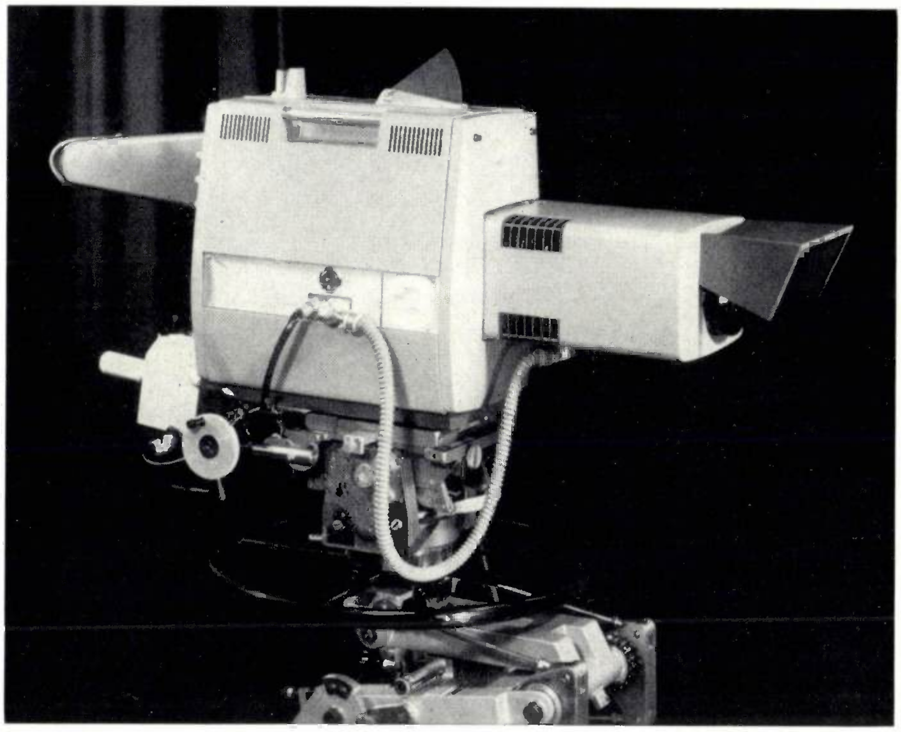
PHILIPS TECHNICAL REVIEW

VOLUME 27, 1966, No. 1



The "Plumbicon", the new type of camera tube developed at the Philips Research Laboratories in Eindhoven, has been greeted with considerable interest from the television world. The tube has all the advantages of the vidicon but none of its fundamental disadvantages, and is therefore suitable for applications previously reserved for image orthicons. This makes it desirable to compare the picture quality and other characteristic features of the various types of camera tube as objectively as possible, a comparison which necessitates careful regard to the differences in noise characteristics, setting of the operating point, the likelihood of over-exposure, etc. The "Plumbicon" gives a very good account of itself for black-and-white television, and for colour television its superiority is unquestionable.

The "Plumbicon" compared with other television camera tubes

A. G. van Doorn

621.397.331.222

The heart of any television camera is the camera tube. In this, an optical image is converted into a pattern of electrostatic charges, which is scanned by an electron beam to produce electrical signals. The camera tubes in most general use at the moment are image orthicons and Sb_2S_3 vidicons. These two types are based on different principles. The image orthicon makes use of photoemission and secondary emission to form a charge pattern; in the vidicon, the conversion is based on photoconduction. The characteristic features of these types of tube also differ considerably.

As may be seen in *fig. 1*, the vidicon is much smaller than the image orthicon. Other features in its favour are its simple adjustment, its stability — and its price.

Image orthicons are nevertheless widely used because Sb_2S_3 vidicons have a number of drawbacks that can seriously impair the quality of the picture. Their speed of response under normal lighting conditions is so slow that pictures of moving objects are blurred, and they may also have a fairly large dark current (signal arising when there is no light incident on the tube). Under certain conditions, and particularly at the normal illumination levels in the television studio, these undesirable effects can be so serious that

*) *A. G. van Doorn is with the Television Design Department, Philips Electro-Acoustics Division, Eindhoven.*

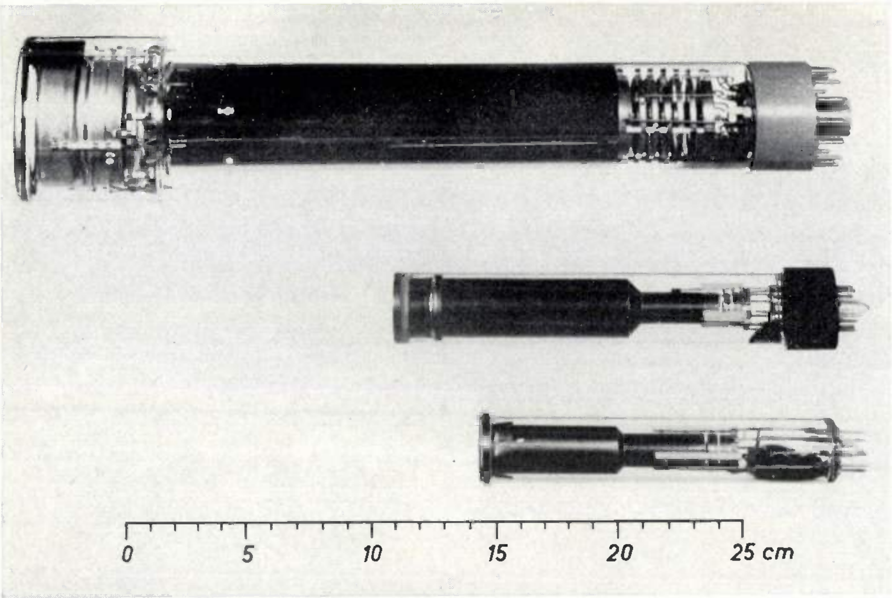


Fig. 1. Three types of camera tube. Top to bottom: a 3″ image orthicon, a "Plumbicon" and a 1″ vidicon. The type often used in studio cameras nowadays is the 4½″ image orthicon, which is considerably larger even than the 3″ type shown. There is also a ½″ model of the vidicon, specially designed for very small, compact cameras.

acceptable picture quality becomes unattainable.

Sb_2S_3 vidicons are, however, unexcelled for many applications outside the television studio. The undesirable effects mentioned above are much less of a nuisance at high levels of illumination, while small size, light weight, simple (possibly automatic) operation, robust construction, low price and long life are all of special importance in such applications.

The development of the "Plumbicon" makes yet another type of camera tube available. This tube, like the Sb_2S_3 vidicon, is based on the photoconduction principle, and is constructed in almost the same way (fig. 1). However, the use of a different photoconductive layer, with an appropriate method of operation, has almost completely removed the drawbacks of the Sb_2S_3 vidicon, i.e. the poor response speed and the high dark current, while retaining the useful features. The "Plumbicon", therefore, can be used for purposes for which so far only the image orthicon has been suitable. The picture quality of the "Plumbicon" compares very well with that of the image orthicon, so that it will also be possible to put to good use in the studio the practical advantages which made vidicon-type tubes so attractive in other fields.

While the "Plumbicon" for these reasons is a very attractive proposition for use in black-and-white television, its special features really come to the fore, as we shall see, in colour television. In fact, one of the main motives behind the development of the "Plumbicon" was the need for a camera tube that was more suitable for colour cameras than either the image orthicon or the vidicon.

Various characteristics of the "Plumbicon" and the explanation of its physical principles have already been dealt with in an earlier article in this journal [1]. We now wish to go more deeply into the use of this new tube and to compare it with image orthicons and vidicons already in use. A detailed description of the operation of these other tubes will not be given, as there is sufficient information on the subject in the literature [2][3]. The only factors to be mentioned will be those necessary for mutual comparison.

The objective comparison of two types of tube differing so widely in their operation and characteristic features as the "Plumbicon" and the image orthicon is a complicated undertaking. It is, in fact, virtually impossible to take into account and set off against each other all the factors that determine the quality of a television picture, such as resolution, tonal gradation, brightness range, signal-to-noise ratio, uniformity, etc., particularly as personal preference and habit also influence the eventual assessment of the picture. It is equally difficult to find a common denominator for features such as warming-up time, stability, life, sensitivity to interference, etc., that do not in the first instance affect the picture quality but are nevertheless important in operational practice.

We shall therefore restrict ourselves to those points which best lend themselves to an objective comparison, including the *illumination required for optimum picture quality*. For such a comparison the conditions of operation and the assumptions made must be rigidly specified. We shall first therefore discuss a few of the concepts involved, including the depth of focus, the light-transfer characteristic, and the signal-to-noise ratio.

Depth of focus

Fair comparison between different camera tubes must relate to conditions in which the cameras pick up exactly the same scene, as regards both foreground and background, and moreover with the same depth of focus in each case. It can easily be shown that the tubes will "see" the same scene only if lenses with focal length proportional to the linear dimensions of the image rectangle are used (*fig. 2*). The image rectangle of the image orthicon is 24 mm × 32 mm, while that of the "Plumbicon" is 12 mm × 16 mm and that of the ordinary Sb_2S_3 vidicon is about 9 mm × 12 mm.

If, moreover, the lenses have the same entrance pupil, the "equal depth of focus" condition is satisfied (*fig. 3*).

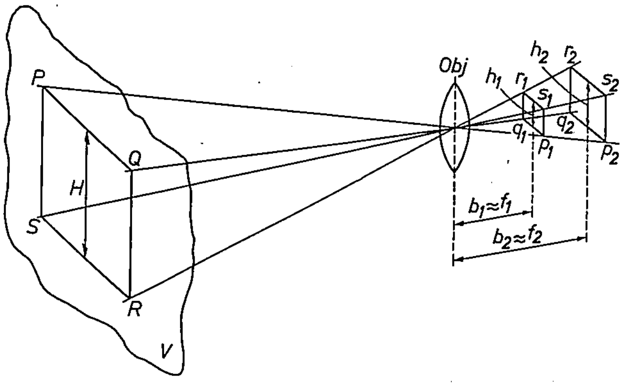


Fig. 2. A television camera with its lens at point *Obj* is assumed to take a scene in which the rectangle *PQRS* in the object plane *V* is sharply reproduced on the image rectangle $p_1q_1r_1s_1$ of the camera tube. If a second television camera having a camera tube with the larger image rectangle $p_2q_2r_2s_2$ is to take exactly the same scene (i.e. with the same foreground and background), its lens must be placed in position *Obj* and the rectangle *PQRS* must be sharply reproduced on the larger image rectangle. The figure shows that this is so if the image distances b_2 and b_1, and hence because of the great reduction the focal lengths f_2 and f_1, are in the same ratio as the heights of the image rectangles: $f_2/f_1 \approx b_2/b_1 = h_2/h_1$.

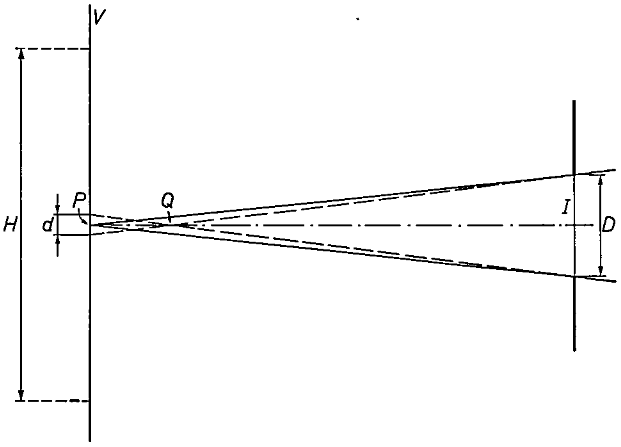


Fig. 3. The aperture *I*, of diameter *D*, represents the entrance pupil of the lens of a television camera. Point *P* of the scene in the object plane *V* is sharply reproduced on the photosensitive layer (not shown) of the camera tube. Point *Q* gives, on this layer, a spot (of diameter *d'*) that coincides with the image of a disc of diameter *d* around point *P*, since each ray through *Q* and *I* behaves like a ray originating from this disc. If *H* is the height of the rectangle in *V* that is reproduced on the photosensitive layer as the image rectangle of height *h*, then $d'/h = d/H$, i.e. d/H is a measure of the relative blurring of points *Q* belonging to the foreground or background. Hence, when one camera is replaced by the other, with *I* in the same position, the relative lack of definition, and thus the depth of focus, remains the same, provided that *D* is the same.

In practice, the focal length *f* and the relative aperture *D*/*f* of every lens are known. Here, *D* represents the diameter of the entrance pupil. The condition of equal entrance pupil means that the product of the focal length and relative aperture must also be equal. If a scene is reproduced with an image orthicon camera which has a 1: 5.6/50 mm lens, then with a "Plumbicon" the corresponding camera must have a 1: 2.8/25 mm lens, as the dimensions of the image rectangle of the "Plumbicon" are half those of the image orthicon. In comparing tube sensitivity, it is usual to assume that the lenses satisfy the above conditions. Questions such as the availability, quality and price of the lenses, and also the effect of the specified conditions on, for example, the dimensions of a lens, are not taken into consideration here [4].

Light-transfer characteristic

In comparing different types of camera tube, it is hardly realistic to think in terms of the usual concept of sensitivity and to attempt to quote one single numerical value. One of the essential factors in the sensitivity of a camera tube, however, is its "light-transfer characteristic". This gives the signal current I_s as a function of the illumination E_f at the photoconductive layer. *Fig. 4* shows examples of such characteristics for a "Plumbicon", an Sb_2S_3 vidicon and three different types of image orthicon. These characteristics are plotted with logarithmic scales for both coordinates, so that a straight line indicates that the signal current is proportional to a power of the illumination: $I_s \propto E_f^{\gamma}$, the constant exponent γ being given by the slope. If the slope is unity, which is the case over a wide range of the characteristics of the image orthicon and the "Plumbicon", I_s is proportional to E_f.

The horizontal axis in fig. 4 shows the reduced illumination E_f (in millilumens), i.e. the product of the illumination and the area of the image rectangle. This is the only way in which camera tubes with different image rectangle can be directly compared, for, provided that the illumination of the scene and the depth of focus are the same this quantity is independent of the size of the image rectangle.

[1] E. F. de Haan, A. van der Drift and P. P. M. Schampers, The "Plumbicon", a new television camera tube, Philips tech. Rev. **25**, 133-151, 1963/64. This article is further referred to as I.

[2] R. Theile and H. Fix, Eine vergleichende Betrachtung der heute verfügbaren Fernseh-Bildaufnahmeröhren, Radio-Mentor **25**, 448-452, 1959.

[3] J. W. Wentworth, Camera tubes for studio use, J. SMPTE **72**, 153-157, 1963.

[4] G. H. Cook; The performance of television camera lenses, J. SMPTE **69**, 406-410 (and 867), 1960.

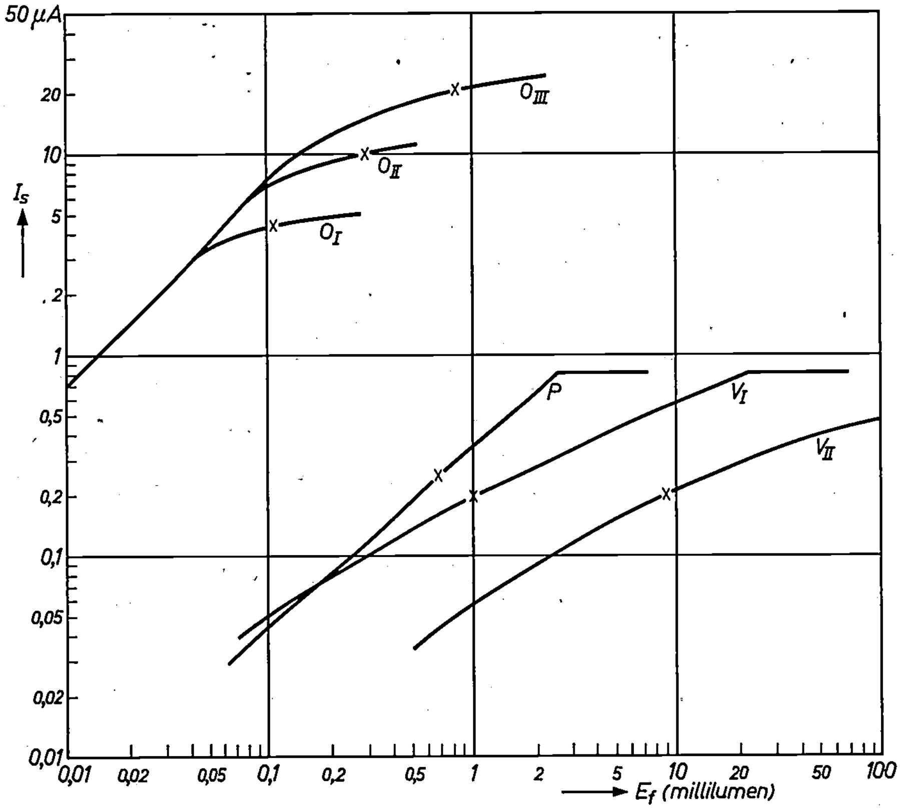


Fig. 4. Light-transfer characteristic of different types of camera tube. The signal current I_s is plotted as a function of the luminous flux on the photosensitive layer, the illumination being uniform. P applies to a "Plumbicon" of average sensitivity (350 μA/lm), O_I to a very sensitive 3″ image orthicon, O_{II} to a 3″ image orthicon with the knee at a higher illumination, and O_{III} to one of the types of 4½″ image orthicon now often used whenever the highest picture quality is required.

V_I applies to an Sb_2S_3 vidicon at a sensitive setting (signal plate voltage V_S = about 40 V, dark current $I_d = 0.02$ μA), and V_{II} to a similar vidicon with a lower signal plate voltage (V_S = about 15 V, $I_d = 0.005$ μA) and increased speed of response.

The operating point for optimum setting is given on each curve by a cross. The illumination required for this setting can be read on the abscissa.

If the tubes "see" exactly the same scene with lenses of the same entrance pupil diameter — the specified condition for fair comparison — the values of the illumination in lux at corresponding points on the photoconductive layer are inversely proportional to the areas of the image rectangles. The product mentioned above is therefore a quantity independent of this area.

The comparison of these characteristics should nevertheless be made with a certain degree of caution. With the image orthicon, for instance, the signal current is, indeed, much greater than in the "Plumbicon" and Sb_2S_3 vidicon, but this does not necessarily imply greater sensitivity. The secondary emission multiplier built into the image orthicon provides this large signal current. The real criterion for the sensitivity is the position of the optimum operating point and the related value of the illumination. It will be seen that one of the important factors in the choice of this operating point is the signal-to-noise ratio in the television signal.

The small crosses in fig. 4 show some of the normal operating points used in practice. Various designs of image orthicon are employed, which will be discussed in greater detail later. Characteristics have been plotted for two different settings of the Sb_2S_3 vidicon (from now on, for convenience, we shall refer to this simply as the vidicon). Only one characteristic has been given for the "Plumbicon", for a tube with an average sensitivity. The factors affecting the choice of the operating point will now be discussed in greater detail.

Signal-to-noise ratio

The signal-to-noise ratio of a television camera signal is generally taken to be the ratio between the useful signal S and the effective value N of the noise (the ratio is often expressed in dB). If S appears as a signal current, N is also considered as a noise current. If the noise in the signal is independent of S, the signal-to-noise ratio is proportional to S, so that a given situation may be characterized by the value S_W/N alone, where S_W is the signal provided by the brightest part of the scene.

N can however be a function of S. This is indeed so in the image orthicon, but with the "Plumbicon" N is even more strongly dependent on S because of the application of non-linear amplification for "gamma-correction". We shall, therefore, make use of an "equivalent noise" N_{eq}, which is independent of S and provides the same visual impairment as the actual noise. With the aid of statistical data and subjective assessments made by observers, Theile and Fix [5] have found that a useful expression for N_{eq} is:

$$N_{eq} = \frac{N_W + 3N_G + 2N_Z}{6} . \quad . \quad . \quad (1)$$

Here, N_W, N_G and N_Z represent the effective values of the noise in the signal corresponding to white, midgrey (the signal equal to 40% of that of white) and black (e.g. 2% of white).

Another factor important in the determination of the signal-to-noise ratio is the frequency spectrum of the noise, referred to as the "noise character". It is well known that high-frequency noise gives much less impairmant than low-frequency noise. Several workers [5][6] have attempted to analyse this visual difference and to reduce it to numerical or graphical terms that could be used as a measure of the impairment given by different noise frequencies. These differences in impairment become even more significant if pictures are to be compared whose noise spectra differ widely, or if noise measurements have to be made on signals having different noise frequency spectra.

In image orthicons, the output current of the multiplier is so large that the noise contribution in the signal current comes entirely from the tube itself. This noise is almost independent of the frequency, so that the noise spectrum is practically a horizontal line. This is called white or "flat" noise.

Both the vidicon and the "Plumbicon" deliver much smaller signal currents and the noise contribution originates almost entirely from the very sensitive input circuit of the electronic signal amplifier. Such input circuits are designed for a minimum of noise, and this has the result that the noise is no longer independent of frequency. In most cases, the effective noise current increases practically linearly with the frequency. Thus, we refer to noise with a "triangular" spectrum.

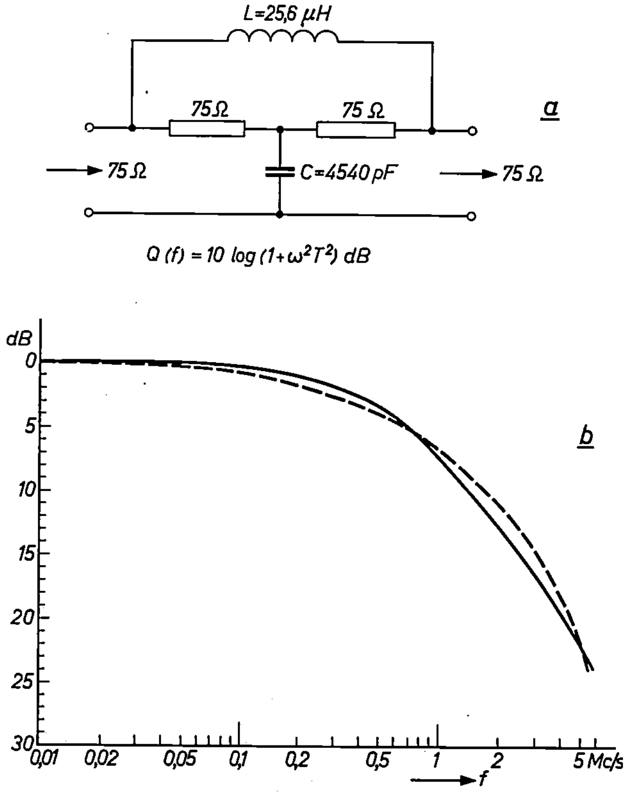


Fig. 5. *a*) A filter that can be used to take the visual impairment effect of noise into account in noise measurement. The component values have been selected so that $T = 0.33$ μs (CCIR recommendation).
b) Attenuation characteristic of this filter (solid line) compared with the statistically determined noise impairment curve (broken line).

In noise measurements, use is often made of a "noise-weighting filter", whose transmission-frequency curve approximates very closely to the average curve for the visual impairment effect of noise. *Fig. 5a* shows a diagram of such a filter and fig. 5*b* the resulting noise-weighting curve. To compare the image orthicon, with its flat noise spectrum, and the "Plumbicon", with its triangular noise spectrum, however, it is simpler to use one single numerical value for the relative impairment (noise weighting factor). This numerical value, which gives the difference between the impairment effect of flat noise and of noise with a triangular spectrum, may be set at 2.25 (7 dB) for a bandwidth of 5 Mc/s. Noise measurements made without a noise-weighting filter will give signal-to-noise ratio values for the "Plumbicon" and the vidicon which, for equal visibility of noise, are 7 dB lower than those for the image orthicon.

Characteristics and operating point of vidicons

The light-transfer characteristic of a camera tube is always affected, to a certain extent, by the various d.c. voltages on the electrodes of the tube. In the vidicon, the characteristic is particularly strongly affected by the signal plate voltage V_S. This is clearly shown in fig. 4, which shows two characteristics for different values of V_S. Because of this dependence, the characteristic can be adapted over a considerable range to the illumination. Fig. 4 shows the vidicon characteristics as slightly curved lines, of slope varying between about 0.8 and about 0.4. It is seen from the characteristics for low values of V_S that, as the illumination E_f increases, I_s will show slight saturation: the characteristics gradually become horizontal. The characteristics for high values of V_S show a sharp transition to a horizontal line. The actual transition

[5] R. Theile and H. Fix, Zur Definition des durch die statistischen Schwankungen bestimmten Störabstandes im Fernsehen, Archiv elektr. Übertr. **10**, 98-104, 1956. See also p. 64 of the second article mentioned in note [6].

[6] J. Müller and E. Demus, Ermittlung eines Rauschbewertungsfilters für das Fernsehen, Nachrichtentechn. Z. **12**, 181-186, 1959. See also H. Fix and A. Kaufmann, Die spektrale Zusammensetzung der statistischen Schwankungen bei zur Zeit üblichen Fernsehkameraanlagen, Rundfunktechn. Mitt. **4**, 60-65, 1960.

occurs at the point where the photocurrent has become so large that the electron current in the scanning beam (the beam current) is no longer capable of stabilizing the photosensitive layer at cathode potential. This is therefore an unstable state. If the beam current is increased, this transition point is shifted towards a higher value of I_s.

This point should never be exceeded, not only because with a horizontal characteristic all contrast is lost, but rather because unstabilized whites in the picture appear as washed-out areas with a tendency to spread, or "bloom", and which disappear only slowly if E_t drops below the transition point. With moving objects, very unnatural after-images and trailing effects are produced. These phenomena, which are bound to occur when scanning with "slow" electrons if the beam current is set at too low a value, arise similarly in the "Plumbicon", and even in the image orthicon a similar effect can come about due to a wrong setting.

In the vidicon, however, the signal plate voltage V_S also determines the dark current I_d, i.e. the signal current supplied by the tube when no light is falling on the photoconductive layer. As V_S increases, I_d increases more than linearly. For instance, when $V_S = 15$ V, the value of I_d is about 0.005 μA, while when $V_S = 40$ V, I_d can have a value between 0.02 and 0.1 μA. At the same time, the dark current depends to a great extent on the temperature. This dark current appears in the television signal as a non-uniform spurious signal, which gives a deterioration in picture quality. If the value of I_d, and hence the non-uniformity of the spurious signal over the image area, becomes too great, it may even become necessary to supply an electronic correction signal to compensate for the non-uniformity.

The speed of response depends almost entirely on E_t: it increases as E_t increases. Two factors limit the speed of response: the beam current lag, which is governed by the electron beam scanning mechanism and also depends on E_t (discharge lag), and the photocurrent lag. In fact, it takes a little time for the current through the photoconductive layer to respond when the value of E_t changes (see I, pp. 146 and 147).

The slow response at low values of E_t, which is a characteristic of the vidicon, is due entirely to the photoconductive lag. We shall see later that the lag in the "Plumbicon" is of a completely different nature. It does not depend on E_t, and is only slightly affected by V_S; above a certain value of V_S, the photoconductive lag is negligible in relation to the discharge lag.

The interdependence of various parameters makes it difficult to specify one suitable operating point for the vidicon. We cannot make a choice without a more thorough analysis of the effects of this on the picture quality. It is therefore simpler to select a few situations and examine the effects of the chosen setting on sensitivity, lag, picture quality, etc. What is evident from such an examination is the tremendous flexibility of the vidicon, since it is possible to deal with a considerable range of illuminations by adjusting V_S alone. This means that completely automatic cameras can quite easily be made.

A situation in which attempts are made to obtain the best picture quality in all respects merits special consideration.

Let us take as an example a setting for a useful signal current $I_s = 0.2$ μA, resulting in a very good signal-to-noise ratio; a dark current $I_d = 0.005$ μA, which is sufficiently low in relation to I_s; and a sufficiently high illumination level at the photoconductive layer to ensure fast response. Under these conditions, V_S will be between 10 and 20 V, depending on the temperature and the differences between individual vidicons, and the appropriate luminous flux on the layer will be about 10 millilumens, corresponding to an illumination of the layer of about 80 lux (8 foot-candles).

Another situation, where the picture quality may still be regarded as acceptable, but where slow response effects may become objectionable and the dark current I_d has increased to about 10% of I_s, is obtained with values of V_S of between 30 and 40 V. For $I_s = 0.2$ μA, and hence $I_d = 0.02$ μA, the luminous flux required is about 10 times lower and an illumination of 8 to 10 lux at the layer is required.

A very *sensitive* setting is obtained with values of V_S above 50 V. Here, I_d is 0.1-0.2 μA and is of the same order of magnitude as the useful signal current, while the slow response is extremely troublesome with moving pictures. Nevertheless, the picture quality at this setting, at which an illumination at the layer of 2 to 3 lux is adequate, is still quite acceptable for many industrial uses.

