained in the article [1]) referred to, considerably higher voltages, usually alternating.

The fact that *P-N* luminescence essentially occurs at low DC voltages makes it attractive as a light source for electro-optical switching devices [4]), and for this reason it is the subject of a great deal of research.

*P-N* luminescence in the visible spectral region has hitherto only been observed in silicon carbide [2]), aluminium phosphide [5]) and gallium phosphide [6]). The efficiency was very low. In recent times, by improving the activation, we have succeeded in raising the efficiency of *P-N* luminescence in GaP crystals by a factor between 100 and 1000 compared with the figures previously achieved. This activation amounts to the creation of both "deep" and "shallow" levels in relation to the neighbouring band (valence or conduction). The deep levels are important in order to give the radiative recombination a good chance against competitive processes. The shallow levels, although they have poor trapping properties, are necessary to obtain reasonable electrical conductivity in the parts of the crystal outside the *P-N* junction.

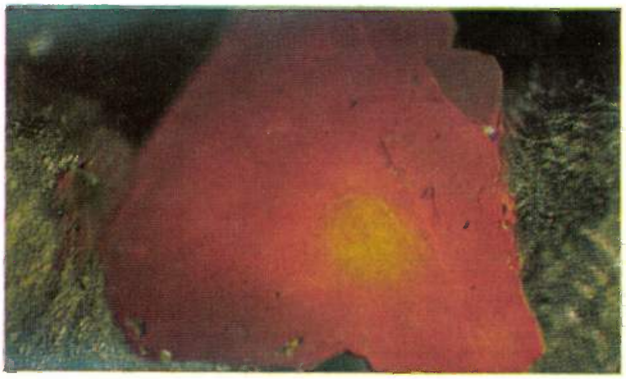


Fig. 1. *P-N* luminescence from a polycrystalline sample of gallium phosphide, on Kodachrome film (artificial light, reversal film), exposure one minute. Magnification approx. 100 ×. To bring out the contours of the sample and the contacts clearly, supplementary lighting was used.

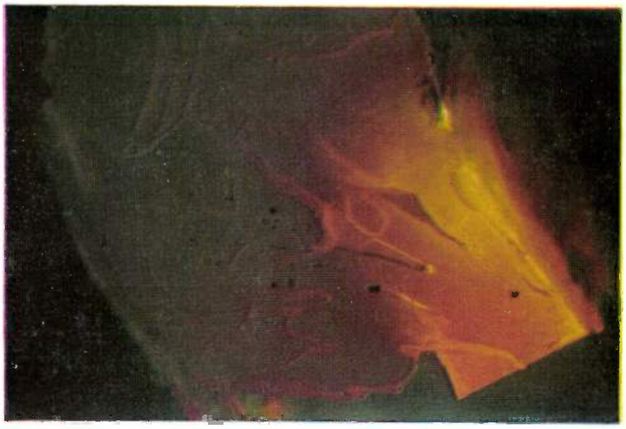


Fig. 2. As fig. 1, with a different GaP sample and without supplementary lighting. Magnification approx. 60×.

The GaP crystals can emit green, yellow and red light. *Figs. 1* and *2* show colour photos of strongly yellow and red emissive samples. The contacts visible in fig. 1 consist of metallic gallium, which melts at about 30 °C. Between the contacts a DC potential of 4 volts is applied. With the intensities of luminescence so far achieved it is possible to envisage practical applications on the lines mentioned above. For example, if one of our GaP samples is combined with a photoresistor of cadmium sulphide [7]), type LDR-03, the *P-N* luminescence makes it possible to reduce the resistance of the cadmium sulphide from $10^9$ Ω to $3 \times 10^3$ Ω.

The GaP used for the photos was polycrystalline. The *P-N* junctions needed for the luminescence occur naturally here at the grain boundaries. The nature and the situation of the *P-N* junctions cannot yet be properly controlled, so that in addition to strongly emissive GaP samples many weakly emissive ones are obtained, and some that are not emissive at all. Further improvement is expected when it becomes possible to make single crystals of GaP in which *P-N* junctions can be introduced in a reproducible manner.

H. G. GRIMMEISS *) and H. KOELMANS.

[4]) G. Diemer and J. G. van Santen, Philips Res. Repts. **15**, 368, 1960 (No. 4).

[5]) H. G. Grimmeiss, W. Kischio and A. Rabenau, Phys. Chem. Solids **16**, 302, 1960 (No. 3/4).

[6]) H. G. Grimmeiss and H. Koelmans, Philips Res. Repts. **15**, 290, 1960 (No. 3).

[7]) N. A. de Gier, W. van Gool and J. G. van Santen, Philips tech. Rev. **20**, 277, 1958/59.

*) Philips Zentrallaboratorium GmbH, Laboratorium Aachen.

# LATTICE IMPERFECTIONS IN METALS, SEMICONDUCTORS AND IONIC CRYSTALS

by H. G. van BUEREN.

548.4

*It may now be taken as common knowledge that disturbances on an atomic scale in the regular ordering of the atoms or ions in a crystal lattice have a very pronounced influence on the macroscopic properties of the crystal, or may sometimes entirely govern them. Lattice imperfections have many aspects, and the consequences of their presence differ very widely from one substance to another. This appears to be bound up with differences in crystal structure and the binding energies between the atoms; the size of the crystals may also play some part. This article discusses some of the more outstanding points of difference as regards their effects on the electrical and mechanical properties of crystals.*

Knowledge of the nature and influence of deviations from exact periodicity of crystal lattices (lattice imperfections) is no longer the domain of only a small group of specialists: it has become an essential part of the whole fund of knowledge required for solid state research. Moreover, it has proved possible in recent years to provide spectacular and convincing visible evidence of the most disputed group of lattice imperfections — dislocations. Among the means employed to this end are the techniques of "decorating" dislocations [1]), and the electron microscopy of thin foils [2]). This work has shown that nearly all the major predictions made in this field in the last fifteen years (e.g. regarding the structure and behaviour of dislocations) were well founded. The theory of the imperfect crystal, which in those years had largely been worked out on paper, has thus been given a firm experimental foundation.

*Figs. 1* and *2* illustrate the striking agreement between theory and observation with regard to the Frank-Read source and the dislocation network. These two concepts of dislocation theory were dealt with at some length in an article published some years ago in this journal [3]) on the principles of the theory of simple physical lattice imperfections — vacancies, interstitial atoms and dislocations. The emphasis then was placed on the influence of lattice imperfections in metals. Since that time there have been innumerable investigations on other types of solids. These investigations have revealed a number of interesting and instructive differences which threw new light not only on the properties of the lattice imperfections themselves, but also on the substance in which they occur. Various such differences will be reviewed in this article. Our choice

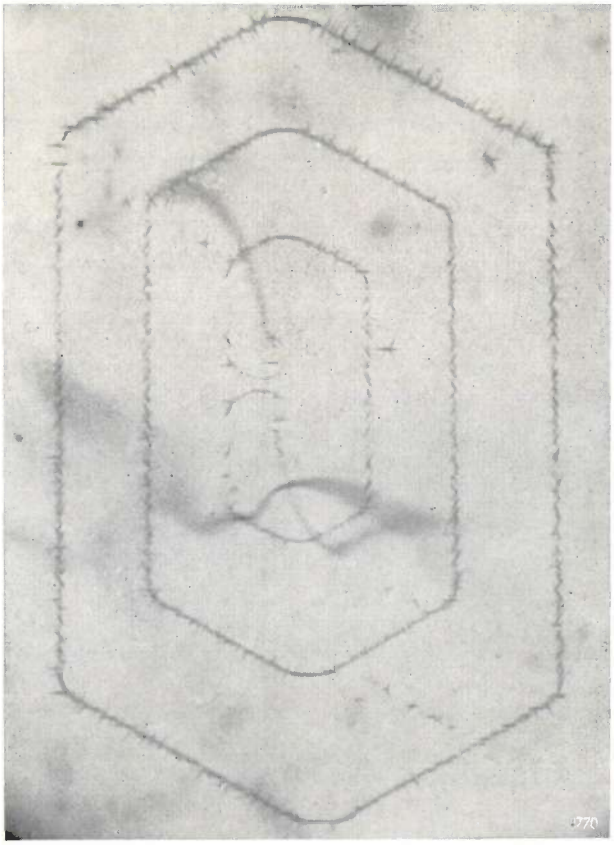


Fig. 1. Frank-Read dislocation source in a silicon crystal. The dislocations are made visible by "decoration" with copper particles precipitated round the dislocation lines. Compare this figure with fig. 13 in reference [3]). (Taken from W. C. Dash, J. appl. Phys. 27, 1193, 1956.)

[1]) A. Hedges and D. Mitchell, Phil. Mag. **44**, 223 and 357, 1953; S. Amelinckx, Phil. Mag. **1**, 269, 1956.

[2]) P. B. Hirsch, R. W. Horne and M. J. Whelan, Phil. Mag. **1**, 677, 1956.

[3]) H. G. van Bueren, Lattice imperfections and plastic deformation in metals, I. Nature and characteristics of lattice imperfections, notably dislocations; II. Behaviour of lattice imperfections during deformation, Philips tech. Rev. **15**, 246-257 and 286-295, 1953/54.

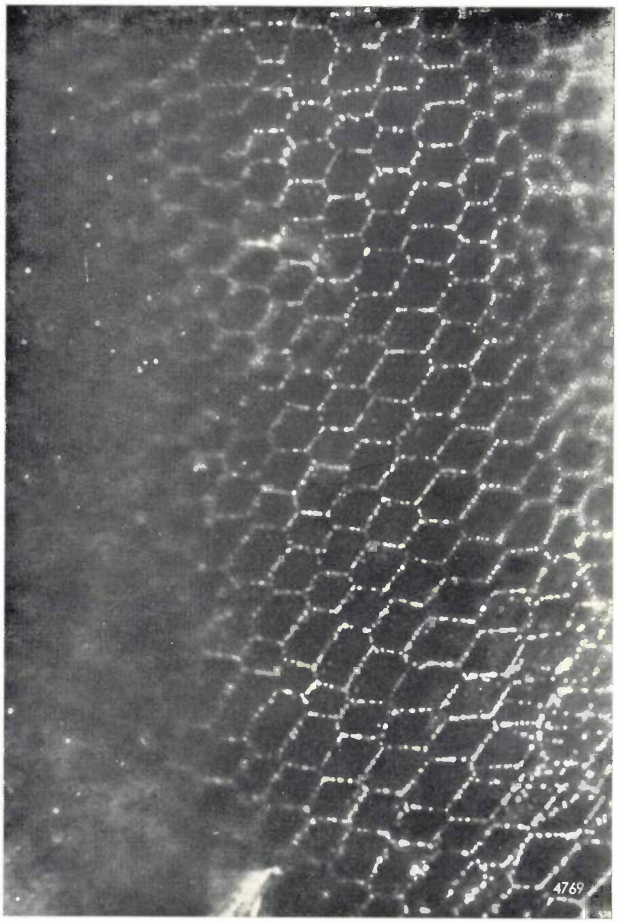


Fig. 2. Hexagonal dislocation network in a KCl crystal, made visible by precipitating silver particles on the dislocation lines. (Taken from S. Amelinckx, Acta metallurgica 6, 34, 1958.)

must necessarily be limited, and we shall be primarily concerned with examining the influence of vacancies and dislocations. Composite lattice disturbances, such as crystal boundaries and chemical imperfections (foreign atoms) will not be discussed.

### Formation of lattice imperfections

During the growth of a crystal, apart from vacancies large numbers of dislocations are formed. These are due either to thermal stresses (e.g. on growing from the melt) or to the condensation of vacancies (see below), or again to the low mobility of the atoms making up a crystal (an atom that has arrived at the wrong site is then unable to leave it). The latter case may arise, for example, when a crystal has grown at a relatively low temperature from the vapour phase. The process by which dislocations are formed may in a certain sense be called autocatalytic: where a few dislocations are present, they normally promote the creation of further dislocations, either because the dislocations act as sources for new ones or as a consequence of the occurrence of the Frank spiral-growth mechanism.

All these possibilities, including the way in which plastic deformation (which may be due to thermal stresses) can give rise to large numbers of dislocations, vacancies and interstitial atoms, have already been discussed in this journal [3][4][5]. Reference has also been made to the way in which rapid cooling (quenching) may freeze-in large concentrations of vacancies that are not in thermodynamic equilibrium. When the temperature has been raised high enough to make these vacancies mobile, they condense into certain dislocation configurations whose shape differs according to the substance in which they occur (*figs. 3* and *4*). This extremely significant effect of *vacancy condensation*, leading to the formation of dislocations, was theoretically predicted just before it was observed with the aid of a new experimental technique — electron transmission microscopy of thin foils [6]).

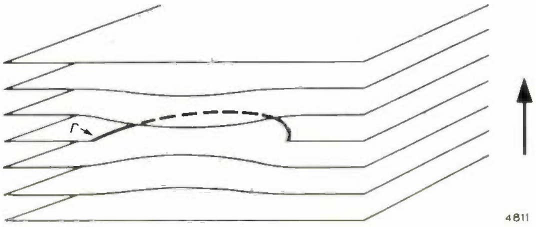


Fig. 3. Formation of dislocation *loop* by the condensation of a disc-shaped accumulation of vacancies and its subsequent collapse. The drawing shows a number of crystal lattice planes (only half-planes are drawn, for clarity). The edge *r* of the inserted "extra" half-plane, which constitutes the dislocation, takes the form of a ring perpendicular to the plane of the drawing. The Burgers vector (arrow) is perpendicular to the plane of the loop.

Large concentrations of lattice imperfections out of thermodynamic equilibrium, particularly of vacancies and interstitial atoms, can also be formed by irradiation with high-energy electrons, neutrons or other particles. Such point defects are usually formed by the direct displacement of lattice atoms, a process which has many interesting facets but with which we shall not be concerned here; an impression of the process is given in *fig. 5*. Imperfections can also, however, be formed indirectly, and

[4]) P. Penning, The generation of dislocations by thermal stresses, Philips tech. Rev. **19**, 357-364, 1957/58.

[5]) B. Okkerse, A method of growing dislocation-free germanium crystals, Philips tech. Rev. **21**, 340-345, 1959/60 (No. 11).

[6]) P. B. Hirsch, J. Silcox, R. E. Smallman and K. H. Westmacott, Phil. Mag. **3**, 897, 1958. This effect had been theoretically predicted some months before by D. Kuhlmann-Wilsdorf.

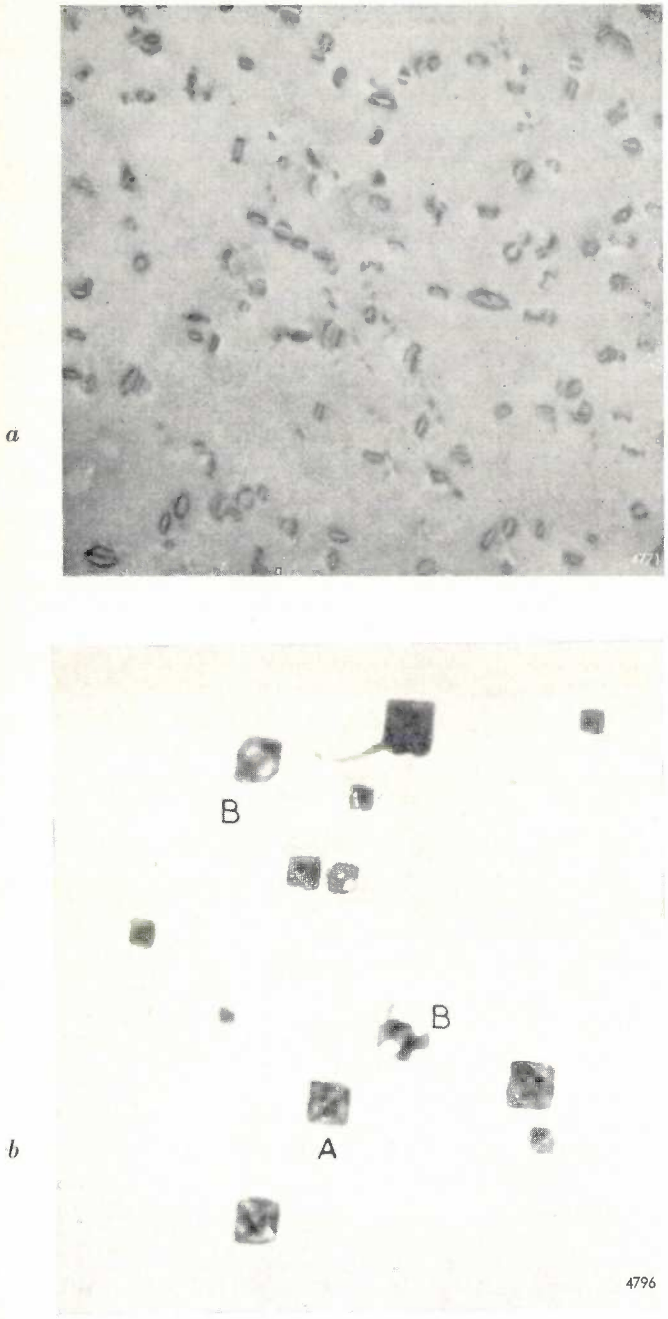


Fig. 4. Two photographs, taken by electron transmission microscopy of thin metal foils. The phenomena represented are related to those sketched in fig. 3.
*a*) Dislocation loops in aluminium foil, quenched from 560° and aged at room temperature. *b*) Tetrahedral condensation figures in quenched gold foil, aged at 180 °C. The latter figure differs from the former because a simple dislocation loop in gold is not stable but changes into a tetrahedral configuration of four stacking faults in four different octahedral planes. Both cases, however, arise in just the same way from vacancy condensation. (Taken from P. B. Hirsch, J. Inst. Metals **87**, 406, 1958/59 and from J. Silcox and P. B. Hirsch, Phil. Mag. **4**, 72, 1959.)

it is here that we find a characteristic difference between two types of substances, in this case between metals and ionic crystals. Whereas in metals all electrostatic disturbances produced during the irradiation are completely screened-off within one interatomic spacing by the electron gas, so that in practice they have little or no influence on the structure, in an ionic crystal a disturbance of the state of charge of an ion endangers the links with the neighbouring ions. For example a singly charged chlorine anion in a crystal of rocksalt may easily be doubly ionized upon irradiation with fast electrons, much more easily than it can be completely displaced. After double ionization, however, it has become a positive ion, occupying a lattice site intended for negative ions (*fig. 6*). Obviously, the surrounding positive sodium ions will exert a repulsion on the positive chlorine ion. As a result the ion is pushed into an interstitial site where it feels less unhappy, leaving behind a chlorine vacancy. In spite of the relatively minor damage caused by the irradiation directly, the lattice *itself* subsequently makes a contribution which converts the small temporary damage into a permanent imperfection. In metals there is no question of any such lattice contribution. In the fundamental study of radiation damage it is therefore not permissible to regard

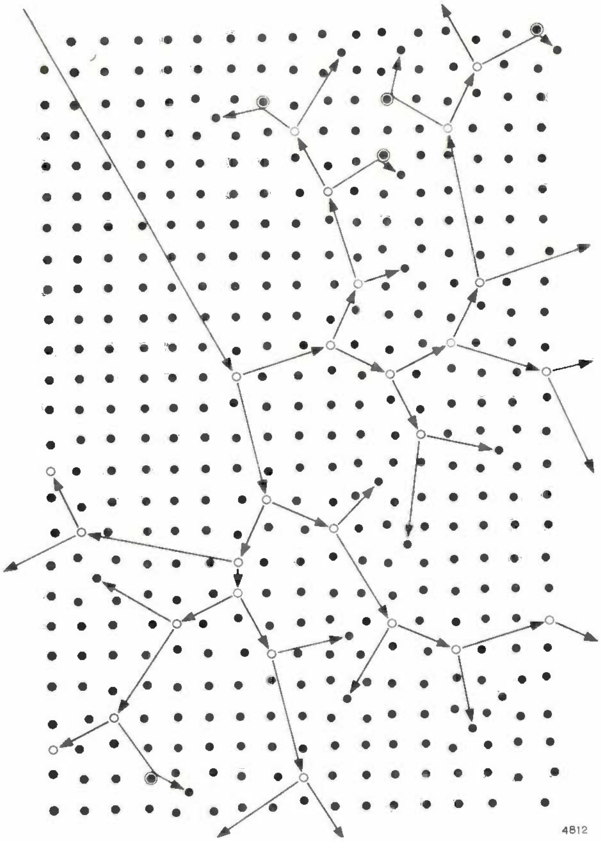


Fig. 5. Schematic representation of the imperfections that may be produced by irradiating a crystal with high-energy neutrons (after D. S. Billington, Scientific American **201**, No. 3, p. 200, 1959). Vacancies (open circles) and interstitial atoms are formed by sequences of collisions in cascade.

metals (hitherto the most investigated group) as the representative class, and to apply the results obtained on metals to other substances. The existence of other indirect consequences of irradiation, which are present in some substances and absent in others, has already been incidentally demonstrated, among other things in organic substances ("cross-linking" in polymers).

After these prefatory comments on the formation of lattice imperfections, we shall now discuss some groups of phenomena in which there is a striking difference between various types of substance. Since we have to be selective, we shall consider in turn two distinct examples, one in the category of electrical phenomena and the other relating to mechanical properties. We shall not be concerned with optical, magnetic or resonance processes, nor with the technically very important influence which lattice imperfections have on diffusion and on the various forms of chemical reactivity.

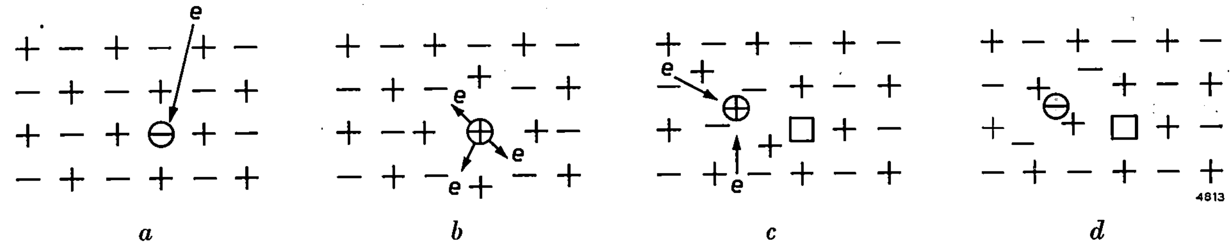


Fig. 6. Schematic representation of the consequences of the double ionization of a negative ion in an NaCl lattice, under the influence of a collision with a high-energy electron (*a*). The ion becomes positive (*b*), is then no longer in a stable position and moves to an interstitial site leaving behind a vacancy (*c*). The ion may then again capture the missing electrons, giving rise to an interstitial negative ion (*d*) together with a vacancy.

This mechanism for the creation of interstitial negative ions was put forward by J. H. O. Varley (Nature **174**, 856, 1954). It should be noted that the analogous formation of interstitial positive ions is not possible.

## Influence of lattice imperfections on electrical conductivity

Dislocations and vacancies reduce the electrical conductivity of metals by scattering conduction electrons, in a manner analogous to the scattering caused by thermal lattice vibrations [3][7]. Whilst the occurrence and influence of the latter are dependent on temperature, this is not so in the case of lattice imperfections (as far as non-equilibrium concentrations are concerned and disregarding recovery processes). To a first approximation, therefore, Matthiessen's rule applies which states that the reduced mobility of electrons in metals as a result of lattice imperfections, is more or less independent of temperature. The *mobility* $\mu$, as used in solid-state physics, is defined by:

$$\gamma = n e \mu , \quad \ldots\ldots\ldots \tag{1}$$

where $\gamma$ is the conductivity, $n$ the number of conduction electrons per unit volume and $e$ the elementary charge.

The reduction of $\mu$, itself an interesting effect, is the only effect that lattice imperfections have on the transport of charge in metals. As mentioned, any electrostatic effects present have no effect on the structure and therefore on the electron density, since all disturbances of the charge equilibrium are screened off within one atomic distance from the dislocation. The situation is entirely different in semiconductors, where there are valence bonds between the atoms and only few charge carriers are present. Lattice imperfections here affect not only the mobility $\mu$, but also as a rule the concentration $n$ of the charge carriers, owing to the fact that the imperfections may act as donors or acceptors [8]. It is therefore in general not possible to say whether the introduction of lattice imperfections will increase or reduce the electrical conductivity of such covalent semiconductors: this depends on the concentration of the conduction electrons. For example, on plastic deformation the resistivity of N-type germanium is increased owing to the introduction of vacancies and dislocations, the reason being that both these imperfections act as acceptor centres and thus reduce the concentration of conduction electrons. The higher the concentration of vacancies the lower the electron concentration, until

[7] See e.g. J. Volger, Philips tech. Rev. **22**, 226, 1960/61 (No. 7).

[8] For the principles of semiconductor theory, and the definition of various terms and concepts used here, reference may be made to the article under [7] and to: F. H. Stieltjes and L. J. Tummers, Simple theory of the junction transistor, Philips tech. Rev. **17**, 233-246, 1955/56.

finally a stage is reached where the material has become "intrinsic", with the conductivity at a minimum. A further increase in the number of dislocations gives rise to conversion; the material becomes $P$-type and the resistivity drops again (*fig. 7*), finally reaching a more or less constant value depending on the concentration of the other centres. At room temperature this change is brought about both by dislocations and vacancies; at very low temperature it is mainly due to the vacancies.

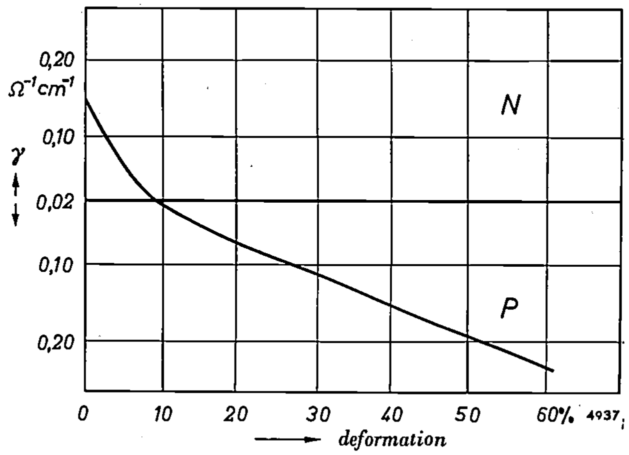


Fig. 7. Change in the conductivity $\gamma$ of $N$ germanium as a result of plastic deformation. After about 10% deformation, conversion to $P$ type germanium occurs owing to the lattice imperfections formed acting as acceptors. (From E. S. Greiner and W. C. Ellis, Bell Lab. Record 34, 403, 1956.)

*Charged dislocations in semiconductors*

The acceptor action of dislocations is generally attributable to the breaking of valence bonds between germanium atoms along a dislocation. Take, for instance, a 60°-dislocation in germanium (*fig. 8*), whose Burgers vector makes an angle of 60° with the dislocation axis. The atoms at the edge of the "extra half plane" that terminates along such a dislocation are not surrounded by four neighbours, as they should be, but by three neighbours. These atoms thus retain one free bond, i.e. they possess an unpaired electron, and they will therefore readily accept a conduction electron in order to fill their electron shell. A dislocation of this kind may therefore be regarded as a row of acceptors.

The filling of the row, i.e. the occupation of all acceptor levels along a dislocation, is hindered — even in pronounced $N$-type material — by the electrostatic repulsion which all these unscreened extra electrons exercise on one another. In this respect dislocation acceptors differ from acceptors formed by foreign atoms, which are distributed arbitrarily throughout the lattice: in the case of dislocations a correlation arises between the acceptors which leads to a degree of occupation differing from the usual one.

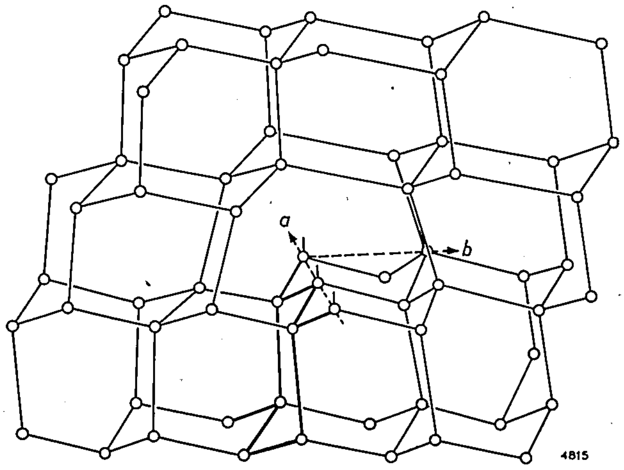


Fig. 8. A 60°-dislocation along the ⟨110⟩ direction in the diamond lattice. The extra half plane is indicated by thick lines; the free bonds are clearly visible. $a$ represents the dislocation axis, $b$ the direction of the Burgers vector. (From J. Hornstra, Physica 25, 409, 1959.)

A dislocation line in $N$-type material, along which some of the acceptor states are occupied, carries a charge. This charge is neutralized by a space charge of opposite sign in a roughly cylindrical region around the dislocation: the donors in this region — the donors are fairly uniformly distributed throughout the lattice — have supplied the electrons that have filled the acceptors (*fig. 9*). The cylindrical region in semiconductors may be quite thick; assuming that 10% of the free bonds along the dislocation line have accepted an electron, the average number of accepted electrons per unit disloca-

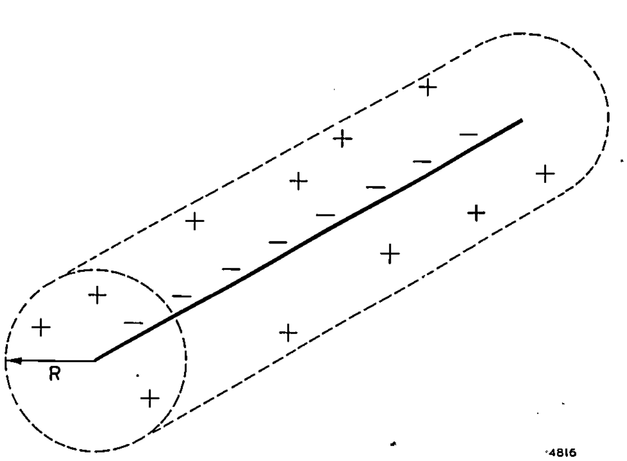


Fig. 9. A dislocation line in $N$ germanium is negatively charged, because electrons are accepted here and there along the line. This charge must be neutralized by the ionization of donors in a more or less cylindrical region around the dislocation, in which a space charge is formed. The radius $R$ of this region depends on the donor concentration: the more donors the narrower the space-charge region.

tion length is $2\times10^6$ cm$^{-1}$. Putting the density of the donors at $10^{15}$ cm$^{-3}$ (a reasonable value for $N$ germanium) we find a value of roughly 0.5 microns for the diameter of the space-charge region in which all donors are ionized. The energy needed to build up such a space-charge region around the charged dislocation is about 0.5 eV per electron. This is of the same order as the energy needed to release an electron from a germanium atom and to convey it to another atom lacking an electron. This makes it understandable that only a fraction of the dislocation levels can be occupied.

To calculate the scattering of charge carriers by dislocations in a semiconductor, it is necessary to take this space-charge effect into account. In the case described it occurred only in $N$-type germanium, since the dislocation acted as a row of acceptors. It is also possible, however, that in other semiconducting elements or compounds an atom with a free bond may prefer to give up an electron rather than accept a second one; the dislocation then acts as a row of donors, and the space charge effects will occur in $P$-type material.

Detailed experimental investigations into the electrical properties of dislocations in covalent semiconductors have not yet been made, owing to the numerous other complications involved, which cannot be discussed here.

*Charged dislocations in polar crystals*

In polar (ionic) crystals analogous electrostatic effects occur to an even larger extent than in covalent substances, owing to the fact that all constituent particles are of course charged and that, due to polarization of the lattice, disturbances in the charge distribution are only slightly screened. In covalent materials a dislocation receives a charge as a result of the interaction with electrons. In ionic crystals, however, the role of the mobile charge carriers is taken over by the ions themselves or, in other words, by the point defects — mainly vacancies — which are necessary for the movement of the ions. These behave in the lattice as if they possessed a charge of opposite sign from that of the absent ion. This "effective" charge is compensated by vacancies of oppositely charged ions (Schottky pairs), or by interstitial ions having the same charge as the absent ions (Frenkel pairs), or again by appropriate foreign ions. In thermodynamic equilibrium, charge neutrality prevails inside the crystal. Expressed in another way, the continuous formation and disappearance of lattice imperfections takes place on the average in pairs, in such a way that there are always as many positive charges present as there are negative charges. The creation and destruction of lattice imperfections occurs preferably at the surface, at crystal boundaries or on dislocations. However, there is no *a priori* reason to assume that the energy needed to form a vacancy of a certain ion will be equal to that needed to form the vacancy that must act as the counterpart of the charge (in the following we shall confine ourselves to Schottky pairs). Near a dislocation the two kinds of vacancies appear and disappear, seen on an atomic scale, entirely independently of each other.

How this takes place will not be discussed here. It will be sufficient to note that "jogs" [3]) in the dislocations play a major part. A vacancy occurs more readily at such jogs than elsewhere in the crystal, the ions then having fewer neighbours, and also because part of the required electrical charge is concentrated around the jog [9]).

The equilibrium concentrations $c_+$ and $c_-$ of the two kinds of vacancy of a Schottky pair will thus as a rule be unequal in the immediate neighbourhood of the dislocation, and a space charge of density $e(c_+ - c_-)$ exists around the dislocation; $e$ represents the magnitude of the ionic charge. The relation between the space-charge potential $\varphi$ and the charge density is given, according to Poisson's equation, by

$$\nabla^2\varphi = -\frac{4\pi e}{\varepsilon}(c_+ - c_-)\,, \qquad . \; . \; . \; (2)$$

where $\varepsilon$ is the dielectric constant of the material. We can now calculate the equilibrium concentrations of both kinds of vacancies separately. If the free energies of formation are $U_+$ and $U_-$, respectively, we find for the concentrations:

$$\left.\begin{aligned} c_+ &= \text{const.} \times \exp\left[-(U_+ - e\varphi)/kT\right]; \\ c_- &= \text{const.} \times \exp\left[-(U_- - e\varphi)/kT\right]. \end{aligned}\right\} \; . \; . \; (3)$$

Insertion of (3) in (2) gives the Debye-Hückel equation, the solution of which far from the dislocation is given by:

$$c_+ = c_- = \text{const.}\times\exp(-U/2kT)\,, \quad . \; (4)$$

where: $U = U_+ + U_-$. This expresses mathematically the fact earlier noted that, because of the charge neutrality, both concentrations inside the crystal must be equal beyond a certain distance from the

[9]) For a more detailed discussion, see H. G. van Bueren, Imperfections in crystals, Part III, North-Holland Publishing Company, Amsterdam 1960.

dislocations and from the surface. Within a certain distance $R$ from the dislocation, roughly given by

$$R = \sqrt{\frac{\varepsilon kT}{4ne^2N}}, \quad \ldots \ldots (5)$$

the equality expressed in (4) no longer holds. (In this formula, $N$ represents the total number of vacancies per cm$^3$, far from the dislocation.) The distance $R$ determines the diameter of the cylindrical region around the dislocation within which space-charge effects are perceptible. Depending on the temperature, $R$ varies in rocksalt, for example, from one hundredth to a few tenths of a micron.

As in the case of covalent semiconductors, the dislocation line here too behaves like a linear charge, now with the potential

$$\psi = \frac{U_+ - U_-}{2e}, \quad \ldots \ldots (6)$$

surrounded by a space-charge region. In a rocksalt crystal $\psi = -0.28$ volt. Both the electrical and the mechanical properties of ionic crystals are influenced by the charge of the dislocations, because it affects both the mobility of the charge carriers (vacancies) and the mobility of the dislocations themselves.

We shall examine the latter point at greater length, since it is related to an interesting experimental illustration of the above theory. As a result of the mechanism described, a space charge will be present not only around the dislocations but also at the surface layer of an ionic crystal. A dislocation approaching the surface will therefore be electrostatically repelled (*fig. 10*). Now it is known that the surface of these crystals has a very marked bearing on their plasticity. In dry air NaCl crystals are very brittle, but if their surface is moist they can undergo plastic deformation up to 20 or 30% (Joffé effect). This may well be due to electrostatic as well as mechanical causes. Similar effects are also clearly apparent in semiconductors, but are much less manifest in metals.

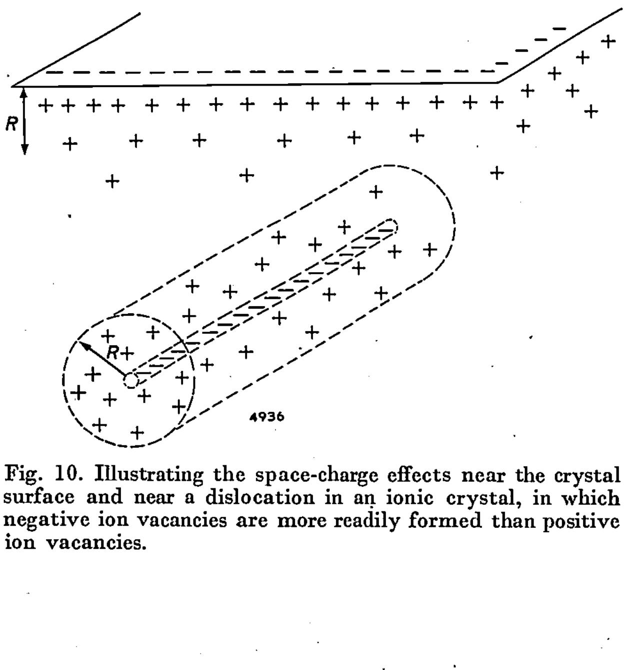


Fig. 10. Illustrating the space-charge effects near the crystal surface and near a dislocation in an ionic crystal, in which negative ion vacancies are more readily formed than positive ion vacancies.

A periodic elastic deformation of an ionic crystal gives rise to small periodic movements of the dislocations, relative to the space-charge zones, and these charge movements induce in the surface a periodically varying electrical potential. A weak alternating voltage has in fact been measured on a crystal subjected to an alternating bending stress. This effect should not be confused with the piezoelectric effect; the latter occurs only in crystals which have no centre of symmetry, whereas the effects of dislocation movements have been found in cubic crystals. Conversely, an electrical potential applied to these crystals can produce a slight dislocation movement, and consequently a slight amount of elastic deformation. Both phenomena have been observed in recent experiments in Belgium [10]).

We have seen that the charge of the dislocations necessarily influences the mobility of the charge carriers, and thus affects their contribution to the electrical conductivity. In quantitative terms this influence is found to be small: the charge carriers — the vacancies or, which amounts to the same thing, the ions of the crystal — have too large a mass to be scattered to any appreciable extent. On the ocher hand, as pointed out above, dislocations do play an important part in the *formation* of charge carriers. Generally speaking the introduction of lattice imperfections causes a drop in the resistivity of ionic crystals. This effect is considerably accentuated by slight plastic defomation, the result normally being a drastic temporary decrease of resistivity. The reason for this is that, when the dislocations move, large numbers of vacancies or interstitial atoms are formed by one of the mechanisms discussed above. This dynamic process of point-defect formation is distinct from the static process which we have been discussing, but the point imperfections formed obviously contribute their share to the conductivity.

The different effects of lattice imperfections on the electrical properties of the three classes of substance considered are summarized in *Table I*.

[10]) R. L. Sproull, Phil. Mag. **5**, 815, 1960.
S. Amelinckx, G. Remaut and J. Vennik, Phys. Chem. Solids **11**, 170, 1959 and **16**, 158, 1960.

Table I. Effect of lattice imperfections on electronic properties (+: enhancing effect; —: reducing effect; 0: no or very little influence).

| | Metals | | Semiconductors | | Ionic crystals | |
|---|---|---|---|---|---|---|
| | Dislocations | Point defects | Dislocations | Point defects | Dislocations | Point defects |
| Mobility $\mu$ | — | — | — | — | 0 | 0 |
| Number of charge carriers $n$ | 0 | 0 | + or — | + or — | + | ++ |
| Conductivity $\gamma$ | — | — | + or — | + or — | + | ++ |

## Plasticity of crystals

### *Governing factors*

We have recalled above that the plasticity of crystals depends on the presence of dislocations: these may move under relatively low stresses and give rise to displacements on certain slip planes. The degree of plasticity is governed by various factors, quantitative information on some of which is still lacking.

The first factor to be mentioned is the dislocation length which the crystal contains per unit volume (dislocation density). If this is small, only minor displacement will occur and the plasticity will be low; if it is large, the dislocations obstruct one another, and this again reduces the ductility of the material. There is in fact a maximum permissible force that can be applied to a crystal, beyond which cleavage or breakage will result. Where the dislocation density is very high, the average interaction force between the dislocations is comparable with this maximum force, and rupture will occur before the dislocations have perceptibly shifted. The highest ductility is found in general at a dislocation density of about $10^8$ cm$^{-2}$. For metal crystals this is a normal density, but for semiconductors it is very high, and rather high for ionic crystals.

A second factor affecting plasticity is the mobility of the dislocations (not to be confused with the mobility of the charge carriers, discussed above). In metals, apart from the above-mentioned mutual interaction of the dislocations and their interaction with point defects, there is in principle hardly any obstruction of the movements of dislocations. On the other hand, in most valence-bond substances the mobility of dislocations is extremely small, at least at moderate temperatures, because of the rigid bonds between the atoms. These materials are therefore very brittle, and show ductility only at high temperatures, where the dislocations then appear to become mobile. In polar crystals the dislocations in themselves are fairly mobile, but here the movement of a dislocation is generally accompanied by electrostatic effects, which counteract the movement and thus make the effective mobility relatively low. Only under certain suitable conditions can ionic crystals be substantially deformed without them breaking.

The third factor to be mentioned is the presence of dislocation sources to replenish dislocations leaving the crystal. Without sources there can be no permanent ductility, but the presence of sources itself depends on various factors, such as the state of the surface, the presence of impurities, the dimensions of the crystal and the dislocation density. The relative influence of these factors differs from one material to another. This, and the differences in density and mobility referred to, will be found summarized in *Table II*. This table can serve as the starting point for a systematic comparative study of the plasticity of various types of substance. Depending on the substance, plastic properties can also be studied under entirely different conditions, so that, conversely, information on the behaviour of dislocations can be obtained from the data on plasticity.

To illustrate the foregoing, we shall examine a little more closely the way in which the study of the plasticity of semiconductors and ionic crystals has yielded information on the mobility of dislocations, information which is more difficult, if not impossible, to derive from investigations of the characteristically plastic metals alone.

Table II. Differences in dislocation structure and mechanical properties.

| | Metals | Semiconductors | Ionic crystals |
|---|---|---|---|
| Dislocation density | high | low to moderate | moderate |
| Dislocation mobility | high | 0 to moderate (at high temperature) | moderate |
| Ductility | high | 0 to moderate (at high temperature) | moderate |
| Dislocation sources | internal | superficial | internal +<br>superficial |
| Effect of dimensions on ductility | slight | very marked | very marked |

### *Direct measurements of the velocity of dislocation movement*

As a measure of the mobility of dislocations we can use the velocity at which a dislocation, at a given temperature, is displaced under the influence of a given mechanical stress. This can only be measured directly in a relatively perfect crystal, possessing few dislocations. Only in such a crystal can anything be seen of the individual dislocation movements, and moreover the interaction between the dislocations is negligible. Perfect crystals (of reasonable dimensions) are seldom if ever found among metals; they are, however, obtainable among covalent substances and also, though less readily, among certain polar materials. It is therefore not surprising that dislocation velocities were first measured in substances representative of these categories, i.e. in germanium and lithium fluoride. The results, which we shall now discuss, cannot of course be directly extended to metals, but some general data can be obtained that are applicable to all solids. These data have recently been confirmed by measurements on metals, e.g. silicon iron.

#### *a) Displacement of etch pits*

With the aid of an etching technique [11]) the movement of individual dislocations has been observed in the transparent material lithium fluoride. When a superficial layer of the crystal is etched in a diluted aqueous solution of ferrichloride, etch pits are produced at the places where the dislocations meet the surface. The application to the crystal of a short mechanical stress causes the dislocations to move over a short distance, so that they break away from their etch pits. Further etching then reveals a new group of etch pits, which are displaced from the first ones by the distance in question. The latter have a different appearance: "fresh" etch pits at which a dislocation terminates are somewhat pointed, whilst the "old" etch pits that have lost their dislocation develop a flat base and larger dimensions on renewed etching (see *fig. 11*). By varying the time and the stress, and possibly the temperature, and measuring the distances between the etch pits, it is possible to calculate the velocity of the dislocations. (It is obvious that this method can only be used on crystals containing only a few, easily distinguishable dislocations.) The result of such measurements is represented in *fig. 12*. The dislocation velocity depends very markedly on the stress, roughly exponentially; if the stress is increased by a factor of 2 the dislocation velocity increases by a factor of about $10^3$.

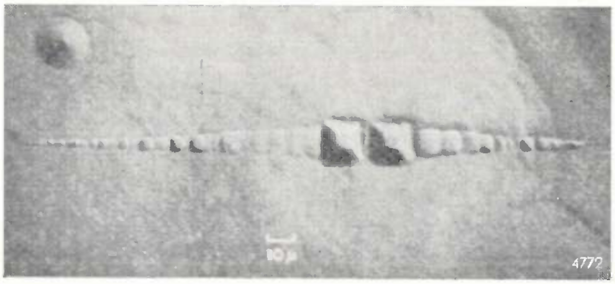


Fig. 11. The movement of dislocations in lithium fluoride can be studied from the etch pits occurring at the points where the dislocations terminate at the surface. The series of small pits indicates the dislocation sites after the crystal has been subjected to successive short-lived bending stresses; the two large pits indicate the positions of the same dislocations before this treatment. Upon the application of a bending stress the two dislocations move in opposite directions and thus have opposite signs. Magnification approx. 500 ×. (Taken from W. G. Johnston and J. J. Gilman, J. appl. Phys. **30**, 129, 1959.)

#### *b) Creep tests*

Results similar to those obtained by these direct "microscopic" investigations have been found from indirect "macroscopic" creep tests, also done on LiF,

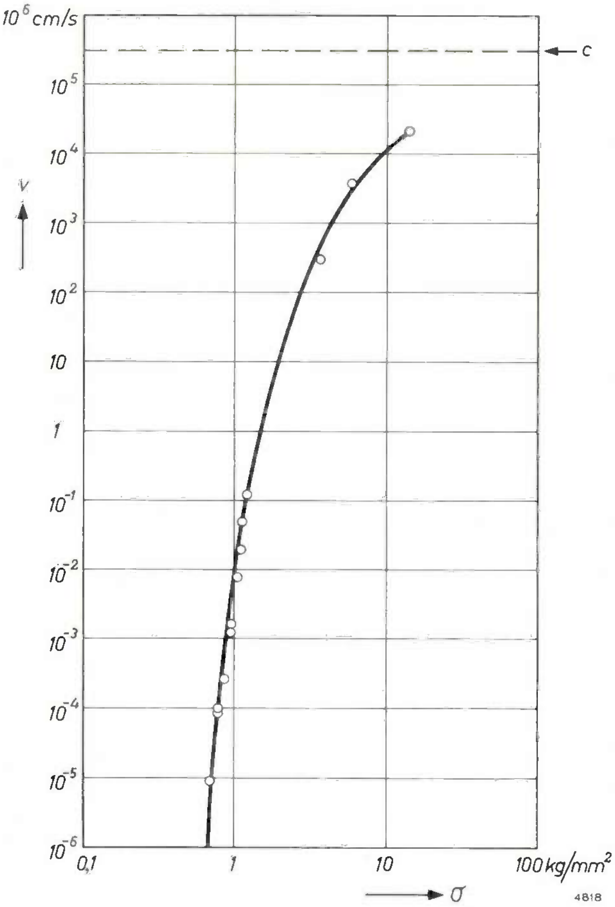


Fig. 12. Velocity *v* of dislocations in lithium fluoride as a function of stress *σ* [11]); *c* is the speed of sound in the material, which represents an upper limit to the velocity of the dislocations.

[11]) W. G. Johnston and J. J. Gilman, J. appl. Phys. **30**, 129, 1959.

but mainly on single crystals of germanium [12][13]. The latter material can be produced with a very low initial density of dislocations [5]. A permanent mechanical stress applied to such an "almost perfect" crystal gives rise at elevated temperature to a gradual deformation (creep), which depends in a very specific way on time. An example is given in *fig. 13*: after an initial, "incubation" period, in which hardly any deformation is to be observed and which under average conditions may last anything from some minutes to half an hour or more, the rate of deformation gradually increases until it reaches a constant value, many times higher than the rate of deformation during the initial period. The analysis of numerous such creep curves has shown that they can all be described using only two parameters, $t_p$ and $a$, whose significance appears from fig. 13: $t_p$ is a measure of the length of the incubation period, and $a$ is the ultimate constant value of the creep rate. In the incubation period the creep curve can be represented fairly accurately by a third-degree function of time: the deformation $z$ increases with time according to

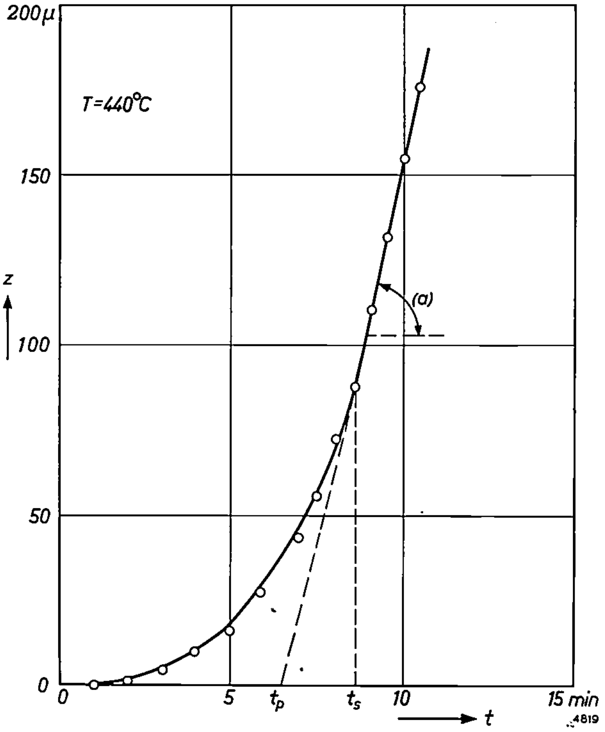


Fig. 13. Creep curve of a germanium crystal subjected to a bending load at 440 °C. The long axis of the crystal rod was in the [111] direction; the stress in the outer layers was 13 kg/mm². The deformation $z$ is plotted versus time $t$. The plotted points are taken from reference [12], the solid curve is derived from the theory [13].

$$z = k\,t^3. \quad \ldots\ldots\ldots \quad (7)$$

Beyond $t = t_p$ the curve becomes a straight line:

$$z = a(t - t_p). \quad \ldots\ldots\ldots \quad (8)$$

The factor $k$ is not an independent parameter: for a given $a$ and $t_p$ it is determined by the condition that the curves must join without any discontinuity.

The curve yielded by formulae (7) and (8) can readily be understood if we represent the creep mechanism as follows.

A dislocation source somewhere in the crystal emits dislocation loops under the action of a stress $\sigma$, roughly as described in reference [3] (*fig. 14*). We assume that these loops expand at a constant velocity $v$, which is the dislocation velocity that we wish to determine in the material. There will thus always be the same time $t_0$ between the emission of two successive loops, and the loops therefore follow one another at equal spatial intervals $\delta = vt_0$. The distance $\delta$ is determined by the elastic interaction between a loop just emitted and the source; it may be estimated at about 1 micron.