Characteristics and operating point of image orthicons

As we have already said, the behaviour of the image orthicon is quite different and is not easy to compare with that of the vidicon or "Plumbicon". The light-transfer characteristic is only slightly dependent on the voltage settings and is determined by other factors. As may be seen in fig. 4, it is possible to obtain different characteristics with the image orthicon, but these are obtained with *different types* of tube. The characteristic of one given type cannot easily be shifted by varying the settings without undesirable side-effects, and in practice different designs of the image orthicon are used for different purposes.

A closer look at the light-transfer characteristic shows that, over a certain illumination range, the signal current varies linearly with the illumination, and that above a certain value, the "knee", it quickly becomes saturated. Unlike that of the vidicon, the situation here is completely stable if the beam current is so set that brightnesses above the knee value can still be handled. This can be an important advantage in practice. The light control, i.e. the setting of the operating point, is not critical, as the knee acts as an automatic signal current limiter. It is therefore possible to handle a wide variety of scenes in which there are marked differences in the brightness of the white portions, including specular high-lights, without altering the setting. It is, moreover, possible to suit most of the scenes to be taken to the limited brightness range of the television system without difficulty. However, to obtain the best picture quality, the illumination of the sensitive layer, the photocathode, must be very carefully adjusted. This is done with the aid of adjustable diaphragms and grey filters. The optimum setting or operating point of an image orthicon is at or just over the knee. As a general rule the best setting is obtained when the illumination on the layer is such that white parts of a scene, in which a certain amount of detail must still be shown, are just above the knee of the given characteristic.

A further advantage, which we shall not discuss in detail here, but which is extremely important in setting up, is the increased resolution obtained when the knee is exceeded. It is rather difficult to define resolution here, but the same physical cause that gives rise to the knee in the characteristic of an image orthicon also gives rise to apparently enhanced contours at black-to-white transients. (The physical cause is in fact a redistribution of the electrons forming the charge pattern.)

Because of the fact, mentioned above, that the optimum setting is at or just above the knee, the signal-to-noise ratio of the most sensitive type of image orthicon (with the knee at the lowest illumination) is relatively poor. Fig. 4 in fact shows that, of the various designs, the most sensitive provides the weakest signal current. The signal-to-noise ratio — entirely determined by the camera tube, and proportional to the root of the signal current — is therefore at its worst for this type. The development of image orthicons over the past few years has been directed towards shifting the knee in order to obtain a better signal-to-noise ratio. The most significant improvement has been the 4½″ image orthicon, which has a higher knee than the 3″ type, a better signal-to-noise ratio, a better resolution and which can pick up a greater range of brightnesses without undesirable side-effects.

The dimensions 3″ and 4½″ in the designations of image orthicons refer to the diameter of the tube. The integral electron-optical image-forming elements make it possible for both types to provide optical images of the same size. However, the area of the target on which the charge pattern scanned by the electron beam is formed is about three times as great in the 4½″ model.

Apart from the size of this target, other factors also affect the position of the knee, so that various designs of both types are possible. Most of the characteristic features of the latest 3″ types approximate very closely to those of an average 4½″ type, and, although the signal-to-noise ratio and resolution are basically better in the 4½″ type, the differences have become so slight that both types are used side by side.

The response speed is limited only by the discharge lag. At the optimum setting, this lag is quite small; it becomes noticeable only under poor lighting conditions, where the operating point is a long way below the knee.

It is a disadvantage that a certain amount of burning-in with stationary pictures, rather than speed of response with moving pictures, sets a limitation. Whenever an orthicon camera has been directed on one scene for some time, say a few minutes, a burnt-in image is formed, which can render the camera useless for a while. If it is used by inexperienced persons, this can even cause irreparable damage to the tube.

Fig. 4 shows that a 4½″ image orthicon at the optimum setting is only half as sensitive as a comparable 3″ design. It has, however, an improved signal-to-noise ratio: where this is 36 dB in a 3″ tube, it may be as much as 39 dB in a 4½″ model.

Characteristic and operating point of the "Plumbicon"

At low signal plate voltages V_S, the light-transfer characteristic of the "Plumbicon" exhibits a shift similar to that of the vidicon. Because of the special properties of the photoconductive layer of the "Plumbicon", however, the signal current rapidly becomes saturated as V_S increases. Above $V_S = 30$ V, sensitivity virtually ceases to increase. In other words, for $V_S > 30$ V the characteristic can be considered as fixed. If the illumination E_f is increased at a constant V_S, then, as with the vidicons, we find a point at which there is a sharp transition to a horizontal line, owing to the beam current being no longer adequate for stabilizing at cathode potential. It is also of course possible, with the "Plumbicon", to shift this transition point by varying the beam current.

One of the main differences, however, between the "Plumbicon" and the vidicon is that the characteristic of the former in the stable part is a straight line of unit slope, which means that I_s is proportional to E_f. Only in this case can the sensitivity be indicated by one single numerical value, expressed in microampères per lumen. This sensitivity of the "Plumbicon" lies between 300 and 400 μA/lm.

The determination of the optimum setting will now be explained and a more quantitative treatment will follow. In the image orthicon, the operating point was entirely determined by the knee of the characteristic, which also determined signal-to-noise ratio, permissible brightness range and sensitivity. For the vidicon, the choice of the operating point was a compromise between speed of response, dark current and the available quantity of light. For the "Plumbicon", it will be seen that the optimum setting likewise is determined by the effect of various factors.

As already mentioned in the introduction and described in detail in I, the "Plumbicon" has two significant advantages over the vidicon. In the first place, the dark current is negligibly low and does not become very much greater at higher values of V_S. Secondly, the photoconductive lag is very small and independent of the illumination E_t. The photoconductive lag is to some extent dependent on V_S, but at the usual values of V_S it is negligible compared to the discharge lag.

Because of these advantages, finding the right setting is simple. If the operating point is made low, the illumination required is also low (i.e. the sensitivity is high) and the signal current is small, while the signal-to-noise ratio is poor. With a higher operating point, more light is needed, but the signal current becomes proportionately larger and the signal-to-noise ratio proportionately better.

The amount of noise in the signal has thus become practically the only decisive factor in the choice of the setting, and all factors affecting this noise therefore directly affect the choice of the operating point.

As already described for the vidicon, an inadequate beam current can cause a sharp transition in the characteristic to a horizontal line, giving rise to an unstable state of "over-exposure". With the linear characteristic of the "Plumbicon" the risk of over-exposure due to a sudden increase in the illumination level, or to specular highlights, is higher than with the vidicon characteristic, which exhibits saturation. This must be given careful consideration, which means that the risk of over-exposure must be reduced by stopping down, using grey filters or adjusting the level of illumination of the scene. With a camera equipped with a "Plumbicon", therefore, it is of the utmost importance to have a fast and accurate light-control. It is, however, extremely difficult, if not impossible, to prevent accidental specular highlights, and these too must not give rise to over-exposure. The knee in the characteristic must, therefore, be far enough above the operating point to leave an adequate safety margin. The knee can be shifted upwards by increasing the beam current, but this can prejudice resolution. In view of the safety margin required, therefore, it is recommended that the operating point should never be set higher than is necessary to obtain an acceptable signal-to-noise ratio.

Quantitative determination of the operating point of the "Plumbicon"

We shall now discuss in more detail the signal current necessary to obtain a given picture quality. The signal-to-noise ratio is our starting point since, as already mentioned, it is an important factor here. Let us assume that the signal-to-noise ratio of the output signal is to be 40 dB for flat noise — a value which can hardly be attained with the best image orthicons and which is generally considered fully adequate for studio use. As already stated (see p. 5), the most significant source of noise in both the vidicon and the "Plumbicon" is the input circuit of the signal amplifier, and this noise is frequency-dependent. If the signal-to-noise ratio is to be at least 40 dB for flat noise, then for triangular noise the ratio S/N between the peak signal current and the effective noise current must be at least 100/2.25 = 45.

In *fig. 6* the elements of a camera signal amplifier that affect the signal-to-noise ratio of the output signal can be seen. These are the preamplifier A_v, the aperture corrector A_a and the gamma corrector A_γ. The preamplifier is always necessary; whether aperture and gamma correction are also required depends to some extent on circumstances, as will be seen later.

The camera tube, whose inherent noise may be neglected, is connected to the preamplifier A_v by means of a signal resistor R_S. Analysis [7] shows that as R_S increases — if we assume the output impedance of the camera tube to be infinitely high — the signal-to-noise ratio at the output of A_v approaches a maximum given by:

$$\frac{S}{N} = 22\times10^8 \frac{I_s}{FG_t\sqrt{FR_t}}. \quad \ldots \quad (2)$$

Here, F is the bandwidth, R_t the "equivalent noise resistance" of the input circuit of the preamplifier, and C_t the total parasitic capacitance in parallel with R_S. The constant 22×10^8 is the value of $\sqrt{3/(16\pi^2kT)}$ at room temperature.

It may be seen from (2) that, at a given bandwidth, C_t and R_t set a limit to the value of S/N, and also that the parasitic capacitance is of prime importance. With practical values: $F = 5$ Mc/s, $C_t = 25$ pF and $R_t = 200\ \Omega$, equation (2) gives:

$$\frac{S}{N} \approx 5.6\times10^8 I_s,$$

so that a minimum signal current of 0.08 μA (for white) is needed in order to obtain at the output of A_V a signal-to-noise ratio of 40 dB (S/N at least $= 45$). In practice, because of the noise contribution from R_S, a slightly worse ratio must be taken into account, and the real value for I_s must be taken as 0.1 μA. (It should be pointed out that C_t and R_t may be made smaller by making use of more complex preamplifier circuits, using selected components and a very accurate setting[8].)

retain the original value of 40 dB assumed for the signal-to-noise ratio, and make full use of the increased resolution obtained by aperture correction, I_s must be increased to 0.13 μA.

Resolution has thus been improved at the expense of sensitivity. If enough light is available, this optimum correction is, of course, desirable, but if the lighting is poor (as for example during the evening), a compromise must be made between resolution and noise, by suitably setting the aperture correction.

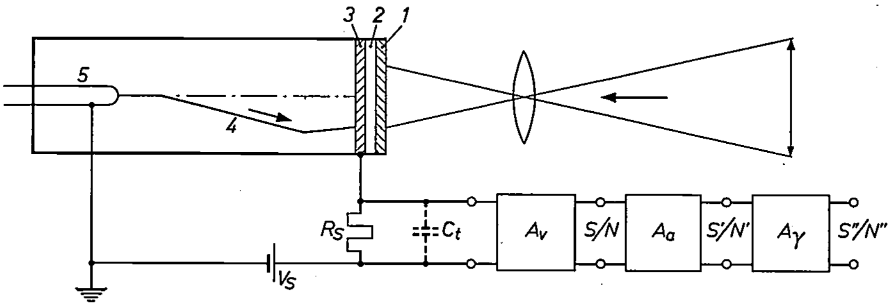


Fig. 6. Connection of a vidicon or "Plumbicon" to preamplifier A_V by means of signal resistor R_S. The equivalent noise resistance of the input stage of A_V and the parasitic capacitance C_t limit the signal-to-noise ratio at the output of A_V. *1* glass window, *2* transparent signal plate, *3* photosensitive layer, *4* electron beam and *5* cathode. V_S signal plate voltage (see I, fig. 2), A_a and A_γ: circuits for aperture and gamma correction.

The effect of aperture correction

The resolution of the picture is determined by a number of factors. It depends to a great extent on limitations in the camera tube itself (see I, p.144), but the optical section can also be significant. Zoom lenses, in particular, i.e. lenses with a continuously variable focal length, which are so often used nowadays, are certainly not ideal at large apertures. Spot correction, or, as it is more often called, aperture correction, which compensates for the effective size of the electron beam scanning spot exceeding the line width, can improve resolution. A simple electronic circuit giving amplification increasing with frequency, without the introduction of phase errors, can give optimum resolution at least in the horizontal direction[9]. Aperture correction gives no amplification at low frequencies, so that the amplitudes of the signal over large areas in the picture do not alter with the insertion of A_a, but the higher signal frequencies, and hence the higher frequencies in the noise spectrum, receive extra amplification.

At 5 Mc/s, the modulation depth of the average "Plumbicon" is only 50% of its value at 0.5 Mc/s, so that the amplification at 5 Mc/s must be at least twice as great (6 dB extra) for optimum correction. Further analysis shows that this causes the signal-to-noise ratio with a triangular noise spectrum to decrease by a factor of 1.3. If, therefore, we want to

The effect of gamma correction

The aperture correction mentioned above is also often used in vidicon cameras. Gamma correction, which we shall now discuss, is, however, only required for "Plumbicon" tubes.

This is because, in a picture tube, the luminance of the screen increases roughly with the square (at least) of the signal voltage applied to the control grid of the tube. To make the screen luminance proportional to the luminance of the original scene, the output signal of the camera must be approximately proportional to the root of the luminance of the scene. This is roughly what happens in the vidicon, as in the light-transfer characteristic $I_s \propto E_f^{0.6}$, while E_f is, of course, proportional to the luminance of the scene. The proportionality in the image orthicon is not so good, but the effect of the knee and the additional effect that the average scene content also has on the characteristic make gamma correction seldom necessary.

[7] V. K. Zworykin and G. A. Morton, Television, Wiley, New York 1954 (2nd edition), pp. 529-532.

[8] K. Sadashige, A study of noise in television camera preamplifiers, J. SMPTE 73, 202-206, 1964.

[9] Vertical aperture correction is possible, e.g. with the aid of delay lines. See: C. F. Brockelsby and J. S. Palfreeman, Ultrasonic delay lines and their applications to television, Philips tech. Rev. 25, 234-252, 1963/64.

In the "Plumbicon", however, where I_s is proportional to E_f and no limitation or saturation occurs right up to the point of instability, the gamma corrector A_γ is essential. This consists of a non-linear amplifier in which the output signal S'' and the input signal S' are connected by the relationship (*fig. 7*):

$$\frac{S''}{S_W''} = \left(\frac{S'}{S_W'}\right)^\gamma . \quad \ldots \ldots \quad (3)$$

The subscript W again indicates the signal for white. The amplification S''/S', dependent here on the signal level, is given by the slope of the chord to the origin (disregarding the constant factor S_W''/S_W'). The noise N' superimposed on S' is also amplified, but this amplification is determined by the slope of the tangent at S'. For a parabolic characteristic ($\gamma = 0.5$ in eq. 3), it can easily be verified that the signal amplification is always twice that of the noise.

In spite of this, gamma correction reduces the normalized signal-to-noise ratio. Although by definition we start from a noise level N' independent of the signal level S, the amplification of the noise does indeed depend upon the signal level. We shall, therefore, calculate the equivalent noise $N_{eq}'' = \frac{1}{6}(N_W'' + 3N_G'' + 2N_Z'')$ (see eq. 1). N_Z'' is determined by the practical consideration that, in a real gamma corrector, the tangent will never be vertical. If the characteristic has a slope 4 at the origin and gradually becomes a parabola, excellent tonal gradation is also obtained in the dark parts of the picture. In this case, $N_Z'' = 4N'$. Moreover, for white, $N_W'' = 0.5N'$ and in the grey (for which it is assumed that $S'' = 0.4S_W''$), $N_G'' = 1.25\,N'$, so that $N_{eq}'' \approx 2N'$.

Gamma correction, therefore, reduces the normalized signal-to-noise ratio by half, and, to maintain the desired signal-to-noise ratio of 40 dB, this must be compensated by a proportional increase in the signal current I_s.

Thus, if both aperture and gamma correction are used in a "Plumbicon" camera, the operating point must be set at $I_s \approx 0.25$ μA for the postulated signal-to-noise ratio. It should also be pointed out that scenes with limited brightness range may allow a different degree of gamma correction: an acceptable tonal gradation may for instance be obtained if the characteristic has a slope of 2 at the origin. In this case, $N_{eq}'' = 1.4N'$, and the operating point may be set at $I_s = 0.18$ μA.

If, on the other hand, the scene to be taken has a very wide brightness range and the most important parts of the picture are the dark ones, the operating point may have to be set at a higher I_s to prevent the noise in the dark parts of the picture from being too pronounced at maximum gamma correction. The extent to which, if at all, this can be done, will depend upon the safety margin required to allow for specular highlights which, as mentioned previously, can cause a transition of the characteristic into that for the unstable state if the beam current is inadequate.

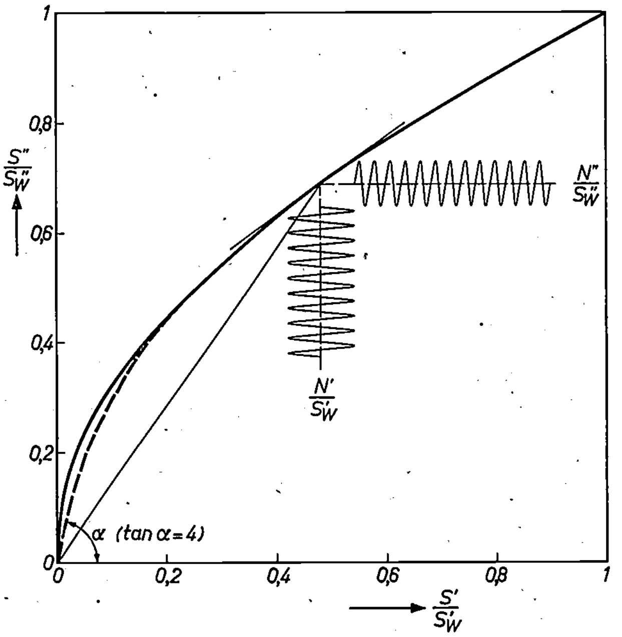


Fig. 7. Characteristic of the gamma corrector for a camera tube in which the signal current is proportional to illumination, as in the "Plumbicon". S' is the input signal, S'' the output signal, S_W' and S_W'' are the signal values for white. The slopes of the chord through the origin and of the tangent at S' determine the relative amplification of the signal S' and of the superimposed noise N' respectively. In theory, the characteristic must always be a curve with infinite slope at the origin (since γ is smaller than unity), but, in practice, a slope of 4 at the origin is sufficient for good tonal gradation (see the broken line).

Comparison between the "Plumbicon" and the image orthicon

As we have said in the introduction, it is impossible to make a purely objective comparison between different types of camera tube. It is in fact quite reasonable not to attempt any comparison between the vidicon and the image orthicon, since these two types of tube are used for quite different purposes. The "Plumbicon", however, as stated in the introduction, does not have the fundamental limitations of the vidicon, and can, therefore, be considered for the same applications as the image orthicon. Thus it is both possible and useful to make a comparison between the "Plumbicon" and the image orthicon.

We shall be restricting ourselves to some of the characteristic features of particular importance: sensitivity, resolution, tonal gradation and the signal-to-noise ratio.

Fig. 4 and the determination of the operating points have already shown that the sensitivity of the "Plumbicon" is roughly comparable with that of the best 3″ image orthicons. Inspection of a large variety of scenes bears this out in practice. The "Plumbicon" is, however, more sensitive than the 4½″ image orthicon, especially if one bears in mind that the sensitivity of image orthicons decreases as the number of operational hours increases, and sometimes even drops to half during their life.

The situation is different for resolution. Without special corrective measures the resolution of the "Plumbicon" is not as good as that of the 4½″ image orthicon. The image orthicon, properly set to the knee of the characteristic, benefits here from the apparent "contour-enhancing" effect mentioned earlier, which accentuates transitions from black to white. The perceptible differences between the two types of tube do, however, diminish considerably if aperture correction is used, which is fairly simple for the horizontal direction. The differences also depend upon the beam current setting. Even the colour temperature of the light has some effect: outdoor scenes with a high colour temperature taken with a "Plumbicon" often give surprisingly sharp pictures.

On the subject of tonal gradation, it must first of all be said that many factors are involved here. "Plumbicon" pictures, with the appropriate gamma correction, are characterized by faithful, "photographic" gradation within a wide (albeit still restricted) range of brightness. In practice, however, the gradation in pictures from the image orthicon is equally acceptable, even though, to prevent the signal-to-noise ratio from being reduced, it is not usual in this case to apply gamma correction. The effect of the knee in the light-transfer characteristic of the image orthicon, which has already been described, makes it possible to find an acceptable setting for a wide variety of scenes. This has quite justifiably been regarded as a considerable advantage. Nevertheless, a better result can generally be obtained with a "Plumbicon" camera equipped with a signal limiter and an adjustable gamma corrector.

It is only in scenes with a very wide brightness range or with specular highlights that situations occur which adversely affect the picture quality of the "Plumbicon". — either by loss of resolution, or by trailing effects in the non-stabilized highlights, or by increase of noise in the image, due to the fact that the operating point has to be reset to a lower signal current I_s.

The signal-to-noise ratio for a given type of image orthicon is determined by the tube itself and, when noise is measured without a noise-weighting filter, is 34-36 dB in a 3″ type and 37-39 dB in a 4½″ type, depending on setting and operational life. The illumination required for the "Plumbicon" in the discussions so far has been calculated on the basis of a signal-to-noise ratio of 40 dB. Here we have one of the most important differences between the image orthicon and the "Plumbicon". The signal-to-noise ratio of a given image orthicon can be improved only very slightly. If more light is available, the operating point can hardly be shifted since otherwise the knee in the characteristic would cause all gradation in the lighter parts of the image to disappear. In the "Plumbicon", on the other hand, the operating point may be displaced along the characteristic, so that if desired the signal-to-noise ratio can be still further improved.

This shows that, for applications where high picture quality is called for and the ordinary vidicon cannot be used because it is not sensitive enough, the "Plumbicon" is an interesting proposition, and can, in many instances, produce pictures of a quality comparable to that of the best image orthicon pictures.

Application in colour television cameras [10]

The requirements made of camera tubes for colour television cameras are more stringent and rather different from those for black-and-white cameras. The "Plumbicon" satisfies many of these requirements particularly well.

For a good colour balance and a correct colour gradation it is extremely important that the tube should have a well-defined light-transfer characteristic dependent on only a few parameters. The "Plumbicon", therefore, because its characteristic is independent of the illumination at the layer and of the average picture content, considerably simplifies the task of giving good, natural colour reproduction with a colour television camera under all kinds of conditions.

Spurious signals causing for instance an uneven and fluctuating black level are particularly undesirable in colour television because such signals are not the same for each of the three primary colours, and therefore give rise to colour impurities. The adjustment of a colour camera then becomes very complicated, especially if the spurious signals are dependent on temperature, focus, beam current setting or beam centreing. The low dark current of the "Plumbicon" has been found to be highly significant in this connection and has made any form of shading (dynamic correction of the black level) superfluous. This considerably widens the brightness range that can be covered while retaining good colour reproduction. A uniform sensitivity over the whole area of the picture is also of importance. Differences in local sensitivity between

[10] For general concepts, see: F. W. de Vrijer, Fundamentals of colour television, Philips tech. Rev. **19**, 86-97, 1957/58.

the various camera tubes produce colour shifts, which are particularly noticeable in scenes where the background is uniform.

The sensitivity of a colour television camera, which will, of course, be less than that of a black-and-white camera, is highly important, since an increase in the illumination in studios involves many other factors. These can include architectural and economic considerations (for instance, air-conditioning may be required). Sensitivity will be discussed separately in the next section, and it will be seen that in this respect the "Plumbicon" compares very favourably.

Moreover, it is obvious that the advantages of small size and smaller picture format, which vidicon-type camera tubes have compared with image orthicons, are more important considerations in the design of a colour television camera than in the design of black-and-white studio cameras.

Quite apart from all these factors, there is yet another, very important, and even perhaps, decisive requirement in the use of a particular type of camera tube in a colour television camera. The three primary colour pictures taken through three tubes must be geometrically exactly similar to one another. Primary colour pictures that do not precisely cover one another ("registration" errors) give rise to undesired coloured edges in the reproduced picture. This will affect not only the quality of the colour picture, but also the display of the picture in black-and-white, which should satisfy the colour monochrome compatibility requirements. It will be clear that "registration" errors will adversely affect the resolution of the picture received by a monochrome receiver, since, in the colour transmission system, the signal for black-and-white reception is built up from the three primary colour signals. With "Plumbicon" tubes no great difficulties are experienced in accurate registration of the three primary colour pictures.

All these factors make the use of the "Plumbicon" in colour cameras very attractive. The Philips Research Laboratories in Eindhoven had already developed colour television cameras incorporating the "Plumbicon" a few years ago. Very good results were obtained, and practical tests have shown that particularly for colour television, the "Plumbicon" has a great many advantages to offer.

Sensitivity of a colour television camera

In colour television, as in monochrome systems, the sensitivity is, to a large extent, determined by the signal-to-noise ratio. A difference is made in colour television, however, between luminance noise and chrominance noise. As the name implies, luminance noise relates purely to fluctuations in the total luminance, i.e. the sum of the luminances of the three primary colours in the reproduced picture, while chrominance noise relates to fluctuations in the particular colour.

Experience has shown that, taking the average of a large number of scenes, the determining factor is the luminance noise.

To compare this with the noise in monochrome television, let us consider a scene taken with a black-and-white camera and a colour camera, both fitted with "Plumbicon" tubes and identical lenses.

For a white area, the three camera tubes provide signals giving fractions f_R, f_G and f_B respectively of the signal from the tube in the monochrome camera. The actual values of f_R, f_G and f_B depend upon the colour-separating system, the spectral response of the "Plumbicon" and the colour temperature of the light sources used to illuminate the scene.

The spectral response curves for the three colour channels are given in *fig. 8*. This shows the ideal relative spectral response curves for optimum colour reproduction, together with those that have been obtained in practice. With the standard "Plumbicon" and a colour temperature of 3200 °K for the light source, f_R, f_G and f_B are 8, 20 and 12% respectively. For the present considerations gamma correction can be ignored, and as a starting point for the simple linear case we can assume a luminance signal E_y for which:

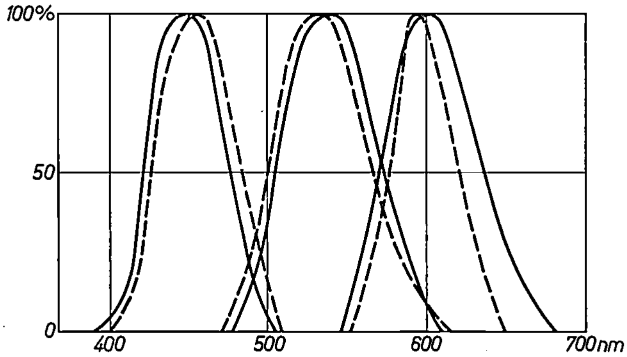


Fig. 8. Relative spectral response curves for the three colour channels in a colour television camera with "Plumbicon" tubes. The solid lines are the theoretical curves that, with certain assumptions, give the best possible reproduction of most colours. The broken lines are measured curves, which must be regarded as the best compromise. The cut-off at longer wavelengths in the red channel is caused by the limited red sensitivity of the "Plumbicon". This limitation could be removed by the use of a "Plumbicon" with an increased red sensitivity [11].

$$E_y = 0.3\,E_R + 0.6\,E_G + 0.1\,E_B.$$

The primary colour signals are amplified in such a way that for "white":

$$E_R = E_G = E_B = E_y.$$

The differences between f_R, f_G and f_B lead to different amplification factors, so that, for the effective value of the noise:

$$N_y = N \sqrt{\left(\frac{0.3}{f_R}\right)^2 + \left(\frac{0.6}{f_G}\right)^2 + \left(\frac{0.1}{f_B}\right)^2}.$$

N_y thus becomes about $5N$, or, in other words, to obtain the same signal-to-noise ratio, the illumination level must be 5 times higher for a colour camera than for a black-and-white camera.

Things are, of course, a little more complicated for colour scenes, but, as long as the colours are not highly saturated and, in particular, large, highly saturated red areas do not occur, it is confirmed in practice that this factor of 5 is about correct.

Because $f_G = 20\%$ and the illumination level has to be 5 times higher, the quite incidental result is that the "green" camera tube is set to the same operating point as that previously determined for a black-and-white camera. If there is ample light available, as is usual in outdoor work, then by making use of correction filters it is possible to set all three tubes to the optimum operating point. This requires 12 to 15 times as much light as for a black-and-white camera, but N_y now becomes smaller than N:

$$N_y = N \sqrt{0.3^2 + 0.6^2 + 0.1^2} = 0.68\,N.$$

Under these conditions, therefore, with settings at the optimum operating point, the signal-to-noise ratio in the luminance signal is 4 dB better than that attained with a black-and-white camera at the optimum setting.

Colour television camera with image orthicons

Apart from the fact that a colour television camera employing image orthicons is rather bulky because of the large size of the tubes, and requires a very complex optical system, there are two other important factors.

Firstly, the image orthicons cannot be fully operated into the knee of the light-transfer characteristic, as the characteristic at that point is not so well-defined. This adversely affects the signal-to-noise ratio and the resolution. Secondly, the image orthicons for the three channels must be accurately set to the same operating point to obtain good colour balance, and this means that the values for f_R, f_G and f_B are equalized. None of them therefore can be made larger than 8 to 10%, with the result that such a camera is 10 to 12 times less sensitive than a black-and-white camera using the same type of image orthicon. A less favourable setting is therefore required.