The dislocation, of radius $r$, causes back stress at the position of the source amounting roughly to 0.1 $Gb/r$, where $b$ is the Burgers vector of the dislocation and $G$ the shear modulus of the material. The back stress becomes smaller than the applied stress $\sigma$ as soon as the loop has expanded to

$$r \geqq 0.1\,Gb/\sigma. \quad \ldots\ldots\ldots \quad (9)$$

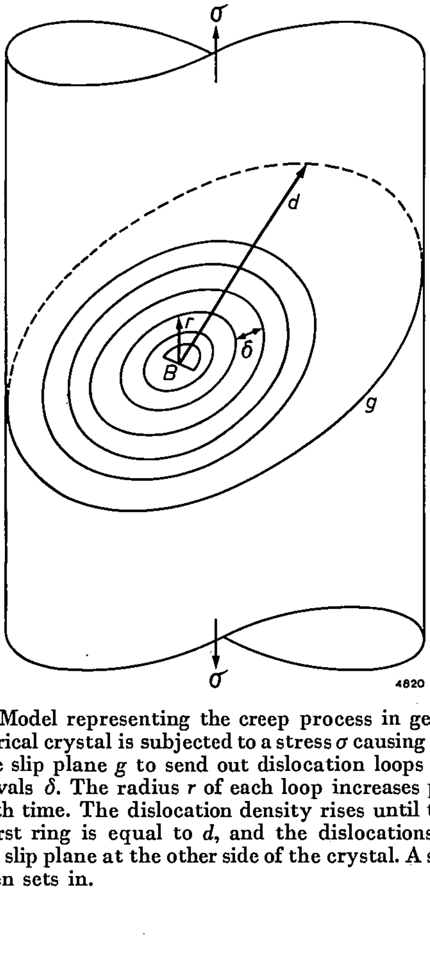


Fig. 14. Model representing the creep process in germanium. A cylindrical crystal is subjected to a stress $\sigma$ causing the source $B$ on the slip plane $g$ to send out dislocation loops separated by intervals $\delta$. The radius $r$ of each loop increases proportionally with time. The dislocation density rises until the radius of the first ring is equal to $d$, and the dislocations begin to leave the slip plane at the other side of the crystal. A stationary state then sets in.

[12] P. Penning and G. de Wind, Physica **25**, 765, 1959.
[13] H. G. van Bueren, Physica **25**, 775, 1959.

This value of $r$ thus roughly corresponds to the distance $\delta$ between the old loop and the new loop being emitted at that moment.

The total length of the dislocation loops emitted by a source at the moment $t$ after the application of a stress $\sigma$ is now simply the sum of an arithmetical progression:

$$\Lambda = 2\pi vt + 2\pi v(t-t_0) + + 2\pi v(t-2t_0) + \ldots \approx \frac{\pi vt^2}{t_0}. \quad . \quad . \quad . \quad (10)$$

All these dislocations move at a velocity $v$, and therefore contribute the amount $\Lambda vb$ to the deformation rate $\dot{z}$ (see reference [3]). If there are $N$ sources per cm$^3$, then

$$\dot{z} = N\Lambda vb, \quad . \quad . \quad . \quad . \quad . \quad (11)$$

and after integration:

$$z = \frac{\pi Nbv^2}{3t_0}\, t^3 = \frac{\pi Nbv^3}{3\delta}\, t^3. \quad . \quad (12)$$

This is the observed cubic dependence in the initial period. However, this dependence cannot continue indefinitely, since the dislocation density stops increasing as soon as the first dislocation loop reaches the opposite wall of the material. That occurs after a time

$$t_s = d/v, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (13)$$

where $d$ is the distance involved, i.e. roughly the diameter of the specimen. At the end of this time a stationary state sets in and a constant dislocation length is reached, approximately equal to:

$$\Lambda_s = \frac{\pi v t_s^2}{t_0} N = \frac{\pi d^2 N}{\delta}. \quad . \quad . \quad (14)$$

The creep rate is then constant:

$$\dot{z} = a = \frac{\pi d^2}{\delta} Nbv, \quad . \quad . \quad . \quad . \quad (15)$$

in agreement with the experiment. The fit between the curve (12) and the straight line of slope (15) is obtained by putting the "extrapolated" time $t_p$, defined in fig. 12, equal to

$$t_p = \frac{2}{3} ts = \frac{2}{3}\frac{d}{v}. \quad . \quad . \quad . \quad . \quad . \quad (16)$$

An essential feature of the creep mechanism in almost perfect germanium crystals, as here described, is the fact that the initial curve is not exponential. If it were, we should have to assume the "multiplication" of dislocations in that stage, that is to say the dislocations formed would themselves act as the source of fresh dislocations.

It is interesting to note how it was ascertained convincingly that the curves found are not exponential but are to be represented by third degree functions. From formulae (7), (8) and (16) the value 6.75 is found for the dimensionless ratio $at_p/z_p$ where $z_p$ is the deformation at the moment $t_p$. (This ratio, incidentally, is not dependent on stress nor on the temperature; see below.) This value is in good agreement with the experimental values (principally derived from the well-established linear part of the curves) which lie between 5 and 7. (Assuming an $n$th degree function, the value generally obtained is $n(\frac{3}{2})^{n-1}$.) On the other hand, if the function were an exponential one, the ratio would have to be e $\approx$ 2.7.

A further experimental fact of importance is that the parameters $a$ and $t_p$ of the observed creep curves are found to depend strongly on the temperature $T$ and the stress $\sigma$ at which the experiment is done (*fig. 15*). This dependence, at least over most of the region of stress variation, can be expressed by the relations:

$$\left.\begin{matrix} a \\ t_p^{-1} \end{matrix}\right\} \propto \exp\left(-\frac{Q-w\sigma}{kT}\right). \quad . \quad . \quad (17)$$

This formula is analogous to those used to describe the temperature-dependence of, for example, diffusion and chemical reactions. The factor of the form $\exp(-q/kT)$ occurring in all these cases is bound up with the fact that every elementary process

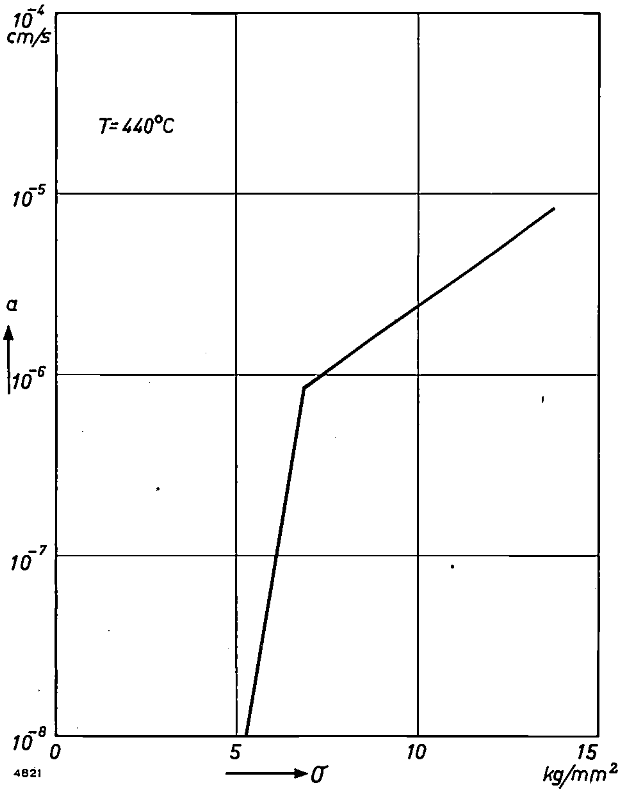


Fig. 15. Stationary creep rate as a function of bending stress $\sigma$ in germanium crystals. The curve represents the average of about 100 experiments. The right portion, which covers by far the largest part of the region of stress variation, obeys the relation (17).

involved in these phenomena requires an activation energy $q$, which must be supplied by thermal agitation. In each case, where the elementary process consists in the repeated displacement of a dislocation over one atomic distance, the required activation energy is seen from formula (17) to be $q = Q - w\sigma$. It is thus a function of the stress $\sigma$, and the higher the stress the less thermal energy has to be supplied. The factor $w$, which has the dimension of a volume, is called the activation volume; $Q$ may be termed the "stress-free" activation energy.

Using formulae (15) and (16) to express the dislocation quantities $N$ and $v$ in terms of the observational data $a$ and $t_p$ we find:

$$\left.\begin{aligned} v &= \frac{2}{3}\, d\, t_{\mathrm{p}}^{-1}, \\ N &= \frac{3}{2\pi}\,\frac{\delta}{bd^3}\, a\, t_{\mathrm{p}}. \end{aligned}\right\} \quad . \; . \; . \; . \quad (18)$$

Combining these expressions with (17) we find that the source density $N$ is *in*dependent of temperature and stress, whereas the dislocation velocity $v$ depends exponentially on both. This is in agreement with the results of the microscopic measurements on LiF crystals (fig. 12). The results are also quantitatively comparable.

*General conclusions*

The good agreement referred to gives reason to believe that the rules derived indicate a general property of almost perfect crystals, namely that the speed at which the dislocations move in those crystals varies exponentially with stress and temperature — although of course the stress-free activation energy $Q$ and the activation volume $\omega$ will differ considerably from one substance to another. In metals, both $Q$ and $w$ will be small, since the plasticity of metals is very high and experiments have shown that it is not strongly dependent on temperature and stress. In spite of this fact, metals too as a rule will satisfy (17), which mainly expresses that the movement of dislocations through a crystal calls for a stress-dependent activation energy. This has been confirmed by very exact experiments carried out at low temperature.

A second and much more general conclusion, which is not confined to the material investigated, relates to sources of dislocations. From a closer analysis of the distribution of the sources it may be concluded that all sources in these materials containing few dislocations are situated at the surface. This throws an interesting light on the important role, mentioned on page 368, played by the surface of a crystal in plastic deformation. The fact that this is not generally so noticeable in metals as in other types of material is due to the dislocation density. The dislocations in metal crystals are normally so dense that the interaction and intersection of the numerous dislocation lines (networks) gives rise to large numbers of internal sources, so large as to overshadow the effect of the surface sources. In very thin metal crystals, however, the surface is of considerable importance. In any attempt to produce a dislocation-free (and hence strong) metal crystal, and to keep it free from dislocations, the most scrupulous attention must be paid to the state of the surface. In the preparation of dislocation-free germanium and silicon crystals [5]) this fact was already known and taken into account.

**Summary.** In this article, devoted to differences in the behaviour of lattice imperfections in various materials, the author discusses the manner in which these imperfections are formed and their influence on the electrical and mechanical properties of the substance concerned. Emphasis is placed on the electrostatic effects that may accompany the formation of lattice imperfections. It is shown that, as a result of these effects, dislocations in non-metals generally possess a charge and are surrounded by a space-charge region to compensate for that charge. The influence of these charges on the mobility and concentration of charge-carriers is discussed, and differences in the effect of plastic deformation on the electrical conductivity of various types of crystals is explained. The plastic properties of nearly perfect polar and covalent crystals are then examined. It is shown, with reference to experiments, that the movement of dislocations in both types of material follows virtually the same laws, and it is concluded that similar laws must also in principle apply to metals. Finally, something is said of the distribution of dislocation sources and the significance of the surface in this respect.

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

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution.

**2821:** J. Hornstra and L. J. van der Pauw: Measurement of the resistivity constants of anisotropic conductors by means of plane-parallel discs of arbitrary shape (J. Electronics and Control **7**, 169-171, 1959, No. 2).

It has previously been shown that the resistivity $\varrho$ of a plane-parallel sample of arbitrary shape provided with four contacts along the circumference, can be derived from the values of two quantities having the dimensions of resistance which are easily measured (see Philips tech. Rev. **20**, 220, 1958/59). The theorem is now extended to the case of anisotropic material with resistivities $\varrho_1$, $\varrho_2$, $\varrho_3$ in the directions $X$, $Y$ and $Z$. Three samples then have to be taken perpendicular to $X$, $Y$ and $Z$, giving values of $\sqrt{\varrho_2\varrho_3}$, $\sqrt{\varrho_3\varrho_1}$ and $\sqrt{\varrho_1\varrho_2}$, respectively, from which $\varrho_1$, $\varrho_2$ and $\varrho_3$ can be found.

**2822:** G. Thirup: Studies on networks with periodically variable elements (thesis, Copenhagen, June 1959).

This thesis deals with perturbations of non-linear networks that are excited by a periodic voltage or current. For the perturbation voltages and currents, the network is considered as a linear network with periodically varying elements. In particular the technique of measuring variable networks is treated. It is shown how problems relating to oscillators and to the measuring technique of constant linear networks can be solved by the methods described. The known theory of variable networks is briefly reviewed and generalized, and the relation between variable networks and mixing and modulation is discussed (Chapter 2). The stability problem of variable networks is considered and the concept of return difference of feedback amplifiers is extended to include general linear networks. In certain cases it is possible to measure the system determinant itself, which then can be used for experimental stability investigations. Stability properties of an autonomous system are derived from the variational equation of the system (Chapter 3). In Chapter 4, modulation and synchronization of oscillators are investigated, making use of the technique of analysing variable networks. Special attention is paid to pure amplitude modulation, pure frequency modulation and the stability of synchronized oscillators. The properties of the coupling network enter the variational equation as three parameters, which are derived from the network elements for a number of coupling networks. Chapter 5 deals with the basic principles of the measuring technique of variable networks, and the practical construction of measuring equipment is described in Chapter 6. A device for measuring complex conversion properties of semiconductor diodes is given in detail. In Chapter 7 it is shown how some problems of the technique of measuring constant networks can be analysed by means of the technique of variable networks; some new methods of measurements are proposed. Results of stability measurements on autonomous and synchronized oscillators are given in Chapter 8. Chapter 9 deals with semiconductor diodes; an equivalent circuit is considered and results of some measurements of complex admittances are given.

**2823*:** M. Avinor: Photoconductivity of activated cadmium sulphide single crystals (thesis, Amsterdam, October 1959).

This thesis deals with the influence of activators such as Cu and Ag on the photoconduction, light absorption and luminescence of single crystals of CdS. A new method of preparation by zonal sublimation is described. It is shown that Cu and Ag each produce two distinct energy levels, depending on the ratio of the activator and coactivator concentrations. A third level was found with trivalent coactivators. The spectral location of photoconductivity bands and emission bands is indicated. Ni is shown to produce not only "killing" effects but also traps having a depth of 0.23 eV. At room temperature Ni prolongs the decay time of the photoconductivity. The models of Rose, Klasens and Lambe and Klick are examined in the light of the experimental evidence. The two-level Klasens model is found to give the best insight into the various electronic processes, though only as a first-order approximation. The model of Lambe and Klick is in disagreement with some of the experimental facts.

**2823a:** J. H. Spaa: A continuously operated instrument for the stepwise measurement of the radioactivity of gas sols with a special background compensation (Progr. nucl. Energy, Series 12 — Health Physics — 1, 219-227, 1959).

Many dangerous radioactive isotopes will generally be present in the atmosphere as dust particles. The tolerable concentrations are so low that dust filtered from several cubic metres of air has to be accumulated to provide reasonable measurements. These measurements are hampered by the presence of the daughter products of the natural radioactive gases radon and thoron, which adhere to the dust particles, and which are always present in noticeable concentrations. Although these daughter products disintegrate fairly rapidly, they make it difficult to detect contamination during the accumulation period. In the instrument described here an $\alpha$-ray detector is mounted immediately above the filter paper, and a $\beta$ counter underneath it, and the measured currents of the corresponding count-rate meters can be so adjusted that they compensate each other when only radon and thoron daughter products are present. A calculation, taking the lifetime of these products into account, shows that this is practically feasible, and it has been confirmed by experiments. If a reading is nevertheless obtained, this must be attributable to contamination. The instrument also contains a second set of counters, which serves for accurate monitoring some time after the accumulation period, after the background radioactivity due to radon and thoron has been sufficiently reduced.

**2823b:** G. Brouwer: The simulation of electron kinetics in semiconductors (Proc. 2nd int. analogue computation meetings, Strasbourg, Sept. 1958, pp. 135-137, published by Presses Académiques Européennes, Brussels 1959).

The analysis of the distribution of electrons and holes in semiconductors under excitation by light pulses leads to a set of rather complicated nonlinear differential equations containing numerous unknown parameters. As the latter have to be determined by a trial-and-error method, it was found that an analogue computer was most suitable for the problem.

The special-purpose simulator built contains 8 multipliers, 4 integrators and a signal generator. It is of the repetitive type and the cycle is started with the zero setting of the integrators, followed by the insertion of the initial values. DC restorers are employed whenever the absolute value of the signal is of importance, as it is in the case of multiplication. A simple master clock synchronizes the various phases of the computation cycle. The time constants of the integrators are adjustable and are usually set on different time scales in one and the same problem. With the simulator it is possible to study the dynamics of photoconductivity, fluorescence and thermal stimulation in semiconductors and the dynamics of the spin system in a maser.

**2823c:** H. Bremmer: The surface-wave concept in connection with propagation trajectories associated with the Sommerfeld problem (IRE Trans. on Antennas and Propagation **AP-7**, spec. suppl., S175-S182, 1959).

The author considers the field generated by a short vertical electric dipole placed above a homogeneously dielectric, flat earth, when the aerial current is an ideal pulse (Dirac $\delta$ function). With the aid of two-dimensional operational calculus, the field is expressed by integrals over the earth's surface. These integrals may be interpreted as composed of contributions due to rectilinear rays leaving the transmitter; such rays strike the earth's surface, spread out over it like surface waves, and, after leaving this surface, arrive at the receiver along another straight trajectory (which may be above or inside the earth). The velocity of propagation $c/n_{12}$ of the surface waves is related to the refractive indices $n_1$ and $n_2$ of the space above the earth and that inside it according to $1/n_{12}{}^2 = 1/n_1{}^2 + 1/n_2{}^2$.

**2823d:** J. Bloem, C. Haas and P. Penning: Properties of oxygen in germanium (Phys. Chem. Solids **12**, 22-27, 1959, No. 1).

The behaviour of oxygen in silicon has been extensively studied, but in spite of much experimental evidence a satisfactory picture is still lacking. The authors have therefore investigated the behaviour of oxygen in germanium, in the hope of throwing more light on this problem. The oxygen was introduced by zone-levelling germanium crystals in an oxygen atmosphere. The oxygen concentration in the crystal, as deduced from infrared absorption at 856 cm$^{-1}$, is proportional to the oxygen pressure when the latter is lower than 20 mm Hg. At higher pressures the oxygen concentration is constant at $7 \times 10^{17}$ atoms/cm$^3$.

On heat treatment at 400 °C, donors are introduced into the oxygen-containing crystal, which disappear again if the crystal is heated to a higher temperature (700 °C). The number of donors depends strongly on the oxygen concentration. A tentative description of the results is given in terms of the equilibrium between isolated oxygen atoms and donor complexes consisting of four oxygen atoms. The increase in the number of conduction electrons with temperature (determined from Hall-effect measurements) indicates the presence of several donor levels. Some of the germanium crystals levelled under oxygen showed two additional absorption bands at 1100 and 1260 $cm^{-1}$.

**2823e:** J. Goorissen and A. M. J. G. van Run: Gas-phase doping of silicon (Proc. Instn. Electr. Engrs. **106 B**, suppl. No. 17, 858-860, 1959).

This article describes a method of making single crystals of silicon which contain a known, small, homogeneously distributed quantity of another element, e.g. phosphorus. The doping technique employed consists in creating a constant flux of phosphorus atoms from the gas phase via the liquid into the solid by decomposing phosphine in the vicinity of the floating liquid zone. The experimental results obtained are readily reproducible. Under certain conditions, nearly all the added phosphorus is taken up.

**2823f:** P. Massini: Uptake and translocation of 3-amino- and 3-hydroxy-1,2,4-triazole in plants (Proc. 2nd United Nations int. Conf. on the peaceful uses of atomic energy, Geneva, Sept. 1958, Vol. 27, pp. 58-62).

Part of a programme of fundamental research into the relation between the chemical structure of substances and the nature of their uptake and translocation in plants. In this article the author compares the behaviour of two substances which are closely related chemically, namely 3-amino-1,2,4-triazole and 3-hydroxy-1,2,4-triazole. The substances are followed in the plant by labelling them with radioactive carbon atoms.

**2823g:** J. Abels, M. G. Woldring, H. O. Nieweg, J. G. Faber and J. A. de Vries: Ethylenediamine tetra-acetate and the intestinal absorption of vitamin $B_{12}$ (Nature **183**, 1395-1396, 1959, No. 4672).

Short communication concerning the intestinal absorption of radioactive vitamin $B_{12}$ in rats, and the effect thereon of ethylenediamine tetra-acetate. The experiments are connected with the study of pernicious anaemia in humans.

**2823h:** M. J. Koopmans: An *in vitro* evaluation of the toxicity of chemicals for erysiphaceae (Meded. Landbouwhogesch. Opzoekingsstat. Gent **24**, 821-827, 1959, No. 3/4).

Description of a method of evaluating the toxic effect of chemicals on the conidia (spores) of powdery mildews (Erysiphaceae) *in vitro*, but without making a spore-germination test. In the latter test a comparison is made at the end of the incubation period (here 20 hours at 20 °C) between the percentage of germinated conidia that have been exposed to the substance whose toxicity is to be assayed and those that have not. Since the percentage of germinated conidia, even under favourable conditions, is small (and almost zero in the case of powdery mildew from apple), the spore-germination test is not entirely effective. The method described here is based on the loss of turgidity rather than on the inhibition of germination, and consists in determining the percentage of healthy, turgid conidia at the end of the test. This percentage is much higher than the percentage of germinated spores, and the test can also be applied to powdery mildew from apple. An investigation into the toxic effect of karathane showed that there is a linear relationship between the dose of karathane and the response of the test organism. The method will be important in the first instance to fundamental research on problems of fungal physiology.

**2824:** J. Volger: Dielectric properties of solids in relation to imperfections (Progress in Semicond. **4**, 205-236, 1960).

Review article on the dielectric properties (permittivity, conductivity, loss angle, as functions of frequency) of insulators containing certain imperfections, and of semiconductors. A list of 110 references is given. After a theoretical introduction and a consideration of models for dielectric relaxation, the author discusses losses due to ionic movements, losses due to trapped electrons in colour centres, donor centres in semiconductors, the effects of irradiation and of electric fields, oxidic semiconductors and microwave effects.

VOLUME 22, 1960/61, No. 12 pp. 377-404 Published 11th September 1961

Philips Technical Review

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

# RECENT DEVELOPMENTS IN ELECTRONIC-FLASH LAMPS

by C. MEYER *).

621.327.9

*Photography with electronic-flash lamps has become such a commonplace in recent years, amongst amateurs as well as professionals, that these lamps now rival in importance the older combustion-type flash bulbs. This is evident from the fact that the annual world production of electronic-flash units has been roughly half a million over the past five years. Assuming that each outfit is used on the average, say, 40 times a year, we see that a total of 100 000 000 photographs will be made this year with electronic-flash lamps. A development on this scale has been made possible by the progress achieved in the quality, economy, and reliability of the flash lamps themselves and the apparatus in which they are used. The principles underlying the design and manufacture of modern electronic-flash lamps are dealt with in the following article.*

## Introduction

Since the era of the explosive, smoky flash powders, the technique of flash photography has developed in two directions. In the first place there are the combustion-type flash bulbs, that can only be used once [1]; the flash unit for these bulbs contains a minimum of circuitry, namely a battery, a resistor and a capacitor. Secondly, there is the electronic-flash lamp, which was a later arrival on the scene. Originally, this type of lamp was only intended for special purposes, but subsequent development has now given it a firm footing in many fields of photography, including amateur photography. The circuit for the operation of the electronic-flash lamp is rather more elaborate than that needed for the combustion flash bulb, but it has the advantage of enabling large numbers of shots to be taken in fairly rapid succession. An article on the development of electronic-flash lamps was published in this journal seven years ago [2].

Progress since that time has been marked by the advent of lamps for operation at lower voltages, making it possible to replace the paper capacitors originally used in the flash units by electrolytic capacitors, which are much lighter and smaller, but are not suitable for voltages higher than about 600 V.

The requirements of the designer of flash units have to be taken into account in the design of electronic-flash lamps, not only in the question of operating voltage but quite generally. This has led to the wide variety in the shape and dimensions of the electronic-flash lamps nowadays being made. As an illustration of this *fig. 1* shows a selection from the range of types manufactured by Philips.

According to their application the numerous types of electronic-flash lamps can be divided into four categories: *a*) flash lamps for voltages between 400 and 500 V, used in outfits with electrolytic capacitors; *b*) flash lamps for voltages between 2000 and 3000 V, used in outfits with paper capacitors; *c*) flash lamps for special scientific and industrial applications; *d*) stroboscopic lamps. In this article we shall be concerned only with flash lamps of categories *a*) and *b*), intended for general use.

Some remarks may be made here about the other two categories. The electronic-flash lamps under *c*), for special scientific and industrial purposes, closely resemble in many respects the types for general use. Apart from the shape of

*) Lighting Division, Eindhoven.

[1] J. A. M. van Liempt and J. A. de Vriend, The "Photoflux", a light-source for flashlight photography, Philips tech. Rev. 1, 289-294, 1936.
L. H. Verbeek, The specific light output of "Photoflux" flash-bulbs, Philips tech. Rev. 15, 317-321, 1953/54.
J. A. de Vriend, Ignition of "Photoflux" flash-bulbs with the aid of a capacitor, Philips tech. Rev. 16, 333-336, 1954/55.

[2] N. W. Robinson, Electronic flash-tubes, Philips tech. Rev. 16, 13-23, 1954/55.
Instead of "electronic flash-tube" the term "electronic-flash lamp" as recommended by the C.I.E., will be used here; where no misunderstanding is possible, we shall often simply refer to flash lamp.

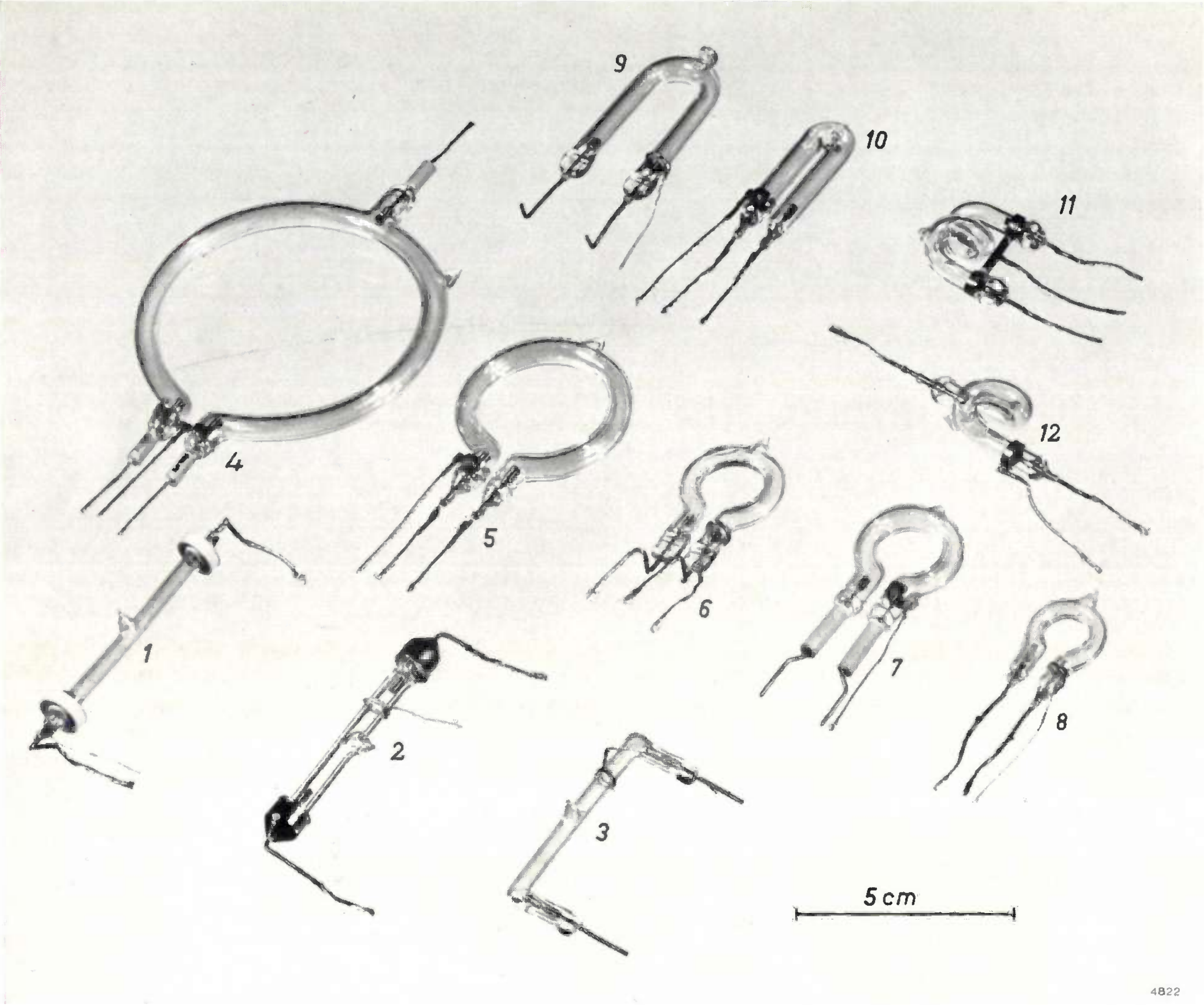


Fig. 1. Some examples of the wide variety of electronic-flash lamps made by Philips. Some types are U-shaped, some Ω-shaped, and others are in the form of spirals or straight tubes. All the types shown have flying leads. Their long life (more than 10 000 flashes) means in effect that they scarcely ever need to be replaced.

the lamps, the voltage, load and spectral distribution of the flash are adapted to the purposes for which the lamps are required. For example, there are special lamps for photographing tracks in Wilson cloud chambers [3]), for photographing the human eye [4]), for approach-warning on police cars, and for making photocopies.

The lamps in categories *a*), *b*) and *c*) are as a general rule suitable only for separate flashes or short successions of flashes, e.g. 10 to 20 flashes within a few minutes, after which they need some time to cool down. The stroboscopic lamps under *d*), however, are required to operate for up to several hours at a time, with a continuously variable flash frequency of anything from 15 to 300 flashes per second [5]). In the case of normal electronic-flash lamps, the permissible loading is given in watt.seconds per flash, while for stroboscopic lamps the average load in watts is usually specified, although of course these lamps, too, are pulse-loaded. It will be clear that the requirements imposed on stroboscopic lamps involve a whole range of fresh problems, with regard to both the design of the lamps and the circuitry.

In dealing with general-purpose electronic-flash lamps we shall first consider the electrical conditions under which the lamps have to operate. After then examining certain characteristics of the flash, we shall deal at somewhat greater length with design and manufacture. We shall then try to form a synthesis of all these questions in order to see how

[3]) N. Warmoltz and A. M. C. Helmer, A flash lamp for illuminating vapour tracks in the Wilson cloud chamber, Philips tech. Rev. **10**, 178-187, 1948/49.

[4]) J. E. Winkelman and N. Warmoltz, Photography of the eye with the aid of electronic flash-tubes, Philips tech. Rev. **15**, 342-346, 1953/54.
H. J. J. van Boort, N. Warmoltz and J. E. Winkelman, Colour photography of the retina and the anterior segment of the eye with the aid of a discharge flash lamp, Medicamundi **3**, 56-65, 1957.

[5]) S. L. de Bruin, An apparatus for stroboscopic observation, Philips tech. Rev. **8**, 25-32, 1946.

we may arrive at optimum designs of flash lamps. Finally, with reference to some typical representatives from the Philips range of flash lamps an idea will be given of the possibilitics offered by this kind of lamp.

## Electrical operating conditions of flash lamps

The energy to be dissipated in an electronic-flash lamp is derived from a capacitor, charged to a certain voltage. The energy stored in the charged capacitor is then converted into light in the flash lamp at the appropriate moment. This is done by discharging the capacitor through the lamp by means of a triggering circuit. There are thus two different electrical processes involved: the first process makes the required energy available, the second process, which is as a rule controlled from the camera, converts the energy into light at the right moment.

We shall now examine these two processes with reference to the circuit shown in *fig. 2.* The main capacitor, $C_f$, is charged to a suitable DC potential $U$; this potential also prevails between the anode $A$ and the cathode $K$ of the flash lamp $B$. In contrast to the arrangement described in reference [2], the cathode is here earthed, that is to say connected to the frame of the apparatus; this has become increasingly the practice in recent years.

In parallel with the lamp is a voltage divider $R_1$-$R_2$, whose function is to reduce the voltage across the initiating contact $S$ (in the camera) to 300 V, which is generally regarded as the maximum permissible [6]. The capacitor $C_i$ and the primary of the trigger-pulse transformer $T$ constitute the triggering circuit. When the contact $S$ is closed, the discharge of the capacitor $C_i$ induces a voltage surge in the secondary of the transformer $T$, and the trigger electrode $D$ receives the high voltage pulse required to trigger the flash lamp.

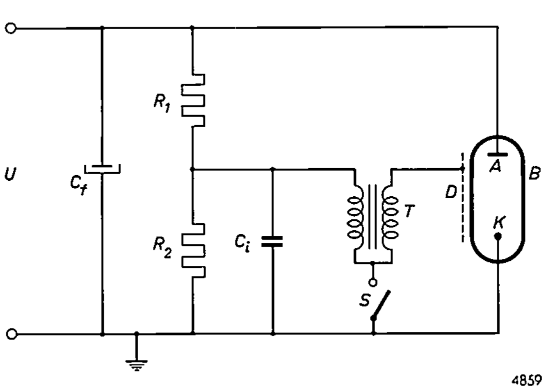


Fig. 2. Basic circuit for an electronic-flash lamp. The main capacitor $C_f$ is charged to a DC potential $U$. $B$ envelope, $A$ anode, $K$ cathode and $D$ trigger electrode. $R_1$-$R_2$ voltage divider for the triggering voltage. $C_i$ capacitor for trigger pulse. $S$ initiating contact, normally built into the camera and synchronized with the shutter. $T$ trigger-pulse transformer.

The potential $U$ can be obtained in various ways. Formerly, an accumulator or dry battery was generally used for this purpose in conjunction with an electromechanical vibrator, a transformer and a rectifier, but in recent years increasing use has been made of transistor circuits.

We shall now indicate the requirements to be met by the power supply, which are of importance in the considerations to follow. In order to produce the maximum number of effective flashes per dry battery or per charge of the accumulator, the load per flash must of course be as small as possible. Moreover the power supply must be so designed as to keep the final voltage which the main capacitor receives as constant as possible, since the energy accumulated in the capacitor varies according to the square of the operating voltage. A drop in the supply voltage $U$ by $12\frac{1}{2}\%$ causes roughly a 25% drop in capacitor energy and hence in flash energy; the light output then drops by about the samc percentage. This can have particularly unfortunate results in colour photography, owing to the relatively small latitude of colour-sensitive emulsions. From the photographic point of view, the above drop in light output means of course that the diaphragm ought to be opened a further half stop. Some improvement in this respect became possible with the advent of the control circuits used at the present time. These circuits meet both the above-mentioned requirements: they switch off the current source as soon as the capacitor is fully charged, and they automatically re-charge the capacitor as soon as the leakage current in the capacitor causes the voltage to drop below a certain preset value.

At a capacitance $C$ and a voltage $U$ the energy accumulated in the main capacitor is $\frac{1}{2}CU^2$. We shall disregard here the fact that the effective capacitance of electrolytic capacitors depends on the speed of the discharge. For an energy of, say, 125 Wsec, we then have the choice between a capacitance of, for example, 1000 μF at 500 V and a capacitance of 40 μF at 2500 V. In the latter case, in view of the high tension a paper capacitor must be used, which, on the given data, would weight about four pounds and have a volume of just over a litre; in the other case (lower voltage) we can use two electrolytic capacitors, having a total weight of about 2 pounds and a volume of about two thirds of a litre. A much smaller power pack is thus possible. Since it has become possible to make electrolytic capacitors capable of withstanding

[6] In Germany, where the most electronic-flash units for the European market are made, this value has been laid down in a standard recommendation (DIN 19 014).

repeated charge and discharge whilst retaining a constant capacitance, there has been a steady trend, in view of the advantages mentioned, towards flash lamps for lower voltages.

Further factors that militate in favour of lower voltages include safety measures (protection from high tensions becomes simpler) and the fact that for a given flash lamp and a given flash energy the duration of the flash can be longer. We shall return to this point presently.

After the main capacitor and the power supply, attention must be turned to the triggering conditions. In the circuit shown in fig. 2 the voltage divider is usually designed with $R_2 \leqq R_1$, so that where the operating voltage is 500 V we must reckon with a primary triggering voltage of less than 250 V (this is the voltage on the contact in the camera).

The decisive factor here is the available triggering energy. To guarantee a reliable discharge, the designer of electronic-flash lamps requires a triggering energy of at least 2 mWsec. For reasons of safety an upper limit of 12 mWsec has been proposed (see the recommendation in reference [6]). The triggering energy $W_i = \frac{1}{2}C_iU_i^2$ is established by the suitable choice of the primary triggering voltage $U_i$ and the capacitance $C_i$ of the triggering capacitor. Present-day flash units remain far below the safety limit, and generally deliver a triggering energy between 3 and 5 mWsec, but not more, in order to avoid loading the contact in the camera more than necessary. The loading of the contact has been investigated, and a value recommended for the product of maximum current and maximum voltage on the contact which should not be exceeded if the contact is to have a long life. This condition leads to the specification of a minimum value for the inductance $L_i$ of the trigger transformer. From $\frac{1}{2}L_i i_{\max}{}^2 = \frac{1}{2}C_iU_i{}^2$ the value of the maximum current is $i_{\max} = U_i \sqrt{C_i/L_i}$, giving the condition $U_i{}^2 \sqrt{C_i/L_i} \leqq 2250$ W. In designing the trigger units, use is commony made of a nomogram such as that in *fig. 3*, from which the value of $C_i$ and the minimum value of $L_i$ can be read off for a given $U_i$ and $W_i$.

The triggering energy is converted into a very short pulse, which covers approximately half a cycle of the sinusoidal oscillation produced by the circuit consisting of $C_i$ and the inductance $L_i$ in the triggering unit. The length of the pulse is thus $T_i \approx \pi\sqrt{L_iC_i}$ sec. From the condition $U_i{}^2 \sqrt{C_i/L_i} \leqq 2250$ W it then follows that $T_i \leqq \pi W_i/1125$ sec, so that with a triggering energy of 2 mWsec the available pulse length should be at least 6 μsec, and with $W_i =$ 12 mWsec at least 35 μsec. To get some idea of the triggering time of a flash lamp, it must be borne in mind that strong ionization occurs only when the voltage of the pulse is near maximum, and therefore the effective ionization time may only be half or one third of the total pulse length.

The flash-lamp designer requires from the triggering circuit not only a minimum energy $W_i$ but also a certain minimum triggering potential, i.e. a minimum secondary voltage at no-load operation of the transformer. The value generally required is 8 kV, which ensures reliable triggering. Oscillographic investigation of the triggering process has shown that the secondary voltage does not usually reach this value in practice, because the triggered flash lamp presents a short-circuit path to the triggering voltage. The oscillograms also show that, after the actual triggering process is completed and the discharge in the lamp has taken place and extinguished, a number of damped oscillations still occur in the triggering circuit; since the main capacitor is discharged, however, these can never cause the lamp to fire again.

A trigger unit based on the circuit of fig. 2 has been designed by Philips [7]), which meets both the energy and voltage requirements mentioned above and is very simple in design (printed wiring is used). It is fitted with an extremely reliable transformer, having polystyrol insulation and a ferroxcube core, see *fig. 4*.

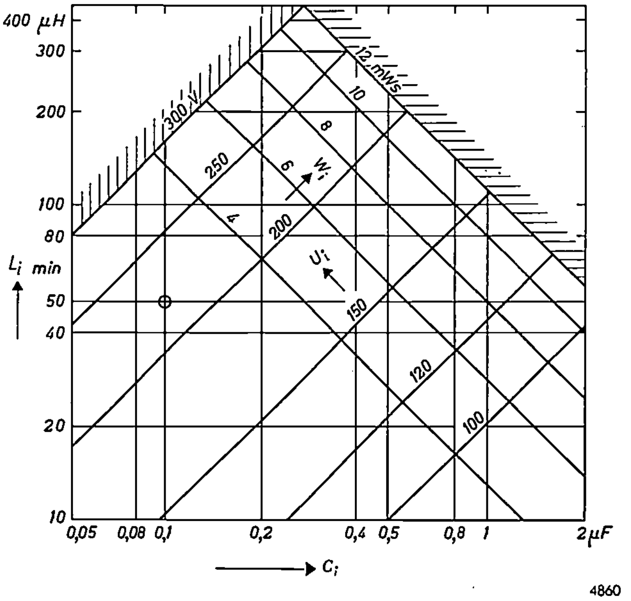
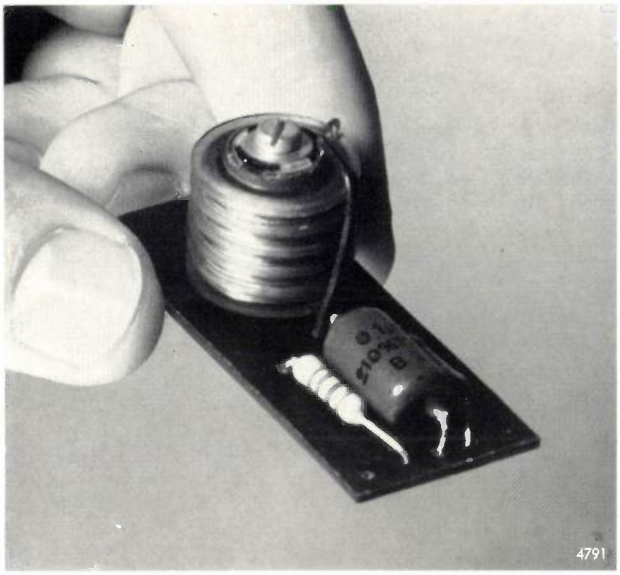


Fig. 3. Nomogram for the design of the triggering circuit $C_i$-$T$ in fig. 2. At a given triggering voltage $U_i$ and energy $W_i$, the diagram gives the required capacitance $C_i$ for the triggering capacitor and the minimum inductance $L_{i\,\min}$ required for the transformer. The point corresponding to the triggering unit shown in fig. 4 is marked with a ring: $C_i = 0.1$ μF, $L_i = 50$ μH.

[7]) Developed by H. E. van Brück and C. Slofstra of Philips Icoma Division (Industrial Components and Materials), Eindhoven.



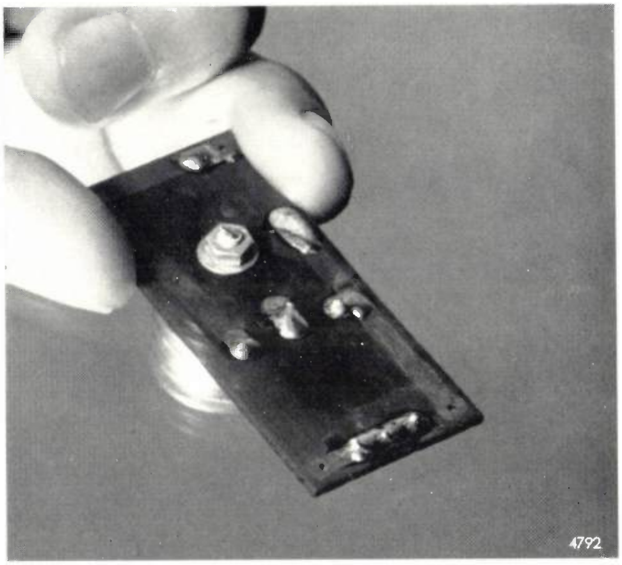


Fig. 4. The Philips triggering unit, with printed wiring. The trigger-pulse transformer has a coil wound on polystyrol round a ferroxcube core.

We shall now briefly consider the relation between the operating voltage and the triggering voltage of an electronic-flash lamp. As the operating voltage is raised, the necessary triggering voltage — which of course is always higher than the operating voltage — decreases. This is represented graphically in *fig. 5*, except that instead of the triggering voltage, which appears on the secondary of the transformer, the primary triggering voltage is shown (this is proportional to the triggering voltage). The regions "ignition" and "non-ignition" are not divided by a sharp line, but by a transitional region of some breadth. In any given flash lamp there is always a certain spread in the triggering process, and a still greater spread between different lamps of the same type. The diagram in fig. 5 was obtained by varying the operating voltage and the triggering voltage independently of one another (in most flash units the available primary triggering voltage is proportional to the operating voltage). Fig. 5 shows that as the operating voltage drops — thereby shifting the operating point roughly in the direction of the arrow — the region is gradually reached where the flash lamp will no longer ignite with certainty, or will not ignite at all. To avoid this it is increasingly the practise to stabilize the primary triggering voltage with a neon lamp, or, as in some modern control circuits, to stabilize the operating voltage itself.

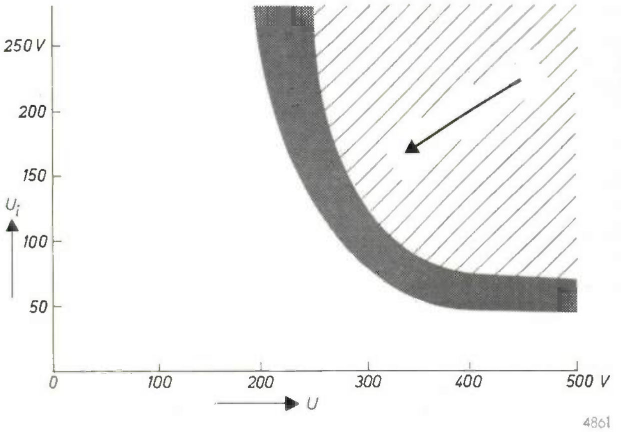


Fig. 5. The required primary triggering voltage $U_i$ of an electronic-flash lamp as a function of the lamp operating voltage $U$. At higher lamp voltages, lower triggering voltages are needed. The regions "ignition" (hatched) and "non-ignition" are not separated by any sharp dividing line. In a given circuit the lamp voltage and the primary triggering voltage are proportional to one another; and if they decrease, the operating point shifts in the direction of the arrow.

## Spectral distribution, integrated light intensity and flash duration

The user of an electronic-flash lamp needs to know the following concerning his flash outfit: the spectral distribution of the radiated light, the light output, and the duration of the flash. Also of importance is the spatial distribution of the radiation achieved with the aid of a reflector, but we shall not be concerned with that here.

The spectral distribution of the radiated light depends in the first place on the type of gas with which the lamp is filled. The inert gas xenon, in a discharge of the kind produced in electronic-flash lamps, delivers a continuous spectrum which largely corresponds to that of "natural" daylight. This is one of the reasons why nearly all electronic-flash lamps today are filled with xenon. *Fig. 6* shows the measured spectral distribution of the radiation emitted by a xenon-filled flash lamp, compared with the spectral distribution of daylight [8]).

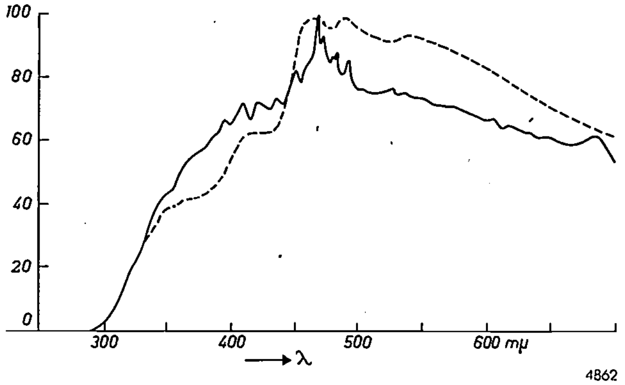


Fig. 6. Spectral distribution of the radiation from a xenon-filled flash lamp (solid line). For comparison, the broken line indicates the distribution of daylight [8]).

The fact that the light emitted by xenon-filled flash lamps so closely resembles daylight can be understood by considering the colour-temperatures of the radiations (i.e. both radiations compared with that of a black body). The colour-temperature attributable to natural daylight, which is composed of sunlight and the diffuse radiation in the firmament [8]), is roughly 6000 °K, and that found in the radiation from electronic-flash lamps is between 5800 °K and 7100 °K. The colour-temperature of flash lamps increases for a given type of gas with the specific loading of the tube wall, that is the electrical load divided by the total area of the discharge-tube walls. Xenon discharges give the best correspondence to daylight, with colour-temperatures of about 6000 °K at specific loadings in the region of 15 Wsec/cm$^2$, and of about 7000 °K at 30 Wsec/cm$^2$.

Because of the high colour-temperature of xenon-filled electronic-flash lamps, it is possible with this light to use normal daylight colour films, unlike the situation with continuously burning photographic lamps, where colour film specially sensitized for artificial (tungsten) light has to be used. The flash unit can also be used in daylight to provide supplementary lighting in dark shadows without causing impermissible colour distortion.

With regard to light output and luminous efficiency, i.e. the quantity of light per watt.second, we shall confine ourselves to the flash lamp itself without taking any account of the influence of the reflector, which is of course essential to the photographically effective use of the light.

In specifying the light output of a photographic light-source it must not be forgotten that the concept "light" by definition relates only to radiation of wavelengths between 380 and 780 mμ, and that every radiation contribution in this spectral region is evaluated in accordance with the spectral sensitivity of the eye. The units commonly used in lighting engineering are therefore based on this spectral sensitivity, for which a certain average curve has been internationally accepted. The photographic use of the light is not concerned with the human eye but with photographic emulsion. Since the spectral sensitivities may differ widely from one emulsion to another — consider, for example, orthochromatic and panchromatic emulsions, or the differing colour sensitivities of negative film and reversal film — any measurement of the "light output" would have to be based on an emulsion of average spectral sensitivity. At one time this was in fact done. Owing to various difficulties, however, including the reproducibility of such a standard emulsion, use is nowadays made of the above-mentioned system of lighting units for evaluating the light output and luminous efficiency of photographic light-sources; the light output, then, is given in lumen.seconds and the luminous efficiency in lumens per watt (or, in our case, in lumen.seconds per watt.second).

Suitable gas fillings, apart from xenon, are the inert gases argon and krypton. Elenbaas has compared these gases in a continuous discharge [9]), and has found that a xenon filling gives the highest luminous efficiency (*fig. 7*). Luminous efficiencies from 40 to 50 lumen/watt are obtained with the xenon-filled electronic-flash lamps now being made. The luminous efficiency of a given flash lamp increases with the loading somewhat as in fig. 7 (which refers to continuous discharges). *Table I* gives such data for the Philips electronic-flash lamp OF 235 Ws (Type No. 103 740) showing the light output in lumens and the luminous efficiency at various loads. We shall return presently to the influence which the *shape* of the lamp has on the luminous efficiency.

Table I. Light output and luminous efficiency of Philips' OF 235 Ws electronic-flash lamp as a function of load. The discharge tube of this type of electronic-flash lamp is made of quartz glass.

| | | | | | | |
|---|---|---|---|---|---|---|
| Load (Wsec) | 40 | 60 | 80 | 100 | 125 | 200 |
| Light output (lm.sec) | 1632 | 2480 | 3410 | 4400 | 5625 | 9380 |
| Luminous efficiency (lm/W) | 40.8 | 41.3 | 42.7 | 44.0 | 45.0 | 49.9 |

[8]) R. Herrmann, Optik 2, 384-395, 1947.

[9]) W. Elenbaas, High-pressure rare-gas discharges, Philips Res. Repts. 4, 221-231, 1949.

It may be said that the duration of the flash is the most important property of an electronic-flash lamp from the point of view of the photographer. In order to get sharp photographs of scenes with moving objects, he needs the shortest possible flash. In this respect the electronic-flash lamps meet all the requirements of normal photography: the flash duration is of the order of 1 msec, which is much shorter than that of flash bulbs of the "Photoflux" type, whose flash may last as long as 30 msec. This implies, too, that a smaller total light output is needed for an exposure with an electronic-flash lamp than with combustion-type flash bulbs, for even when the fastest shutter speeds are used, the entire flash takes place while the shutter is fully open (provided the synchronization is correct): the total light output radiated is thus usefully employed. (In the case of focal-plane shutters it is of course necessary to ensure that the very short flash occurs while the shutter blind has uncovered the whole field.)

It should be pointed out, however, that extremely short flash times may be photographically unfavourable owing to the Schwarzschild effect. At given values of the product of luminous intensity and exposure time, the blackening of a photographic emulsion in the case of very long exposure times, and with the flash lamp, very short exposure times, is less than at average exposures. Where colour-sensitive emulsions are used a similar effect occurs at very short exposures, and may already be noticeable at exposures not much shorter than 1 millisecond. From a photographic point of view, therefore, the flash duration should not be shorter than 1 msec.