In addition, extremely stringent requirements are made on the precision of the tube construction and the stability of the control voltages in order to keep "registration" errors to a minimum. All these factors make it clear that such cameras in practice are not altogether an attractive proposition. Attempts have been made to overcome the difficulties by using colour cameras with *four* camera tubes, the fourth being used solely to provide the luminance signal [12].

A colour television camera with "Plumbicon" tubes

The title photograph shows a colour television camera using "Plumbicon" tubes now in production. The colour-separation system is made up of prisms (*fig. 9*), following the method described in an earlier article in this journal [13]. Such a colour-separation system has various optical advantages and allows the colour television camera to be very compact. The camera has a variable-focus lens, and the adjustments,

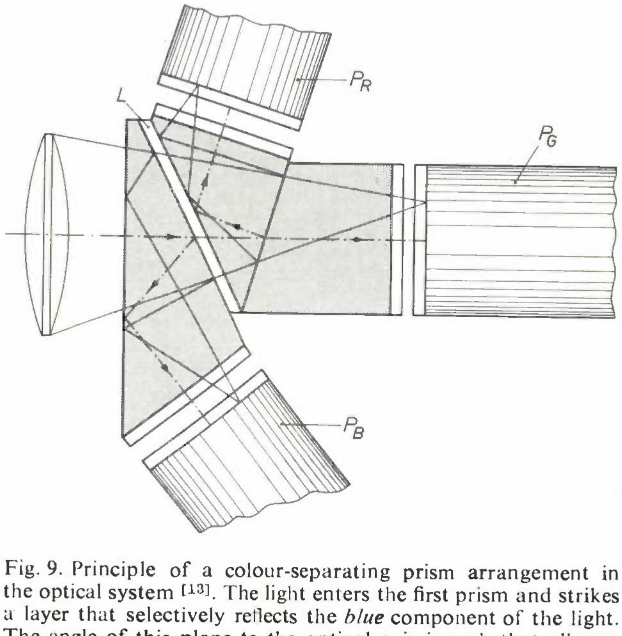


Fig. 9. Principle of a colour-separating prism arrangement in the optical system [13]. The light enters the first prism and strikes a layer that selectively reflects the *blue* component of the light. The angle of this plane to the optical axis is such that all rays reflected from it are totally reflected at the boundary of the prism. The blue component is thereby reflected downwards, towards the "blue" "Plumbicon" P_B. A second layer, which selectively reflects the *red* component of the light, is cemented between the second and third prism. By correct choice of the angles involved, the red light is reflected a second time by total reflection at a glass-air boundary between the first and second part of the prism (L is an air-gap), and thus reaches the "red" "Plumbicon" P_R. The *green* component of the light travels straight through and provides the image for the "green" "Plumbicon" P_G. Special trimming filters in each channel ensure that the spectral transmission curves are as close as possible to the ideal.

[11] E. F. de Haan, F. M. Klaassen and P. P. M. Schampers, An experimental "Plumbicon" camera tube with increased sensitivity to red light, Philips tech. Rev. 26, 49-51, 1965.

[12] W. Dillenburger, Gesichtspunkte für die Entwicklung einer Farbfernsehkamera, Radio-Mentor 30, 974-981, 1964.

[13] H. de Lang and G. Bouwhuis, Colour separation in colour-television cameras, Philips tech. Rev. 24, 263-271, 1962/63.

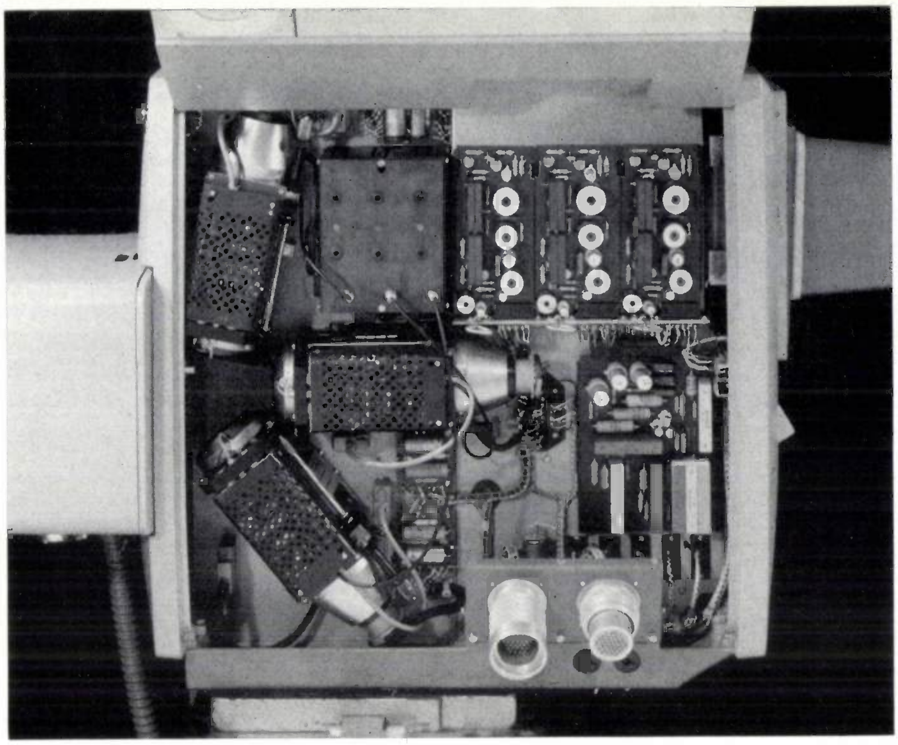


Fig. 10. A view inside a colour television camera equipped with "Plumbicon" tubes. Three identical, adjustable units are arranged around the colour-separating prism system. Each unit comprises a "Plumbicon" tube with its focusing and deflection coil and magnetic screening, as well as a signal amplifier. This side of the camera also contains the horizontal deflection circuits.

such as focusing, focal length setting and iris control, are effected by servo systems. The focal length may be varied between 18 and 180 mm, with a maximum relative aperture of 1 : 2.2. The lens is optically corrected for the prism system inserted in the path of the beam. It is even possible, with an alternative currently available lens, to vary the focal length from 50 to 600 mm.

The compact construction of this camera is also illustrated in *fig. 10.* The arrangement of the three "Plumbicon" tubes in a vertical plane is somewhat unusual; this is determined by the colour-separating system. With such an arrangement, magnetic screening is very important to prevent undesired magnetic fields from reaching the camera tubes and upsetting the relative geometry and the precise registration.

Some idea of the sensitivity of such a camera with "Plumbicon" tubes is readily obtained from the table below, which gives the conditions in a normal scene, the settings to be used for it and the signal-to-noise ratio attained:

incident light level	1500 lux (150 foot-candles)
reflection coefficient (of the white parts of the scene)	60%
colour temperature of the lighting	3200 °K
relative aperture	1 : 2.8
aperture correction	6 dB at 5 Mc/s
gamma correction	$\gamma = 0.5$ in a 1 : 40 range
normalized signal-to-noise ratio in the luminance signal	42 dB
signal-to-noise ratio (determined with a noise-weighting filter)	50 dB

At full lens aperture and with no spot correction, however, faithful colour pictures can still be made at an illumination of 200 to 250 lux with an acceptable signal-to-noise ratio in the luminance signal.

Summary. Because the "Plumbicon" does not have the drawbacks of the Sb_2S_3 vidicon, i.e. poor response speed and excessive dark current, it is suitable for studio use. The factors to be taken into account in comparing its picture quality with that of the image orthicons used up to now include resolution, tonal gradation, brightness range, signal-to-noise ratio, uniformity, etc. A number of conditions for proper comparison are discussed, including the required focal length and diameter of the lens, the selection of the operating point on the light-transfer characteristic, and the measurement of the noise with a noise-weighting filter. Closer examination of the light-transfer characteristic of the different types of tube shows that the signal-to-noise ratio of the "Plumbicon" can be improved by shifting the operating point upwards, but there must be a safety margin to safeguard against "over-exposure" from specular highlights. If a signal-to-noise ratio of 40 dB (for flat noise) is required — a value hardly achieved by the best image orthicons — then in view of aperture correction and the gamma correction desirable for the "Plumbicon", the operating point must be set at a signal current of 0.25 μA (and maintained there by controlling the light level). The sensitivity is then comparable to that of the best 3″ image orthicons. The 4½″ image orthicon is not so sensitive, but gives better resolution. Better tonal gradation can usually be obtained with a "Plumbicon"-equipped camera than with an image orthicon camera. The "Plumbicon" is particularly suitable for colour television cameras, not only because of its small physical size and small image size, but also because its characteristic is independent of the illumination at the layer and of the average picture content, and further because in this case it is not very difficult to prevent annoying "registration" errors in the three primary colour pictures. Its sensitivity is such that a colour camera using "Plumbicon" tubes can easily pick up a normal scene where the illumination is 1500 lux (150 foot-candles); the normalized signal-to-noise ratio in the luminance signal is then 42 dB. Acceptable pictures can still be obtained at 200 to 250 lux.

Printed in the Netherlands

An eddy-current coupling employed as a variable-speed drive

W. Bähler and W. van der Hoek

538.541:62-578.3

The idea of the eddy-current coupling is not new; the device was in fact used long ago for the transmission of high power. In recent years advancing mechanization has stimulated fresh interest in the eddy-current coupling, particularly in the homopolar type, in view of its suitability as a variable-speed drive for low power applications. Electronic control of the coupling offers many new possibilities in process control and other industrial control applications.

In 1825 Arago discovered that a rotating copper disc tends to communicate its motion to a magnetic needle suspended above it (*fig. 1*). Arago was unable to explain this phenomenon and called it "rotational magnetism". It was not until 1831, when Faraday discovered the phenomena of induction that it was understood that the field of the magnetic needle generated eddy-currents in the rotating disc. The forces produced between these electric currents and the magnetic field caused the needle to rotate with the disc.

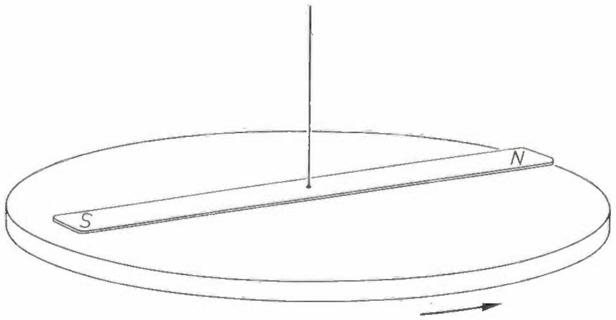


Fig. 1. Arago's experiment. A copper disc rotating in its own plane tends to communicate its motion to a magnetic needle suspended above it.

These eddy-currents, which always arise in a conductor subjected to a varying magnetic field, have been intensively studied since electric generators first came into use. Such studies were not only directed towards limiting the losses caused by the eddy-currents (for example in the iron core of a transformer), but also towards possible uses. The eddy-current coupling is one of the possible applications, and Arago's experiment can be regarded as its first demonstration.

Construction of an eddy-current coupling

Eddy-current couplings, as later developed in many different designs, always contain two members which can rotate freely with respect to one another (*fig. 2*). The magnetic field is generated by the inductor *1*. The second member, the eddy-current cylinder *2* (it may also be a disc) consists of conducting material in which the eddy-currents are induced.

Frequently the inductor is provided with one or more coils, excited by means of slip rings and wound in such a way that opposite poles are alternately produced on the circumference of the inductor. The coils are excited by d.c. current. When the inductor is driven the magnetic field rotates with it. The situation shows much resemblance to that found in a three-phase induction motor, where a rotating magnetic field is produced by electrical means. Just as in the three-phase induction motor, there must, to obtain a torque at the output shaft, be a certain slip between the rotating magnetic field and the eddy-current cylinder of the coupling.

The absence of mechanical contact between inductor and eddy-current cylinder makes the eddy-current coupling suitable for transmitting power under irregular or shock loading, such as found in rolling mills

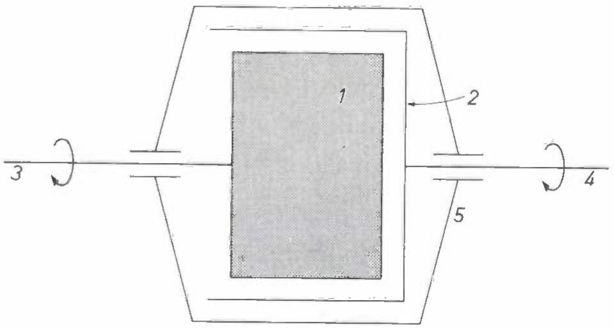


Fig. 2. Schematic representation of an eddy-current coupling. *1* inductor, *2* eddy-current cylinder. The input shaft *3* and the output shaft *4* are supported in bearings in the housing *5*.

Ir. W. Bähler is with Philips Research Laboratories, Eindhoven; Prof Ir. W. van der Hoek is with the Works Mechanization Department of Philips Radio, Gramophone and Television Division, Eindhoven, and an associate professor of Mechanical Engineering at the Technical University of Eindhoven.

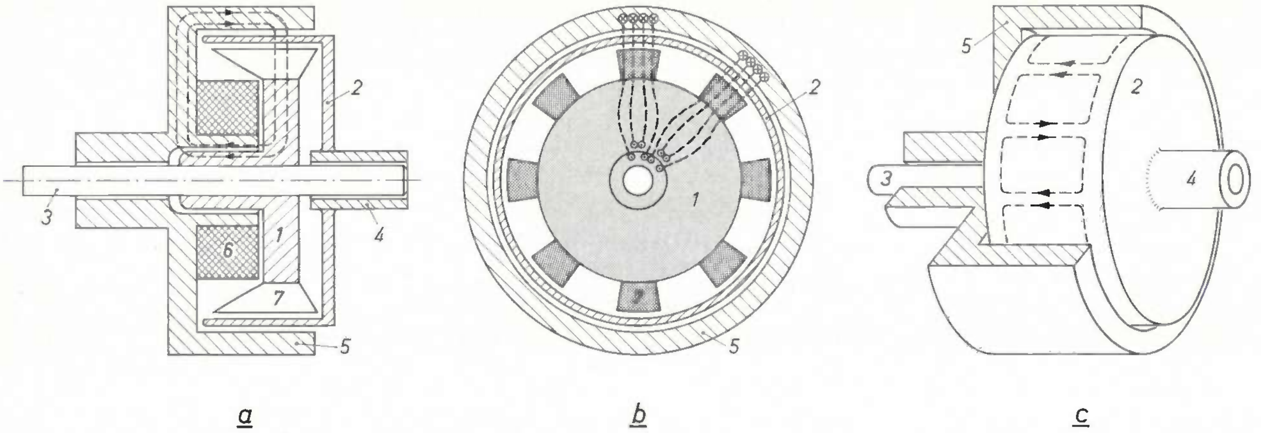


Fig. 3. Sketch of a homopolar eddy-current coupling. *1* inductor, here magnetized by a single stationary exciting coil *6*. *2* eddy-current cylinder. The input shaft *3* is supported in bearings in housing *5*, made, like the inductor, of a ferromagnetic material. The output shaft *4* is supported in bearings on the input shaft. *7* "teeth" of the inductor.
(*a*) and (*b*) show the lines of magnetic field. In (*c*) the eddy current loops are shown.

and cranes. Other advantages are that the coupling can be controlled by a small direct current and that the torque is entirely ripple-free, due to the absence of slots or bars in the eddy-current cylinder.

The eddy-current coupling is called heteropolar when the inductor has both north and south poles. The heteropolar type is widely used for the transmission of high power.

The operation of the coupling does not depend, however, on reversal of the polarity of the magnetic field. It is sufficient if the field varies between a minimum and a maximum value. A drawback of this construction (referred to as "homopolar"), is that for a given maximum amplitude of the magnetic field one can obtain only a quarter of the torque obtainable with a heteropolar type; set against this drawback there is the advantage of certain constructional simplifications.

Some homopolar types, for example, need only one central coil which does not rotate with the inductor, and this means that the exciting current can be supplied by very simple and reliable means. In such couplings the local variation of the intensity of the magnetic field is produced by designing the inductor in the form of a disc with "teeth", which rotates in the space between the coil and the cylinder (*fig. 3*).

The homopolar coupling with stationary exciting coil has proved particularly useful as a low-power variable-speed drive.

A simplified model

For any given application the eddy-current coupling must meet specific requirements. In order to produce a design that meets the requirements, the influence of certain factors on the characteristics of the coupling has to be considered. Typical factors are the relative angular velocity of the inductor in relation to the eddy-current cylinder, the dimensions, the magnitude of the exciting current and also the choice of material, they are dealt with here by considering a simplified model.

It is assumed that an infinitely large flat plate of thickness δ_1 is situated in a flat air-gap of height δ (*fig. 4*). The air-gap is bounded above and below by magnetic material which has zero electrical conductivity, $\sigma = 0$, and infinitely high magnetic permeability, $\mu = \infty$. The conductivity of the plate is $\sigma = \sigma_1$ and its magnetic permeability $\mu = \mu_0$, the same as that of free space.

At the plane $z = +\frac{1}{2}\delta$, which bounds the upper space, there is an infinitely thin layer containing a sinusoidal pattern of currents flowing in the positive and in the negative y direction. The current does not depend on y. The "pole pitch", i.e. the length of half a period of this sinusoidal pattern, is $\lambda/2$. The "current layer" moves with a velocity v in the positive x direction. The plate is stationary.

If we compare this model with the actual eddy-current coupling we see that the inductor is represented by the current layer, and the eddy-current cylinder by the plate. The velocity v of the current-layer is equivalent to the difference in speed between the inductor

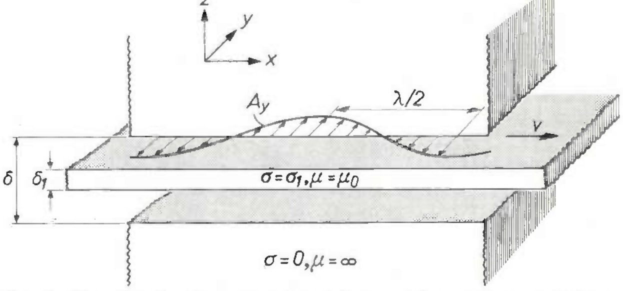


Fig. 4. Simplified representation of an eddy-current coupling, used in deriving the torque equations.

and the eddy-current cylinder of a real coupling.

Since the currents in the exciting current layer flow only in the positive and negative y direction, the magnetic field which they generate in the air-gap has no y component. Since, moreover, the x and z components of the magnetic field are independent of y, the electrical field in the air-gap contains a y component only.

On the basis of Maxwell's equations we can set down a differential equation for the electrical field strength in the air-gap, the "wave equation", which we have to solve for the steady state.

The A_y of the exciting current is given by the expression

$$A_y = A_{y0} \exp \left\{ \mathrm{j}\, \frac{2\pi}{\lambda} (vt - x) \right\}, \quad . \; . \; . \quad (1)$$

where A_{y0} is the amplitude of the current.

Because of the periodicity of the moving pattern of the exciting current the electrical field strength at any given point must be given in the steady state by a solution of the wave equation that has the same time dependence as the exciting current. For E_y, the y component of this electrical field, we therefore write:

$$E_y = E_{y0}(z) \exp \left\{ \mathrm{j} \frac{2\pi}{\lambda} (vt - x) \right\}, \quad . \; . \quad (2)$$

where $E_{y0}(z)$ is a complex variable which denotes both the amplitude and the phase shift of the electrical field with respect to the exciting current.

If, in the solution (2) of the wave equation, we take into account the boundary conditions relating to the model (fig. 4), we find that both in the eddy-current plate and in the space $E_{y0}(z)$ is approximately independent of z and equal to:

$$E_{y0}(z) = - \frac{\mu_0\, v\, A_{y0}}{v \sigma_1 \delta_1 \mu_0 - \mathrm{j} 2\pi\delta/\lambda}. \quad . \; . \quad (3)$$

We therefore have the relation between the electrical field strength in the plate and the exciting current. What we are particularly interested in, however, is the force exerted on the stationary plate by the moving magnetic field.

This force can very easily be found with the aid of theoretical considerations of the energy involved. The Joule heat, which is generated in the plate by the eddy-currents, averaged over a pole pitch, is equal to the work which the tangential force K, averaged over half a period, performs per unit area and per unit time. The Joule heat per unit volume element $\delta_1 \mathrm{d}x$ is $E_y{}^2 \sigma_1 \delta_1 \mathrm{d}x$, and therefore:

$$Kv = \frac{1}{\lambda/2} \int\limits_0^{\lambda/2} E_y{}^2 \sigma_1 \delta_1 \mathrm{d}x\,. \quad . \; . \; . \; . \quad (4)$$

Substitution of eq. (2) in eq. (4), together with eq. (3), yields after some manipulation:

$$K = \frac{\mu_0\, A_{y0}{}^2 \lambda}{4\pi\delta \left(\dfrac{v}{v_0} + \dfrac{v_0}{v} \right)}, \quad . \; . \; . \; . \; . \quad (5)$$

where:

$$v_0 = \frac{2\pi\delta}{\lambda \sigma_1 \delta_1 \mu_0}. \quad . \; . \; . \; . \; . \; . \quad (6)$$

For $v = v_0$ the force reaches a maximum equal to:

$$K_{\max} = \frac{\mu_0\, A_{y0}{}^2 \lambda}{8\pi\delta}. \quad . \; . \; . \; . \; . \quad (7)$$

Substitution of eq. (7) in eq. (5) finally yields:

$$\frac{K}{K_{\max}} = \frac{2}{\dfrac{v}{v_0} + \dfrac{v_0}{v}}. \quad . \; . \; . \; . \quad (8)$$

Equation (3) is arrived at as follows [1]. In our model a part of the air-gap is taken up by the eddy-current plate. The electrical field strength in the eddy-current plate is given by the three-dimensional wave equation:

$$\nabla^2 \mathbf{E} - \mu\sigma \frac{\partial \mathbf{E}}{\partial t} = 0\,, \quad . \; . \; . \; . \; . \; . \; . \quad (9)$$

and the electrical field strength in air is given by:

$$\nabla^2 \mathbf{E} = 0\,. \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (10)$$

Substitution of eq. (2) in (9) and (10) gives expressions for $E_{y0}(z)$ of the form:

$$E_{y0}(z) = P \exp \left\{ z \sqrt{\left(\frac{2\pi}{\lambda}\right)^2 + k^2} \right\} + \\ + Q \exp \left\{ - z \sqrt{\left(\frac{2\pi}{\lambda}\right)^2 + k^2} \right\}, \quad . \; . \quad (11)$$

where $k^2 = \mathrm{j}(2\pi/\lambda) v \mu\sigma$ holds for k. (In air $\sigma = 0$, which means that $k = 0$ in eq. 11.)

P and Q are integration constants which differ in each of the three layers of the air-gap — air, eddy-current plate, air. These integration constants can be found with the aid of the boundary conditions that apply at the surfaces of the three layers. In the first place, the magnetic permeability of the eddy-current plate is equal to that of the air-gap. This means that both at the upper and the lower surface of the eddy-current plate ($z = \pm\frac{1}{2}\delta_1$) the tangential components and also the normal components of the magnetic field, H_x and H_z, are equal on each side of the boundaries. Owing to the infinite magnetic permeability of the material outside the air-gap, the tangential component of the magnetic field at the lower boundary ($z = -\frac{1}{2}\delta$) is zero. Because of the presence of the current-layer this does not apply to the upper boundary of the air-gap ($z = +\frac{1}{2}\delta$). If we take the contour integral of the magnetic field in the plane of a flat loop perpendicular to the y axis and close to the boundary (Maxwell: $\oint H_s \mathrm{d}s = i_{\mathrm{true}}$), we find there $H_x = -A_y$.

[1] This calculation may be found in: R. Rüdenberg, Energie der Wirbelströme in elektrischen Bremsen und Dynamomaschinen, Enke, Stuttgart 1906.

Using the boundary conditions we find six equations from which the integration constants can be calculated. Before solving these equations, however, we replace the exponential powers which they contain by the first term of the series expansion. This is permissible (see eq. 11) when the height δ of the air-gap is small both in relation to the pole pitch $\lambda/2$ and in relation to the skin depth δ_{skin} [2] of the eddy-current plate.

Solution of the simplified equations results in eq. (3).

The torque-speed characteristic

The relation between K and v, as found for the simplified model, also applies in principle to the actual eddy-current coupling. Instead of the force and velocity we introduce two other quantities: the torque M exerted on the output shaft, and the relative speed n of the input shaft with respect to the output shaft.

Corresponding to eq. (8) we can write:

$$\frac{M}{M_{\max}} = \frac{2}{\dfrac{n}{n_0} + \dfrac{n_0}{n}}, \quad \ldots \quad (12)$$

and equations can be written down for $M_{\max}$ and n_0 corresponding to (6) and (7). The curve found when M is plotted as a function of n is referred to as the torque-speed characteristic (*fig. 5*).

Whether an eddy-current coupling is suitable for a particular application depends on the position and height of the maximum in the torque-speed characteristic. In designing an eddy-current coupling it is also possible, with the aid of the equations for $M_{\max}$ and n_0, to choose the dimensions, exciting current and the materials in such a way as to produce the required characteristic. For example, increasing the thickness of the plate or its conductivity gives a lower speed for maximum torque.

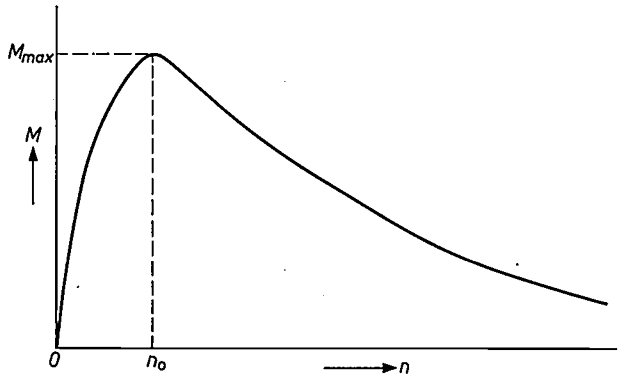


Fig. 5. Torque-speed characteristic of an eddy-current coupling. Maximum torque $M_{\max}$ is reached at a speed n_0.

It can also be seen that increasing the gap-width δ causes the maximum to shift towards a higher speed. When δ is increased, however, it is necessary at the same time to increase the exciting current in proportion to δ in order to keep the maximum torque $M_{\max}$ constant. Theoretically the required exciting current is at a minimum when the gap-width δ is equal to the plate thickness δ_1.

The homopolar type

From now on we shall be solely concerned with the homopolar version of the eddy-current coupling. Before applying to this version the results found so far, it is necessary to consider the effects of the differences between the model and the actual coupling.

In the model we first assumed a heteropolar version in which the field distribution was purely sinusoidal. As a first approximation we can treat the field distribution in the actual homopolar coupling as a sinusoidal field superimposed on a field which is constant around the whole circumference. The constant field component has no influence on the behaviour of the coupling as the potential difference that it produces between the two ends of the eddy-current cylinder is constant over the whole circumference, so that no currents can flow. In practice, however, the field differs from that of the model in a further respect: the alternating component, as measurements show, is not sinusoidal but has a shape between a sine wave and a square wave. We can think of the alternating field as the resultant of a sinusoidal field and higher harmonics, whose amplitudes can be calculated by means of numerical Fourier analysis. An investigation has shown that the influence of the higher harmonics on the shape of the torque-speed characteristic is not negligible. Compared with the characteristic for the purely sinusoidal field distribution, the maximum torque is larger and displaced towards a higher speed; and the initial slope of the curve becomes steeper.

Another difference concerns the gap-width δ. In the model this was assumed to be constant. In the version of the eddy-current coupling which we consider (see fig. 3) the inductor is in the form of a disc with teeth around the circumference, so that the gap-width varies. To obtain agreement between the torque speed characteristic of the coupling and the theoretical characteristic we have to introduce a fictitious gap-width δ', which is greater than the gap-width at the teeth.

In the model we assumed an eddy-current plate of infinite length in the y direction and a current layer independent of y. This meant that the eddy-currents flowed only in the y direction, that is to say perpendicular to the direction of motion. In actual eddy-current couplings, including the homopolar type, the finite dimensions of the eddy-current cylinder require that the eddy-currents at both ends of the fingers have to form closed loops (see fig. 3*c*). Because of this, extra resistance is presented to the eddy-currents, causing a further discrepancy between the actual torque-speed characteristic and the theoretical one. The discrepancy can be found by estimation and by measurements on actual eddy-current couplings [3].

A variable-speed drive

We have already mentioned that the homopolar version of the eddy-current coupling has proved particularly useful as a variable-speed drive for low-power transmission. This is due not only to the simple construction of the coupling, but more especially to the possibility of electronic control of the exciting current.

Fig. 6 shows schematically a speed-control system in which the eddy-current coupling is used as a variable speed drive. The input shaft of the eddy-current coupling K is driven by an electric motor M which runs at constant speed. The speed of the output shaft is measured by a tachometer which gives a voltage proportional to the speed. This voltage is compared with a reference voltage which corresponds to the desired speed and which can be adjusted with a potentiometer P. An amplifier A energizes the exciting coil of the eddy current coupling with a current proportional to the difference between the two voltages.

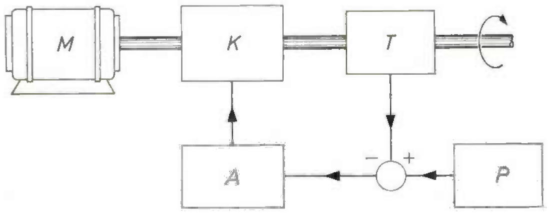


Fig. 6. Diagram of a speed-control system employing an eddy-current coupling K. The input shaft of the coupling is driven by a motor M at constant speed. The speed of the input shaft is measured by tachometer T, whose output voltage is compared with a reference voltage, adjustable by means of potentiometer P. Amplifier A energizes the exciting coil with a current proportional to the difference between the two voltages.

As is often the case in control systems the actual value of the controlled quantity — in our example the speed — has to be lower than the desired value before current can be supplied to the exciting coil. The difference between the two speeds decreases as the gain of the amplifier is increased. The magnitude of the gain (amplification factor) has however to be limited to a certain maximum depending on the dynamic characteristics of the coupling.

Many variable-speed drives of the type described above have been in use at Philips for some years now in a wide variety of production machines, including machines for coiling wire and tape after rolling, drawing or lacquering. Recently both the eddy-current coupling and the complete drive were put on the market [4]. *Fig. 7* shows an exploded view of a Philips eddy-current coupling. The device incorporates a tachometer, which indicates the speed of the output shaft.

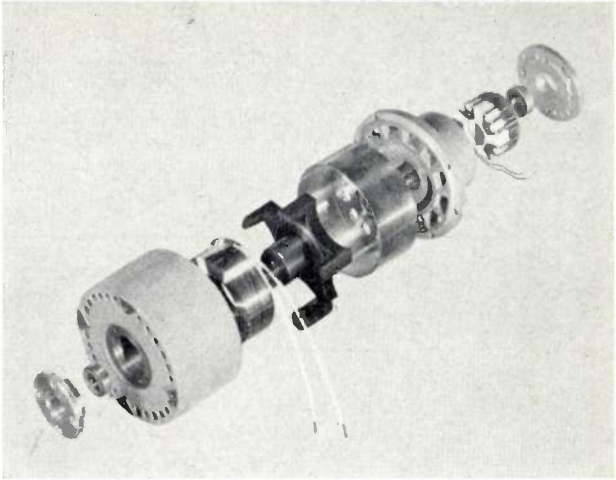


Fig. 7. Exploded view of a Philips eddy-current coupling. The large cylindrical member on the left is the stationary housing, made of ferromagnetic material. To the right can be seen in order the exciting coil; the inductor, mounted on the input shaft; the copper-eddy-current cylinder, mounted on the output shaft; the cap and the tachometer. At each end bearings are visible.

Torque-current characteristic of the variable-speed drive

For a variable-speed drive it is desirable to be able to transmit the maximum torque in the widest possible range of speeds between zero and the speed of the motor. We have stated earlier that the form of the torque-speed characteristic of the eddy-current coupling can be influenced by the appropriate choice of materials and dimensions. The choice is not, however, entirely free, because the maximum exciting current, the width of the air-gap and the thickness of the eddy-current cylinder wall are also governed by constructional and other conditions. Therefore, in a good design the maximum in the torque-speed characteristic does not always lie in the speed range employed, as might be supposed.

[2] The skin depth of the material of the eddy-current plate is given by: $\delta_{\text{skin}} = \sqrt{\lambda/\pi\mu_0\sigma v}$.
At this depth $|\mathbf{E}| = (1/e)|\mathbf{E}_{\text{surface}}|$.

[3] The discrepancy has been exactly calculated for a situation as in fig. 4, but with a sinusoidal field in both the x and the y direction. See the article quoted [1].

[4] The electronically controllededdy-current coupling was developed for use in Philips establishments by the Works Mechanization Department of the Radio, Gramophone and Television Division. The further development, production and marketing have been taken over by Philips Industrial Equipment Division.

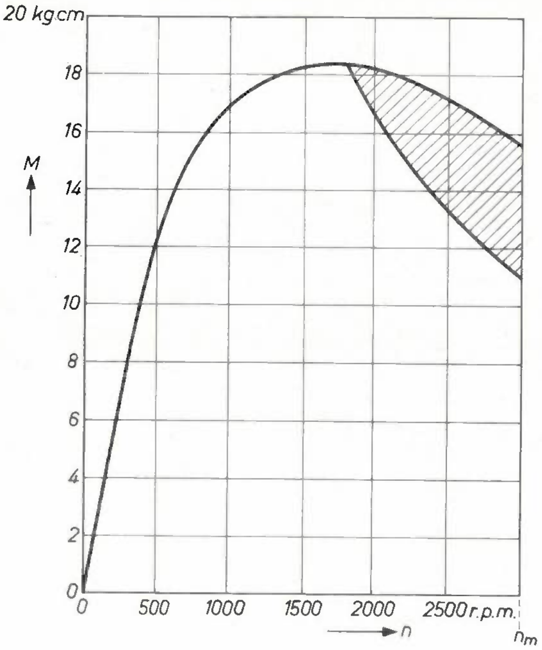


Fig. 8. Measured torque-speed characteristic of a Philips eddy-current coupling type PE 2245. In the hatched area continuous operation of the coupling is not recommended, as the temperature — particularly at the bearings — then exceeds the maximum permissible value.

Fig. 8 shows the torque-speed characteristic derived from measurements on the Philips PE 2245 eddy-current coupling. In the figure n is the relative speed of the input shaft with respect to the speed of the output shaft of the coupling. At the point $n = 0$ the speed of the output shaft is thus at a maximum: at the point $n = n_m$ (n_m is the speed of the motor) the output shaft is stationary and the difference between the speeds of the two shafts is at a maximum. For a given speed the curve gives the maximum torque that can be delivered to the output shaft.

The value of n can also be greater than n_m, i.e. the speed of the output shaft may in certain cases be "negative". This is found for example in servo systems using *two* eddy-current couplings whose output shafts are coupled together and whose input shafts are driven in opposite directions. Compared with the drive described in this article, a servomotor of such a type has the advantage that the speed is variable in *two* senses of rotation.

A system of this kind is used for driving the radiotelescopes at Dwingeloo (Netherlands) and at Malvern (England) — an application in which an important feature is the exceptionally smooth torque of the eddy-current coupling at low revolutions.

Another advantage already mentioned, which the eddy-current coupling has when compared with servomotors of other types, is that it requires only very low driving power.

The power delivered by the output shaft is equal to the product of the torque and the shaft speed. Apart from this useful output, there is a certain amount of power which is converted into heat in the eddy-current cylinder: this power is equal to the product of the torque and the difference between the speeds of the input and output shafts. The heat is partly dissipated by the air, due to the fan-like action of the inductor. The transfer of heat to the air is assisted by the strong air currents close to the eddy-current cylinder which are set up by the teeth on the inductor. In spite of this effective heat removal, so much heat can be generated in the eddy-current cylinder (when the output shaft is

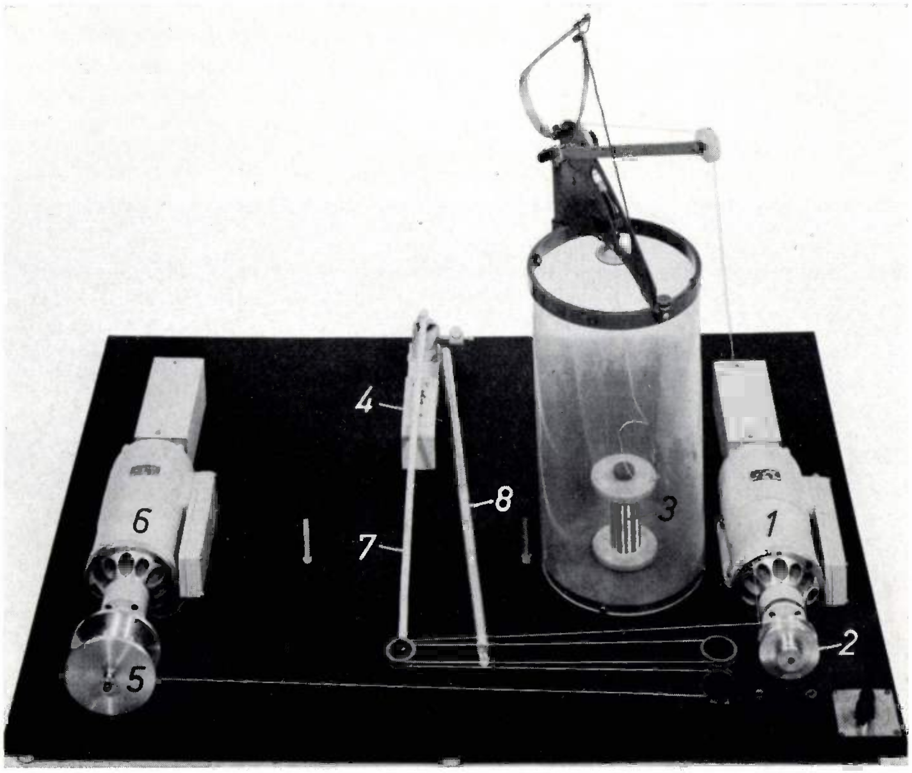


Fig. 9. Demonstration arrangement with two eddy-current couplings, Coupling *1* drives drum *2* at constant speed, so that the wire is unwound at constant speed from reel *3*. The wire runs from the drum over a dynamometer mechanism *4* to the take-up reel *5*, which is driven by eddy-current coupling *6*. The wire is guided by a number of rollers in such a way that the arm *7* of the dynamometer is drawn in one direction only. A spring holds the arm in balance.

turning at low revolutions, and is delivering a high torque), that in places, particularly in the bearings, the temperature of the drive can become higher than is strictly permissible.

The hatched area inside the curve in fig. 8 refers to such a situation. If the coupling is used continuously in this range, the temperature will exceed the maximum permissible value.

Applications of the variable-speed drive

One application of the coupling as a variable-speed drive has already been described. A few other examples are mentioned below.

In driving a take-up reel for wire (or tape) it is usually the speed of the wire, rather than the speed of the reel that has to be kept constant. Here, instead of a tachometer on the output shaft of the eddy-current coupling, we can use a tachometer driven directly by the wire.

The speed of the wire can also be kept constant by continuously measuring the diameter of the take-up reel. This can be done by means of an arm, fitted to the reference voltage potentiometer, and which rests on the wire being wound on to the reel. As the winding becomes thicker, the potentiometer is turned by the arm so that the angular velocity of the reel is reduced, and the wire speed is kept constant.

A further possibility is to drive the wire itself instead of the take-up reel. *Fig. 9* shows a demonstration arrangement with two eddy-current couplings (fixed here to the front plate of the associated electric motors), which winds a wire on to a reel at constant speed and under constant tension. A few turns of the wire are looped around the drum *2* which is driven at a constant speed by the eddy-current coupling *1*. The eddy-current coupling *6* drives the take-up reel *5*, and in such a way that the tension of the wire remains constant. This is achieved by passing the wire through a dynamometer mechanism *4* and *7*, whose output voltage is compared with a reference voltage.

In addition to the speed of a shaft, the speed of a wire and the tension in the wire, other quantities can be controlled, for example the torque exerted on a stationary shaft. The only condition is that it must be possible to represent the controlled quantity by a voltage, which can be compared with the reference voltage in the control amplifier. There is thus a considerable field of application for variable-speed drives in process control, since in many processes performance is monitored electrically, and very often performance is governed by a speed.

Summary. With the aid of a simplified theoretical model of a (heteropolar) eddy-current coupling the relation is derived between the transmitted torque and the relative speed of the input and output shafts of the coupling. This relation is also valid for a homopolar eddy-current coupling. The homopolar version, the design of which is particularly simple as the exciting coil is stationary, is finding increasing application as a variable-speed drive for low-power transmission. An important feature is that the exciting current can be electronically controlled, so that with simple circuits a shaft or other speed or a torque can be kept constant at a desired value.

New phosphors for colour television

A. Bril and W. L. Wanmaker

535.371.07:621.397.132

Until recently, the phosphor most commonly used for "red" in colour television tubes has been a sulphide phosphor. The location of the colour co-ordinates has proved very satisfactory, but its light output leaves something to be desired. In the Philips laboratories at Eindhoven it has been found that red phosphors of very high efficiency can be made by activating suitable substances with trivalent ions of certain rare-earth metals. The best results have so far been obtained with phosphors activated with europium. Europium phosphors have already found fairly wide application.

Nearly all the colours found in our environment can be faithfully reproduced on the screen of a modern colour television tube. Colour tubes are enabled to do this through the fact that some of the substances that fluoresce when bombarded with cathode rays, i.e. the phosphors, have colour co-ordinates in one of the corners of the chromaticity diagram. For example, the colour co-ordinates of silver-activated zinc sulphide (ZnS-Ag) lie well in the "blue" corner, those of (0.75Zn, 0.25Cd)S-Ag lie well in the "green" and those of (0.2Zn, 0.8Cd)S-Ag lie well in the "red" corner (*fig. 1*). Thus, by suitably exciting these three phosphors and mixing the light obtained, it is possible to reproduce any colour whose colour co-ordinates in fig. 1 lie within the triangle *BGR* [1].

It will be evident that to obtain the required mixing of the three colours of light on a television screen it is necessary to use a combination of small luminescent areas. In colour television tubes of the most common type, shadow-mask tubes, the phosphors are therefore arranged in regular configurations of neighbouring dots. These tubes have three electron guns, each of which bombards one type of dot.

Although the colour co-ordinates of the various phosphors are satisfactorily situated, the suitability of a phosphor for use in a colour television tube is not solely dictated by its colour co-ordinates. This is immediately apparent from the following. In order, using the three phosphors mentioned, to produce white light with chromaticity coordinates $x = 0.310, y = 0.316$ (called the white *C*, see fig. 1), the electron beam that bombards the red phosphor dots must be considerably stronger than the two others. This is because the fluorescence spectrum of (0.2Zn, 0.8Cd)S-Ag is extremely wide, extending from the yellow into the near infra-red. A good part of the fluorescent radiation, therefore, because of the nature of the sensitivity curve of the eye, contributes very little to the perception of the light; see *fig. 2*. In others words, the *lumen equivalent*, i.e. the ratio of luminous flux to the corresponding radiant flux, of (0.2Zn, 0.8Cd)S-Ag is fairly low. This appears in the fact that the wavelength at which the fluorescent radiation is strongest (about 675 nm) differs considerably from that of monochromatic radiation with the same colour co-ordinates (about 611 nm).

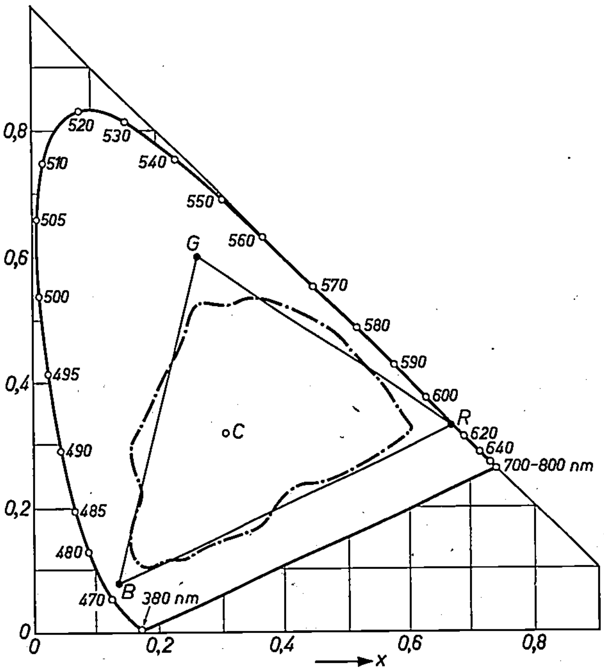


Fig. 1. Chromaticity diagram in *x*-*y* co-ordinates. The wavelength of the spectral colours is expressed in nm (1 nm = 10 Å). The colour co-ordinates of the three phosphors now most commonly used for colour television are *R*, *G* and *B*: all colours whose colour co-ordinates lie within the triangle *BGR* can therefore in theory be faithfully reproduced. The colour co-ordinates of all reflection colours in nature, and also those of all known dyes and printing inks, lie within the dot-dash contour.

Dr. A. Bril is with Philips Research Laboratories, Eindhoven; Dr. W. L. Wanmaker is with Philips Lighting Division, Eindhoven.

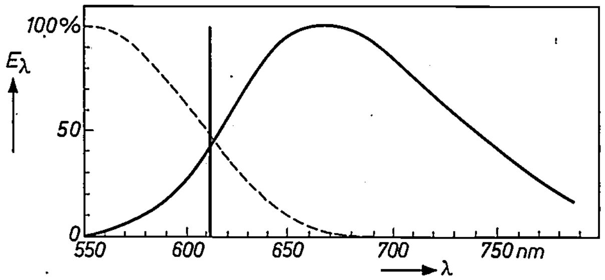


Fig. 2. Spectral energy distribution (schematic) of the fluorescent radiation of the conventional "red" phosphor (0.2Zn, 0.8Cd)S-Ag. Since the spectrum is very wide and falls largely within a region in which the eye is relatively insensitive, the fluorescent radiation of this phosphor has a fairly low luminosity. If the same power were radiated in the form of monochromatic light with a wavelength of 611 nm (represented by the vertical line), a much greater amount of light of the same colour would be obtained. The dashed line is the spectral sensitivity curve of the eye. (A maximum of 100 % is assumed for both curves.)

It has now been found [2] that a red fluorescence with very nearly the same colour co-ordinates and a very high lumen equivalent can be obtained by activating certain substances with trivalent *europium*. The emission of these phosphors takes place mainly in a wavelength interval of only ten nm or less, — a very nearly ideal situation [3]. As an example *fig. 3* shows the fluorescence spectrum of the europium-activated gadolinium oxide (Gd_2O_3-Eu) developed in our laboratory [4].

Because of the very high lumen equivalent of this europium-activated phosphor, its *luminous efficiency*, i.e. the ratio of the emitted luminous flux to the power of the incident electrons, is substantially higher than that of (0.2Zn, 0.8Cd)S-Ag. This is in spite of the fact that the *radiant efficiency* — i.e. the ratio of the radiant flux to the power consumed, i.e. that of the incident electrons — is lower. Because of the high luminous efficiency of the Eu phosphors it is now possible to obtain considerably brighter television pictures. To produce the brightest white, for example, it is no longer necessary to forgo a large part of the "capacity" of the green and blue phosphors because the red has already reached its maximum. For the white colours the luminance of the screen can be roughly 1½ times higher [5].

In the following we shall first explain how it comes about that such a narrow fluorescence spectrum is obtained by using europium as an activator. We shall then briefly consider the method of preparing Eu phosphors and the requirements which the activated matrix compound has to meet in order for Eu^{3+} to be effectively incorporated. Finally we shall deal in more detail with the properties of Gd_2O_3-Eu and compare them with those of some Eu phosphors prepared by other investigators. It will be shown that Eu phosphors have, apart from a high luminous efficiency, further attractive features of interest in other applications then in television tubes.

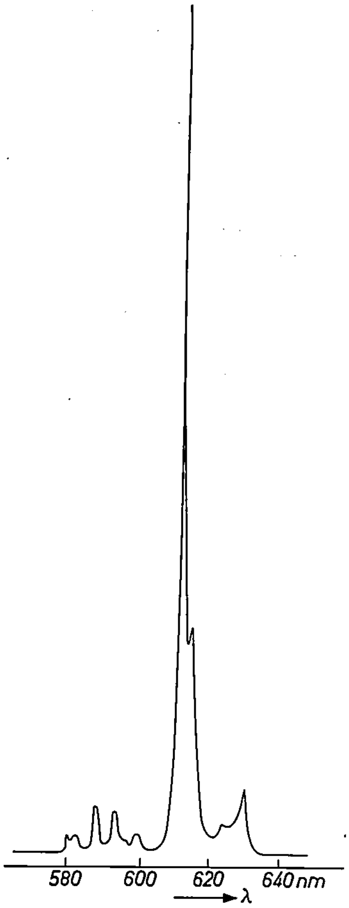


Fig. 3. The spectral energy distribution of europium-activated Gd_2O_3 excited by electrons of 20 keV. There are two extremely narrow spectral lines very close together (wavelengths 611.3 and 614.2 nm). Because of this property Gd_2O_3-Eu has a relatively very high lumen equivalent.

Fluorescence of substances activated with rare-earth metals

The atoms of the rare-earth metals differ from those of the other elements in that not only the outer electron shell but also a deeper shell is incompletely filled. In this latter shell (the *N* shell) the 4*f* group is incomplete

[1] For a more extensive treatment of the principles of colour television see F. W. de Vrijer, Philips tech. Rev. **19**, 86-97, 1957/58.

[2] A. Bril and W. L. Wanmaker, J. Electrochem. Soc. **111**, 1363, 1964 and A. Bril, W. L. Wanmaker and C. D. J. C. de Laat, J. Electrochem. Soc. **112**, 111, 1965.

[3] See A. Bril and H. A. Klasens, Philips Res. Repts. **10**, 305, 1955, where the requirements to be met by the ideal red phosphor are formulated on page 317.

[4] See the articles referred to in [2]. The fluorescence of phosphors activated with rare earths was found many years ago by G. Urbain (Ann. Chim. Phys. **18**, 293, 1909); he obtained spectra, however, in which not only the red lines, but the orange, green and blue lines as well, had a high intensity.

[5] For details on this subject see: A. Bril and C. D. J. C. de Laat, Light output of color-television screens with europium-activated phosphors as a red component, Electrochem. Technol., 4, 21-24, 1966 (No. 1/2⁴).

(see *Table I*). Electron transitions in this shell, caused by the absorption or emission of radiation, are little disturbed by the environment owing to the depth at which they take place; they are effectively screened by the electrons in the *O* shell. One result of this is that the spectral emission lines of the trivalent ions in rare-earth metals — here the outer three electrons are absent — are very narrow. Their width at room temperature is between 0.5 and 1 nm, and at the temperature of liquid nitrogen (— 196 °C) the line width can be as small as 0.3 nm. In this respect substances activated with trivalent rare earths differ considerably from most other fluorescent substances, whose fluorescence spectrum usually consists of broad bands.

Table I. Electron configuration of the rare-earth metals ($Z = 58$ to 71). In these elements the 4*f* group of the *N* shell is incomplete, although this is not a peripheral shell. (These data are from: B. G. Wybourne, Spectral properties of rare earths, J. Wiley, New York 1965.)

Shell *n*	*K* 1	*L* 2		*M* 3			*N* 4				*O* 5			*P* 6
Element *l*	*s* 0	*s* 0	*p* 1	*s* 0	*p* 1	*d* 2	*s* 0	*p* 1	*d* 2	*f* 3	*s* 2	*p* 6	*d* 1	*s* 0
57 La	2	2	6	2	6	10	2	6	10	—	0	1	2	2
58 Ce	2	2	6	2	6	10	2	6	10	1	2	6	1	2
59 Pr	2	2	6	2	6	10	2	6	10	3	2	6	—	2
60 Nd	2	2	6	2	6	10	2	6	10	4	2	6	—	2
61 Pm	2	2	6	2	6	10	2	6	10	5	2	6	—	2
62 Sm	2	2	6	2	6	10	2	6	10	6	2	6	—	2
63 Eu	2	2	6	2	6	10	2	6	10	7	2	6	—	2
64 Gd	2	2	6	2	6	10	2	6	10	7	2	6	1	2
65 Tb	2	2	6	2	6	10	2	6	10	9	2	6	—	2
66 Dy	2	2	6	2	6	10	2	6	10	10	2	6	—	2
67 Ho	2	2	6	2	6	10	2	6	10	11	2	6	—	2
68 Er	2	2	6	2	6	10	2	6	10	12	2	6	—	2
69 Tm	2	2	6	2	6	10	2	6	10	13	2	6	—	2
70 Yb	2	2	6	2	6	10	2	6	10	14	2	6	—	2
71 Lu	2	2	6	2	6	10	2	6	10	14	2	6	1	2
72 Hf	2	2	6	2	6	10	2	6	10	14	2	6	2	2

Furthermore, because of this screening, the wavelength and width of the spectral lines are virtually independent of the nature of the matrix compound. The spectra of the various phosphors activated with a particular rare-earth metal therefore show little variation in the positions of the spectral lines. Roughly speaking, they are similar to the spectra found upon excitation in a non-perturbing environment (i.e. in a spark discharge; *fig. 4*). Owing to the electric field in the crystal, however, the energy levels exhibit a more or less pronounced splitting (Stark effect), so that each line is replaced by a number of neighbouring lines.

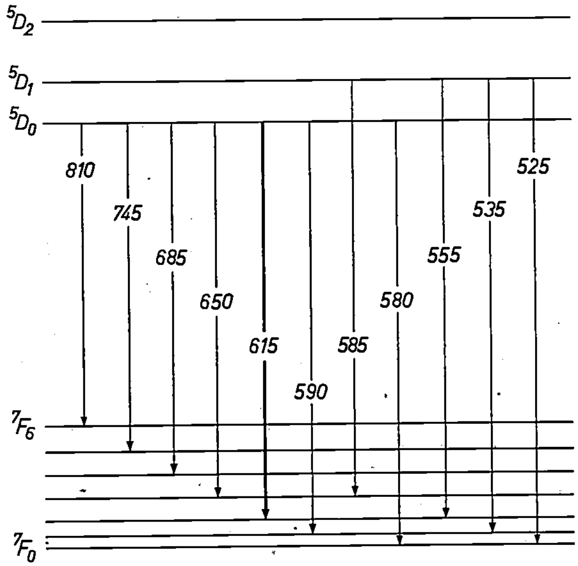


Fig. 4. Part of the energy level scheme of trivalent europium, indicating the transitions accompanied by the emission of light[6]. The figures denote the wavelengths of the relevant spectral lines (rounded off to 5 nm).

The situation is different for the relative intensities of the lines; these can be markedly different, even where there is not much difference in the environment of the ions. An example is to be seen in *fig. 5*, which gives the fluorescence spectra of Gd_2O_3-Eu, $Gd_2O_3.B_2O_3$-Eu and $GdVO_4$-Eu under excitation by short-wave ultraviolet radiation. In the first phosphor the red lines at 611.3 and 614.2 nm, which correspond to the transitions from the $^5D_0$ to the $^7F_2$ level, are much stronger than the orange lines at 590 nm, which relate mainly to transitions from the $^5D_0$ to the $^7F_1$ level. In the second phosphor the situation is exactly the opposite.

Owing to the negligible interaction with the environment, the efficiency of the fluorescence is often high and depends very little on the temperature: some phosphors can be heated up to 300 to 400 °C without any appreciable drop in efficiency (*fig. 6*). In this respect as well the Eu-activated phosphors are superior to the sulphide phosphor, whose intensity has already dropped to one half at about 100 °C.

Owing in particular to the great variations in the intensities of the various spectral lines that can be emitted, it is by no means possible as yet to predict the combination of matrix compound and activator that should be chosen to obtain a phosphor that meets rigorously defined requirements; a good deal of research in this field remains as yet to be done.

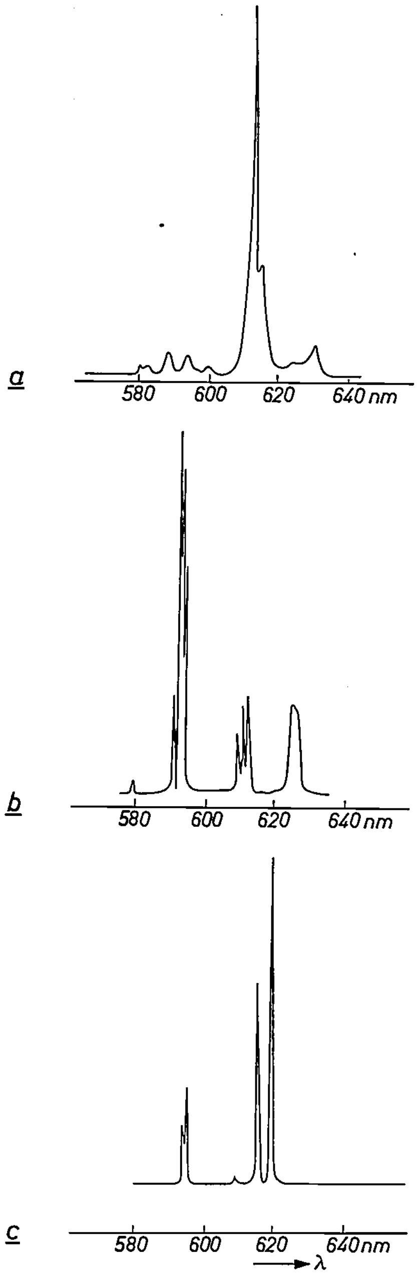


Fig. 5. The wavelengths of the lines of the fluorescence spectrum are hardly influenced by the nature of the matrix activated by europium. The spectra shown here are those of *a*) Gd_2O_3-Eu, *b*) $Gd_2O_3.B_2O_3$-Eu and *c*) $GdVO_4$-Eu. In all three phosphors a cluster of lines is found between 590 and 600 nm and one between 610 and 620 nm. The intensity distribution, however, differs substantially. (The ordinate scales have been chosen so that the highest peaks in the three figures have about the same height.)