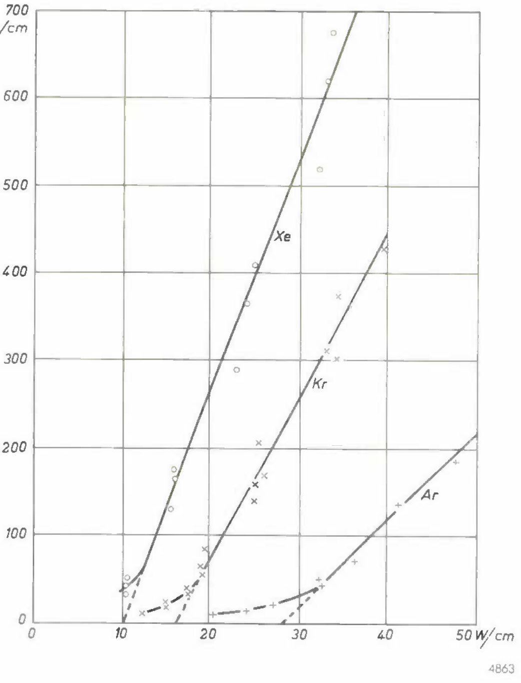


Fig. 7. Luminous flux as a function of the power in the arc of discharge tubes filled with argon, krypton and xenon, under continuous burning [9]). For the purpose of comparing discharge tubes of different dimensions, the luminous flux and power are reduced to correspond to one centimetre length of arc.

The variation of luminous flux with time can be displayed on an oscilloscope. An example of such an oscillogram is shown in *fig. 8*: a sharp rise in the luminous flux is followed by a relatively slow decline.

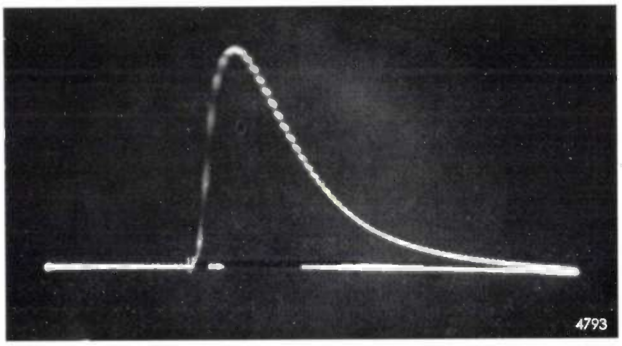


Fig. 8. Oscillogram of the luminous flux of an electronic-flash lamp. The electron beam of the oscilloscope is modulated at a frequency of 20 kc/s, so that each spot in the waveform corresponds to 0.05 millisecond.

The flash duration of various lamps, even where the time variation of the luminous flux differs from one lamp to another, may be compared on the basis of the half-width of the pulse, defined as the time during which the luminous flux is greater than half the peak value. Sometimes the 10% width is also given, the definition of which is analogous to the above. Since the triggering delay in electronic-flash lamps is negligible compared with the flash duration, it may be said that the beginning of the flash coincides with the beginning of the discharge (i.e. the triggering). In this respect electronic-flash lamps differ considerably from combustion flash bulbs.

The length of the flash depends on the design of the flash lamp, on the capacitance of the main capacitor and, as mentioned, on the operating voltage. As the capacitance increases the flash duration also gradually increases, but it drops as the operating voltage increases. The change to flash lamps with a lower operating voltage, using correspondingly larger capacitances, therefore amounted in fact to prolonging the flash duration. *Fig. 9* shows the dependence of the flash half-width on the capacitance and on the operating voltage, at a constant load of 40 Wsec. Also, for increasing load, the operating voltage remaining constant, the flash duration increases; this appears from *Table II*, which refers to a typical electronic-flash lamp. It can be seen that at the nominal value of the operating voltage the flash durations (half-widths) are still shorter than a millisecond. They are thus shorter than are really desirable from a photographic standpoint. This is a consequence of the geometrical limitations imposed on the flash lamps: considerations of simple optics set a limit to the arc length and to the diameter of the discharge tube. At the present time it is not yet technically possible to achieve half-widths of 1 millisecond or more with the desired convenient dimensions of flash unit and reflector.

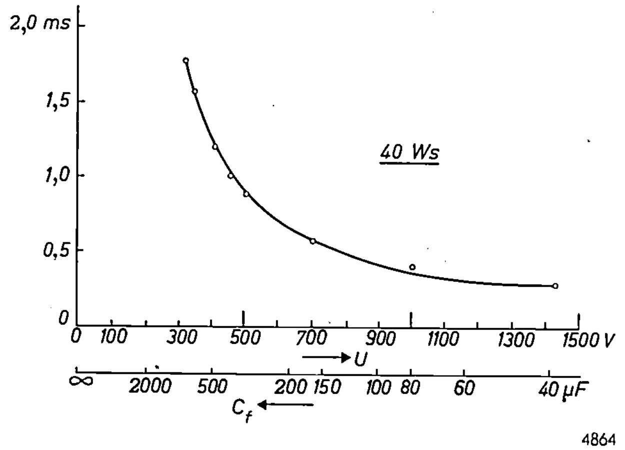


Fig. 9. Relation between the half-width of the pulse and the capacitance $C_f$ and the operating voltage $U$, under constant load. Longer flash times are obtained the lower the voltages used and the higher the capacitances.

For some purposes, for example for optical warning signals on police cars, a particularly long flash is required. This is possible by sacrificing something of the luminous efficiency, the compactness and the cheapness of the flash apparatus; one can, for example, connect a resistance in series with the flash lamp, or modify the circuit in accordance with the considerations given earlier.

Table II. Flash duration (measured as half-width and 10% width) of a typical electronic-flash lamp as a function of load, at a constant operating voltage of 485 V.

| Flash energy (Wsec) | 60 | 125 | 200 |
|---|---|---|---|
| Half-width (msec) | 0.50 | 0.67 | 0.84 |
| 10% width (msec) | 1.11 | 1.62 | 2.29 |

## Construction and manufacture of electronic-flash lamps

In broad lines the construction of electronic-flash lamps may be described as follows. Sealed into the glass tube are two electrodes between which the discharge takes place. The tube is filled with xenon (usually) to a pressure of a few hundred torr (1 torr = 1 mm Hg). The trigger electrode is mounted on the outside of the tube. Pins are usually fitted, serving the dual purpose of electrical connections and mounting legs. The pins may be fixed in one or more bases. These details of the construction will be further discussed presently.

The discharge tube proper is usually of hard glass, sometimes quartz glass. The latter is used where the tube wall is to be subjected to particularly heavy loading. If a great deal of heat is generated during the discharges, the glass may be locally heated to its softening point, giving rise to stresses upon cooling; in the course of time hair-cracks may form ("sintering"), which may finally lead to breakage. There is much less danger of this happening with quartz glass, owing to its higher melting point and lower expansion coefficient.

Although formerly hard glass was usually found to be adequate in most cases, increasing use has recently been made of quartz glass, in view of the growing trend towards small, heavily loaded lamps. Since a point source is optically preferable, this trend is understandable. However, as can be seen in *fig. 10*, the discharge tube in practice takes the most various forms. The simplest shape for round reflectors, which were initially very widely used, is the U shape (*a*). Somewhat more complicated, but better adapted to the round reflector, is the Ω form (*b*) or the helical loop (*c*). Efforts were later made to adapt the reflector to the rectangular form of the picture, with the idea of ensuring that the picture would receive uniform overall illumination. For this purpose a linear shape of tube is more suitable: (*d*) or (*e*), the latter with the electrodes fitted perpendicular to the tube. Finally there is the spiral or helix form (*f*), which has been popular for high-tension apparatus right from the beginning. Since a long discharge path was needed for high operating voltages, the obvious method of producing a compact light source was to spiralize the tube.

In conjunction with a suitable reflector, the shape of the lamp can influence the spatial distribution of the light. The luminous efficiency, on the other hand, depends on the dimensions of the tube (apart from the electrical operating conditions). For a given flash energy, the dimensions of importance are the inside diameter of the tube and the distance between the electrodes, i.e. the length of the tube. We shall return to this question when we come to consider the optimum design of a flash lamp.

Turning now to the electrodes, which are sealed into the ends of the discharge tube, there are two processes that take place at the electrodes and therefore determine their design. The first is the triggering process (ignition), the second is the actual burning

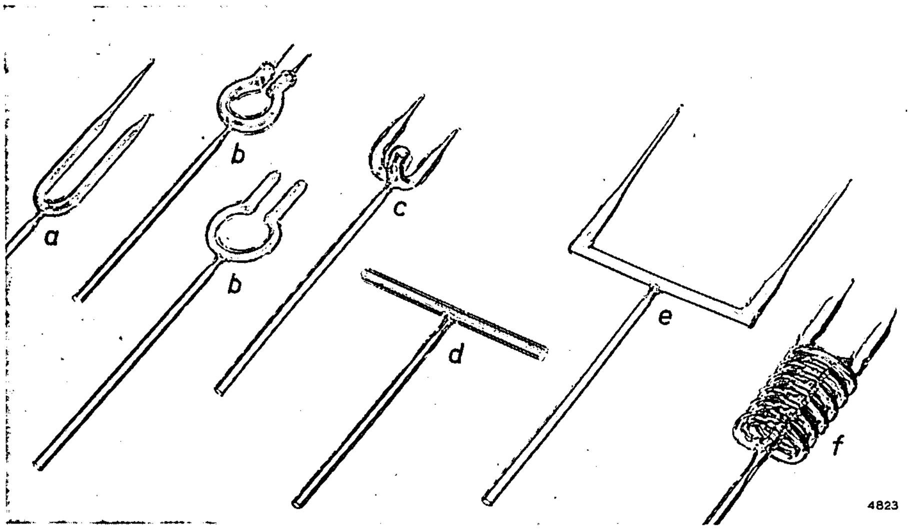


Fig. 10. Various shapes of discharge tube.

of the lamp which, though short-lived, can cause considerable heating of the electrodes as a result of the high discharge current. The electrodes must be designed in such a way as to enable the tube to ignite at given values of operating and triggering voltages, and they must continue to do so during the whole life of the lamp. Furthermore, the evaporation of electrode material should be minimized to prevent deposits forming on the wall and reducing the light output during the life of the lamp.

In order not to be unduly restricted in the choice of the other parameters, a low triggering voltage is aimed at, which implies that the electrode material must have a low work function. To this end the electrodes, which consist of a tungsten wire core with a tungsten or molybdenum wire wound around it, are coated with a highly emissive substance. In most cases the emitter substance consists largely of thorium oxide and barium oxide. The emitter also acts as a getter. It has been found that the anode, too, can advantageously be given an emissive coating. The two electrodes are therefore usually made identical. It is not advisable, however, to operate flash lamps using one electrode first as cathode and later as anode, or vice versa. The electrodes must therefore be distinguishable. In the type of flash lamp described here the electrode used as cathode during the burning-in and aging periods in manufacture can be recognized from the fact that the connection to the trigger electrode is situated at the cathode end of the flash lamp (see e.g. figs. 11 and 15).

The kind of trigger electrode used also has its influence on the ignition of the lamp. In electronic-flash lamps the triggering is invariably capacitative, i.e. the trigger electrode is mounted on the *outside* of the discharge tube. As can be seen in *fig. 11*, there are three common types of trigger electrode: a spirally wound wire; a conducting strip usually laid parallel to the discharge along the inside of the loop; and a transparent, conductive layer with which the entire discharge tube is coated. The discharge is initiated along the trigger electrode, and for this reason the conductive layer, introduced in recent years, has the advantage over the others.

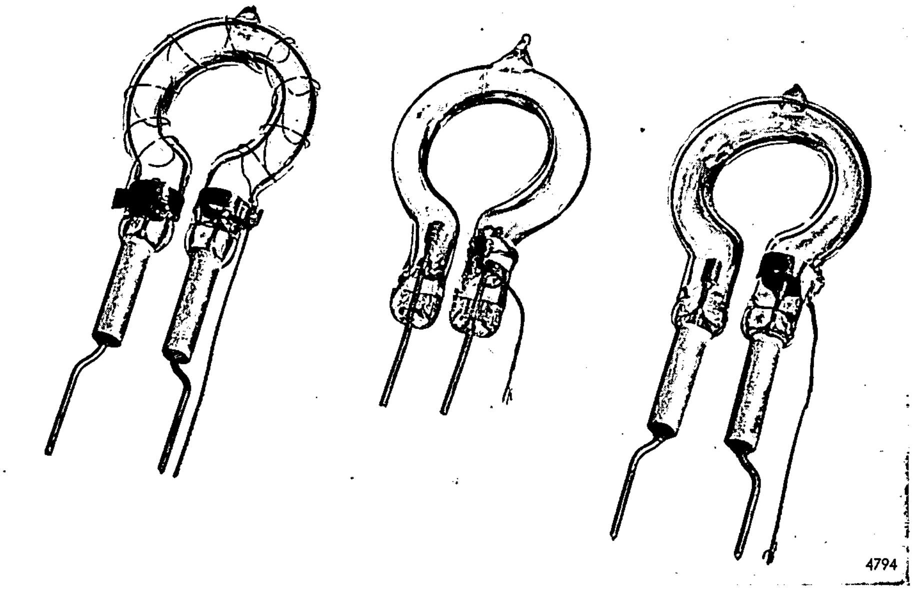


Fig. 11. Three types of trigger electrode for electronic-flash lamps; from left to right: trigger wire wound around the tube; strip of conductive material (not distinctly visible, owing to reflections in the glass); transparent conductive layer (not visible). The latter type has certain advantages.

It gives the lowest triggering voltage and leads to a uniform discharge throughout the tube.

A word here about the sealing-in of the electrodes. Generally speaking the seals present no difficulties, in spite of the high current surges of several hundred amperes. In quartz-glass tubes, however, where the usual pinch seal with thin molybdenum foil is to be used, the current surges can cause considerable heating at the seals and may destroy the thin foil used. This problem can be solved by using somewhat thicker molybdenum foil. In some particularly troublesome cases, tungsten seals with intermediate glass had to be adopted [10]).

The economic manufacture of electronic-flash lamps is somewhat of a problem owing to the relatively small production runs required. For this reason, flash lamps are frequently hand-made. Nevertheless, even in the case of small production runs, some mechanization is an advantage in that it leads to a more uniform quality of the product. To what extent is mechanization possible even in the manufacture of small series of flash lamps? In many cases the discharge tube can be bent to the required shape mechanically. A simple machine for this purpose is shown in *fig. 12*. The glass tube is heated by a burner (*A*), rough-shaped around a mandrel (which projects from the bottom of pipe *B*) and introduced into a mould (*C*), after which the two halves (*D*) of the mould are closed and air is blown into the tube (the other end of the tube being sealed off). This method of shaping has the advantage of keeping the outer dimensions of all tubes within very narrow limits, a point of importance as regards assembly in the reflector. After a pump stem has been sealed to it, the discharge tube is thoroughly cleaned and dried ready for further working.

Parallel with these operations, the electrodes are made. To begin with, the electrode wire core is provided with a bead of the intermediate glass (*fig. 13a*), which serves to compensate for the difference in expansion coefficient between the electrode wire and the glass tube (we shall discuss here only the case of hard-glass tubes). This being done, the helix of W or Mo wire is slid onto one end of the glazed core, and the complete assembly is again thoroughly cleaned and dried (fig. 13*b*). The electrodes are then coated with emitter paste, in such a way that the paste also penetrates into the space between the core and the helix wire (fig. 13*c*). After the electrodes have been sintered in a tungsten-strip furnace, their surface is carefully brushed (fig. 13*d*). The discharge should not spring directly from the emitter substance, since this might cause evaporation of the latter and lead to the formation of light-absorbent deposits on the glass wall. On the other hand, the complete absence of emitter material on the outside of the electrode would imply a higher triggering voltage. A compromise is therefore adopted, some emitter material being left

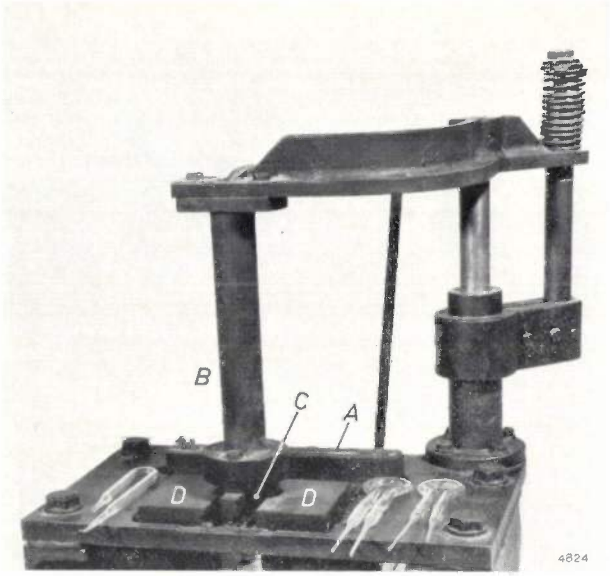


Fig. 12. Simple machine for shaping the envelopes of discharge tubes, in this case Ω-shaped. *A* burner for heating the glass tube. *B* pipe, through the base of which a mandrel projects for pre-shaping the glass tube. *C* mould into which the tube, bent into a U shape, is introduced. *D* mould halves for final shaping.

Fig. 13. Stages in the fabrication of electrodes for hard-glass flash lamps.
*a*) Wire core with intermediate glass.
*b*) Wire helix fitted over one end of the core.
*c*) Assembly coated with emissive paste.
*d*) After sintering in a tungsten-strip furnace, the superfluous paste is brushed away from the spiral.

[10]) See in this connection page 81 of the article by P. Hoekstra and C. Meyer, Motion-picture projection with a pulsed light source, Philips tech. Rev. 21, 73-82, 1959/60 (No. 3).

on the outside of the electrodes, which vaporizes when the lamp is burnt in and forms a slight deposit on the glass wall — a deposit which, however, causes no significant drop in light output.

The two electrodes can now be sealed into the discharge tube. This too is done on a small machine (*fig. 14*), making it possible — since both electrodes are sealed in at the same time — to achieve the specified dimensions, in particular the electrode spacing, more accurately and faster than by hand. Next, the tubes are evacuated on the pump and carefully degassed. Gases or vapours released from the electrodes or glass wall may contaminate the gas filling and make the tube unreliable in operation. With this in mind, the tube is evacuated to a pressure lower than $10^{-4}$ torr. If necessary, the electrodes can then be activated, and the tube is filled with xenon to the specified pressure and sealed off. The subsequent operations consist primarily in applying the trigger electrode. If this is a conductive layer, it is sprayed on in an oven whose temperature is nearly at the softening point of the glass. The connection wires are then soldered to the tube and a base is fitted if required. This completes the actual production process (*fig. 15*).

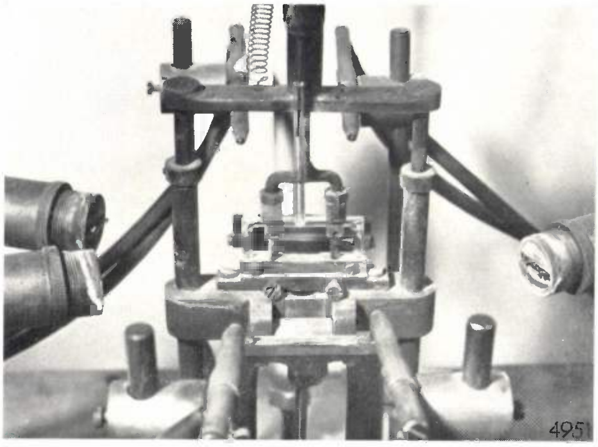


Fig. 14. Machine on which the two electrodes of an electronic-flash lamp are simultaneously sealed into the tube.

The flash lamps have still to be aged and inspected, however. Both steps are of especial importance for obtaining a product of good quality. Aging is necessary to bring the electrodes into the best condition for operation. We have seen that the emitter is not only important as regards reliable ignition but also acts as a getter. Although it is not possible to say exactly which part of the electrode is responsible for ignition and which part for gettering, it has been found that, after a certain number of flashes under normal load, the flash lamp becomes stable in operation, which means in particular that the triggering voltage becomes fairly constant. In the inspection stage the principal dimensions of the lamps are examined, the triggering voltage of each lamp separately is measured, and sample inspections are made to determine the light output, and also the decline in light output after a specific number of flashes.

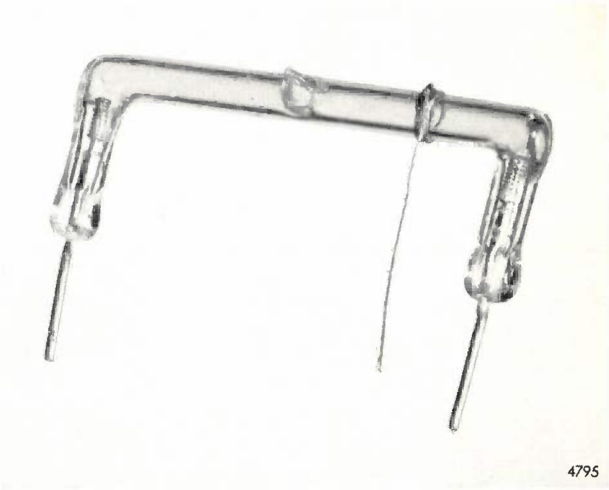


Fig. 15. Philips electronic-flash lamp, type OF 50 Ws (No. 3 in fig. 1).

### The optimum design of electronic-flash lamps

So far various questions that arise in the development of electronic-flash lamps have been examined separately. We shall now consider the question of how to produce "optimum" flash lamps, and to do this we must investigate the relation between the electrical and lighting requirements and the constructional details.

The three essential requirements to be met by a flash lamp have already been mentioned: reliable triggering, maximum light output under given conditions, and minimum decline in light output during the life of the lamp. The electrical circuit to be used is generally established. As regards triggering, then, we may assume that the following data are given: the minimum operating voltage at which the lamp must ignite, the minimum primary triggering voltage available, the primary triggering energy available, and the type of trigger-pulse transformer to be used. These data are needed, and the triggering transformer itself must be available, in order to determine the ignition behaviour of a flash lamp. As regards the voltage data, it must also be borne in mind that a certain safety margin is necessary in the production process. In ascertaining whether the specification has been met, we must therefore take values a few per cent lower than the minimum values mentioned.

Calculation of luminous efficiencies from measurements of the light output must be done on the basis of the nominal values of the operating voltage and the capacitance of the main capacitor.

Life tests, for determining the number of flashes that the lamp can withstand, must be done at the maximum load likely to be encountered in practice, that is to say the highest encountered values of capacitance and operating voltage. The decline in light output during the life of the lamp is also determined under these conditions, and used as the basis for assessing the quality of the lamp. On the other hand, the light-output measurement necessary for determining this decline is carried out under nominal loading.

It is also necessary to know the maximum operating voltage because of the fact that, if the voltage used is too high (close to the self-breakdown voltage), the flash lamp may ignite without being triggered. It is particularly necessary to take this into account in the case of high-tension flash lamps, where the voltage of the high-tension supply may vary considerably.

We have seen that marked differences between actual and nominal voltages can cause a considerable spread in the light output of the lamp. The same applies to differences in capacitance and resistance values in individual flash units of the same type. Capacitors, for example, can normally only be made with a tolerance of —10% to +20% in their capacitance value. This again entails a spread in the light output by more than 25%. Because of this capacitance spread and the differences already mentioned in the charging potential of the capacitor, the user of a flash unit has hitherto had to make a series of test exposures to determine the right stop to be selected or the guide number to be used. In this respect the advent of voltage-stabilized circuits has brought some improvement, but there is still the need to have all electrical data accurately specified without unduly wide tolerances. As regards the capacitance tolerances of electrolytic capacitors there has latterly been some slight improvement.

Once the electrical data are established, the behaviour of the lamp in respect of triggering and light output can be controlled by varying the dimensions of the discharge tube, i.e. the electrode spacing $l$ and the inside diameter $d$, and also by varying the pressure $p$ of the gas filling.

The relation between these quantities has been experimentally investigated on linear discharge tubes; inclusion of the effects of tube shape would have endlessly prolonged the investigation. In any case, the data for bent tubes are not reliable, inasmuch as bending changes the diameter of the tube in a manner that is difficult to define satisfactorily. The experiments were set up according to statistical principles. The various parameters were varied within the following limits: electrode spacing from 20 to 80 mm, diameter of tube from 2 to 6 mm and gas pressure from 100 to 700 torr. Preliminary measurements had shown that the maximum light output was to be expected between these limits. The entire investigation was done in three experimental runs, namely for three different loads of 40, 62.5 and 100 Wsec. Although of course the light output from an electronic-flash lamp increases with increasing load, the load itself (i.e. the flash energy) is laid down by the flash-unit designer and is therefore not a freely variable parameter.

The light output, with constant tube diameter and constant load, is found to depend on the electrode spacing $l$ and the gas pressure $p$ in the manner shown in *fig. 16*. We see that there is indeed a cer-

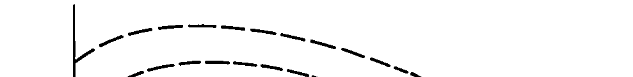


Fig. 16. Curves of constant light output of a flash lamp showing dependence on the gas pressure $p$ and the electrode spacing $l$, for constant discharge-tube diameter $d$ and constant load. The figure on each curve represents the light output expressed as a percentage of the maximum. (The dashed portions of the contours were obtained by extrapolation [11].) If the diameter $d$ of the discharge tube is varied, the point of maximum light output shifts along the chain line $M$; at a certain point along this line (i.e. at a certain tube diameter) there is an optimum maximum light output, i.e. a maximum of the various maxima.

[11] Statistical analysis of the results carried out by J. W. Sieben (Lighting Division, Eindhoven).

tain optimum combination of $p$ and $l$ at which the light output is a maximum. When the tube diameter is varied, the position of the peak shifts along the chain line $M$, and the height of the peak itself becomes maximum at a particular point along this line. Of course, when the tube diameter is varied, the whole family of curves shifts together with the position of the peak. Since this involves no fundamentally new phenomena, however, it is not necessary to reproduce all these diagrams here. The same applies to the diagrams for the three different loads.

It follows from the measurements just described that it is indeed possible, with given electrical conditions, to produce an optimum flash lamp, provided the tube diameter, the electrode spacing and the gas pressure may be freely chosen. If some other shape of flash lamp is required, e.g. a U shape or Ω shape — we assume that the results obtained on linear discharge tubcs can be applied to other shapes without serious error — the lamp designer can make one optimum type of flash lamp for any required load. The dimensions may be adduced from fig. 16 or from the corresponding diagrams for other tube diameters and loads.

It may happen that a flash lamp thus designed is found to be too large and that for optical reasons, i.e. with an eye to the reflector, a smaller and more compact lamp is required (in most cases shorter). Furthermore, a small tube diameter may be wanted in order to lengthen the duration of the flash. In that case, only the gas pressure can be freely chosen, and herc too the best choice may be adduced from a diagram as in fig. 16.

We still have to ascertain the way in which triggering is affected by the variation of parameters $l$, $d$ and $p$. As a rule, the minimum operating voltage at which the flash lamp can still just be ignited increases with rising $l$ and $p$ and with decreasing $d$. This relation can be represented by contours of constant minimum triggering voltage in an $l$-$p$ diagram as shown in *fig. 17* for the same case as in fig. 16 (i.e. for the same tube diameter $d$). The figure shows, for example, that maximum light output cannot be achieved with a flash-unit circuit where the lower limit of the available primary triggering voltage is 260 V: to ensure reliable ignition in that case, we must make $l$ and/or $p$ smaller than the values needed to produce maximum light output.

We have not yet considered the decline in light output during the life of the lamp. It is mainly governed by the quality of the electrodes, and is not much affected by the three parameters $l$, $d$ and $p$.

Finally, an idea of the present situation in the development of electronic-flash lamps is given in *Table III*, which gives data for a few representative types. A few years ago, most flash units for amateurs were still equipped with lamps for a flash energy of about 80 to 120 Wsec. Nowadays a light output sufficient for most photographic purposes is obtained with a flash energy of 45 Wsec. Examples of such flash lamps are type OF 45 Ws (linear shape) and OF 50 Ws (fig. 15; electrodes perpendicular to discharge tube to reduce length). The U-shaped lamp OF 165 Ws is intended for higher-performance apparatus, and can be subjected to a maximum load of 165 Wsec. The light output from this lamp is so high that objects larger than normal can be given good overall illumination; for close-up shots of smaller objects it is then sometimes necessary to change to a lower flash encrgy.

The two types OF 80 Ws and OF 100 Ws approach closely to the optimum design described above, at loads of 80 and 100 Wsec, respectively. Type OF 235 Ws, with Ω-shaped quartz-glass discharge tube, is intended for press photographers. Type OF 500 Ws is a universal flash lamp, of which only

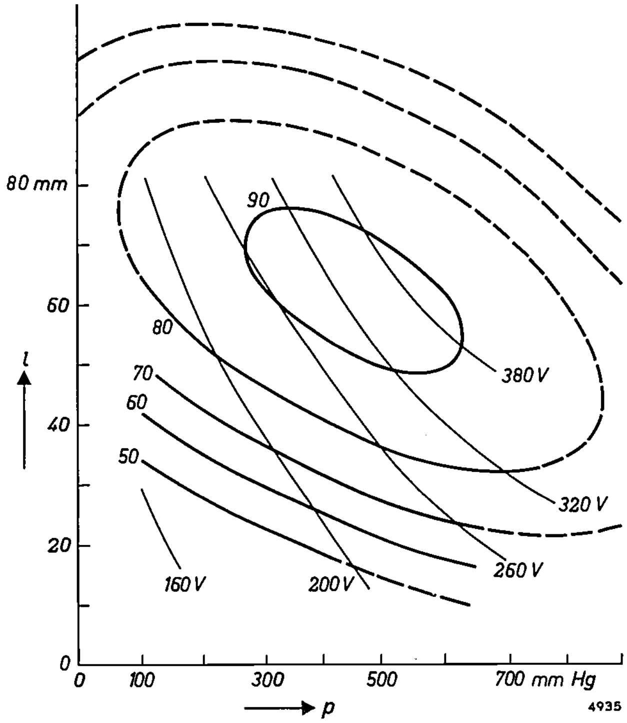


Fig. 17. At a given tube diamete r the minimum triggering voltage can be specified for any electrode spacing $l$ and gas pressure $p$. The set of thin curves shown here is found by joining together the points of minimum triggering voltage in fig. 16. If the power-supply circuit in a flash unit is so designed that the (minimum) available triggering voltage is only 260 V, the maximum light output cannot be achieved in the case shown here: electrode spacing and gas pressure would then have to be chosen small enough to remain on the left of the curve for 260 V.

Table III. Data on some Philips electronic-flash lamps; in each case the type designation indicates the maximum permissible flash energy.

| Type designation and number | Operating voltage V | Shape | Envelope | Electrode spacing mm | Light output lm.sec | Luminous efficiency lm/W | Wall load per flash Wsec/cm$^2$ | Colour-temperature °K |
|---|---|---|---|---|---|---|---|---|
| OF 45 Ws<br>103 909 | 500 | fig. 1,<br>No. 1 | glass | 44 | 1900 | 47.6 | 13 | 6400 |
| OF 50 Ws<br>103 931 | 500 | fig. 15 | | | | | | |
| OF 165 Ws<br>103 798 | 500 | fig. 1,<br>No. 9 | glass | 70 | 6550 | 46.8 | 12 | 6500 |
| OF 80 Ws<br>103 952 | 500 | fig. 1,<br>No. 12 | glass | 65 | 2735 | 45.6 | 7.3 | 5800 |
| OF 100 Ws<br>103 958 | 500 | fig. 1,<br>No. 10 | | | 2830 | 47.2 | | |
| OF 235 Ws<br>103 740 | 500 | fig. 11 | quartz glass | 63 | 9380 | 46.9 | 23 | 7100 |
| OF 500 Ws<br>103 965 | 500<br>to<br>2500 | fig. 1,<br>No. 5 | quartz glass | 110 | 15 000 | 50 | 22 | 6600 |
| OF 1100 Ws<br>103 752 | 2700 | fig. 10*f* | quartz glass | 500 | 39 500 | 49.4 | 18.5 | 5800 |
| OF 1500 Ws<br>103 730 | 2700 | fig. 1,<br>No. 4 | quartz glass | 205 | 45 600 | 45.6 | 38.8 | 6900 |

experimental versions have been made, and which operates reliably in a very wide range of voltages: the operating voltage can be selected in this case between 500 and roughly 2500 V. Because of the relatively large dimensions of the discharge tube, this type provides the relatively long flash duration required for photography: at lower voltages half-widths of a few milliseconds are obtained.

For comparison, the table also gives data on two representative types of high-voltage electronic-flash lamps. The spiralized lamp OF 1100 Ws can be used in flash apparatus which, though heavy, is nevertheless portable. Type OF 1500 Ws is intended for studio use; shaped like a large ring, it can be fitted around the lens of the camera.

Summarizing, the major development in the design of electronic-flash lamps may be said to have been the reduction of the flash energy for amateur equipment to 45 Wsec, made possible by the improvement of luminous efficiency. The introduction of these 45 Wsec lamps has made it a practical proposition to produce electronic-flash units weighing less than 2 pounds. The general trend is in the direction of still lower flash energies with a view to producing even smaller and lighter flash equipment. Present indications, however, are that the luminous efficiency of the lamps will then be lower.

**Summary.** Electronic-flash lamps are made in a wide variety of types to meet the requirements of the manufacturers of flash equipment. An important feature of developments in this field in recent years has been the advent of flash lamps for operation at the relatively low voltage of 500 V. This has made it possible to produce readily portable flash equipment fitted with light-weight electrolytic capacitors. The designer of flash lamps must work on the basis of both the electrical operating conditions (governed by the triggering and discharge circuit) and the photographic lighting requirements (spectral distribution of light, light output and flash duration). Lamps filled with xenon (at a pressure of some hundreds of torrs) give a spectral distribution closely resembling natural daylight. The light output at a given flash energy (luminous efficiency) has been so improved in recent years that a flash energy of 45 Wsec is now sufficient for most photographic purposes. As a result it is now for the first time possible to make flash apparatus with a total weight of less than 2 pounds. The duration of the flash is short enough for most practical cases; indeed, in view of the Schwarzschild effect, it was even desirable to make it somewhat longer, and the flash duration has now been brought close to the desired value of 1 millisecond. After a description of the basic design and methods of manufacture of flash lamps, details are given concerning the appropriate choice of the electrode spacing, the diameter of the discharge tube and the gas pressure to produce an optimum design. Tabulated data are given for a few typical electronic-flash lamps made by Philips.

# AN EXPERIMENTAL NOISE GENERATOR FOR MILLIMETRE WAVES

621.373.432.029.65:621.327.52

A microwave noise source is needed for purposes such as measuring the noise factor of millimetre-wave equipment (e.g. radar receivers), and also as a standard noise source in plasma research. A suitable noise source of this nature is the positive column of a discharge in an inert gas, provided it is so dimensioned as to give a high equivalent noise temperature. External factors, like the magnitude of the discharge current, the filament voltage and the ambient temperature, generally have little influence on the noise temperature and the matching.

In the millimetre-wave region the positive column can be regarded as a black-body radiator of very high temperature, closely approximating to the electron temperature of the plasma. The electron temperature is primarily governed by the gas used — at least under the conditions chosen for the discharge column in a noise generator — and is higher the lighter the gas atoms.

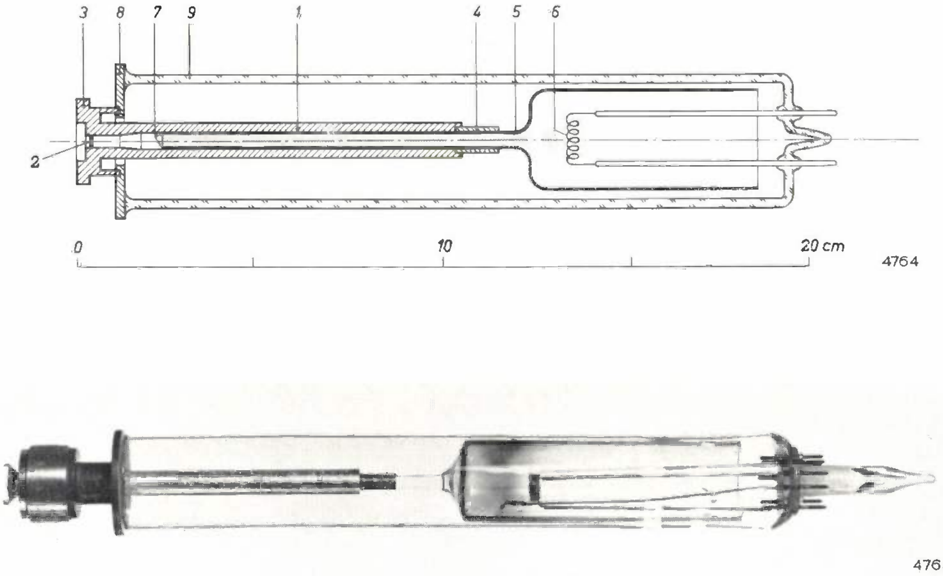


Fig. 1. Cross-section and photograph of experimental noise generator for the 4 mm waveband. *1* circular waveguide. *2* mica window. *3* flange for coupling to circuit. *4* spring clips. *5* quartz-glass tube. *6* oxide cathode. *7* part of inside wall of waveguide acting as anode. *8* molybdenum disk. *9* neon-filled glass envelope (pressure 10 cm Hg).

For frequencies up to about 40 Gc/s (wavelengths down to 7.5 mm) a noise generator can be made by simply taking the glass tube in which the gas discharge takes place and passing it obliquely through a waveguide in such a way that the part of the column inside the guide is properly matched and thus delivers the maximum noise power to the guide [1]). At higher frequencies, however, the dimensions of the waveguide are too small to make this system practicable. More suitable in this case is the design illustrated in *fig. 1*, which gives good results in the 4 mm band. A circular copper waveguide *1* is closed at one end (left) by a mica window *2*, and provided at that end with a flange *3* for coupling to the rest of the circuit. Inside the waveguide a thin-walled tube *5* of quartz glass is introduced. At the right this flares out into a widened section which contains an oxide cathode *6* for the gas discharge; the tube is filled with neon. The left-hand end of the tube is open and the anode is formed by the inside of the waveguide near *7*, immediately beyond the end of the tube. The positive column is contained

[1]) W. W. Mumford, A broad-band microwave noise source, Bell Syst. tech. J. **28**, 608-618, 1949.
K. S. Knol, Determination of the electron temperature in gas discharges by noise measurements, Philips Res. Repts. 6, 288-302, 1951.

inside this tube, i.e. in the axial direction of the waveguide. The column can be made long enough for the equivalent noise temperature to approach closely to the electron temperature, so that the maximum noise power is delivered to the waveguide and is transmitted through the mica window to the rest of the circuit. To minimize reflection losses, the window is provided with a tuned diaphragm.

In principle it would be possible to seal the discharge tube at the cathode side hermetically by using a quartz-glass base with lead-ins for the cathode. Since the tube has to be very thin-walled, however, the assembly would then be too vulnerable. For this reason a design as shown in fig. 1 was adopted: soldered to the outside of the waveguide is a molybdenum disk *8*, sealed to which is a glass envelope *9* carrying the cathode leads.

The experimental tube built in this way has a neon pressure of 10 cm Hg. The discharge current is 75 mA, the burning voltage 150 V, and the noise temperature $T$ is 21 000 °K, the maximum error of measurement being $\pm 1000$ °K. The noise power available in a narrow frequency band $\Delta f$ is $kT\Delta f$; $k$ is Boltzmann's constant $= 1.38 \times 10^{-23}$ J/°K.

In *fig. 2* the standing-wave ratio $s$, measured on the experimental noise generator, is plotted as a function of frequency $f$. From $s$ the percentage by which the equivalent noise temperature is lower than 21 000 °K at a given frequency can be derived. This percentage is shown on the scale at the right of fig. 2.

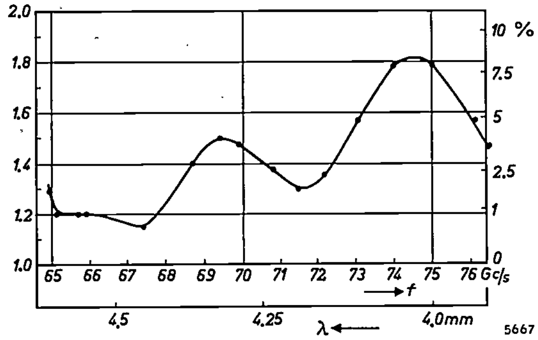


Fig. 2. Standing-wave ratio $s$, measured on the noise generator shown in fig. 1, as a function of frequency $f$ (or wavelength $\lambda$). The scale on the right indicates by what percentage the equivalent noise temperature at a given frequency is lower than 21 000 °K.

It is certain that this type of noise generator may be extended to wavelengths shorter than 4 mm. To produce a noise source for the 2 mm band it will probably be sufficient to add a transition section from 4 to 2 mm and to compensate for the mismatch.

The tube discussed here will be dealt with in more detail in an article on standard noise sources to be published in this journal.

P. A. H. HART and G. H. PLANTINGA.

# MULTIPATH TRANSMISSION EFFECTS IN FM RECEPTION AND THEIR SIMULATION IN THE LABORATORY

by J. KOSTER *).

621.391.826.2:621.376.33

*In mountainous regions the waves from a broadcasting transmitter may reach the receiver along multiple paths of different length as a result of reflections from mountain ridges. The consequence of this in frequency-modulated broadcasts may be severe distortion of the sound. The author has designed a signal generator for simulating and studying this interference in the laboratory. This makes it possible to check at any time, irrespective of receiving conditions, the effect of measures taken in the receiver to reduce this distortion.*

In FM broadcasting (frequency-modulated VHF transmissions) use is made of waves in the metre bands. These waves as a rule reach the receivers along the direct path from transmitting to receiving aerial. Not infrequently, however, one or more other transmission paths may exist at the same time, owing to the waves being reflected from some natural obstacle, such as a mountain ridge. The various paths will generally differ in length, which means that two or more waves having different transit times and hence a phase difference arrive at the receiving aerial. The consequence, particularly if the reflected waves are not much weaker than the direct wave, is a peculiar distortion of the detected signal. The impression one receives is as if the output amplifier were overloaded, or as if something were loose in the loudspeaker. The cause, however, is of quite a different nature, as will appear from the analysis given below.

The distortion in question was very soon noticed when frequency modulation first began to be used. Its cause was also correctly ascertained, and measures for improvement were proposed [1]) [2]) [3]) [4]). In this connection the investigations led by Arguimbau made an especially useful contribution [3]). In order to study the effect of these measures, a signal generator is needed which is capable of delivering a signal corresponding to that which an aerial receives under the conditions mentioned. A relatively simple solution of this problem is described in the present article. To make clear the requirements to be met by such a signal generator, we shall first give a simplified analysis of the effects involved.

## The FM receiver

*Fig. 1* shows the familiar block diagram of a normal FM broadcast receiver. The audio-frequency section ($A_3$-$L$) need not be considered here. The radio-frequency amplifier ($A_1$), the mixer ($M$) and the intermediate-frequency amplifier ($A_2$) can be regarded for our purposes as linear networks, in other words, we may apply to them the superposition theorem. This states that in the simultaneous presence of more than one signal the total effect is the sum of the effects of the individual signals, provided only that the amplitude characteristic is sufficiently horizontal and the phase characteristic sufficiently straight. These conditions are reasonably satisfied if the bandwidth of the amplifiers is not less than about three times the maximum frequency deviation of the transmitted signal. In the case of FM broadcasting stations the maximum frequency deviation is fixed at 75 kc/s by international agreement.

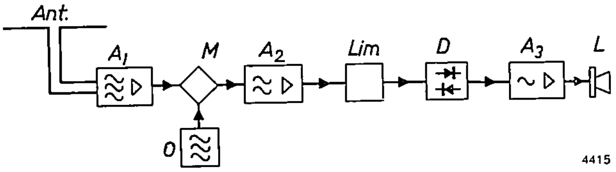


Fig. 1. Block diagram of a conventional FM broadcast receiver. *Ant* aerial. $A_1$ radio-frequency amplifier. *M* mixer. *O* local oscillator. $A_2$ intermediate-frequency amplifier. *Lim* limiter. *D* FM detector (discriminator). $A_3$ audio-frequency amplifier. *L* loudspeaker. (For clarity the limiter is shown as a separate block; in reality, limiting occurs partly in the last stage of $A_2$, partly in the discriminator.)

*) Radio, Television and Record-player Division, Eindhoven.

[1]) M. S. Corrington, Frequency-modulation distortion caused by multipath transmission, Proc. Inst. Radio Engrs. **33**, 878-889, 1945.
M. S. Corrington, Frequency modulation distortion caused by common- and adjacent-channel interference, R.C.A. Rev. 7, 522-560, 1946.

[2]) F. L. H. M. Stumpers, Interference problems in frequency modulation, Philips Res. Repts. 2, 136-160, 1947.

[3]) L. B. Arguimbau and J. Granlund, The possibility of transatlantic communication by means of frequency modulation, Proc. Nat. Electronics Conf., Part III, 644-653, 1947.
L. B. Arguimbau and J. Granlund, Interference in FM reception, Tech. report No. 42, Research Lab. of Electronics, Massachusetts Inst. of Technology, 1947.

[4]) L. W. Johnson, F.M. receiver design, Wireless World 62, 497-503, 1956.

The limiter (*Lim*) and the discriminator (*D*) are essentially non-linear systems. It is therefore not enough to consider each input signal individually; we must also take their resultant into account. We shall see presently that in certain circumstances the instantaneous frequency of the resultant signal may make excursions far beyond the band within which the instantaneous frequencies of the constituent signals remain. To meet these unfavourable circumstances it is necessary to give the non-linear part of thc receiver a bandwidth larger than is needed for a normal FM signal.

## Analysis of multipath transmission effects

In an article in the previous issue of this journal it was shown that an FM receiver to which two signals are simultaneously applied will detect thc stronger of the two, whilst the weaker one will act as an interfering signal [5]). Multipath reception of FM signals is to be treated as a special case of this.

To avoid unneccssary complication of the problem, we assume that the transmitter is modulated by a sinusoidal audio signal (frequency $p = \Omega/2\pi$) and that between the transmitting and the receiving aerial there are only two transmission paths, with a transit-time differencc of $\tau$. Of these two paths one may be the direct path and the other indirect, though both may also be indirect.

In the radio-frequency part of the receiver there will then be two signals present, both sinusoidally modulated in frequency and one lagging behind the other by a time $\tau$. Instead of these RF signals we can better consider the corresponding intermediate-frequency signals, both of which are lower than their corresponding RF frequencies by the same amount (which is equal to the frequency of the local oscillator). The instantaneous frequencies $f_a$ and $f_b$ of the two intermediate-frequency signals are given by:

$$\left.\begin{aligned} f_a &= f_0 + \Delta f \sin \Omega t, \\ f_b &= f_0 + \Delta f \sin \Omega(t-\tau)\,. \end{aligned}\right\} \quad . \; . \; . \; . \quad (1)$$

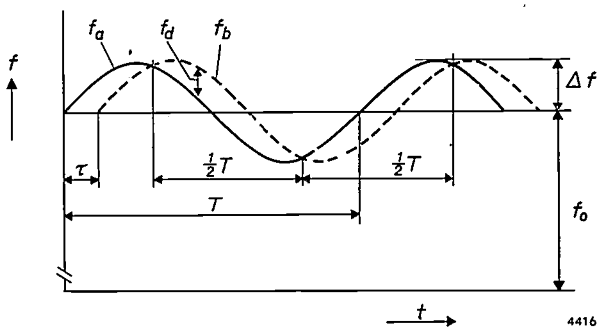


Fig. 2. Instantaneous intermediate frequencies $f_a$ and $f_b$ of the direct and indirect signals, respectively, received from an FM transmitter modulated by a sinusoidal audio signal (frequency $p = 1/T$). The fixed centre intermediate frequency is $f_0$ (usually 10.7 Mc/s). The difference in the transit times of the two transmission paths is $\tau$.

Here $f_0$ is the centre intermediate frequency (often fixed at 10.7 Mc/s) and $\Delta f$ is the frequency deviation. *Fig. 2* shows $f_a$ and $f_b$ as functions of time *t*. As can be seen, during one half of the period *T* of the audio signal the frequency $f_a$ is lower than $f_b$, and during the other half it is higher. The difference $f_d$ of the two instantaneous frequencies is found by simple calculation from (1) to be a cosine function of time with the audio frequency $\Omega/2\pi$:

$$f_d = f_a - f_b = 2\,\Delta f \sin \tfrac{1}{2}\Omega\tau \cos \Omega(t - \tfrac{1}{2}\tau)\,. \quad (2)$$

We represent the two IF signals by the vectors **a** and **b** (*figs. 3* and *4*) of length *a* and $b \leqq a$, respectively [6]). We keep thc vector **a** stationary; in accordance with eq. (2) the vector **b** then rotates for one half period $\frac{1}{2}T$ anticlockwise a number of times ($f_b > f_a$) and in the next half period the same number of times clockwise ($f_b < f_a$), with an instantaneous angular vclocity of $\omega_d = 2\pi f_d$.

The resultant of **a** and **b** is the vector c, and it is the signal corresponding to **c** that is detected by the

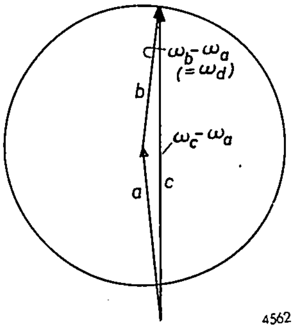


Fig. 3

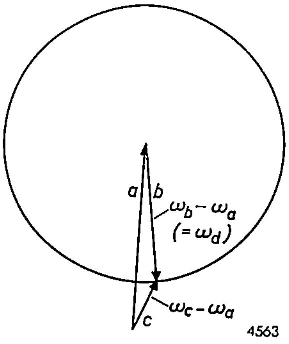


Fig. 4

Figs. 3 and 4. Vectors a and b represent the intermediate-frequency signals corresponding respectively to direct and indirect reception of an FM signal. Vector a is stationary, vector b rotates alternately anti-clockwise and clockwise with an angular velocity $\omega_d = 2\pi f_d$ in accordance with eq. (2). The resultant c represents the signal to which, after limiting, the discriminator responds. Vector c shows amplitude modulation and an irregular angular velocity ($\omega_c$ is small at the moment represented in fig. 3; it is large and in the opposite sense to $\omega_d$ at the moment to which fig. 4 refers). In both figures $b/a = 0.8$.

[5]) J. van Slooten, FM reception under conditions of strong interference, Philips tech. Rev. **22**, 352-360, 1960/61 (No. 11).

[6]) In variable reception conditions it may happen that *b* becomes greater than *a*. In that case *b* becomes the desired and *a* the interfering signal, and in the following considerations, *a* and *b* change places. See article cited under [5]).

discriminator. The function of the latter is to deliver an audio-frequency signal whose instantaneous amplitude is proportional to the frequency deviation of the input signal, i.e. in this case proportional to the instantaneous value $\omega_c$ of the angular velocity of the vector c.

It is easily seen from figs. 3 and 4 that $\omega_c$ shows marked variations during one revolution of **b**, particularly if $b$ is not much smaller than $a$. When **b** points roughly in the opposite direction from **a** (fig. 4), then c rotates faster (and in the opposite sense) than when **b** is more or less in line with **a** (fig. 3). There is thus no simple relation between the angular velocity of c and that of **b**. The fact that this must cause distortion of the audio signal is evident. We shall presently examine this phenomenon in quantitative terms.

It also appears from figs. 3 and 4 that the vector c changes in length during every revolution of **b**, varying from the maximum value $a + b$ to the minimum value $a — b$, which may be zero. The effect of this amplitude modulation will be considered separately.

### *Distortion due to frequency modulation of* c

Let $\psi$ be the phase of vector **b**, and $\varphi$ the phase of vector c (*fig. 5*), then

$$\varphi = \tan^{-1} \frac{b \sin \psi}{a + b \cos \psi}.$$

By differentiating $\varphi$ with respect to time and writing $x$ for the ratio $b/a$, we find for the angular velocity $\omega_c$ of the vector c:

$$\omega_c = \frac{x + \cos \psi}{x + x^{-1} + 2 \cos \psi} \omega_b , \quad . \quad . \quad (3)$$

where $\omega_b = d\psi/dt$.

To arrive at the instantaneous frequency $f_c$ of the resultant signal, which corresponds to the vector c, it must be remembered that as a result of keeping vector **a** stationary we must now add $f_a$ to the frequency of c, and also that the calculated angular velocity $\omega_b$ corresponds in reality to the difference frequency $f_d = f_a - f_b$. We then find from (3):

$$f_c = f_a + \frac{x + \cos \psi}{x + x^{-1} + 2 \cos \psi} f_d .$$

In *fig. 6*, $f_c$ is plotted as a function of $\psi$ for various values of the signal ratio $x$; the angle $\psi$ runs from zero to $2\pi$ (one revolution of **b**) and we assume that in this single period there is no significant change in the frequencies $f_a$ and $f_b$. At $x = 0$ ($b = 0$, interfering signal absent) $f_c$ is obviously equal to $f_a$; at $x = 0.25$ a distinct fluctuation occurs in $f_c$; at $x = 0.5$ the fluctuation is more pronounced, and at $x = 0.75$ it has become a sharp peak at $\psi = \pi$. In the limiting case $x = 1$ (indirect signal just as strong as the direct signal) $f_c$ is always equal to $f + \frac{1}{2} f_d$, except at $\psi = \pi$, where $f$ is discontinuous and equal to $-\infty$. It appears, then, as mentioned above, that the instantaneous frequency $f_c$ of the resultant signal may cover a much wider band than the instantaneous frequency of the constituent signals individually.