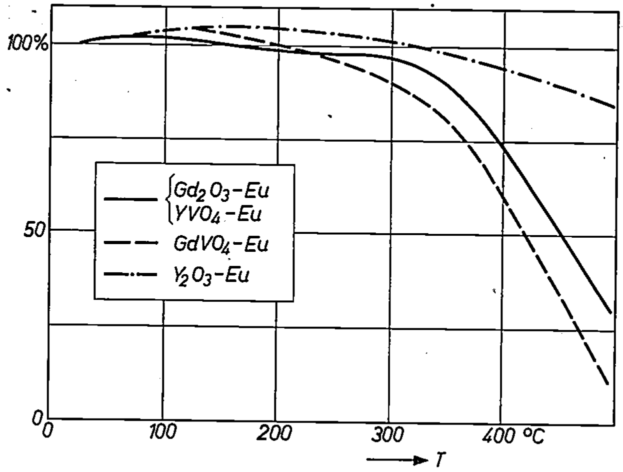


Fig. 6. Intensity of the fluorescent radiation of some Eu phosphors plotted against temperature. The intensity at room temperature is assumed to be 100%. Up to 300 °C the temperature has hardly any effect.

The preparation of Eu-activated phosphors

The Eu-activated phosphors are prepared by means of chemical solid-state reactions. First, a highly homogeneous mixture of the starting materials is made. This mixture is then fired in air and kept for a certain time at an elevated temperature (1000 to 1300 °C). The preparation of Gd_2O_3-Eu, for example, can start with a mixture of very pure Gd_2O_3 and Eu_2O_3, or with europium and gadolinium salts which dissociate upon heating and yield an oxide. Since solid-state reactions are seldom entirely complete, the product is finely ground after heating, homogenized and then fired again.

A starting material that has proved to have very suitable properties is europium oxalate. This dissociates at a relatively low temperature forming an extremely fine and hence highly reactive powder. This oxalate is obtained by dissolving europium oxide in HCl, adding oxalic acid, and then filtering off, washing, and drying the precipitate.

We have also found that the efficiency is improved by adding to the starting mixture a certain quantity of a fluoride, preferably europium fluoride.

There are limits to the amount of europium that can be added to the matrix compound. At a certain europium concentration the efficiency reaches a maximum; it then starts to decrease (an effect known as concentration quenching). With Gd_2O_3-Eu we

[6] From M. A. El'yashevich, Spectra of the rare earths, book 2, Off. Tech. Serv., Dept. of Commerce, Washington D.C. 1961.

found [7] that the maximum is at about 0.06 gramatoms of europium per mole Gd_2O_3.

To obtain high efficiency phosphors one has to choose rare earths that can easily find a place in the matrix lattice. It is therefore necessary to choose substances in which the metal ion has the same valency and roughly the same radius as europium; examples are Gd_2O_3, Y_2O_3 and $GdVO_4$. It is probably advisable as well to use a matrix compound whose crystal structure is similar to that of the corresponding europium compound: the crystal structure will not then be affected by the incorporation of the activator. Gd_2O_3 and Eu_2O_3 meet this requirement, but La_2O_3 and Eu_2O_3 do not. This may explain why the radiant efficiency of La_2O_3-Eu with cathode ray excitation is lower than that of Gd_2O_3-Eu and Y_2O_3-Eu.

Properties of technical importance

As we saw from fig. 3, Gd_2O_3-Eu mainly emits light with wavelengths of 611.3 and 614.2 nm. The colour point has the co-ordinates $x = 0.66$ and $y = 0.34$ These are almost the same as those of the conventional red phospor (0.2Zn, 0.8Cd)S-Ag and therefore differ very little from the chromaticity co-ordinates standardized by the NTSC (National Television System Committee) in the United States ($x = 0.67$ and $y = 0.33$), corresponding to a spectral line with a wavelength of 611 nm.

In *Table II* the efficiency of Gd_2O_3-Eu is compared with that of (0.2Zn, 0.8Cd)S-Ag. Because of its exceptionally high lumen equivalent, the luminous efficiency of Gd_2O_3 is much higher than that of the sulphide, although its radiant efficiency is lower. This applies however for all oxide phosphors; Gd_2O_3-Eu even has the highest radiant efficiency in this group.

It is interesting to compare the above data with those of two other Eu phosphors developed elsewhere: Y_2O_3-Eu [8] and YVO_4-Eu [9]; the latter phosphor is now used on a fairly wide scale. The spectrum of Y_2O_3-Eu is almost identical with that of Gd_2O_3-Eu, but our measurements show that the efficiency is lower. For YVO_4-Eu the two strong red spectral lines have a somewhat longer wavelength (614 and 619 nm); in addition there is a strong line at 700 nm. Of course, this spectrum makes it possible to reproduce the dark red colours rather better than when Gd_2O_3-Eu is used — the colour co-ordinates lie farther over in the red corner — but the lumen equivalent is not nearly so good: we found only 245 lm/W, compared with 300 lm/W for Y_2O_3-Eu and Gd_2O_3-Eu. The question of which phosphor is to be preferred for practical applications, quite apart from the subjective element in the judgement, cannot be answered on the grounds of these properties alone. Apart from efficiency and colour co-ordinates there are other factors, which are outside the scope of this article, such as the cost and workability of the phosphors..

Table II. Comparison of the radiant efficiency η, the lumen equivalent L and the luminous efficiency l of the red phosphor previously used in ordinary television tubes, of the red phosphor used for large-screen projection television, and of the europium phosphor on the basis of Gd_2O_3. The values refer to normal loading in ordinary television tubes. Under very high loading, as in projection television, the sulphide phosphor shows saturation and is inferior to $(Zn, Be)_2SiO_4$-Mn. The relation between the three given quantities can be expressed as: $l = \eta L$.

	η (%)	L (lm/W)	l (lm/W)
(0.2Zn, 0.8Cd)S-Ag	17	75	13
$(Zn, Be)_2SiO_4$-Mn + red filter	3	185	5.5
Gd_2O_3-Eu	9	300	27

A further important advantage of the Eu phosphors is the fact that they are white when viewed in daylight. Because of this property the dominant wavelength of the colours in a picture does not change when lighting is switched on near the screen: the colours only become somewhat less saturated. For example, on the screen of a tube coated with a Eu phosphor the black parts turn dark grey, whereas on a screen coated with a conventional sulphide phosphor, which is orange, they turn rather orange-brown (*fig. 7*).

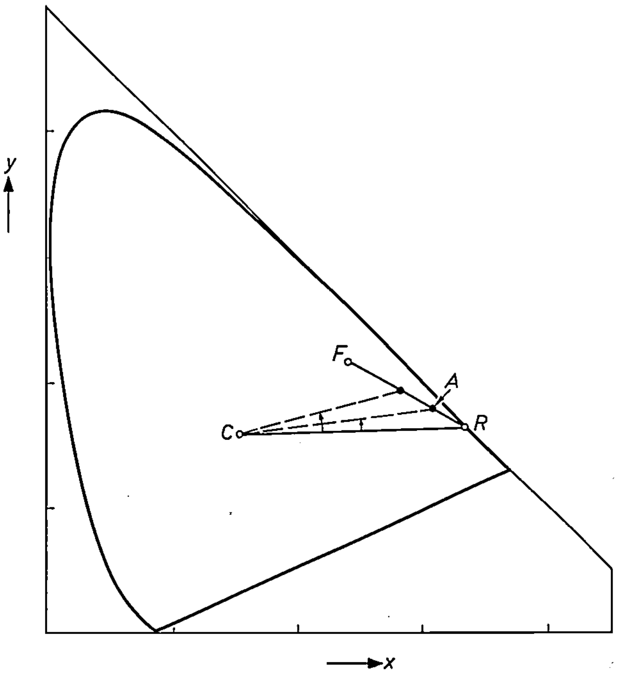


Fig. 7. When artificial illumination is switched on, the eye receives reflected light (colour co-ordinates taken as F) as well as the fluorescent radiation from the tube screen (colour co-ordinates taken as R). The colour co-ordinates A of the combined light lie somewhere on the line RF: the stronger the artificial lighting, the nearer A is to F. In the case represented here a fairly marked colour shift takes place. If the phosphor is white, we have the special case where F coincides with C. Point A is then always on the line CR. When artificial lighting is switched on there is now no colour shift, the only effect is that the colour becomes less saturated.

Europium-phosphor screens are also ideally suited for large-screen projection tubes, where the images from a red, a blue and a green fluorescent tube are projected one over the other greatly magnified. The tubes have to be very heavily loaded to obtain sufficient light on the projection screen [10]. Eu phosphors are very suitable for such applications. In the first place the load can be raised very considerably before any saturation becomes noticeable in the light emission. In the second place, as we have seen (fig. 6), the emission is hardly affected by temperature over a very wide range, so that the efficiency is not impaired by the high temperature resulting from heavy loading. Moreover, the efficiency is five times higher than that of the phosphor now most widely used for large-screen projection television (manganese-activated zinc-beryllium silicate; see Table II). When a europium phosphor is used, 16 lm/W is obtained on the screen with white light, compared with 11 lm/W for a silicate phosphor. This considerable improvement only partly illustrates the high efficiency of the Eu phosphor, because the limiting factor is now no longer the red phosphor, but the blue one.

The practical application of Eu phosphors need by no means be confined to television tubes. Owing to the presence of a small number of narrow emission lines combined with a high efficiency, they are very suitable as a material for lasers, although they then have to be prepared in the form of single crystals. The preservation of their favourable properties at high temperature also makes them attractive for use in those types of discharge lamps in which the phosphor can become very hot, particularly in high-pressure mercury vapour lamps with a fluorescent coating.

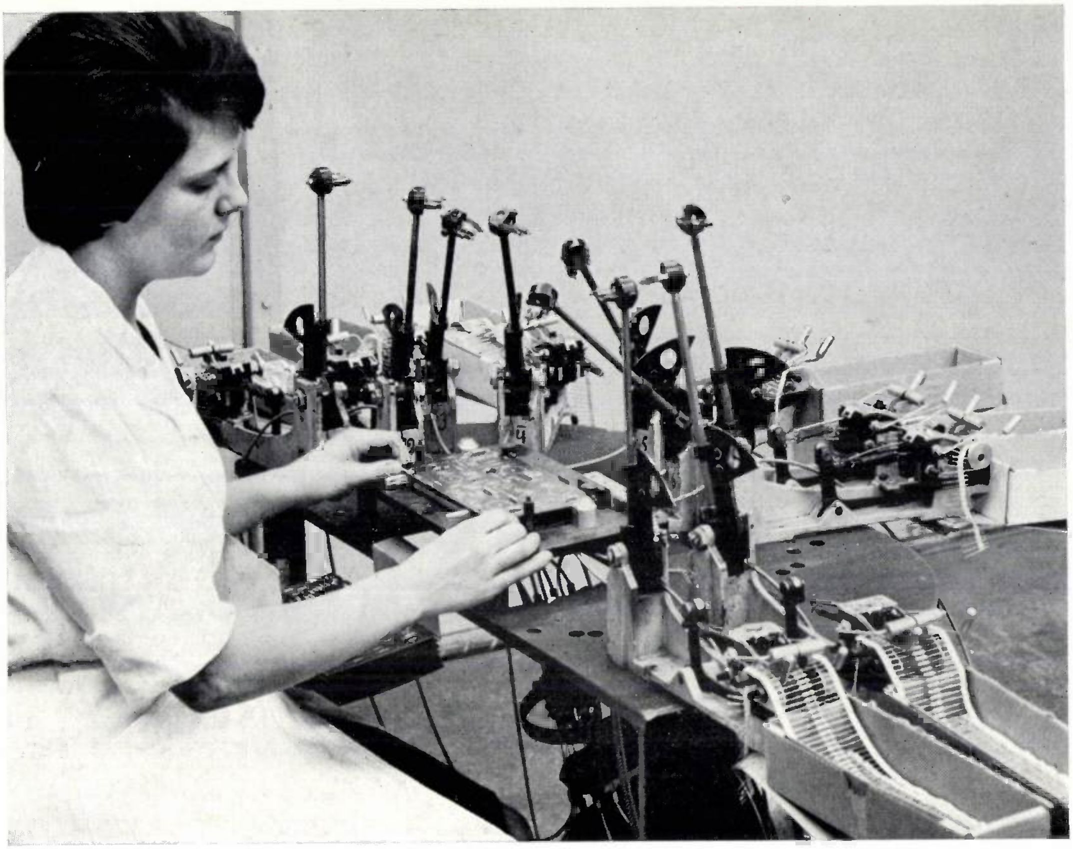
[7] See the second of the articles in reference [2].

[8] The characteristics of this phosphor have been investigated by K. A. Wickersheim and R. A. Lefever (J. Electrochem. Soc. **111**, 47, 1964), by N. C. Chang (J. appl. Phys. **34**, 3500, 1963) and by R. C. Ropp (J. Electrochem. Soc. **111**, 311, 1964 and **112**, 181, 1965).

[9] Characteristics of this phosphor have been investigated by L. G. Van Uitert, R. C. Linares, R. R. Soden and A. A. Ballman (J. chem. Phys. **36**, 702 and 1793, 1962). The phosphor was proposed and developed for use in colour television tubes by A. K. Levine and F. C. Palilla (Appl. Phys. Letters **5**, 118, 1964).

[10] For further details see T. Poorter and F. W. de Vrijer, Philips tech. Rev. **19**, 338-355, 1957/58.

Summary. The conventional red phosphor in picture tubes for colour television has a fairly low efficiency, because its fluorescence spectrum is very wide and the lumen equivalent of the radiation therefore rather low (75 lm/W). It has been found that phosphors activated with trivalent rare earths (in particular europium) can have extremely narrow spectral lines and hence a much higher lumen equivalent ($\approx$ 300 lm/W). In these metals the $4f$ group of the N shell is not completely filled. Owing to the depth of this shell the wavelength of the europium lines is very little affected by the nature of the matrix compound; their intensity ratio is however affected. The Gd_2O_3-Eu phosphor developed in the Philips laboratories emits two strong red spectral lines (611 and 614 nm, colour co-ordinates $x = 0.66$, $y = 0.34$) and has a luminous efficiency of 27 lm/W. This is more than twice as high as that of the conventional phosphor. The luminance of white on the screen of a picture tube can therefore be roughly $1\frac{1}{2}$ times higher. Because its fluorescence is almost temperature-independent up to 300 °C, the phosphor is also ideally suited for large-screen projection television tubes and for use in discharge lamps that become very hot.



Impact-mounting of components on printed wiring panels

H. Groenhuis and W. L. L. Lenders

621.397.62.049.75.002.72

In the mechanized mounting of components on printed wiring panels the aim is to design machines that can insert the maximum number of components in the shortest time, at the minimum capital outlay. Not long ago, Philips in Eindhoven brought a new machine into use which can mount ten components in a panel simultaneously. The mechanism is interesting, consisting basically of a swinging arm which moves in much the same way as the human forearm.

In recent years the mechanized mounting of components on printed wiring panels has made interesting progress. After the cutting-and-bending device and the assembly line, previously described in this journal [1], a new technique has been developed, called *impact-mounting*. Before dealing with this system, we shall briefly recall the other methods in use.

Cutting-and-bending machine and assembly line

The cutting-and-bending machine is very simple in design and versatile in operation, but calls for a considerable amount of manual manipulation. The connection wires of the components have to be previously bent so that they can be inserted into the holes in the panel. The panel is set up on the machine by hand, and each component is separately inserted by hand; the machine then cuts off and bends over the ends of the wire projecting through the panel.

On the assembly line all these operations are fully

H. Groenhuis and W. L. L. Lenders are with the Works Mechanization Department of Philips Radio, Gramophone and Television Division, Eindhoven.

mechanized. The machine consists of a series of insertion heads, each of which can insert one component at a time in a panel under the head. An intermittently moving conveyor belt feeds 100 panels under the heads. If the assembly line contains 20 heads, then 20 components are mounted in each panel in one complete pass of the conveyor belt. For example, to mount 90 components on the panels, they have to travel five times under the heads. After each pass the insertion heads have to be supplied with another 100 components and their positions in relation to the panel have to be changed. The heads can be reset for this very quickly, so that small runs of panels can also be handled efficiently. There are insertion heads for different kinds of component, — resistors, valve holders, and so on.

Impact-mounting

The impact-mounting machine is much simpler than the assembly line. There is only one position in the machine for the panel, which is introduced by hand. The new aspect of the method is that different components, up to ten in number, can be inserted in the panel *simultaneously*. This has been achieved by arranging the insertion heads not, as in the assembly line, above the panel, but around it, and by giving the mechanism a semicircular instead of a linear motion. Each head is therefore fixed to an arm which has a swinging movement. When the arm is at one end of its swing the component is picked up by the head, and at the other end of the swing it is inserted in the panel. In both positions the component is turned so that its connection wires are pointing downwards. This is made possible by a second movement of the arm, a rotation of 180° around its axis. The arm rotates as it swings, and indeed the whole motion of the arm can best be compared with that of the human forearm.

Since the actual insertion of a component does not take much longer than in the assembly line, the simultaneous mounting of different components gives a considerable saving of time. If, for example 90 components have to be inserted, it can be done with 9 impact-mounting machines. Depending on the number of panels to be produced, these 9 machines could be attended by 1 to 9 operators. For small runs, however, fewer machines are sufficient, since the arm units can be rearranged on the bench top. It is then advantageous to have one spare machine so that while one is being reset by the mechanic, production can continue at the other machines. The arm units are not fed with separate components but with strips each containing 1000 components (see title photograph) taken straight from the box in which the component manufacturer packs his products. Some possible layouts for the impact-mounting machine are shown in *fig. 1*.

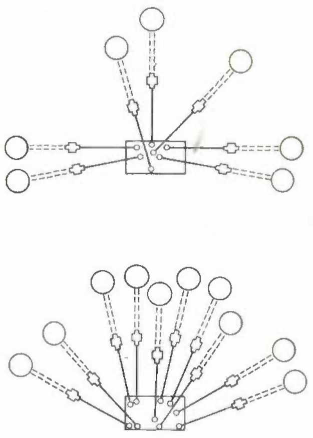


Fig. 1. Two possible arrangements of arm units on a work bench. The rectangle represents the panel. Up to ten arm units can be used.

Construction of the arm unit

The simplified sketch in *fig. 2* gives a rough idea of how the mechanism of the arm unit works. More details can be seen in a photograph below. The arm *3* to which the head *7* is fixed swings from position *I* to position *II* and back. The strip *10* containing the components is fed into the machine by the mechanism *C*, the lever *6* being controlled by a cam *5* fixed to the arm-holder *4*. Working in conjunction with this mechanism the head cuts the component from the strip, simulta-

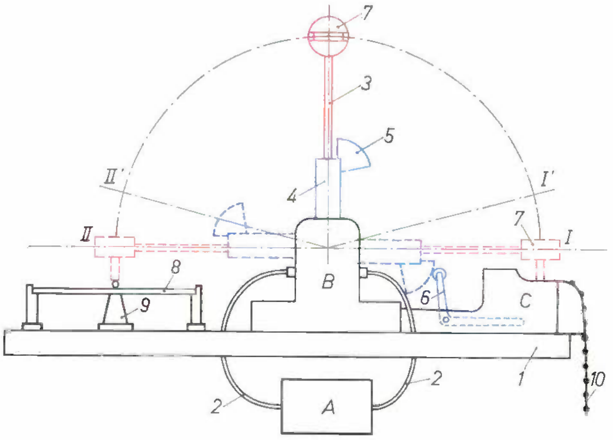


Fig. 2. Sketch of the mechanism of an arm unit. *1* is the bench top, and beneath it is located the driving mechanism *A* which, by means of Bowden cables *2*, imparts a swinging movement to the arm *3* with its inserting head *7* (red). The mechanism *B* ensures that during its 180° swing from "grip position" *I* to "insert position" *II* the arm also rotates 180° around its own axis. The cam *5*, fixed to the arm-holder *4*, operates, through the lever *6* (blue), the mechanism *C*, which controls the feed of the component-strip *10*. On coming down in position *I* the head cuts the connection wires of the component, bends them over and grips the component. In position *II* the connection wires are punched into the holes in panel *8* so that they bend over against the stops *9* and are thus fixed firmly to the panel.

[1] R. van Beek and W. W. Boelens, Printed wiring in radio sets, Philips tech. Rev. **20**, 113-121, 1958/59.
R. van Beek and A. J. Halbmeyer, Mechanized mounting of components on printed-wiring panels, Philips tech. Rev. **24**, 41-57, 1962/63.

neously bending the connection wires and gripping the component tight. After a swing of 180° the component is inserted at position *II*. The connection wires are punched through holes in the panel *8* against concave stops *9*, under the panel, which bend them in the appropriate direction. The component is now attached firmly to the panel. During its swinging motion the arm, as mentioned, rotates 180° about its axis, so that the component arrives in the proper position above the panel.

Fig. 3 gives a somewhat schematic section at the long axis of the component-feed mechanism (blue) and the cutting mechanism (green). (The components themselves are shaded.) The components are pulled forward by a rake *12*, which has two teeth that engage the connection wires of the components. As our starting point we take the situation in which the arm is in position *I* (see fig. 2) and a component is gripped in the head. The spring *17* is prevented from pulling the lever *16* to the left, because in position *I* lever *6* is held back by the cam *5*. When the arm moves upwards, lever *6* can turn to the left, so that lever *16* is released and plate *22* with the rake attached to it moves to the left. While the arm is swinging towards the panel to insert the component, the rake shifts the strip with components over the distance required to bring the next component into the appropriate position. (This situation is represented in fig. 3.) When, after completing a cycle, the arm returns to position *I*, the cam and rod system pushes the rake to the right again, so that it is ready for the next cycle. While this is happening the strip with components is held by the teeth *11*. When the arm has completed its swing to the right, the head descends on to the cutters *18* and forces them down against the action of the springs *23*.

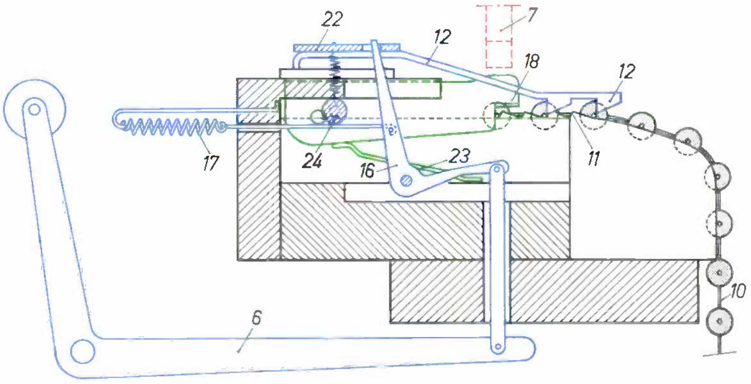


Fig. 3. Mechanism for feeding in and cutting off the components. (This sketch shows only the most important parts; the moving parts responsible for the feed are marked blue, those for cutting the components are green.)
11 teeth on which the strip of components rests. *12* rake which, upon each stroke of the arm, pulls the strip to the left so that a new component comes to rest under the cutters *18*. The rake is fixed to a plate *22* which is drawn to the left by lever *16* and spring *17*. This happens as soon as lever *6* (see fig. 2) is able to turn anti-clockwise after arm *3* with cam *5* has been raised. Plate *22* is held by a spring and a roller *24* which permit movement in the horizontal direction only. The components are severed by cutters *18* which are pushed down by the head *7* against the action of a spring *23*.

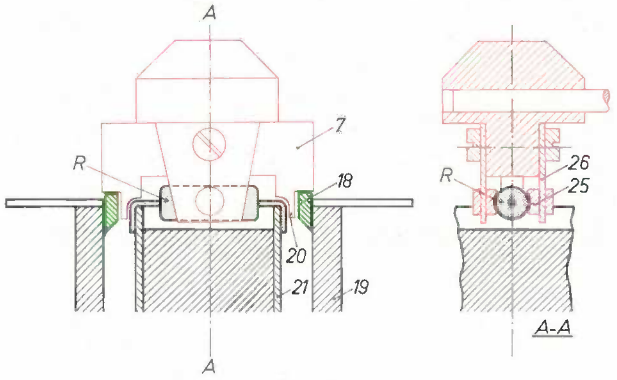


Fig. 4. Mechanism for cutting and bending the connection wires and for gripping the component. The head *7* (red) pushes the cutters *18* (green) down the inside surface of plates *19*, thus cutting off the wires. Studs *20* then bend the wires around plates *21*. The body of the component *R* is gripped by two nylon pads *25* which are pushed against the component by two springs *26*.

The mechanism for cutting off and gripping the component is illustrated in *fig. 4*. The impact of the head forces the cutters *18* (green) down the sides of the plates *19*, so that the wires are cut off. Immediately afterwards two projecting studs *20* on the head bend the wires over the plates *21* and the head grips the component bodily by means of two spring-loaded nylon pads. *Fig. 5* illustrates the mechanism shortly after the head has gripped a component. The component-feed system is clearly visible in this photograph.

The arm mechanism

The movements described by the arm will be discussed with reference to fig. 2. During the swing from *I* to *I'* the head remains in the position for gripping a component. Between *I'* and *II'* the arm rotates 180°, so that at *II'* the head has taken up the correct position for inserting the component. It maintains this position as far as *II*, and in the return swing as far as *II'*. The arm then turns round again so that the head at *I'* is back in the "grip" position, and so on.

Fig. 6 illustrates the mechanism which produces all these movements. The holder *4* (blue) in which the arm (red) is fixed can pivot around two tapered horizontal pins *35* fixed in a housing *34*. The pivoting of the holder allows the swinging motion of the arm; the Bowden cables *2a* and *2b*, which lie in grooves in the lower part of the holder give the backwards-and-forwards movement. The arm can rotate around its own axis in the holder; the drive consists of two bevel gears, gear-

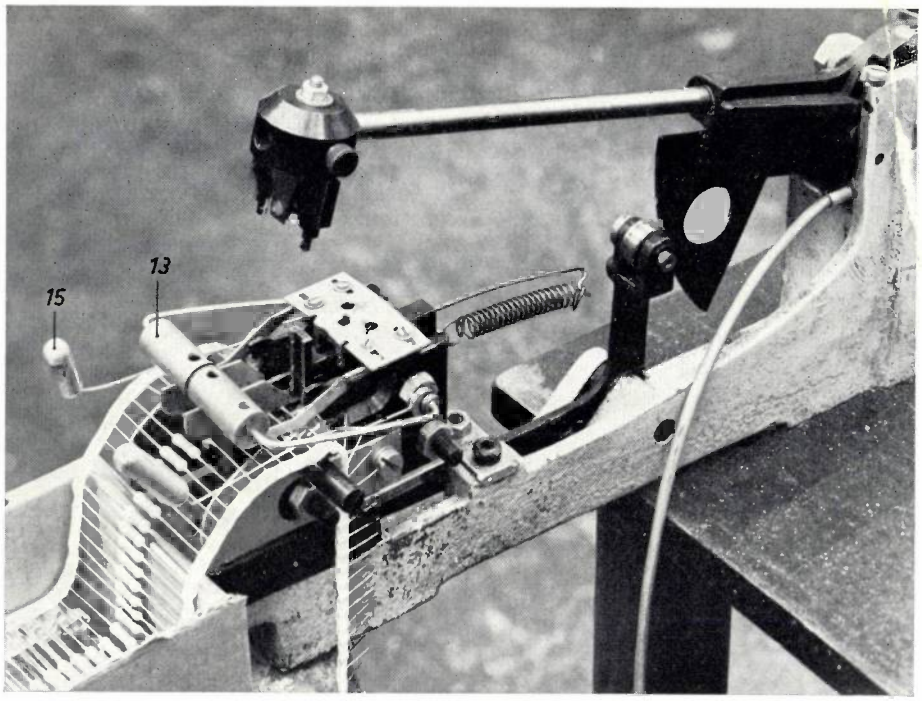


Fig. 5. The feed system and cutting-and-bending mechanism. The cutters and the rake can clearly be seen. Springs with weights *15* hold down the component connection wires, and weights *13* with vertical plates keep the components in the right position.

wheel *27* (red) which is fixed rigidly to the arm, and gearwheel *30* (green) which rotates on a stub on the holder. The latter gear is flexibly connected to the housing by a pin *32* which fits in a groove *33* in the housing and is pressed towards the centre of the groove by springs *31* (cross-section *A-A*). During the movement from *I'* to *II'*, gearwheel *30* is held stationary with respect to the housing by the pin and springs, and gearwheel *27* rolls on *30* so that the shaft rotates. When the arm has completed the required rotation, stud *29* comes up against one of the stops *28* (cross-section *C-C*), preventing further rotation of the arm and hence of gearwheel *27*. The arm continues its swinging motion, however, so that gearwheel *30* has to rotate, in which process pin *32* is pushed along its groove against the action of springs *31*. When the arm swings back again

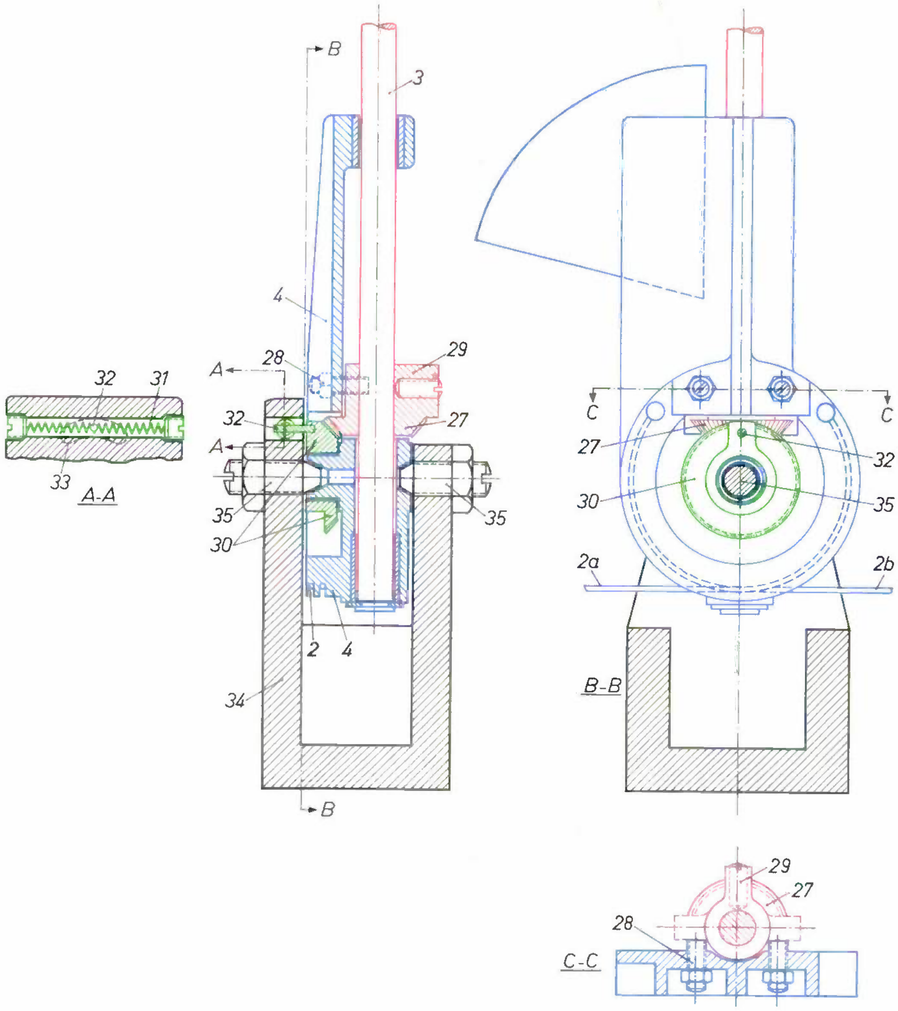


Fig. 6. Cross-sectional views of the arm mechanism. The housing *34* is fixed to the bench top. The holder *4* (blue) in which the arm *3* (red) is fitted can pivot around two pins *35* (for the swinging motion of the arm). The arm can rotate in its holder. Gearwheels *27* (red) and *30* (green) produce this rotation during the swinging of the arm.

after reaching the end of its travel, the pin first goes back to the centre of the slot, so that *30* returns to its original position, gearwheel *27* rolls on *30* again and the arm rotates in the other direction.

The drive for the whole of this rather complicated series of movements thus calls for only two cables which have to be pulled back and forth over a certain fixed distance for each movement of the arm. This does not take place at the same speed in the two directions, more work being needed to grip a component (including the cutting and bending of the wires) than for inserting the component. The mechanism responsible for moving the cables will not be discussed here. It is situated under the bench (see *fig.* 7) and consists mainly of two cams which give an up and down movement to a rod to which the cables are attached. The cams are driven by an electric motor with a "one revolution" clutch. This allows the cam to complete one revolution when a contact is closed. The machine is very simple to operate, all that is needed being to place the panel in the right position on the bench and to press it against the above-mentioned contact. The switchbox at the the bottom right of the work bench in fig. 7 contains the main switch and fuses and also a simple rectifier which supplies the voltage for operating the clutch.

Finally, let us again compare the impact-mounting method with the completely mechanized assembly-line system. We have already mentioned the fact that impact-mounting offers a higher assembly speed. Another virtue is that the simpler construction of the mounting units reduces the number of faults that can occur. An advantage of the assembly line, on the other hand, is that during production the operators can quickly and easily reset the line for the assembly of a different type of panel, whereas with the impact-mounting machine this calls for more time and skill. For a given production capacity impact-mounting involves a lower capital outlay than mechanized line assembly, although the line can still be more profitable because of the simplicity of resetting the heads when different types of panel have to be assembled soon one after another. Each method, then, has its own field of application: impact-mounting is indicated for large or medium production runs of only one or two types of panel, and line assembly is indicated for the quantity production of widely varying types.

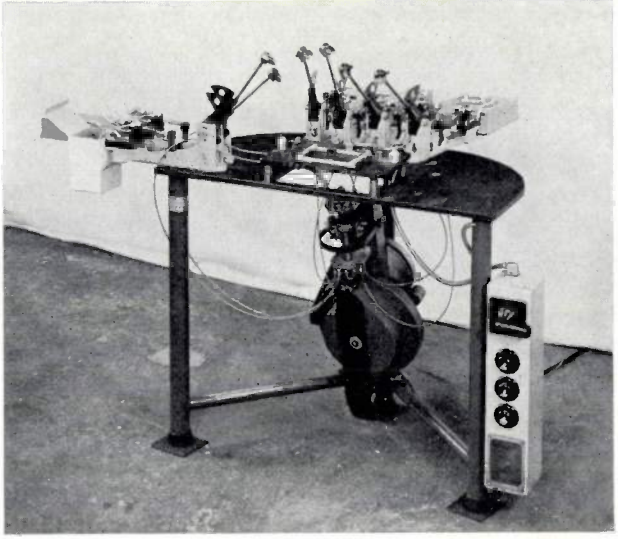


Fig. 7. Impact-mounting machine with six arm units. The drive mechanism, using Bowden cables, which give the arms their backwards-and-forwards movement, can be seen under the bench. The drive comes into action when the panel is fitted in place and pushed against a contact (see title photograph).

In the methods described here components of conventional form are used. It would be possible to simplify the assembly machines considerably and thus substantially cut down on costs if the design of the components were specially adapted for mechanized assembly.

Summary. Impact-mounting is a new technique for inserting components in printed wiring panels. Up to ten components can be *simultaneously* mounted on a panel which is placed in the machine by hand. The components are fed into the machine in strip form. A number of arm units are arranged around the panel, each unit consisting of an arm fitted with an inserting head. The arm swings through an angle of 180°. At one end of the swing the head severs a component from the strip and grips it, and in the other extreme position it inserts the component in the panel. During the swing, the arm also rotates 180° about its own axis so that the component arrives in the right position above the panel. The rotation is derived from the swinging motion by means of a system of gears. Since the rotation has to be completed before the swing, this involves a rather complicated mechanism. The components are fed into the machine by a system of levers also operated by the arm. A comparison is drawn between impact-mounting and the fully mechanized assembly line method previously described.

Volume 27, 1966, No. 1 pages 1-32 Published 27th June 1966

PHILIPS TECHNICAL REVIEW

VOLUME 27, 1966, No. 2

Colour television transmission systems

F. W. de Vrijer

621.397.132

Few will dispute the economic and technical usefulness of standardization. Indeed, at the present time, considerable efforts are being made to bring about a degree of orderliness in the variety of measures, parameters, definitions, symbols, etc., that have grown unchecked in many branches of technology. International standardization is, in fact, absolutely essential as a preliminary *to new developments that are only just under way, especially where huge investments are involved, as is now the case with colour television in Europe. It is to be hoped that, in the imminent decision in the struggle between the rival systems, the need for uniformity will prevail. In this article, the author attempts to set out as objectively and as clearly as possible the questions under discussion.*

Additive colour-mixing of three basic colours is used for colour television. In order to obtain as wide a range of colours as possible [1] red, green, and blue are used as the basic colours (*fig. 1*). Nearly all naturally occurring colours can in principle be faithfully reproduced with the phosphors currently used to give the basic colours in the picture tube.

If, for example, a shadow-mask tube [2] is used as the picture tube, then, to obtain a colour picture, the three electron guns of this tube have to be controlled by the appropriate video signals. This means that the "red" signal must be applied to the "red" gun, the "green" to the "green" gun and the "blue" to the "blue" gun. These three television signals, representing the red, green and blue components of the picture, and which originate in the studio from, for instance, a colour television camera with three camera tubes, have to be supplied to the receiver at the same time (the simultaneous system). There are various ways of transmitting the three signals from the transmitter to the receiver, each one leading to a different colour television transmission system. One such system, the NTSC system, has already been introduced in the United States and Japan. In many other countries a decision on the transmission system to be used is imminent. There is a great deal of lively international discussion — and disagreement — on this subject.

The three most important systems under discussion

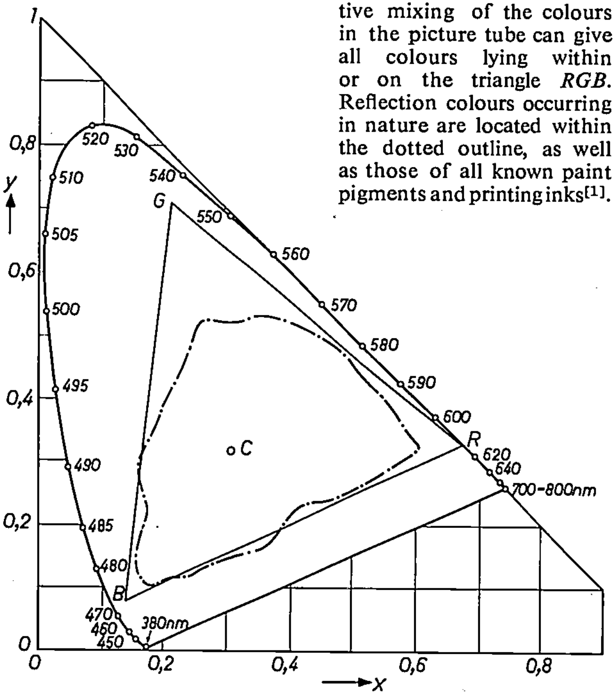


Fig. 1. Colour diagram, with the spectral pure colour curve, (wavelengths shown in nanometres), and the "standard white" colour at C (colour co-ordinates $x = 0.310$, $y = 0.316$). R, G and B are the basic colours chosen for colour television. If the three phosphors used provide these colours, additive mixing of the colours in the picture tube can give all colours lying within or on the triangle RGB. Reflection colours occurring in nature are located within the dotted outline, as well as those of all known paint pigments and printing inks[1].

Dr. F. W. de Vrijer is with Philips Research Laboratories, Eindhoven.

[1] F. W. de Vrijer, Fundamentals of colour television, Philips tech. Rev. 19, 86-97, 1957/58.

[2] H. B. Law, A three-gun shadow-mask color kinescope, Proc. IRE 39, 1186-1194, 1951. See also: R. R. Bathelt and G. A. W. Vermeulen, An experimental fluorescent screen in direct-viewing tubes for colour television, Philips tech. Rev. 23, 133-141, 1961/62.

are NTSC, PAL and SECAM. Before discussing the differences between them we shall deal with some of their similarities. A more detailed discussion of the points of similarity can be found in the article quoted [1].

General principles of NTSC, PAL and SECAM

The "red", "green" and "blue" signals R, G and B from the camera (or other signal source) are first subjected to a non-linear process called gamma correction, to compensate for the non-linear characteristic of the average picture tube. This characteristic, i.e. the relationship between the luminous flux Φ and the control voltage V (measured from the cut-off point) is approximately a power law for the conventional picture tubes:

$$\Phi \propto V^{\gamma} ,$$

the exponent γ being about 2.5. If the camera tubes give signals proportional to the illumination of the photosensitive layer (as with the "Plumbicon" camera tubes), the relationship between the output voltage V_0 and input voltage V_i of the gamma correctors must be:

$$V_0 \propto V_i^{1/\gamma}.$$

Gamma correction, then, ensures that the intensity of the red, green and blue at every point of the reproduced picture is proportional to the intensity of the red, green and blue at the corresponding point of the original scene. This is very important for good colour reproduction.

The gamma-corrected signals R', G' and B' are linearly combined in a "matrix" to form three other signals (see *fig. 2*). One of these, the "brightness" or "luminance" signal:

$$Y' = 0.30\,R' + 0.59\,G' + 0.11\,B', \quad \ldots\ldots \quad (1)$$

modulates the carrier-wave in the same way as in black-and-white television. This ensures that the colour transmission is "compatible", i.e. that it can be received by a standard black-and-white receiver as a monochrome picture. The two other linear combinations of R', G' and B' constitute the colour information signals S_1 and S_2. These modulate a "sub-carrier" lying within the frequency band covered by the luminance signal Y' (*fig.3*). The combinations:

$$\left.\begin{aligned} S_1 &= \alpha(R' - Y') \\ S_2 &= \beta(B' - Y') \end{aligned}\right\}, \quad \ldots\ldots\ldots\ldots \quad (2a)$$

which may also be written:

$$\left.\begin{aligned} S_1 &= \alpha(0.70\,R' - 0.59\,G' - 0.11\,B') \\ S_2 &= \beta(-0.30\,R' - 0.59\,G' + 0.89\,B') \end{aligned}\right\}, \ldots \quad (2b)$$

are generally chosen as the colour information signals. These are "colour-difference signals": the *sum* of the coefficients of R', G' and B' is zero, and therefore for neutral parts of the picture (white and grey), for which $R' = G' = B'$, both S_1 and S_2 are zero. (We shall return later to the choice of the constants α and β.) Before these signals are used to modulate the sub-carrier, their bandwidths are limited, to about 1Mc/s, say. As explained in the reference quoted [1] this leads to blurred colour transitions in the picture: but this is quite permissible in normal pictures, as long as the change in brightness at such transitions remains sharp.

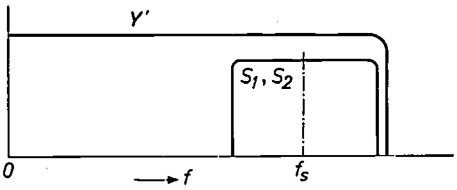


Fig. 3. Frequency spectrum of the complete colour television video signal for a system like that shown in fig. 2. f_s is the sub-carrier frequency.

It has already been agreed internationally that those countries using a 625 line black-and-white system, including Britain and France, will use a sub-carrier frequency of about 4.43 Mc/s.

The modulated sub-carrier is, in fact, an "interference" factor in the luminance signal. Certain measures have to be taken to keep the effect of this interference on the picture within reasonable bounds, especially with black-and-white receivers. This will be explained further when the different systems are dealt with. Conversely, the cross-talk on the colour information signals due to the components of the luminance

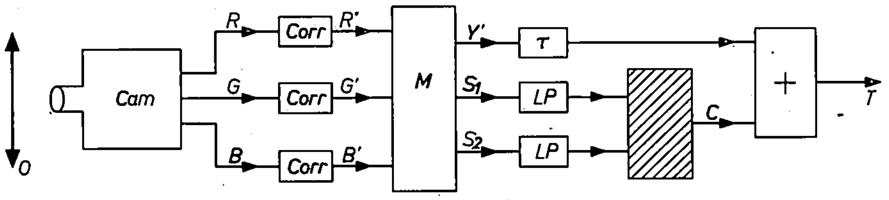


Fig. 2. Block diagram of a compatible transmission system for colour television. O is the scene, *Cam* the camera, *Corr* are gamma correctors, *LP* low-pass filters, and Z is the transmitter. The linear combinations Y' (the luminance signal) and S_1 and S_2 (colour difference signals) are produced from the gamma-corrected colour signals R', G' and B' by the "matrix" M. Up to this point, NTSC, PAL and SECAM are identical, but they differ in the manner in which the sub-carrier is modulated by the colour information signals S_1 and S_2 (in the cross-hatched area) in order to give the chrominance signal C. The luminance signal Y' is passed through a delay network τ to make its time delay equal to that of S_1 and S_2.

signal present in the relevant part of the video band signals ("cross-colour") also poses a problem.

The differences between the various transmission systems that will now be discussed relate to the way in which the two colour information signals modulate the sub-carrier.

Colour information modulation in the three systems

The NTSC system [3]

In the NTSC (National Television System Committee) system, the total video signal that modulates the carrier-wave is in the form:

$$Y' + 0.88(R' - Y')\cos \omega_s t + 0.49(B' - Y')\sin \omega_s t. \quad (3)$$

Here, $\omega_s/2\pi = f_s$ is the frequency of the sub-carrier. It can be seen from (3) that both colour information signals amplitude-modulate the sub-carrier orthogonally, i.e. one colour signal modulates the cosine signal and the other modulates the sine signal. These two modulations result in amplitude and phase modulation of the sub-carrier (*fig. 4*). The modulated sub-carrier, i.e. the sum of the last two terms in expression (3), is also called the chrominance signal. This may be written as

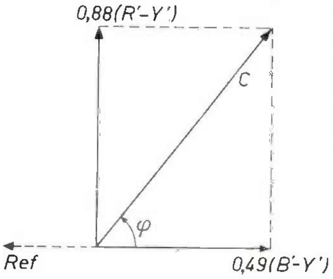


Fig. 4. Vector diagram of the chrominance signal in the NTSC system. The phase angle φ is determined by the hue, the amplitude C (relative to Y') by the saturation.

$$C \sin (\omega_s t + \varphi), \quad \ldots \ldots \quad (3a)$$

where

$$C = \sqrt{[0.88(R' - Y')]^2 + [0.49(B' - Y')]^2} \quad (3b)$$

and

$$\tan \varphi = \frac{0.88(R' - Y')}{0.49(B' - Y')}. \quad \ldots \ldots \quad (3c)$$

Since R'—Y' and B'—Y' can vary independently of each other and can be either positive or negative, all possible values of the phase angle φ can occur. To a first approximation, φ is determined by the "hue" (the dominant wave-length) in the colour, and the amplitude C (relative to the value of Y') by the saturation.

Synchronous detection has to be used to separate S_1 and S_2 at the receiver. In this method, the chrominance signal

$$S_1 \cos \omega_s t + S_2 \sin \omega_s t$$

is multiplied in one channel by $2 \cos \omega_s t$ and in another channel by $2 \sin \omega_s t$, by means of a suitable circuit. We thus obtain, in the first channel:

$$2S_1 \cos^2 \omega_s t + 2S_2 \sin \omega_s t \cos \omega_s t$$
$$= S_1 + S_1 \cos 2\omega_s t + S_2 \sin 2\omega_s t.$$

Provided that the sub-carrier frequency f_s is high enough, i.e. greater than the band-width used for S_1, then the only part of this signal left after passing through a suitable low-pass filter is the signal S_1. In a similar way, the signal S_2 is obtained from the other channel. R', G' and B' are recovered from Y', S_1 and S_2 by linear combination.

For synchronous detection, the auxiliary signals $\cos \omega_s t$ and $\sin \omega_s t$, in the correct phase, are required in the receiver. In order to be able to obtain these auxiliary signals, a "burst", consisting of about ten oscillations of the sub-carrier frequency, in a given phase, is inserted into the transmitted signal after each line synchronization pulse. A phase angle of $\varphi = 180°$ is chosen for this reference phase (see fig. 4). These bursts are used to synchronize an oscillator in the receiver to obtain the auxiliary signals, one directly, and the other after a 90° phase shift. Fig. 4 shows the reference burst in its correct phase in broken lines, while *fig. 5* shows the complete video signal for one line of the picture, the picture consisting of a bar pattern of different colours.

Figs. 6 and *7* show the block diagrams of the encoder and decoder in which modulation and detection of the chrominance signal take place.

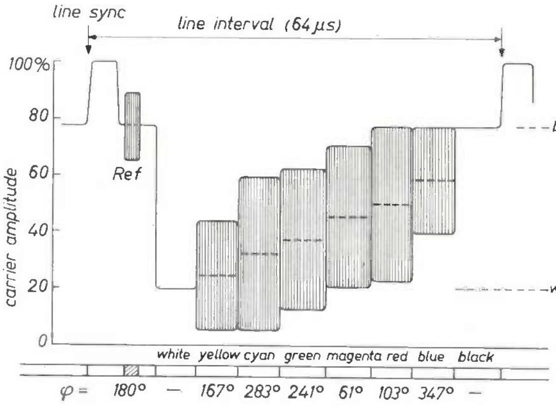


Fig. 5. Complete cycle of the NTSC video signal for one line of the picture, the picture consisting of a bar pattern of the given colours (not at maximum saturation). The strip beneath gives the values of the phase angle φ of the modulated sub-carrier in the successive colour bars. The reference burst (*Ref*) of the sub-carrier, given after the line synchronization signal, may also be seen.
N.B. This is not the normal colour bar signal often used in test measurements.

To reduce the visibility of the sub-carrier in the

[3] Color television standards — selected papers and records of the National Television System Committee (ed. D. G. Fink), McGraw-Hill, New York 1955.

monochrome picture on a black-and-white receiver as much as possible, the frequency of this signal in the NTSC system is linked to the line frequency f_L by the relationship:

$$f_s = (n + \tfrac{1}{2})f_L, \qquad (4)$$

where n is an integer. Thus, the maxima and minima due to the sub-carrier signal are shifted through 180° in relation to one another in successive lines, provided that the colour in these lines is approximately the same. This holds for most lines. Since the total number of lines for a complete picture is odd, the phase of the sub-carrier is, moreover, shifted by 180° when the same line is scanned in the next frame, provided that the picture has not altered much during this time. So, at each point where there is a maximum at one scan, there will be a minimum at the next, and vice-versa (*fig. 8*). Because of the persistence of vision in the eye, the average visible effect due to the existence of the sub-carrier signal is very slight. A fairly high amplitude may be permitted without any adverse effects. This accounts for the excellent compatibility of the NTSC system. The chosen frequency relation also reduces crosstalk effects between brightness and colour signals of the colour picture to a very low level.

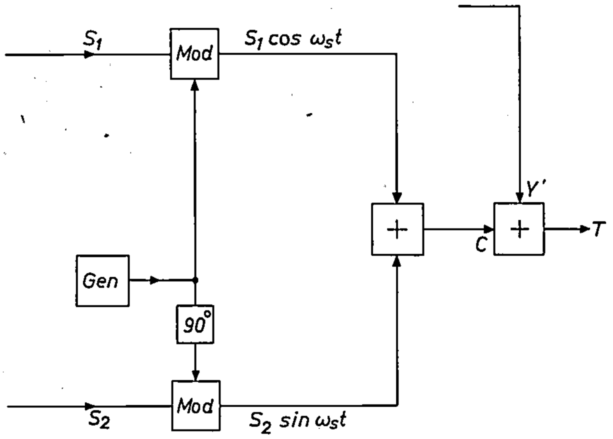


Fig. 6. Block diagram of an NTSC encoder. *Gen* sub-carrier generator. *Mod* modulators. The colour difference signals S_1 and S_2 modulate the sub-carrier of frequency f_s "orthogonally". The chrominance signal C is obtained by addition and the result is added to the luminance signal Y' for modulating the vision carrier.

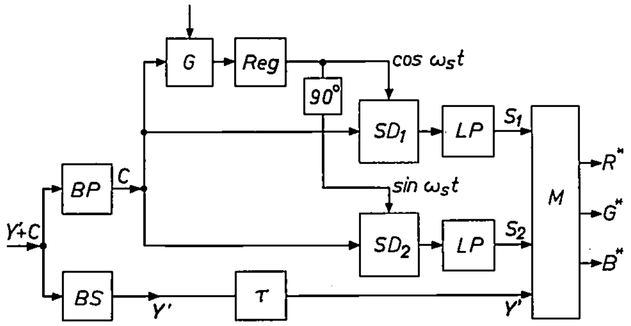


Fig. 7. Block diagram of a decoder for NTSC. Chrominance signal C and luminance signal Y' are separated by means of a band-pass filter *BP* and a band-stop filter *BS*. C is detected in two synchronous detectors SD_1 and SD_2 with a 90° phase difference. The sub-carrier required for this is obtained in the correct phase (regenerated) because the oscillator *Reg* is synchronized with the reference burst present in signal C. This burst is passed to the regenerator each time through the gate circuit G, which is controlled by the line synchronization pulses. After synchronous detection and after the high frequencies have been suppressed (in the low-pass filters *LP*), S_1 and S_2 are recovered. These signals, together with the luminance signal Y', which has again been passed through a delay network τ to equalize its delay time with that of S_1 and S_2, are supplied to the matrix M which, by linear combination, produces red, green and blue signal for the operation of the picture tube.

The asterisks at R, G and B indicate that these signals are equal to R', G' and B' in fig. 2 only as far as the low frequencies are concerned. The high frequencies for all three colour signals are provided from the Y' signal (the "mixed-highs" principle; see the article quoted [1]).

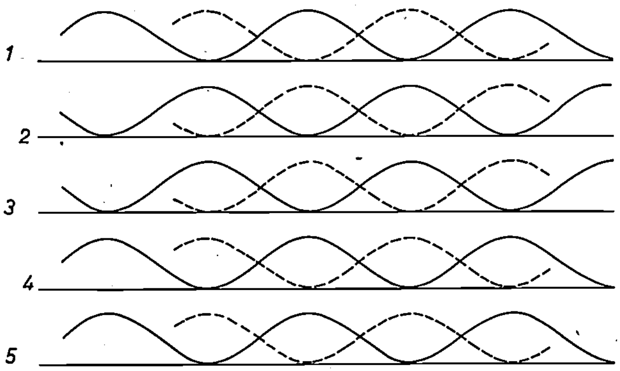


Fig. 8. Crosstalk from the sub-carrier (of frequency f_s) on the luminance signal in NTSC, if $f_s = (n + \frac{1}{2})$ times the line frequency. The brightness maxima that the sub-carrier causes on one line (e.g. line 3) lie vertically beneath the minima of the immediately preceding line (in this case line 1). The resultant averaging-out of the interference in the vertical direction is improved still further because each maximum is replaced by a minimum at the following *frame* scan (shown in broken lines).

In most of the experiments on the NTSC system, with 625 lines, including the test transmissions that have been sent out since 1955 from Philips Research Laboratories in Eindhoven [4], n has been taken as 283. The sub-carrier frequency is then 4 429 688 c/s ± 10 c/s. The tolerance is so close because of requirements for the sub-carrier regenerator in the receiver (low "noise band-width").

The compatibility of the NTSC system also benefits from the fact already mentioned that the colour-difference signals S_1 and S_2 are zero for white and grey, and have correspondingly low values for unsaturated colours. Practically all naturally occurring colours are, in fact, of low saturation. Statistically it has been shown that, with normal picture material, the average amplitude of the chrominance signal in the NTSC system is only 10% of the maximum possible amplitude [5].

These investigations have had an effect on the choice of the factors α and β in eq. (2). For compatibility, small values for α and β are best (this makes the chrominance signal weak), but the colour information is then highly sensitive to interference and crosstalk from the luminance signal. Due to the favourable statistics, the fairly high values given in (3) could be chosen for α and β. The ratio of the chosen values of α and β is such that the sensitivity to interference is much the same for all the colours in the colour diagram [6].

The PAL system [7]

In the PAL (Phase Alternation Line) system, the total video signal has the form:

$$Y' \pm 0.88(R'\text{—}Y') \cos \omega_s t + 0.49(B'\text{—}Y') \sin \omega_s t, \quad (5)$$

the sign of the term $0.88(R'\text{—}Y') \cos \omega_s t$ being different in successive lines. For half of the total number of lines, therefore, this signal is the same as that in the NTSC signal dealt with in the previous section. In the other lines, the chrominance signal is reflected in the $(B'\text{—}Y')$ axis (see *fig. 9*).

Because of the periodic inversion, both the encoder and the decoder for PAL are more complex than for NTSC (see block diagrams in *figs. 10* and *11*). Moreover, besides a reference phase, the signal transmitted must also contain an indication which shows whether the plus or minus sign of expression (5) applies. In the latest PAL proposals [8], this indication is combined with the burst. Its phase in fact is not, as in NTSC, always $\varphi = 180°$, but alternately 135° and 225°, so that the component of this signal in the $(R'\text{—}Y')$ direction always corresponds to a positive value of $(R'\text{—}Y')$. In this connection, therefore, the term "alternating burst" is used (abbreviation AB).