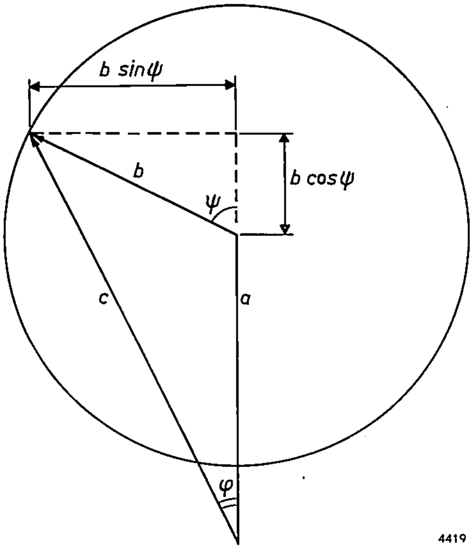


Fig. 5. Illustrating the derivation of the instantaneous frequency $f_c$ and the amplitude modulation of vector c.

The discriminator has to deliver an output signal whose magnitude at any given instant is proportio-

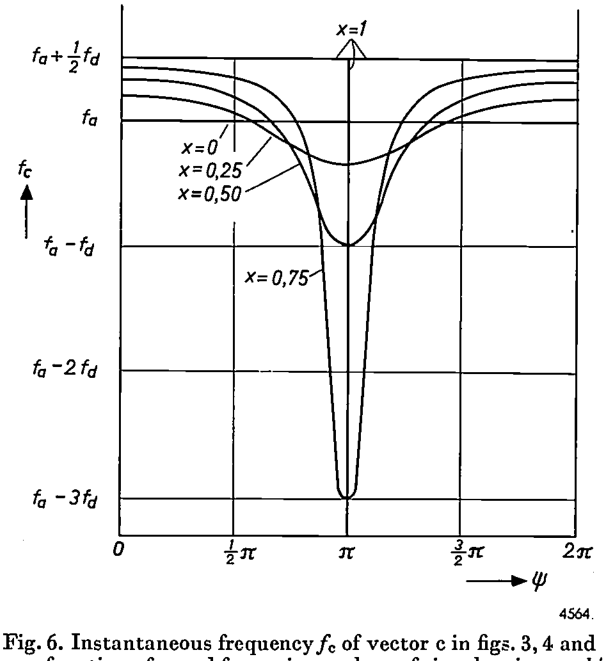


Fig. 6. Instantaneous frequency $f_c$ of vector c in figs. 3, 4 and 5, as a function of $\psi$ and for various values of signal ratio $x = b/a$.

nal to the instantaneous frequency deviation of the input signal. For example, if the instantaneous frequency $f_c$ of the input signal has the form shown by the curve for $x = 0.75$ in fig. 6, the output voltage will have a form that corresponds to the fluctuation of $f_a$ (i.e. the original audio signal) except for a superimposed peak at $\psi = \pi$ in each period $2\pi/\omega_b$ of the vector **b**. The output signal will then have an appearance such as that in *fig. 7*.

The number of peaks per period $T$ is equal to the number of complete revolutions of vector **b** (the vector **a** being stationary) in the time $T$. This number can be determined as follows.

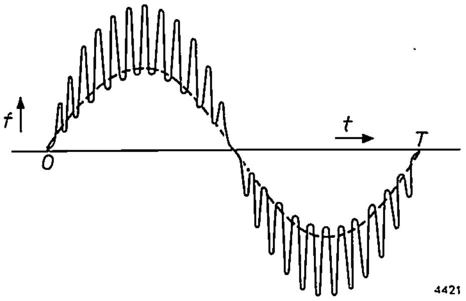


Fig. 7. Broken line: sinusoidal audio signal in undistorted reception. Solid line: with reception via two transmission paths, a peak occurs whenever the vector c (fig. 5) has a high angular velocity ($\psi \approx 180°$). Such a waveform occurs when the phase difference of the modulations of the two received signals is 180°.

We represent the phase of the signal that follows the shorter transmission path by

$$\Phi_a = m \sin \Omega t + \omega_0 t$$

and that of the other signal by

$$\Phi_b = m \sin \Omega(t-\tau) + \omega_0(t-\tau)\,.$$

The instantaneous phase difference of the two signals is then:

$$\Phi_d = \Phi_a - \Phi_b = m\,\{\sin \Omega t - \sin \Omega(t-\tau)\} + \omega_0\tau =$$
$$= 2m \sin \tfrac{1}{2}\Omega\tau \cos \Omega(t-\tfrac{1}{2}\tau) + \omega_0\tau\,. \quad . \quad . \quad (4)$$

This phase difference, then, varies with time in accordance with a cosine function, with the frequency $\Omega/2\pi$. During the quarter period in which the cosine increases from zero to 1, the phase difference changes according to (4) by an amount of $2m \sin \frac{1}{2}\Omega\tau$ radians. In a complete period $T$ the variation therefore amounts to $8m \sin \frac{1}{2}\Omega\tau$ radians. The number of times that the vector **b** in the time $T$ sweeps an angle of $2\pi$ radians (giving rise to one peak) is thus $\frac{4}{\pi}\, m \sin \frac{1}{2}\Omega\tau$, and this occurs per second

$$\overline{f_r} = \frac{4}{\pi}\,\Delta f \sin \tfrac{1}{2}\Omega\tau \text{ times}\,, \quad . \quad . \quad . \quad . \quad (5)$$

since $m = T\Delta f$; the quantity $\overline{f_r}$ is the average repetition frequency of the peaks.

Since the phase angle swept per period is generally not an exact multiple of $2\pi$, some peaks will not be completely formed.

Because the frequency difference $f_d$ varies (see fig. 2 or eq. (2)), the repetition frequency $f_r$ of the peaks fluctuates about the mean value $\overline{f_r}$ given by (5). The peaks are farthest apart at the moments when $f_a = f_b$ (i.e. at the points where the two sine waves in fig. 2 intersect) and are closest together midway between these moments. A high pulse rate is associated with a large amplitude. During the time $\frac{1}{2}T$ in which $f_b$ is greater than $f_a$, the frequency difference $f_d$ is negative and the peaks point upwards; during the other half cycle they point downwards (see fig. 7).

Concerning the numerical value of $\overline{f_r}$ (eq. (5)) it is difficult to be definite, since the factor $\sin \frac{1}{2}\Omega\tau$ with varying $\Omega$ can assume any value between zero and 1. The frequency deviation $\Delta f$, which may go up to 75 kc/s, amounts on an average to no more than about 15 kc/s. In the most favourable case ($\sin \frac{1}{2}\Omega\tau = 1$) the value of $\overline{f_r}$ at the average deviation may lie near the threshold of audibility (20 kc/s), and exceeds it only when the deviation increases. The repetition frequency $f_r$ fluctuates, as mentioned, around the mean value $\overline{f_r}$, thereby varying from zero to a value greater than $\overline{f_r}$. Evidently, therefore, $f_r$ can only be above the threshold of audibility during a *part* of the period $T$. In this part the peaks are generally crowded together. A peak represents a frequency deviation, the mean value of which, $\overline{f_c}$, follows from eq. (3) (if we regard the angular velocity $\omega_b$ as constant during the short time in which a peak is formed):

$$\overline{f_c} = \frac{\overline{\omega_c}}{2\pi} = \frac{\omega_b}{2\pi^2}\int_0^{\pi} \frac{x + \cos\psi}{x + x^{-1} + 2\cos\psi}\,d\psi\,. \qquad (6)$$

It can be shown that for $x$ smaller than unity this integral is zero. This means that the peaks in the time interval in question are practically inaudible.

What is the situation in practice? Apart from on the transit-time difference $\tau$, the phase difference depends on the audio frequency $p$, for at a low audio frequency the phase varies less per unit time (and thus also in the time $\tau$) than at a high audio frequency. Since, in the spectrum of music and speech, numerous frequencies occur which are fairly uniformly distributed over the audio range, there will always be a great many phase differences, some of them small. This, together with the fact that the average frequency deviation is only about 15 kc/s, accounts for the presence of an audible interference in music and speech whenever the strength ratio $x$ of the signals is relatively large.

In the foregoing we have tacitly assumed an ideal discriminator, i.e. one whose bandwidth is unlimited. In reality the bandwidth of a discriminator is limited, and if it is not large enough to deal completely with the peaks of $f_c$, the situation is even more unfavourable than described. The integral in (6) is then no longer equal to zero, so that every clipped peak gives rise to an audible component in the audio signal. Since the peaks representing the largest frequency deviations suffer most from the limited bandwidth, distortion occurs in the audio signal even when $f_r$ is above the audio limit.

To prevent this, it is necessary to ensure that the bandwidth $B$ of the non-linear portion of the receiver — i.e. the discriminator and the preceding limiter or limiters — is above a certain minimum. This minimum follows from the fact that the maximum frequency at the top of the peak is:

$$\frac{b}{c}(f_{\mathrm{d}})_{\max} + (f_{\mathrm{a}})_{\max} = \frac{b}{a-b} \times 2\Delta f + \Delta f = \frac{1+x}{1-x}\,\Delta f.$$

The required bandwidth is twice as large, i.e.

$$B \geqq \frac{1+x}{1-x} \cdot 2\Delta f. \quad . \; . \; . \; . \quad (7)$$

By solving (7) for $x$, we find the signal ratio $x$ at which a given bandwidth $B$ can only just deal with the peaks:

$$x = \frac{B - 2\Delta f}{B + 2\Delta f}.$$

In good FM receivers 400 kc/s is a usual bandwidth for the limiter and discriminator section. At a deviation $\Delta f$ of 15 kc/s, $x = 0.86$ is the value at which the peaks can still just be handled and at which the additional distortion is just suppressed. At a deviation of 75 kc/s the limit lies at $x = 0.455$. Where $x$ is in the neighbourhood of 1 the bandwidth has to be very large indeed, e.g. almost 6 Mc/s at $x = 0.95$ (and $\Delta f = 75$ kc/s).

*Distortion due to amplitude modulation of* c

The foregoing section related to the part contributed to the distortion by the irregular angular velocity of the vector c (fig. 5). A second contribution is due to the variations in the magnitude of the vector c, i.e. to the amplitude modulation (AM) present in the signal corresponding to c.

It can be seen from fig. 5 that

$$c = \sqrt{a^2 + b^2 + ab\cos\psi}\,.$$

In *fig. 8* the variation of the amplitude $c$ as a function of $\psi$ is plotted for a single revolution of the vector **b**, for different values of the signal ratio $x$. Each of these curves has a modulation depth equal to $x$.

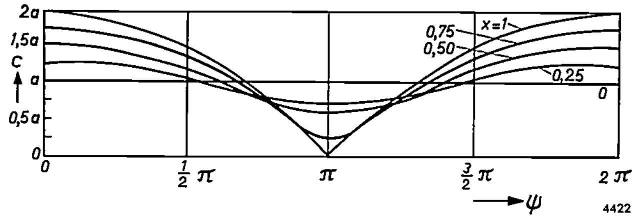


Fig. 8. Amplitude modulation of vector c (fig. 5). The variation of the amplitude $c$ is shown as a function of the angle $\psi$ between a and b, for various values of the signal ratio $x = b/a$.

It is the task of the limiter to suppress this amplitude modulation in order that the output signal from the discriminator shall not have a component possessing the distortion shown in fig. 8. FM broadcast receivers usually have two limiters. The function of the first limiter is performed by the last valve in the intermediate-frequency amplifier; this only acts as a limiter — due to the flow of grid current — when the signal reaches a certain strength. The second limiter — which operates also on weaker signals — is the discriminator itself, if the latter is a ratio detector [7]. The limiting action here is based on the variable damping effect, dependent on the signal amplitude, which a diode exerts on a resonant circuit. Obviously, this damping cannot go lower than zero. If the signal amplitude changes still further in the same direction, the limiting action fails, the diodes being temporarily cut-off. This effect will certainly appear at modulation depths of about 0.7 upwards.

If the signal is strong, the first limiter provides at the least for a 10-fold reduction of the modulation depth; there is then no chance of the detector failing, and the amplitude modulation is effectively neutralized. Where the signal is weak, however, the first limiter is inoperative, and if $x$ reaches the value at which the detector begins to fail, strong AM distortion results.

It should be noted that the fundamental frequency of the AM component is determined by the angular velocity of the vector **b**. As regards the audibility of an insufficiently suppressed AM component, the same remarks apply as to the peaks due to the frequency modulation of the vector c.

[7] The operation of the ratio detector as a discriminator and limiter is briefly explained in Philips tech. Rev. 17, 346, 1955/56, and treated in more detail by F. E. Terman, Electronic and radio engineering, McGraw-Hill, New York 1955, 4th impression, p. 610 *et seq.*

The great importance of good AM suppression may be understood by considering the case where AM is not suppressed at all. The low-frequency voltage generated in the discriminator by the *amplitude* modulation of c is then found to be roughly $15x$ times higher than the undistorted signal that the discriminator should give (for a frequency deviation of 15 kc/s). Even if $x$ is only 0.1, the AM interference is still 1.5 times stronger than the desired audio signal. On the other hand, where the receiver bandwidth is adequate and the values of $x$ are small, but otherwise under the most unfavourable conditions, the interference due to *frequency* modulation of c is only $0.2x$ times the undistorted audio signal. To reduce the total interference it is therefore of paramount importance to have adequate AM suppression.

### Measures for reducing FM distortion due to multipath transmission

From the fact that the distortion discussed increases in severity the closer the signal ratio $x = b/a$ approaches unity, it follows that our first countermeasure must be to try to reduce $x$. For this purpose an aerial having a sharp directional effect should be used, positioned in such a way that $x$ is minimum.

The measures to be taken in the receiver itself consist, as follows from the above considerations, in giving the limiter and discriminator section a large bandwidth, in ensuring rigorous limiting, and in designing a discriminator capable of handling signal ratios close to unity.

This article is not concerned with the means by which these requirements can be met. In the next section, however, we shall describe a signal generator designed to simulate the aerial signal resulting from multipath transmission. With this signal generator it is possible to study the effect of the above-mentioned measures to reduce distortion, even in places where there are no suitable reflecting obstacles in the neighbourhood.

### A signal generator for simulating multipath transmission effects in FM reception

The investigators cited in footnotes [1]) and [3]) studied these effects experimentally as well as theoretically. They conducted the radio-frequency FM signal, produced in a generator, along two paths to the receiver under investigation: along a short direct path and along a path having an appreciable delay time. For this second path Corrington used a coaxial cable which had to be more than 3 km long to give a transit time of 16 μsec — corresponding to a detour in the "ether" of less than 5 km. Arguimbau and Granlund, by piezo-electric means, first converted the radio-frequency signal into an ultrasonic vibration; they passed this through a mercury column, at the end of which they converted it back again into an electrical signal. Neither of these methods is convenient if the apparatus is required to be compact and portable. For these reasons we have adopted a different approach to the problem.

Two radio-frequency signals have to be generated, namely a "direct" signal of instantaneous frequency $F_a$:

$$F_a = F_0 + \Delta f \sin \Omega t ,$$

and a "reflected" signal of instantaneous frequency $F_b$:

$$F_b = F_0 + \Delta f \sin \Omega(t - \tau) .$$

Here $F_0$ is the centre frequency, and there is only one audio frequency ($= \Omega/2\pi$). Introducing an angle $a$ which satisfies

$$a = |\Omega t - 2\pi n| < 2\pi ,$$

where $n$ is an integer, we can write the formula for $F_b$ as

$$F_b = F_0 + \Delta f \sin (\Omega t - a) .$$

The two radio-frequency signals can thus in principle be obtained by using the same audio signal to frequency-modulate two signal generators having the same centre frequency $F_0$, the "delayed" signal being made to lag the "direct" signal by a phase angle $a$. The latter is produced by a phase shifter which will presently be discussed.

In actual FM reception the transit-time difference $\tau$ gives rise not only to a phase difference $a$ between the modulations, but also to a phase difference $\omega_0 \tau$ between the radio-frequency signals; see eq. (4). In order to simulate the latter phase difference, use might be made of a second phase shifter, now for high frequencies. We have not done this, however, since the ouly effect of the phase angle $\omega_0 \tau$ consists in a displacement of the peaks in fig. 7 over less than the width of one peak. This displacement is of no importance to the study of multipath distortion.

With two separate oscillators it is not possible to satisfy the condition that the two modulated signals shall have the same centre frequency. One common oscillator must therefore be used. Frequency modulation, however, can only be introduced in the oscillator itself; modulating this (in frequency) by two different audio signals would be no use whatsoever, for it would not yield the required two radio-frequency signals each modulated by one of the audio signals.

A system free from this limitation is phase modulation, for here the modulation can be applied *after* the oscillator. Phase modulation was therefore the system we decided to adopt. A phase-modulated signal, however, differs from a frequency-modulated signal in that the frequency deviation is proportional not only to the amplitude of the audio signal but also to the audio frequency. In order to make the result of phase modulation identical with that of frequency modulation, the audio signal is passed through an integrating network; this delivers an output voltage which is inversely proportional to the audio frequency.

### *Block diagram*

The block diagram of the signal generator is shown in *fig. 9*. The oscillator $O_1$, whose frequency is controlled by a quartz crystal, gives an output having a constant frequency of 3 Mc/s. In the phase modulators $PM_a$ and $PM_b$ the oscillator output is modulated in phase by an audio signal which is applied to the modulators through the above-mentioned integrating networks, $I_a$ and $I_b$. Two buffers $B_{a1}$ and $B_{b1}$, each consisting of a simple pentode stage, prevent feedback from the phase modulators to the oscillator.

Fig. 9 refers to the case of a *sinusoidal* audio signal. This is delivered by the signal generator $G$, and a phase shifter $P$ preceding the integrator $I_b$ provides the variable phase angle $a$. We shall return in a moment to the circuit arrangements to permit modulation by music and speech.

Following the normal practice in FM transmitters, the maximum frequency deviation is brought to the required 75 kc/s by frequency multiplication. This is done in the stages $Mu_{a1}$, $Mu_{a2}$, $Mu_{b1}$ and $Mu_{b2}$.

For this purpose the total multiplication factor needed is more than 400. A factor of about 30, however, is sufficient to raise the frequency of $O_1$ (3 Mc/s) to a value within the FM broadcast band (87.5-100 Mc/s). For this reason 18-fold multiplication is applied in $Mu_{a1}$ and $Mu_{b1}$, after which, by mixing with an auxiliary frequency of 50 Mc/s, the centre frequency is reduced from 54 to 4 Mc/s, the 18-fold frequency deviation being retained. In $Mu_{a2}$ and $Mu_{b2}$ a 24-fold multiplication is then applied, which raises the centre frequency from 4 to 96 Mc/s and brings the total multiplication of the deviation to $18 \times 24 = 432$.

The auxiliary frequency of 50 Mc/s is generated by the crystal oscillator $O_2$, and mixing is done in the stages $M_a$ and $M_b$. The buffers (pentode stages) $B_{a2}$ and $B_{b2}$ prevent undesired coupling between the channels via $O_2$.

The mixing process would not be necessary if the frequency of $O_1$ were low enough (e.g. 220 kc/s) to allow the same high multiplication factor used for the frequency deviation to be applied for the central frequency. A frequency as low as this for $O_1$, however, would have entailed considerable difficulties with the bandwidth.

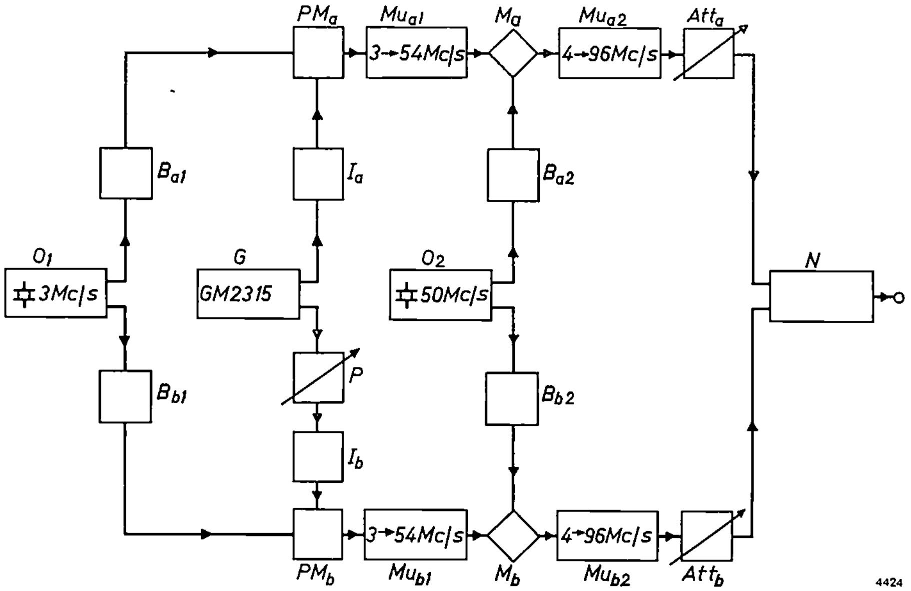


Fig. 9. Block diagram of apparatus for simulating FM multipath reception. Subscripts *a* relate to the "direct" signal, subscripts *b* to the "reflected" signal.
$O_1$ 3 Mc/s crystal oscillator. $B_{a1}$, $B_{b1}$ buffer stages. $G$ signal generator (type GM 2315). $P$ phase shifter. $I_a$, $I_b$ integrating networks. $PM_a$, $PM_b$ phase modulators. $Mu_{a1}$, $Mu_{b1}$ frequency multipliers (18×). $O_2$ 50 Mc/s crystal oscillator. $B_{a2}$, $B_{b2}$ buffer stages. $M_a$, $M_b$ mixing stages. $Mu_{a2}$, $Mu_{b2}$ frequency multipliers (24×). $Att_a$, $Att_b$ continuously variable "ladder" attenuators. $N$ matching network. The aerial terminals of the receiver are connected to the output of $N$.

At the outputs of the multipliers $Mu_{a2}$ and $Mu_{b2}$ a continuously variable attenuator is connected ($Att_a$ and $Att_b$, respectively); this is a so-called "ladder" attenuator, which has the property that the impedance remains constant (here 50 ohms), whatever the attenuation. Through coaxial cables, whose characteristic impedance is likewise 50 ohms, both signals are conducted to a matching network $N$, which terminates the cables with 50 ohms and shows the same impedance at its output. Here the terminals of the receiver are connected.

The circuit shown in *fig. 10* makes it possible to shift the sinusoidal modulation of the "reflected" signal by a variable phase angle $\alpha$ without changing the amplitude of the audio signal. When the resistance $R$ is raised from zero to $R_{max}$, the phase difference $\alpha$ between the output and input voltages increases from zero to almost $\pi$, provided that $\Omega CR_{max}$ is large compared with unity and the output terminals are not loaded. In order to vary $\alpha$ from zero to almost $2\pi$, two of these circuits are connected in cascade. The phase modulators will be dealt with in the next section.

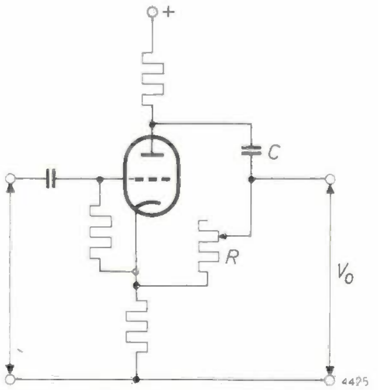


Fig. 10. One of the two cascade stages forming the phase shifter. When the resistance $R$ is varied from zero to $R_{max}$, the phase angle $\alpha$ between output voltage $V_o$ and input voltage $V_i$ increases from zero to 2 $\tan^{-1} \Omega CR_{max}$, that is, to nearly $\pi$ if $\Omega CR_{max}$ is large compared with unity. Since there are two stages in cascade, the total phase shift runs from zero to nearly $2\pi$. The magnitude of $V_o$ is not thereby affected.

A photograph of the equipment is shown in *fig. 11*. Two oscillograms obtained with it can be seen in *fig. 12*. The oscillogram in fig. 12*a* (analogous to fig. 7) relates to a receiver with a good AM limiter; that of fig. 12*b* refers to a receiver with poor AM limitation. In the second case the interference was much more audible than in the first.

### *The phase modulators*

For phase modulation a network is needed where the change of the phase difference between the output and input voltage is proportional to the instantaneous amplitude of the audio signal. A network with this property is a bandpass filter consisting of two coupled $LC$ circuits in which the capacitances $C$ are voltage-dependent. When the capacitances are varied, the tuning of the bandpass filter changes accordingly, and so therefore does the phase $\Theta$ of the output voltage. *Fig. 13* shows the variation of $\Theta$ as a function of $\beta Q$ for the case of a critically-coupled bandpass filter of identical primary and secondary $Q$. Here $\beta$ is the relative detuning, $\approx 2(f_{res} - f)/f$ (where $f$ is the signal frequency and $f_{res}$ is the resonant frequency of the filter). As can be seen, the response characteristic is virtually straight from $\beta Q = -1.7$ to $+1.7$ ($\Theta$ from $-100$ to $+100°$). If $Q$ has the easily achievable value of 85, for example, then $\beta$ can remain small. In this case, then, the phase modulation is practically linear if $\beta$ does not exceed the value $1.7/85 = 0.02$. For such slight detuning, $\beta$ is approximately equal to $\Delta C/C$, so that the capacitance variation $\Delta C$ may amount to a maximum of about 2% of $C$.