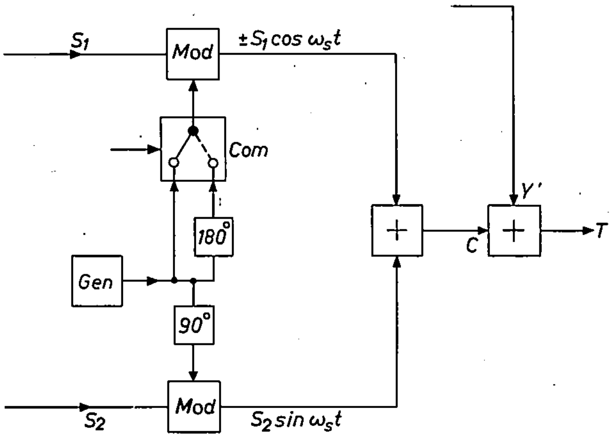


Fig. 10. Block diagram of encoder for PAL. It differs from fig. 6 (NTSC) in the addition of the 180° phase-shifting network and the switch *Com*, which is switched over after every line scan. The colour difference signal S_1 thus modulates the vision carrier alternately for each line as $+ S_1 \cos \omega_s t$ and $-S_2 \cos \omega_s t$.

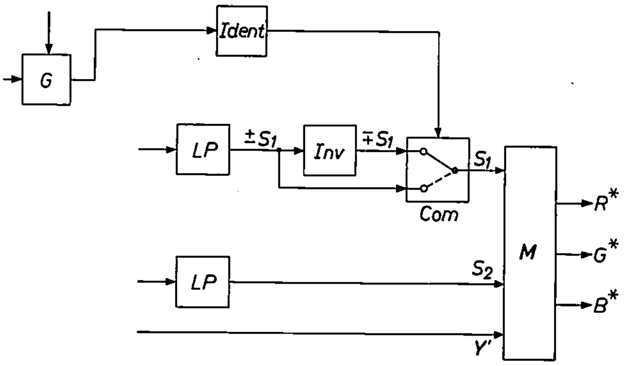


Fig. 11. The circuit diagram of a PAL decoder is, in the main, similar to the diagram of an NTSC decoder shown in fig. 7. Only the components shown here are inserted into the path of the signal S_1 (obtained by synchronous detection) to the matrix *M*: *Inv* is an inverter stage; switch *Com* is reversed after every line scan. Because this reversal is controlled by an identification signal derived from the alternating reference bursts by means of a circuit *Ident* (connected to gate circuit *G* in fig. 7), the signal $+ S_1$ is supplied to the matrix for every line.

The correct choice of the sub-carrier frequency is more difficult in the PAL system than in the NTSC system. The NTSC frequency cannot be used in PAL. The maxima and minima of the sub-carrier signals for colours where $B'\text{—}Y' = 0$ would lie directly one beneath the other for all the lines in the picture, because of the periodic reversal of $R'\text{—}Y'$, thus giving rise to a very annoying stationary vertical bar structure. For these colours, therefore, the right choice would an integral multiple of the line frequency. This, however, is a poor choice for colours where $R'\text{—}Y' = 0$. A better one is $f_s = (n \pm \frac{1}{4})f_L$, and, in the 625 line system, the result is even better if $f_s = (n \pm \frac{1}{4})f_L \pm \frac{1}{625} f_L$. The additional shift of $\frac{1}{625} f_L = 25$ c/s in the

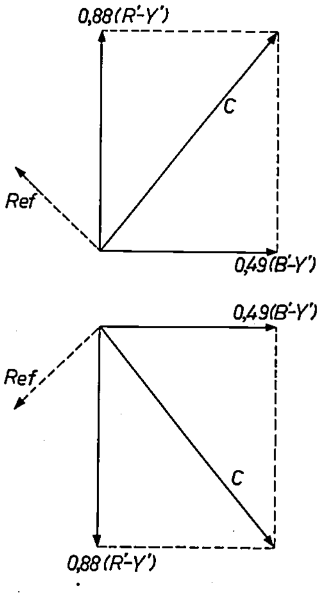


Fig. 9. Vector diagram of the chrominance signal in PAL, for two successive lines. The phase of component $0.88\ (R'\text{—}Y') \cos \omega_s t$ is arranged to be 180° different in successive lines.

[4] J. Davidse, Versuche über die Anpassung des NTSC-Farbfernsehsystems an die europäische 625-Zeilen-Norm, Nachrichtentechn. Z. **11**, 461-466, 1958.

[5] J. Davidse, NTSC colour-television signals, Electronic & Radio Engr. **36**, 370-376 and 416-419, 1959.

[6] F. W. de Vrijer, The choice of chrominance signals in the NTSC system with a view to the differential sensitivity of the human eye to colour, Acta electronica **2**, 103-109, 1957/1958.

[7] W. Bruch, Das PAL-Farbfernsehsystem — Prinzipielle Grundlagen der Modulation und Demodulation, Nachrichtentechn. Z. **17**, 109-121, 1964.

[8] EBU Ad-hoc Group on Colour Television, Document Com. T(E) 197, Report of sub-group 1 meeting (Hanover, June 1965), and Document Com. T(E) 208, Report of sub-group 1 (Rome, Dec. 1965).

sub-carrier frequency ensures that the interfering pattern is averaged out better in successive frames. The current choice for the PAL sub-carrier is:

$$f_s = 283\tfrac{3}{4} f_L + \tfrac{1}{625} f_L = 4\,433\,619 \text{ c/s } (\pm\ 10 \text{ c/s}). \quad (6)$$

In spite of these measures, the interfering pattern in the PAL system is nevertheless more annoying than that in the NTSC system. The principal reason for this is that, in NTSC, the pattern is averaged out within two complete pictures, i.e. within 80 ms, whereas, in PAL, this takes four complete pictures, or 160 ms. This therefore means that the stroboscopic effects that occur if the eye scans the picture at certain speeds are of a coarser structure in PAL. The compatibility of PAL is therefore not as good as that of NTSC.

The SECAM system [9]

The luminance signal Y' in the SECAM system is also transmitted in the normal way, together with a sub-carrier modulated by the colour information. The colour difference signals S_1 and S_2, however, are transmitted not simultaneously, but sequentially, i.e. first line R'—Y', then B'—Y', etc. Furthermore, the method of modulation used in the currently proposed SECAM system (SECAM III) is not amplitude modulation but frequency modulation.

When a SECAM signal is received, therefore, the signals actually available at any one moment are the luminance signal Y' and one of the two colour difference signals R'—Y' or B'—Y'. However, to recover the signals R', G' and B' that are needed for the operation of the picture tube, Y', R'—Y' and B'—Y' are required at the same time. In the SECAM receiver, the signal transmitted precisely one line interval previously is used to fill the gap left by the absent colour difference signal. This means that the receiver must contain a memory that will retain the signal for one line interval (64 μs). Hence the name SECAM, an abbreviation of "Séquential couleur à mémoire". An ultrasonic delay line is used as the memory [10].

The method is based on the assumption that there will generally be sufficient correlation between the picture content in successive lines for this substitution to be permissible. However, in every picture there will of course be some places with vertical colour transitions. At such places, the colour difference signal transmitted for the first line of the new colour is combined with the other colour difference signal from the previous line, i.e. the one belonging to the former colour. The result is a wrong colour. Moreover, in successive frames, the (R'—Y') and (B'—Y') signals for the line concerned are alternately wrong. The errors thus caused are, therefore, different in successive frames, giving rise to a flicker at a frequency of 12½ c/s, i.e. half the frame frequency. This effect, which can be a nuisance particularly where a transition occurs from one saturated colour to another, is inherent in the operating principle of SECAM.

Compatibility is also a considerable problem in the SECAM system. Because frequency modulation is used, the phase of the sub-carrier is not determined by the modulating signal at any one moment. In fact, with frequency modulation of a carrier of frequency $\omega_s/2\pi$, a modulating signal $S(t)$ gives:

$$A \cos \left\{ \omega_s t + \int_0^t S(u)\mathrm{d}u \right\}.$$

The phase difference of the modulated sub-carrier at points on two successive lines lying one above the other (or at the same point in successive pictures) is, therefore, governed by the picture content. The selection of a suitable ω_s, as in NTSC or PAL is therefore impossible. It is possible to choose the sub-carrier phase at the beginning of each line in such a way that the results there, i.e. on the left of the picture, are satisfactory. Nevertheless, depending on the picture content there may be poor areas elsewhere in the picture where, especially with a black-and-white receiver, the sub-carrier wave becomes obtrusive. This often occurs. Moreover, with moving subjects, these areas also move, which accentuates the unwanted effect.

It is, of course, possible to select a low amplitude for the frequency modulated sub-carrier in order to improve compatibility. This has the result, however, that the colour information signal deteriorates particularly in its high-frequency components, because of crosstalk from the luminance signal, noise, and other possible forms of interference. It has been found impossible to achieve a useful compromise without making the system more complicated. The following measures have been taken:

1) Pre-emphasis of the colour difference signals before modulation, i.e. relative amplification of the high frequencies. (This is done by a simple RC network with a time constant of 2.24 μs). To avoid too great a frequency deviation (i.e. too great in view of the available band-width) in transients at an abrupt colour transition, the top and bottom of the signal are clipped after pre-emphasis, so that $|\Delta f|$ is never greater than 500 kc/s.

2) After modulation, the signal is fed through a filter with a characteristic like that shown in *fig. 12*

[9] H. de France, Le système de télévision en couleurs séquentiel-simultané, Onde électr. **38**, 479-483, 1958.
R. Chaste, P. Cassagne and M. Colas, Sequential receivers for French color TV system, Electronics **33**, No. 19, 57-60, 1960.

[10] C. F. Brockelsby and J. S. Palfreeman, Ultrasonic delay lines and their applications to television, Philips tech. Rev. **25**, 234-252, 1963/64.

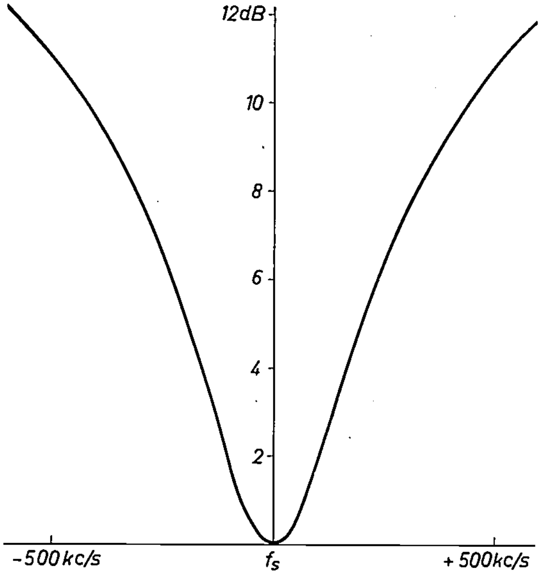


Fig. 12. Transmission characteristic of the "shaping" or "mise en forme" filter in SECAM. This filter has its attenuation peak at the frequency f_s of the unmodulated sub-carrier. The amplitude of the signal transmitted through the filter at instantaneous frequency f_s must be 10% of the black-white spacing.

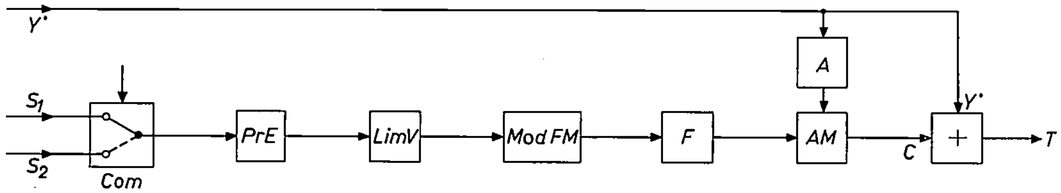


Fig. 13. Block diagram of an encoder for SECAM. Switch *Com* is switched over after every line scan. The pre-emphasis filter *Pre-E* and the video-signal limiter (*Lim V*) are used for measure (1) described in the text and the "shaping" filter *F* for measure (2). Block *A* supplies information about the strength of the high-frequency components of the luminance signal *Y'* to a circuit that correspondingly controls, to a certain extent, the amplitude of the *FM* modulated sub-carrier (*AM*, measure (3)).

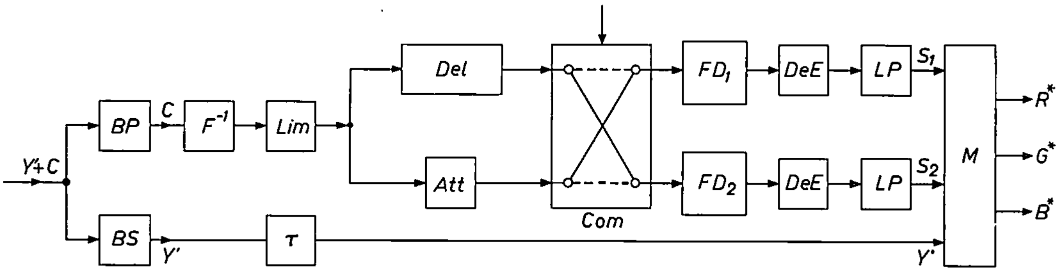


Fig. 14. Block diagram of a decoder for SECAM. The filters *BP* and *BS* separate the chrominance and luminance signals. The filter F^{-1} ("reshaping") cancels out the shaping provided at the transmitter. The amplitude of the signal is then limited (in *Lim*), which helps to suppress noise. The signal then travels, through a delay line *Del* with a delay of one line interval (64 μs) and an equivalent attenuator *Att* to the double changeover switch *Com*, which is switched over after every line scan and is so synchronized that the frequency discriminator FD_1 always provides the $(R'-Y')$ signal and FD_2 the $(B'-Y')$ signal. After compensation of the pre-emphasis, performed at the transmitter, in the de-emphasis stages *DeE* and band-limiting by means of the low pass the signals obtained are filters *LP* combined in matrix *M* with *Y'* to give the signals R^*, G^* and B^*.

("shaping" or "mise en forme"). At an average low sub-carrier amplitude (unsaturated colours) this gives better protection against crosstalk from the luminance-signal, noise and other forms of interference. We mentioned above that in NTSC some benefit is obtained from the fact that $|R'-Y'|$ and $|B'-Y'|$ are statistically rather small with respect to the maximum possible values (the same is true in PAL). In SECAM, in the first instance, no such benefit would be expected, as with frequency modulation the carrier has constant amplitude. However, because of the shaping, SECAM also benefits from this statistical result, though not to the same extent as NTSC and PAL.

3) If the luminance signal has strong components in the frequency band of the chrominance signal, the amplitude of the sub-carrier is temporarily increased by a maximum of 6 dB.This lessens the effect of cross-talk from these components on the colour difference signals. The arrangements provided for this, together with those for (1) and (2), are indicated in the block diagram in *fig. 13.*

We have seen that measures (2) and (3) introduce again a certain amount of amplitude modulation of the sub-carrier. Here however, this modulation is not used for providing information: it is used purely to improve the properties of the system.

In spite of these measures, the compatibility of the SECAM III system turns out to be not as good as that of PAL, and *a fortiori* worse than that of NTSC.

The effect of measures (2) and (1) has to be undone in the receiver ("re-shaping" and "de-emphasis"); see the circuit diagram in *fig. 14.* It is, however, not possible to do this for the clipping at $\Delta f > 500$ kc/s. This has an adverse effect on the quality of horizontal colour transitions.

The effect of differential phase errors

What, then, are the technical reasons that have lead a number of experts to prefer the PAL or SECAM system to the longer established, further developed and, as practice has shown, quite satisfactory NTSC system? The main reason is the concern about the phase modulation of the NTSC sub-carrier. As we have already said, the phase of the sub-carrier in the NTSC system determines the hue. What can go wrong with this? A phase shift of the complete NTSC sub-carrier signal cannot in itself give trouble, since the hue is determined by the phase difference from the reference signal (the burst) which is also present in the signal, and not by the absolute phase. Only if the phase shifts for burst and chrominance signal are different will colour errors occur. Such distortion in the signal can arise in certain unperfected equipment because the chrominance signal is superimposed on the luminance signal, while the reference signal always is given at black level (see fig. 5). If a phase shift dependent on the level now occurs, there will be colour errors in the NTSC image.

Although it has been found quite easy to keep these "differential phase errors" to a minimum with modern equipment, a more attractive idea appeared to be to make the system *insensitive* to such errors. For both PAL and SECAM, the basic concept was the avoidance of difficulties due to phase shifts dependent on the level. However, "what you gain on the swings you lose on the roundabouts", and so this can only be done at the expense of something else. In the SECAM system, the desired result is achieved by the use of frequency modulation: in fact in areas where colour is uniform the frequency remains constant and the phase has no effect. Differential phase errors can give rise to colour errors only when the brightness varies, but this effect is only slight and rarely visible in practice. The price that has to be paid for this advantage has already been partly explained above: the possibility of annoying errors at vertical colour transitions, and poorer compatibility.

In the PAL system, differential phase errors are in the first instance less of a hazard because the colour variations are averaged out. If the chrominance signal contains a differential phase error α, then instead of the desired signal, e.g. $C \sin \varphi$, the circuit shown in fig. 12 gives the signals $C \sin(\varphi+\alpha)$ and $C \sin(\varphi-\alpha)$ for alternate lines in the $(R'-Y')$ direction, see *fig 15.* On average, this is $\frac{1}{2}C[(\sin(\varphi+\alpha) + \sin(\varphi-\alpha)] = C \sin \varphi \cos \alpha$. Likewise, the result will be on average $C \cos \varphi \cos \alpha$ for the $(B'-Y')$ direction. The colour is therefore incorrect for every line, because the colour components have the phase $\varphi \pm \alpha$ instead of φ, but the alternate colour errors are in opposite directions. The colour yellow, for example, will be alternately too green in one line and too orange in the next. A large area of colour, however, seen from far enough away, will nevertheless give the right hue, since the ratio between $R'-Y'$ and $B'-Y'$ is on average exactly right. The only ill-effect will be a degree of saturation that is slightly too low (since $\cos \alpha < 1$), and this in general is not very troublesome.

The above sets out the basic principle of PAL, but unfortunately, in practice things may be a little different. Because of the non-linearity of the picture tube characteristic ($\gamma \neq 1$), a phase error gives rise to brightness as well as colour errors and the brightness errors are also alternate on successive lines. This gives a fairly coarse pattern of stripes which moreover appears to move in the vertical direction, because of the interlacing. This effect has the result that the sensitivity of the PAL receiver to differential phase errors is as great as with NTSC, although the effects on the picture are quite different. In NTSC they appear as colour errors, but in PAL as a moving pattern of stripes ("Venetian blind effect" or "Hanover bars").

Nevertheless it is possible to improve the performance of PAL receivers in this respect. Instead of leaving the averaging out of errors to the eye, a decoder can be used which does this for the viewer by directly averaging the electrical signals. This can be done with the aid of a delay line, of the same type as that used in the SECAM receiver, giving a delay of one line interval (64 μs) (see *fig. 16*). The difference between this scheme and SECAM is, of course, that the delay line is essential to the operation of SECAM, whereas it is introduced here purely for correction. The simple PAL decoder is now known as PALs, that with a delay line being referred to as PALd.

If the PALd decoder is used, the receiver is not very

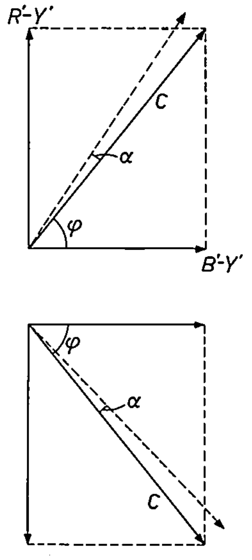


Fig. 15. Vector diagram of the chrominance signal of PAL for two successive lines, where there is a differential phase error α. The vectors shown in broken lines occur instead of the solidly-drawn vectors C. Because φ changes its sign each time, while α does not, there is a certain amount of phase error compensation in PAL.

sensitive to differential phase errors. Averaging over adjacent lines does involve a slight loss in vertical definition in the colour difference signals, but this is generally not noticed in the picture. The very troublesome effects that sometimes occur at vertical colour transitions in SECAM do not occur in this system.

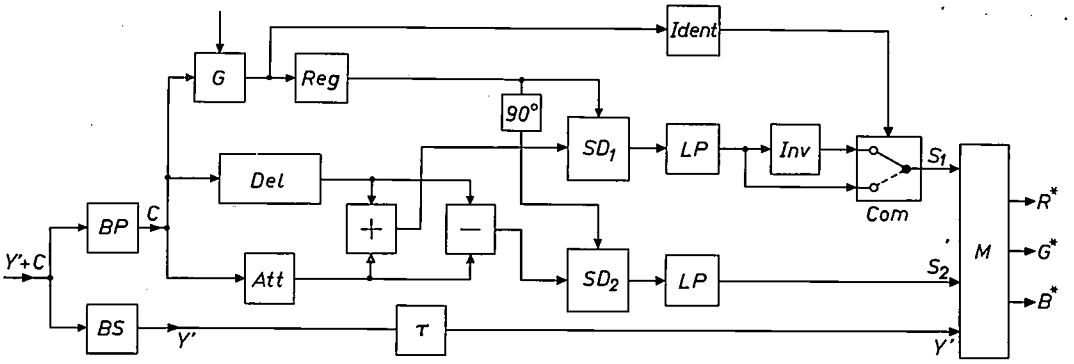


Fig. 16. Block circuit diagram of a decoder for PAL with a delay line (PALd). The chrominance signal passed through filter *BP* is supplied to an addition and subtraction circuit through the delay line *Del* (delay equal to one line interval) and the attenuator *Att*. The sum and difference signals are synchronously detected and further processed in the same way as in the simple PAL decoder without a delay line (PALs). The components shown in the diagrams in figs. 7 and 11 can also be found here.

From a large number of measurements it appears that the maximum permitted differential phase errors occurring in the complete transmission chain are for the different systems:

NTSC	± 12°
PAL with PALs decoding	± 12°
PAL with PALd decoding	± 40°
SECAM III	± 40°

Closer comparison between the three systems

A systematic comparison of the characteristic features of the three colour television transmission systems described shows that no one of them is the best in all respects. Similarly, no one of them is the worst in all respects. To arrive at a final assessment, it is therefore necessary to assess the relative importance of the advantages and disadvantages. It is impossible to find a completely objective method of comparison. We now give a survey of the most important features that were considered for comparison, and also the results of the comparative tests carried out on a large scale by us [11] and by others. Within the scope of this article, of course, the discussion can be only of a summary nature, but all relevant points will be given.

1) *Compatibility*

The degree of compatibility, i.e. the quality of the monochrome picture received by a black-and-white receiver from a colour television transmission, has already been mentioned in the discussion of the three systems. The conclusion that must be drawn from this is that the NTSC system is very good in this respect, PAL not so good, but still reasonable, while SECAM is clearly the worst. Many people, particularly in France, think that it is nevertheless adequate.

2) *Quality of the colour picture under good conditions*

The reproduction of individual colours is equally good in each of the systems when a comparatively noise-free signal with little interference is received and good and properly adjusted equipment is used. There are differences in the quality of the transitions from one colour to another and in the degree of visibility of the modulated sub-carrier on the picture. Once again, NTSC emerges as the best system in both respects. With the visible effect of the sub-carrier in the picture, the situation is much the same as for compatibility. There is, indeed, a band-stop filter in the luminance channel of the colour receiver to attenuate the sub-carrier, but the stopband of this filter must not be too wide, as otherwise the definition of the picture would deteriorate. There is still therefore an unwanted effect from the side-bands of the chrominance signal, and this manifests itself at transitions in the picture, where the pattern due to the sub-carrier is again noticeable. Regarding these effects, NTSC is the best, PAL slightly worse, and SECAM the worst. The sharpness of the horizontal colour transitions is much the same in both NTSC and PAL. This is determined only by the band-width of the chrominance signal. For SECAM the situation is much worse, particularly at transitions between saturated colours, as the frequency sweep used is rather large compared with the bandwidth available for the chrominance

[11] Report of EBU Ad-hoc Group on Colour Television, 2nd ed., Feb. 1965.

signal. The higher order side-bands of the FM signal thus fall outside the band at sharp transitions and this leads to a loss of sharpness after detection. The limiting of the frequency sweep mentioned in the discussion of the SECAM system makes the transitions even poorer [12] *(fig. 17)*. This argument against SECAM carries particular weight in countries where the difference in frequency between the vision and sound carriers is 5.5 Mc/s, as in most Western European countries using the 625 line standard. In France, this difference is 6.5 Mc/s, which makes the situation somewhat less critical in this respect.

For the vertical transitions, we need only refer to what has already been said on the subject. The order of preference is: NTSC, PAL, SECAM.

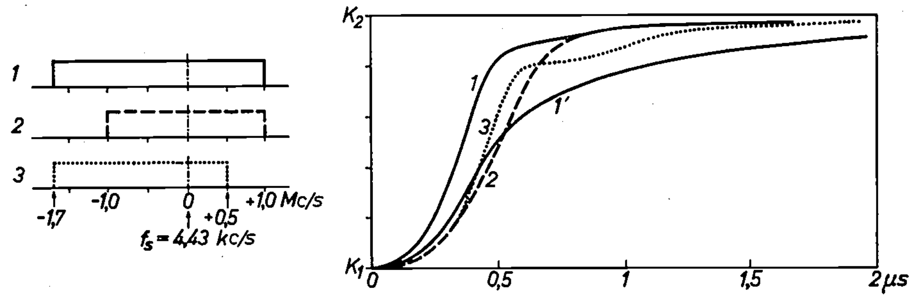


Fig. 17. The reduction in definition that occurs in SECAM in the reproduction of an abrupt horizontal colour transition as a result of the limited bandwidth for the chrominance signal (the clipping of higher side-bands from the frequency-modulated signal). The variation with *time* of the transition from a colour K_1 to a colour K_2 is plotted. Curves *1*, *2* and *3* apply to the case in which the frequency deviation for one of the two colour difference signals (S_1 or S_2) in the coloured areas K_1 and K_2 is -200 kc/s and $+$ 100 kc/s respectively, with three different limitations of the chrominance band, as given in the small graphs on the left. In curve *1'*, with the same bandwidth limitation as in *1*, the frequency deviation for colours K_1 and K_2 was -300 and $+$ 300 kc/s respectively. Here also the limitation of the frequency sweep to 500 kc/s can be seen. In fact, this limitation primarily concerns the high-frequency video components occurring at abrupt transitions, since they are amplified by the pre-emphasis and, on modulation, produce even greater frequency deviations.

It will be seen that the colour transition from one colour to another in curve *1'* takes more than 1 μs, which, in a 625 line system, corresponds to about ten picture elements (a horizontal distance of more than ten line-widths).

3) *Picture quality for signals with noise and interference*

The way in which the picture quality is affected by noise and interference in the signal becomes important particularly where the receiver is in a "fringe area", i.e. at some distance from the transmitter. The luminance signal undergoes the same deterioration under these circumstances as does a black-and-white transmission, and the effects of this on the colour picture are very similar to those which occur in black-and-white television. In addition, in the colour receiver, there are also the picture effects which arise in the channels for the colour difference signals. These effects are much the same in both NTSC and PAL. On the subject of noise, one might perhaps expect an improvement of 3 dB in the signal-to-noise ratio in the PAL system using a decoder with a delay line — as the signals for successive lines add, whereas the noise, which is uncorrelated in successive lines, only adds quadratically. Nevertheless, this improvement cannot be seen in the colour picture itself. This is probably because the noise in the picture does receive a certain degree of correlation from line to line because of the staggered addition. This makes the noise more visible, which apparently cancels out the 3 dB gain. Furthermore, it should be pointed out that, where there is "white" noise in the received signal, most of the interference effects occur in the luminance channel.

In SECAM III the situation is much the same as in NTSC and PAL if there is not too much noise. If the noise exceeds a certain threshold level, the impairment effects suddenly become much worse (the "silverfish effect"). This is a common occurrence in FM systems [13]. The actual location of this threshold depends to a large extent on the equipment used. The properties of the limiter that precedes the FM detector are especially important here. In practice, this effect mainly occurs when, besides noise, there is also a degree of attenuation of the higher video frequencies. The latter can be produced in very long connections or where there are echo signals (e.g. in hilly country).

4) *Distortion of the signal*

Certain types of distortion have a more serious effect on the colour television signal than on a black-and-white signal. Only these types of distortion will be discussed here.

The attenuation just referred to, of the higher video frequencies, only gives rise to slightly reduced definition in black-and-white television. In NTSC and PAL, a reduction in the subcarrier amplitude causes a reduction in the saturation of the colours. If the attenuation is not too high, this can easily be compensated. Many types of receiver even have automatic control for this. Frequency-dependent attenuation in the chrominance band leads to poor quality in the horizontal colour transitions. There is a certain amount of compensation for these errors in a PAL receiver with a delay line, and this is one of the advantages of the PAL system. In SECAM also the horizontal transitions become worse if the bandwidth is limited. For these effects, the order of preference is: PAL with delay line, NTSC, PAL without delay line, SECAM.

No extensive research has yet been carried out into the effect of phase errors independent of the luminance signal level. In this respect no great difference has so far been found between the systems.