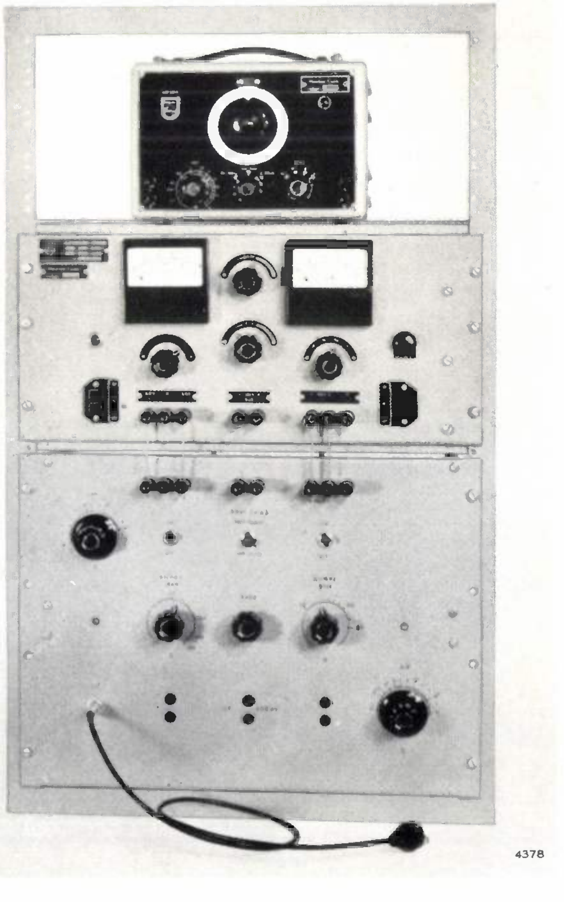


Fig. 11. The complete simulator. Above, the signal generator type GM 2315 (which can be replaced by a delay device with magnetic recording: see fig. 15). The panel below it is the power pack, and the bottom panel is the actual apparatus.

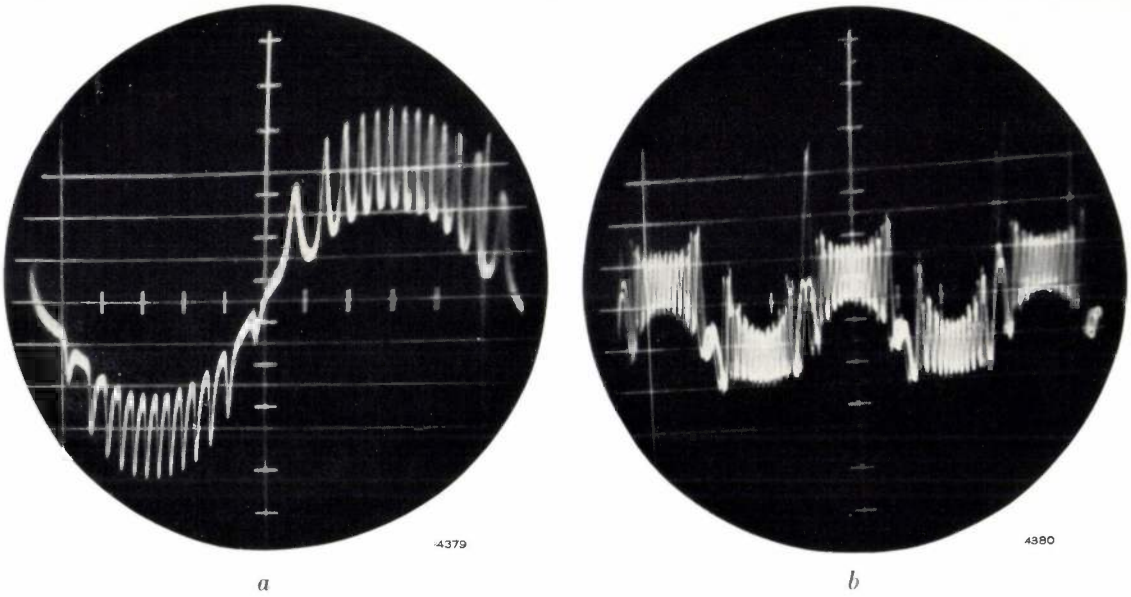


Fig. 12. Oscillograms of the output voltage from the discriminator in an FM receiver connected to the simulator described. The modulation was sinusoidal. *a*) Receiver with satisfactory AM limiting (cf. fig. 7). *b*) Receiver with inadequate AM limiting.

Semiconductor diodes in the cut-off state behave as voltage-dependent capacitances. The capacitance $C$ depends in the following way on the applied reverse voltage $V$:

$$C = \frac{K_1}{\sqrt{-K_2-V}},$$

where $K_1$ and $K_2$ are diode constants ($K_2$ is negative). *Fig. 14* shows the variation of $C$ with voltage $V$ measured on a silicon diode.

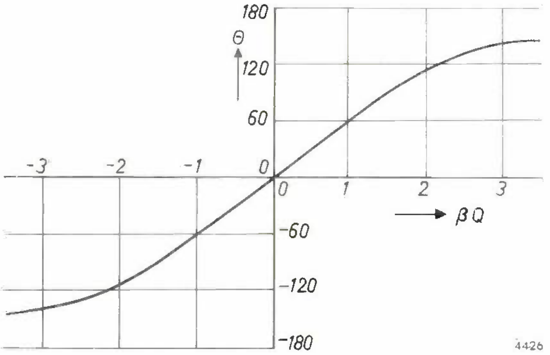


Fig. 13. Phase variation $\Theta$ of the output voltage from a critically-coupled bandpass filter having an identical primary and secondary $Q$ factor, as a function of $Q$ times the relative detuning $\beta$.

If a small alternating voltage is superposed on the direct voltage $V$, the change of $C$ is roughly proportional to the alternating voltage. This principle is applied in the phase modulators $PM_a$ and $PM_b$ (fig. 9).

*Modulation by music or speech*

In order to make the distorted output signal from the receiver visible in an oscillogram, the obvious method is to modulate with a sinusoidal audio signal; this case is represented in the block diagram in fig. 9.

After the signal generator had been thus designed, however, the need arose for some means of judging by ear the quality of received music or speech when the distortion described is present. For this purpose the apparatus was extended with a device for modulating by music or speech. The signal generator and the phase shifter are then put out of action — the latter because the phase angle $\alpha$ it delivers is not proportional to the audio frequency, which means

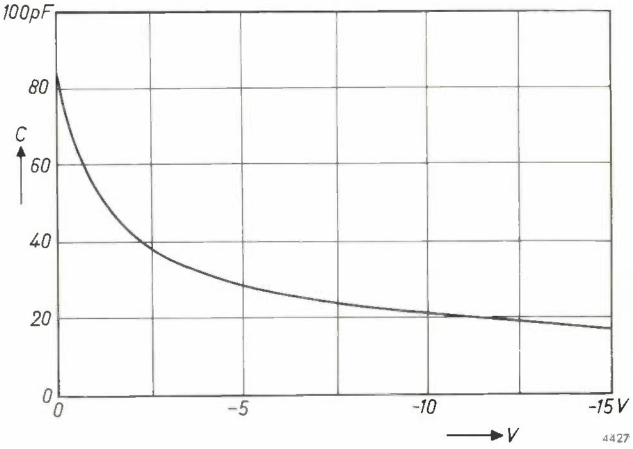


Fig. 14. Measured capacitance $C$ of a silicon diode as a function of the DC reverse voltage $V$.

that the time delay $\tau$ would be dependent on the audio frequency. To obtain a variable and frequency-independent time delay, we record the speech or music on a magnetic tape. When the apparatus is working, the tape passes over two playback heads spaced a distance $s$ apart. This distance, which corresponds to the desired time delay $\tau$, must be variable (at normal tape speed) from 0 to about 0.2 mm; $s$ is therefore always much smaller than the dimensions of the playback heads, and this calls for a special design.

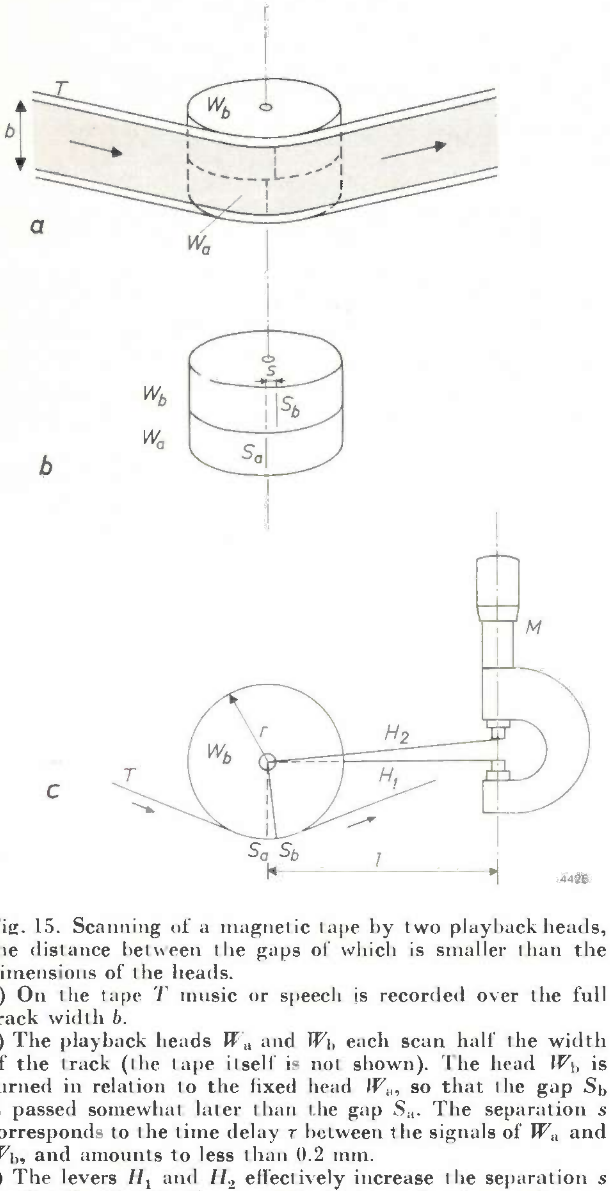


Fig. 15. Scanning of a magnetic tape by two playback heads, the distance between the gaps of which is smaller than the dimensions of the heads.
*a*) On the tape $T$ music or speech is recorded over the full track width $b$.
*b*) The playback heads $W_a$ and $W_b$ each scan half the width of the track (the tape itself is not shown). The head $W_b$ is turned in relation to the fixed head $W_a$, so that the gap $S_b$ is passed somewhat later than the gap $S_a$. The separation $s$ corresponds to the time delay $\tau$ between the signals of $W_a$ and $W_b$, and amounts to less than 0.2 mm.
*c*) The levers $H_1$ and $H_2$ effectively increase the separation $s$ by the (fixed) ratio $l/r$ to a value that can be measured accurately by the micrometer $M$.

A design which satisfactorily meets the requirements is illustrated schematically in *fig. 15*. The audio signal is recorded over almost the entire width of the tape, and the two playback heads each scan only one half of the width. The heads are mounted one above the other; the lower head is stationary, and the upper head is rotatable in relation to the other about a common axis. The distance $s$ between the gaps of the heads varies with the angle of rotation. This distance is increased in a fixed ratio by levers $H_1$ and $H_2$, and can be read from the micrometer $M$.

At a tape speed of 19 cm/sec, a time delay up to 1 millisecond can be achieved in this way, which simulates a difference of 300 km in the length of the radio transmission paths. This is in fact more than in the cases ever encountered in practice. The distance $s$ can be adjusted to an accuracy of 2 µ, which corresponds to a path-length difference of about 3 km. Small differences in path-length can thus equally well be simulated.

**Summary.** In mountainous regions, two or more FM transmission paths of different length may exist between the transmitting and receiving aerials, as a result of reflections from mountain ridges. This can cause distortion in reception, particularly if the received signals are of roughly the same strength. The discriminator then receives the resultant of the signals, and the vector representing this resultant exhibits an irregular angular velocity, which is not directly related to the modulation of the transmitter. Furthermore, the resultant is subject to amplitude modulation, which also causes distortion. Means of improving reception are the use of a sharply directional aerial, increasing the bandwidth of the limiter and discriminator, rigorous limiting, and the use of a discriminator capable of handling a signal ratio close to unity.

In order to simulate the effects in the laboratory, irrespective of terrain or conditions of reception, an apparatus has been designed which delivers two RF signals, one delayed and the other not, which are modulated in frequency by an audio signal (sine wave; music or speech). The time delay is continuously variable, and simulates a maximum path difference of 300 km.

# CIRCULAR OPTICAL ABSORPTION WEDGES

535.345.62

The level of illumination in nature increases from about 0.5 lux at dusk to 50 000 lux in bright sunlight, i.e. by a factor of $10^5$. For outdoor work a television camera should preferably be fitted with a camera tube sensitive enough to respond to the lower of these two levels. In stronger illumination the iris diaphragm in the camera lens must then be stopped down considerably, as the camera tube has a latitude of only about a factor 10 in the average intensity of illumination it can handle. If the aperthat the time delay $\tau$ would be dependent on the audio frequency. To obtain a variable and frequency-independent time delay, we record the speech or music on a magnetic tape. When the apparatus is working, the tape passes over two playback heads spaced a distance $s$ apart. This distance, which corresponds to the desired time delay $\tau$, must be variable (at normal tape speed) from 0 to about 0.2 mm; $s$ is therefore always much smaller than the dimensions of the playback heads, and this calls for a special design.

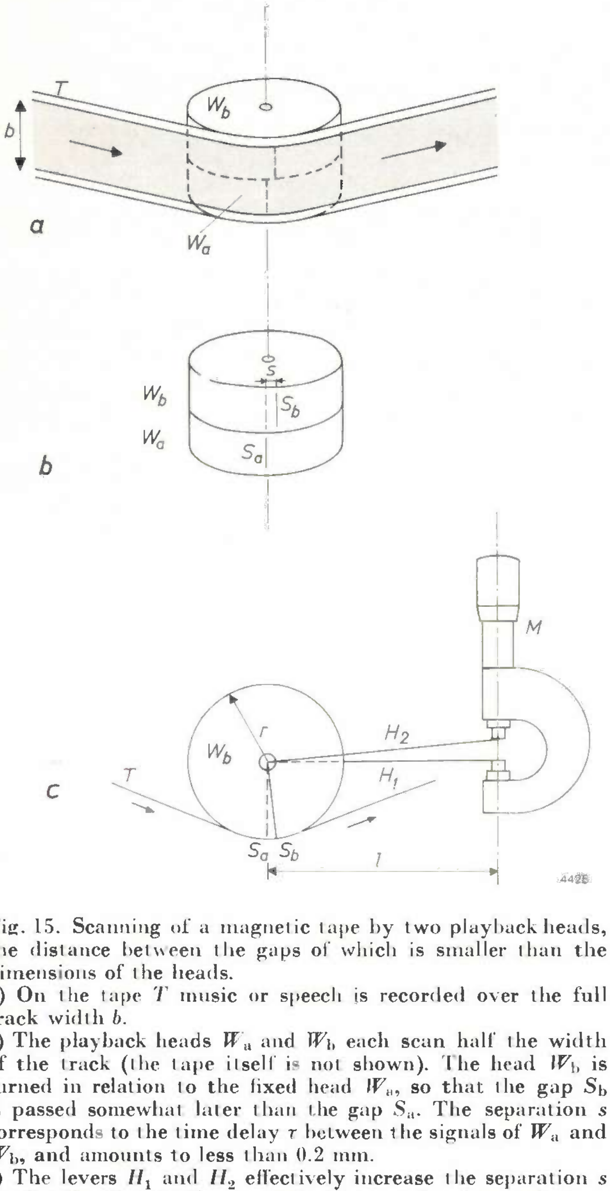


Fig. 15. Scanning of a magnetic tape by two playback heads, the distance between the gaps of which is smaller than the dimensions of the heads.
*a*) On the tape $T$ music or speech is recorded over the full track width $b$.
*b*) The playback heads $W_a$ and $W_b$ each scan half the width of the track (the tape itself is not shown). The head $W_b$ is turned in relation to the fixed head $W_a$, so that the gap $S_b$ is passed somewhat later than the gap $S_a$. The separation $s$ corresponds to the time delay $\tau$ between the signals of $W_a$ and $W_b$, and amounts to less than 0.2 mm.
*c*) The levers $H_1$ and $H_2$ effectively increase the separation $s$ by the (fixed) ratio $l/r$ to a value that can be measured accurately by the micrometer $M$.

A design which satisfactorily meets the requirements is illustrated schematically in *fig. 15*. The audio signal is recorded over almost the entire width of the tape, and the two playback heads each scan only one half of the width. The heads are mounted one above the other; the lower head is stationary, and the upper head is rotatable in relation to the other about a common axis. The distance $s$ between the gaps of the heads varies with the angle of rotation. This distance is increased in a fixed ratio by levers $H_1$ and $H_2$, and can be read from the micrometer $M$.

At a tape speed of 19 cm/sec, a time delay up to 1 millisecond can be achieved in this way, which simulates a difference of 300 km in the length of the radio transmission paths. This is in fact more than in the cases ever encountered in practice. The distance $s$ can be adjusted to an accuracy of 2 μ, which corresponds to a path-length difference of about 3 km. Small differences in path-length can thus equally well be simulated.

**Summary.** In mountainous regions, two or more FM transmission paths of different length may exist between the transmitting and receiving aerials, as a result of reflections from mountain ridges. This can cause distortion in reception, particularly if the received signals are of roughly the same strength. The discriminator then receives the resultant of the signals, and the vector representing this resultant exhibits an irregular angular velocity, which is not directly related to the modulation of the transmitter. Furthermore, the resultant is subject to amplitude modulation, which also causes distortion. Means of improving reception are the use of a sharply directional aerial, increasing the bandwidth of the limiter and discriminator, rigorous limiting, and the use of a discriminator capable of handling a signal ratio close to unity.

In order to simulate the effects in the laboratory, irrespective of terrain or conditions of reception, an apparatus has been designed which delivers two RF signals, one delayed and the other not, which are modulated in frequency by an audio signal (sine wave; music or speech). The time delay is continuously variable, and simulates a maximum path difference of 300 km.

# CIRCULAR OPTICAL ABSORPTION WEDGES

**535.345.62**

The level of illumination in nature increases from about 0.5 lux at dusk to 50 000 lux in bright sunlight, i.e. by a factor of $10^5$. For outdoor work a television camera should preferably be fitted with a camera tube sensitive enough to respond to the lower of these two levels. In stronger illumination the iris diaphragm in the camera lens must then be stopped down considerably, as the camera tube has a latitude of only about a factor 10 in the average intensity of illumination it can handle. If the aperture can be varied with the diaphragm between, say, *f*.2 and *f*.20, this gives an attenuation factor of 100 at the most. In order to use the camera at the higher of the illumination levels mentioned, additional attenuation by a factor of 100 is therefore needed.

This attenuation can be achieved using a neutral absorption filter. However, as every amateur photographer knows, the iris diaphragm of a camera serves not only to regulate the light entering the lens but also to regulate the depth of focus. If an extra attenuating element is to be introduced in the camera, it too should be variable (preferably continuously), so that the depth of focus can be independently selected in as wide a range of illumination levels as possible. This also applies where automatic mechanisms are concerned. A continuously variable attenuating element can be made by designing the absorption filter in the form of a movable density wedge.

We have devised a method of making absorption wedges in the form of a circular disk, the density varying with azimuth. This enables the light transmission to be varied by rotating the disk. It is not practicable to make such a circular wedge by grinding absorbent glass to a continuously tapering thickness — the method often adopted for straight wedges. Our method is therefore to effect the absorption through a non-scattering layer of material vapour-deposited on a glass disk, the thickness of the layer being given the azimuthal density variation required (if necessary non-linear). The set-up used for this purpose is illustrated in *fig. 1*.

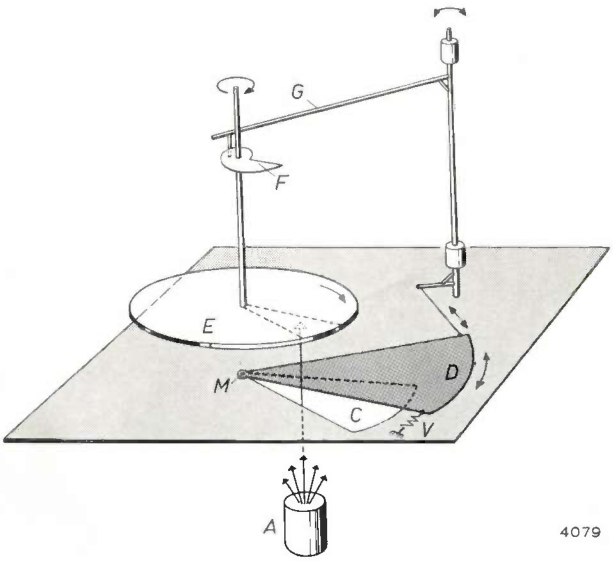


Fig. 1.

The source *A* of the material to be vapour-deposited is situated below a fixed plate in which an aperture *C* is cut in the form of a sector. The width of the aperture can be varied by means of a plate *D* which pivots about the point *M*. Mounted above this assembly is the round glass plate *E* on which the layer is to be deposited and which is rotated at a uniform speed about a vertical spindle through *M*. Fixed to this spindle is a cam *F* which, in turning, displaces a lever *G* and thereby moves the cover plate *D* against the action of a spring *V*. Thus, in every azimuthal position of the glass plate a sector of a certain size is exposed to the vaporized material; the density law of the circular wedge is thus governed by the shape of the cam. The deposition of the layer can be continued over any arbitrary number of complete revolutions of the glass plate, so that — provided the source *A* operates constantly — any desired maximum attenuation can be achieved.

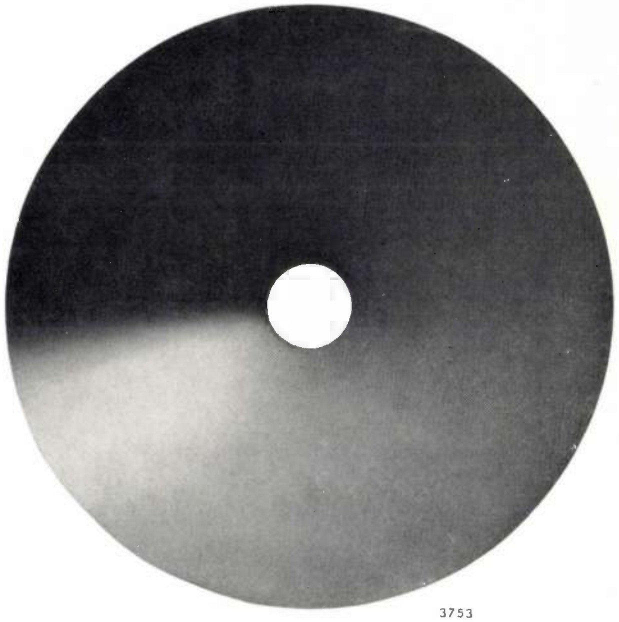


Fig. 2.

Suitable materials for vapour-deposition are metals or mixtures of metals with SiO; various organic dyes may also be used.

*Fig. 2* shows a photograph, taken in transmitted light, of a circular absorption wedge made by the method described. Since the deepest black and the brightest white obtainable with conventional reproduction techniques have a brightness ratio of no better than 15 to 20, it is not possible to do full justice here to the total range of transmission (factor 100) covered by this wedge.

J. van der WAL.

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

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution.

**2825:** H. Bremmer: Some theoretical investigations on fading phenomena (Statistical methods in radio wave propagation, Proc. Symp. Univ. Calif., Los Angeles, June 1958, edited by W. C. Hoffman, pp. 37-39, Pergamon, London 1960).

The author first discusses a statistically-fluctuating signal $h(t) = A(t) \cos \{\omega_0 t + \varphi(t)\}$, where $A$ and $\varphi$ are slowly varying functions of time $t$. Under very general assumptions it possesses the property that $N_A$, the average number of times the amplitude passes through its median value per unit time interval, is about three times greater than $N_\varphi$, the average number of crossings of the phase through any special value. If such a fluctuating signal is superposed on a much larger constant signal of fixed amplitude, frequency and phase, it is shown that $N_A \approx N_\varphi$. The author considers the application of these results to the fading of radio signals due to tropospheric turbulence effects.

**2826:** Th. G. Schut and W. J. Oosterkamp: Die Anwendung elektronischer Gedächtnisse in der Radiologie (Elektron. Rdsch. **14**, 19-20, 1960, No. 1). (The application of electronics to radiology; in German.)

The information presented in a fluoroscopic image can be retained by the eye for only about 0.1 sec. Full observation of such an image, however, calls for a much longer time, e.g. 10 sec. If it is possible to store the momentary image in some form of "memory" and subsequently make it visible for a sufficient length of time, the X-ray dose to the patient can be considerably reduced. The radiograph is one such memory, but has the drawback of not being immediately available. The authors discuss other methods that overcome this drawback, in particular the recording of X-ray images on a magnetic wheel store. See also Philips tech. Rev. **22**, 1-10, 1960/61 (No. 1).

**2827:** L. A. Æ. Sluyterman and J. M. Kwestroo-Van den Bosch: Sulphation of insulin and electrophoresis of the products obtained (Biochim. biophys. Acta **38**, 102-113, 1960, No. 1).

In connexion with investigations concerning the chemical modification of proteins, the $SO_3$ complexes of a few tertiary amines were tested for their ability to introduce $SO_3$ groups into insulin. Pyridinium sulphonic acid was found to be the most suitable one. By variation of the reaction conditions, insulin preparations of various sulphate content were prepared and subjected to paper electrophoresis at $p$H 1.7. A total number of 13 well defined, approximately equidistant bands could be observed, corresponding to insulin molecules carrying different electrical charges and covering a range from +6 units (native insulin) to −6 units (completely sulphated insulin). The biological activity of the preparations decreased with increasing sulphate content.

**2828:** J. S. C. Wessels: Photoreduction of 2,4-dinitrophenol by chloroplasts (Biochim. biophys. Acta **38**, 195-196, 1960, No. 1).

The author had formerly found (see Abstract No. **2776**) that 2,4-dinitrophenol (DNP) is able to catalyse the synthesis of adenosine triphosphate (ATP) by spinach chloroplasts. He now reports that illuminated chloroplasts of spinach are capable of reducing DNP to 2-amino-4-nitrophenol, and that thc latter compound can serve as a co-factor of photosynthetic phosphorylation.

**2829:** L. A. Æ. Sluyterman: The effect of oxygen upon the micro-determination of histidine with the aid of the Pauly reaction (Biochim. biophys. Acta **38**, 218-221, 1960, No. 2).

The colour obtained upon the addition of diazosulphanilic acid to histidine in alkaline medium (Pauly reaction) is bleached rather suddenly after a certain lag. This lag is shorter the more oxygen is present in the alkaline reaction medium.

A method of determining histidine on a micro scale, consisting of an improved Pauly reaction after paper-chromatographic separation, is described in detail.