This is not so, however, with phase errors dependent on the level ("differential phase"). This has already been discussed in detail above. SECAM and PAL with delay line have the advantage here over NTSC and PAL without delay line.

For completeness, we should say something about level-dependent *amplitude* errors, i.e. changes in the amplitude of the sub-carrier signal, which depend on the level of the luminance signal. In NTSC and PAL, such errors produce variations in colour saturation, but in SECAM, the effect is less because frequency modulation is used. The tolerances for a good picture quality are of the order of 30% in NTSC and PAL and 65% in SECAM. This wide tolerance in SECAM is not very significant, as such large variations can generally be avoided in practice. If they do occur, this indicates that the non-linearity is so high that even monochrome pictures would be adversely affected.

Differential phase errors are undoubtedly the type of distortion most discussed. Modern equipment has been shown to satisfy the most rigorous requirements (i.e. those of NTSC). In the receivers, it is, in fact, fairly simple to remain within the tolerances. Sometimes, excessively large errors occur in older professional equipment, such as beam links and television broadcast transmitters. A special correction method for NTSC in such difficult cases has been developed and used with much success in Britain. Picture quality as good as that attainable with PAL has been obtained with NTSC over, for instance, a very poor link between London and Moscow. On this link both systems performed better than SECAM. The poorer result obtained with SECAM was caused by the simultaneous effects of band-width limitation, noise and interference: these interfered with NTSC and PAL to a lesser extent [14].

5) *The effect of echoes*

The effect of echo signals in black-and-white television is well-known. "Ghost pictures" appear, at a certain horizontal displacement from the true picture (see also the article quoted [10]). If this distance is small (echoes with a small time-delay), troublesome effects are observed only at brightness transitions (overshoot, relief). The same phenomena also arise in colour television, and may be put down to the effect of the echoes on the luminance signal. In general, this is also the most marked interfering effect of the echo in the colour picture. Colour changes can also occur, and these are caused in a complex way by the effect of the echoes in the chrominance signal, due for example to attenuation of the higher video frequencies as mentioned above, or due to some effect on the reference burst. These latter effects differ in the different systems. Extensive tests, largely carried out in Switzerland, show that, under very difficult conditions, where the picture quality varies from fairly poor to poor, PAL with delay line often gives better results than NTSC and SECAM. This was also considered to have been shown statistically from tests made in several cities. The differences are not very great, but this feature has been considered particularly important in a number of mountainous countries, and has led to a preference for the PAL system in these countries (in particular Scandinavia, Switzerland, Austria, Italy and also Germany).

6) *Studio technique*

The camera (and most of the rest of the studio equipment) is the same for all the systems. There are, of course, differences between the encoders, as already described. Appreciable differences in the characteristic features of the systems are noticeable in the studio only in magnetic recording, fading, or mixing of colour signals from different signal sources. We now consider these operations separately.

a) Magnetic recording

Since, in NTSC and PAL, the phase of the sub-carrier contains important information, this phase must be properly reproduced when the signal recorded on the magnetic tape is scanned. For NTSC the toler-

[12] H. Schönfelder, Signalverzerrungen bei Fernsehsystemen mit frequenzmoduliertem Unterträger, Archiv elektr. Übertr. **15**, 273-284, 1961.
H. Schönfelder, Der Einfluss von System- und Übertragungsfehlern bei einer Farbfernsehübertragung nach dem SECAM-Verfahren, Archiv elektr. Übertr. **16**, 385-399, 1962.
D. A. Rudd, internal report, Mullard Research Laboratories.

[13] J. van Slooten, FM reception under conditions of strong interference, Philips tech. Rev. **22**, 352-360, 1960/61.

[14] EBU Ad-hoc Group on Colour Television, Document Com. T(E) 164, Memorandum on the Paris/London/Moscow colour transmission tests, Dec. 1964.

ance is of the order of ± 5°, i.e. the time errors must remain within 3 ns. This is a very stringent requirement, and it cannot be satisfied by mechanical means. An answer has been found in correction, with the aid of a controllable electrical delay line, of any time errors that may occur. It is thus possible to fulfil the NTSC requirements, and also therefore the less exacting PAL requirements.

Another not unrelated problem is that of obtaining uniform reproduction from the four heads of the video tape recorders currently used in the studio. If the heads are not adequately matched, horizontal bars appear in the reproduced picture ("head banding").

In the most modern equipment a successful solution to this problem has been found.

An advantage of the SECAM system is that the requirements which are made of the tape recorder, to avoid time errors, are somewhat less stringent. In some cases it was even found possible to use a completely unmodified good black-and-white recorder for SECAM. It must not be thought, however, that this will always be possible. The very fact alone that there is a high amplitude sub-carrier of about 4.4 Mc/s gives rise to problems (e.g. moiré effects), that generally call for special provisions. Furthermore, even if the colour information in SECAM is not seriously adversely affected by time errors, there is still nevertheless a deterioration in the compatibility.

b) *Fading and mixing of pictures*

With NTSC and PAL, the mixing of pictures in a desired intensity ratio presents no difficulties. The ordinary mixing methods as used for black-and-white television can be used, provided that care is taken to ensure that there is no phase difference between the bursts for the signals to be mixed. For SECAM however this is not possible, as the direct additive mixing of two FM signals produces no meaningful result. SECAM mixing therefore requires partial decoding, separate simultaneous mixing of luminance signal and colour difference signals, and finally, recoding (see *fig. 18*). This is an undesired complication in the studio, and is also necessary in fading a signal, since any diminution in the amplitude of an FM signal, of course, is ineffective. Furthermore, an unfortunate deterioration in picture quality occurs in the procedure described. SECAM therefore compares unfavourably with NTSC and PAL in this respect.

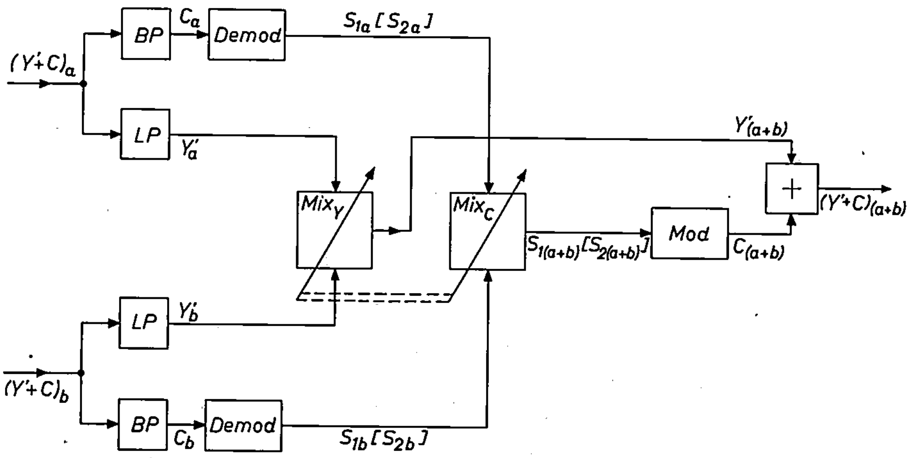


Fig. 18. A mixing circuit for two SECAM signals *a* and *b*, for studio use. The chrominance signal *C* must first be separated from each signal and demodulated to S_1 and S_2 (alternately present during one line interval). Mixing then takes place simultaneously for the luminance signal Y' and the colour difference signal S_1, S_2 present, in the mixing stages Mix_Y and Mix_C. Frequency modulation of the output signal from Mix_C is then necessary before it can be added to the resultant luminance signal.

A disadvantage of this complicated procedure is that the bandwidth of the luminance signal is limited by the low-pass filters *LP*. This is quite unavoidable, as the original chrominance signals must be kept well away from the mixing stage Mix_Y and final addition stage.

In PAL and SECAM the composition of the signal is not the same for every line, as it is in NTSC. As stated, an identification signal is required at the beginning of every line in PAL and SECAM. This gives more problems in the studio. Where two pictures are to be mixed, not only must the picture and line scan be isochronous, but also switching over after each line must be in phase. This again has a few unfortunate consequences, particularly in the mixing of signals originating from magnetic recordings. Space prevents further discussion here of this subject.

7) *Receiver features*

The main characteristics of the receivers are their stability, simplicity of operation — and price. A colour television receiver, for whatever system it is designed, is a complicated piece of equipment. Much of it, including the picture tube and the circuits directly associated with it, is independent of the transmission systems. In principle, the only differences are in the part of the receiver in which the chrominance signal is decoded.

Most experience has been obtained with NTSC receivers, and it has been found that good stability and simple operation can be achieved. Apart from the ordinary controls found on black-and-white receivers, an NTSC set also has two extra knobs for colour control, i.e. one for saturation and the other for hue. These adjustments can be made automatic, thus dispensing with these knobs. It is questionable, however, whether this is desirable. It often appears useful to be able to adjust the receiver manually so as to obtain the best setting for actual viewing conditions, such as the ambient illumination in use, and also to suit personal taste.

As shown in the diagrams in figs. 12 and 16, the PAL receiver is more complicated and, therefore, more expensive than a comparable NTSC receiver. The delay time of the delay line for PALd may only vary at the most by 5 ns from the rated value if undesirable effects in the picture are to be avoided. Glass delay lines, which satisfy these requirements over a sufficiently wide temperature range [10], are currently available, but are not yet in large-scale production. Estimates of the difference in price between a PAL receiver with a delay line and a comparable NTSC receiver vary between 3.5 and 6%. The PAL set without a delay line is cheaper, but it is still more expensive than an NTSC receiver.

A delay line is essential in the SECAM receiver. The tolerances in the delay time are, however, larger here (± 50 ns) which perhaps means that the delay line can be somewhat cheaper than that used for the PAL system. The price of a SECAM receiver is also generally estimated to be somewhat higher than that of a comparable NTSC receiver (1.5 to 5%).

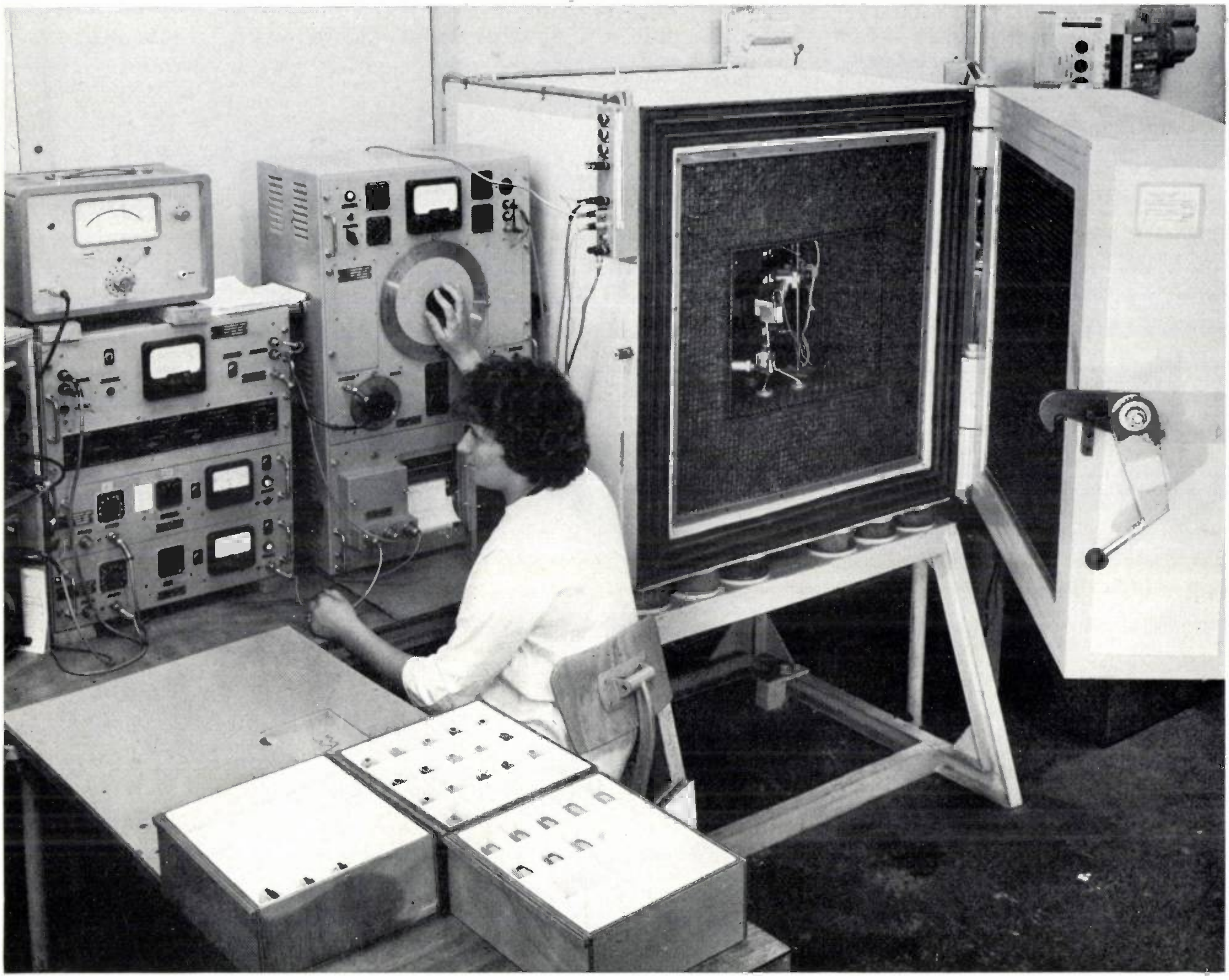
Relatively little experience has so far been gained with SECAM receivers (and this is indeed also true of PAL receivers). One of the difficult factors is the stability of the FM discriminators. If the variation from the centre frequency of 4437.5 kc/s is about 10 kc/s, this already appears in the picture as discoloration. "Reshaping" (see above) must also be carried out very accurately, as otherwise the colour transitions in the picture will be poor. "Contrast" control is also a problem, because, in the composite SECAM signal, the luminance signal is present as amplitude modulation while the chrominance signal is present as frequency modulation (cf. the difficulties, already discussed, encountered with fading). It is also difficult to provide manual colour control, and this is therefore left out in most designs for SECAM receivers. This feature is generally referred to as an advantage, but may well in the long run turn out to be a disadvantage.

Advantages versus disadvantages

As already stated, it is very difficult to assess all the advantages and disadvantages. Apart from its sensitivity to differential phase errors, the problems of magnetic recording and the effect of echoes, the NTSC system is the best in all respects. The first two of these three drawbacks present no great problems with modern equipment. In various countries the echo problem is considered very important, although the differences found in this respect are not particularly great. In Britain and the Netherlands, countries where a fairly great deal of experimental work has been done, the conclusion has been reached that the NTSC system presents the best solution. Many other Western European countries, however, are inclined to favour PAL. British and Dutch experts at present consider PAL a practicable alternative. Many people in Western Europe have very strong objections to the SECAM system, particularly with regard to the criteria discussed above, i.e. 1 (compatibility), 2 (the quality of the colour picture, particularly with a difference of 5.5 Mc/s between vision and sound carriers), 3 (noise), 6b (studio techniques) and 7 (receivers). These objections are not regarded as serious in France, and the official policy there is to adopt SECAM. Russia is also in favour of the SECAM system, although there are indications from that quarter that fairly extensive modifications to it are being considered. As some countries (in any case Britain and West Germany) want to start regular colour television transmissions in the course of 1967, the next CCIR conference, to be held in Oslo in June/July 1966, will be very important. If no agreement is reached then, the prospects of there being one single system in use throughout Europe will be very remote. Adoption of different systems in different countries would have very unfavourable technical and economic effects on the further development of colour television in Europe.

Summary. In the imminent introduction of colour television into Europe a choice will have to be made between three proposed transmission systems: NTSC, PAL and SECAM. After briefly describing the differences between these systems, the author attempts to give as objective a survey as possible of their advantages and disadvantages. To this end, the properties of the systems are compared in all relevant respects, an order of preference being given for each criterion. In the NTSC system, for instance, compatibility is very good, while it is not so good in PAL and clearly the worst in SECAM. SECAM is largely insensitive to differential phase errors. This feature can also be obtained with PAL if a decoder with a delay line is used, while, with NTSC, the differential phase errors must be made sufficiently small. This has been found possible even in the most unfavourable cases if the equipment is properly designed. Where echo effects are very prevalent, e.g. in hilly country, a statistical examination has shown PAL to give slightly better results than the other two systems, but the differences are very small. As far as studio techniques are concerned, SECAM has an advantage in magnetic recording because of its only slight sensitivity to time errors, but complications arise in the operations of fading and mixing that are so common in studio work. In all the other points as well, such as the quality of the colour picture with and without interference, the effect of distortion on the signal, the price and convenience of the receivers, etc., NTSC is to be preferred, while PAL may be regarded as its equal in many respects. The final result of any comparison between these systems will largely depend on the importance one may attach to the various factors.

Acoustic measurements on hearing-aids



Every hearing-aid is checked before it leaves the works, to make sure that its frequency characteristic lies within the permissible tolerances; the tests are made with the aid of a vibration-free, soundproof and reflection-free acoustic box (a small "quiet room"). The hearing-aid is placed in the box in front of a loudspeaker in the rear wall (not visible in the photo). An audio generator drives the loudspeaker at a frequency varying continuously from 100-9000 c/s. The sound intensity is held constant by means of a feed-back system that receives its input signal from a microphone placed next to the hearing-aid. The sound is received by the hearing-aid microphone and the frequency characteristic at the output is recorded.

In the photo the box is shown open and the hearing-aid is visible. Next to the box are the feed-back apparatus, the audio generator and an automatic plotter which records the characteristics.

Physical principles of photoconductivity

II. Kinetics of the recombination process; sensitivity and speed of response

L. Heijne

537.312.5:537.311.4

This second article in the series on photoconductivity is entirely devoted to the question of what happens to light-excited charge carriers before they return to their original quantum state, and to the particular way in which this return takes place: it is considered here as a kind of chemical reaction between holes and electrons. Particular attention is paid to the influence of impurity centres on the behaviour of charge carriers — and hence on the sensitivity and speed of response of a photoconductor — an aspect which is especially important both for scientific research and practical applications.

The situation that arises in a photoconducting material when it is illuminated might in a sense be compared with that in a tank where a liquid is run in with a constant flow i_1 (see *fig. 1*) and simultaneously run off with a flow i_2, the rate of which depends in one way or another on the level H to which the tank is filled. Expressed as an equation: $i_2 = \mathrm{f}(H)$. The flow i_1 in our example corresponds to the number of charge carriers which, at a given intensity of illumination, is released by the light per unit time in one cm^3, the level H in the tank corresponds to the charge carrier concentration, and the outflow i_2 is the number that disappears per unit time from one cm^3. The equilibrium value which H in our model reaches at a given i_1 is of course the value where $i_2 = i_1$, and is thus determined by the dependence of i_2 on H. At the same time $\mathrm{f}(H)$ determines the way in which, upon a sudden change in i_1, the outflow i_2 and the level H move to their new equilibrium value. Applied to a photoconductor: the manner in which the disappearance of the charge carriers depends on their concentration governs both their equilibrium value and the way in which this concentration changes to a new equilibrium value when the intensity of illumination is changed (the speed of response).

In fact, the model in fig. 1 is valid only for a substance which contains no impurity centres and moreover whose conduction band contains in the dark no electrons [1]. In practice, the situation is more complicated. In the first place, no substance ever is so pure and at the same time so perfect in crystal structure that one can assume there are no impurity levels at all in the (forbidden) zone between valence and conduction bands. Moreover, as well as optical transitions, there are usually thermal transitions between the energy levels. Nevertheless, the processes taking place can also be simulated by means of tanks with inflowing and outflowing liquid [2]. In a model of this kind there must be separate tanks for the energy bands and for each group of impurity levels corresponding to the same kind of impurity centre. Again, in the steady state, for each of the tanks, the inflow is equal to the outflow, and the level to which the tank is filled — i.e. in reality the concentration of charge carriers present in a band or in a group of similar impurity levels — is governed by the way in which these flows depend on the head of liquid.

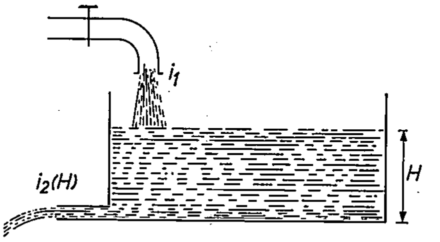


Fig. 1. When a constant flow i_1 of water is run into a tank and a flow i_2, depending on the level H to which the tank is filled, is simultaneously run off, the value of H in the stationary state is such that $i_2 = i_1$. A similar dynamic equilibrium is found in an illuminated photoconductor.

An important aspect of the disappearance of the freed charge carriers is that they can only disappear as a result of the uniting of two charge carriers of opposite sign (recombination). As we shall see presently, the

Dr. L. Heijne is with Philips Research Laboratories, Eindhoven.

[1] Basic concepts of semiconductor physics and the energy band scheme are dealt with in the first article of this series, Philips tech. Rev. **25**, 120-131, 1963/64. This article will here be referred to as I.

[2] See for example G. Brouwer, The simulation of electron kinetics in semiconductors, Proc. Second int. analogue computation meetings, Strasbourg 1958, pp. 135-137, publ. Presses Académiques Européennes, Brussels 1959.

extent to which the charge carriers disappear therefore depends in general on the concentrations of *both.*

The foregoing aspects of what takes place in an illuminated photoconductor — usually summarized under the heading "charge-carrier kinetics" — will be discussed in this article. We begin with the simplest situations. We shall then go somewhat deeper into the way in which the speed of response is affected by impurity centres that act as "traps", and finally we shall describe a method by means of which it is possible to introduce system into the very large numbers of behaviour patterns shown by photoconductors in which different types of impurity centre exist at the same time.

The recombination process: activation

The disappearance of charge carriers as a result of the recombination of two charge carriers of opposite sign is identical in kinetic terms with a chemical reaction involving the combining of two molecules, a *bimolecular reaction.* Just as in the chemical case, the rate of disappearance (reaction rate) is proportional to the concentration of each of the components [3]. As an equation:

$$\mathrm{d}n/\mathrm{d}t = -bnp. \quad . \quad . \quad (\mathrm{II},1)$$

In this expression n and p are the concentrations of electrons and holes respectively, b is a proportionality factor and t the time. When one of the concentrations, for example p, changes relatively little during the reaction, equation (II,1) can be written with good approximation as

$$\mathrm{d}n/\mathrm{d}t = -an, \quad . \quad . \quad . \quad (\mathrm{II},2)$$

where the constant $a \approx bp$. The reaction, essentially bimolecular, then behaves apparently as a *monomolecular* one. A chemical example is the decomposition of a substance: the rate can obviously only depend on one concentration.

To calculate the equilibrium concentrations that occur under constant illumination we must add to the right-hand side of equations (II,1) or (II,2) the constant term G representing the excitation density, i.e. the number of charge carriers freed in unit time and in unit volume. In the following we shall consider two extreme cases. In the first, all charge carriers are due to the photo-excitation so that, in addition to (II,1), we have $n = p$. In the other extreme case equation (II,2) applies. The relevant equations are:

bimolecular: $$\mathrm{d}n/\mathrm{d}t = G - bnp, \quad . \quad . \quad (\mathrm{II},3\mathrm{a})$$

$$n = p; \quad . \quad . \quad . \quad . \quad (\mathrm{II},3\mathrm{b})$$

pseudomonomolecular: $$\mathrm{d}n/\mathrm{d}t = G - an. \quad . \quad . \quad (\mathrm{II},4)$$

In the stationary state $\mathrm{d}n/\mathrm{d}t$ is zero, so that the concentrations are given by:

bimolecular: $$n_{\mathrm{stat}} = \sqrt{G/b}; \quad . \quad (\mathrm{II},5)$$

pseudomonomolecular [4]: $$n_{\mathrm{stat}} = G/a. \quad . \quad . \quad (\mathrm{II},6)$$

The density j_{f} of the photocurrent i_{f} that flows when a voltage is applied is proportional to n_{stat} (see eq. I,7) and is thus in the one case proportional to the excitation density G and in the other proportional to the square root of G. This explains why in the expression $i_{\mathrm{f}} \propto L^{\gamma}$, often used to describe the experimentally found relation between i_{f} and the incident luminous flux L, the constant γ sometimes has the value 1 and sometimes the value 0.5 — as G is proportional to L. Let us return for a moment to the relation (I,5), derived in part I, between G and the lifetime τ of the freed charge carriers, $\tau = n/G$. According to this, for monomolecular recombination $\tau = 1/a$, and for bimolecular recombination $\tau = 1/bn$. We see therefore that in bimolecular recombination the lifetime is not constant but inversely proportional to the instantaneous value of the electron concentration.

The manner in which the electron concentration goes to zero when the illumination is switched off can be found by integrating equations (II,1) and (II,2). This yields (see *fig. 2*):

bimolecular, for the case $n = p$:

$$n(t) = \frac{n_0}{1 + tbn_0} = \frac{n_0}{1 + t/\tau_0}\ ; \quad . \quad . \quad (\mathrm{II},7)$$

monomolecular:

$$n(t) = \frac{G}{a}\,\mathrm{e}^{-at} = n_0\,\mathrm{e}^{-t/\tau}. \quad . \quad . \quad . \quad (\mathrm{II},8)$$

Here n_0 and τ_0 are the values of n and τ at $t = 0$.

We shall now discuss some examples of photoconductive semiconductors and insulators, and consider

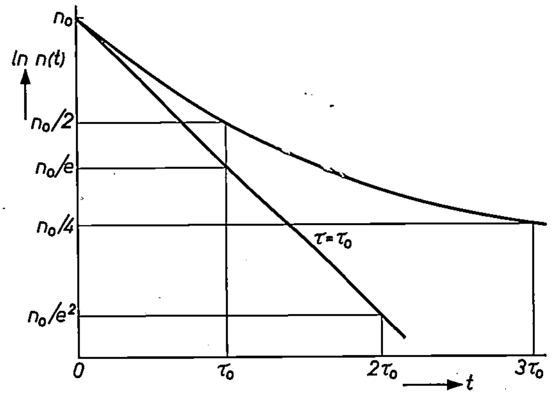


Fig. 2. Variation with the time t of the concentration n of the electrons in the conduction band after the illumination is switched off. If the recombination is a monomolecular process, n decreases exponentially and with log-linear co-ordinates the curve is a straight line of negative slope. With bimolecular recombination a curve is found whose slope with increasing t becomes steadily less steep. The curves apply to cases where the initial concentrations n_0 and the initial slopes are equal (cf. II,7 and II,8).

the manner in which the recombination takes place in these. We begin with four cases which are simple to analyse. ,

1) A substance with a large energy gap ΔE between valence and conduction bands; there are no impurity levels in the forbidden zone (see *fig. 3a*). A substance of this kind is an insulator in the dark. Photo-excitation produces an equal number of charge carriers in the two bands. The recombination process is therefore bimolecular.
2) A substance with a large ΔE and with impurity levels which are so deep that in the dark they are all occupied by an electron (fig. 3*b*). Further, the energy $h\nu$ of the light quanta is such that only the electrons from the impurity levels can reach the conduction band (cf. fig. 3 in I). At first this case is identical with the previous one, but if the illumination intensity is allowed to increase to very high values, the photocurrent can show a saturation effect, because at a certain illumination intensity nearly all impurity levels are empty. The excitation density G is then no longer proportional to L. This does not, however, affect the validity of (II,5).
3) A substance with a large ΔE and with impurity centres which are exclusively either donors or acceptors, so that the substance is a (purely extrinsic) semiconductor of type N or P (*fig. 4*). If the concentration of the photo-electrons (or holes) is small compared with that of the charge carriers originating from the impurity levels, the electron concentration will change only slightly and the recombination process is pseudomonomolecular (cf. II,2). Under strong illumination (high excitation density) the above-mentioned restriction is not of course fulfilled, and the process is again a bimolecular one. Between the two situations there is a transition region.

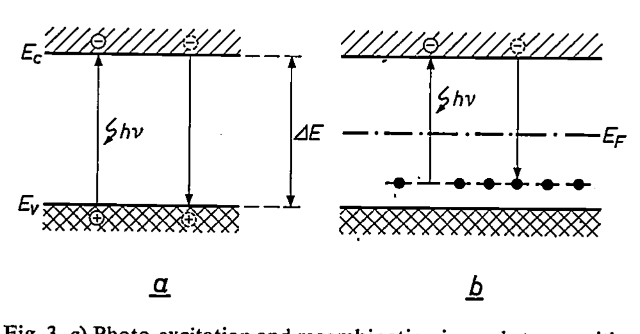


Fig. 3. *a*) Photo-excitation and recombination in a substance with a large energy gap ΔE (an insulator) which contains no impurity centres. The arrows indicate the energy jump of the electrons. E_c electron energy in the lowest level of the conduction band, E_V electron energy at the top of the valence band.
b) Photo-excitation and recombination in an insulator containing impurity centres for the case $h\nu < \Delta E$, so that excitation can only take place from the impurity centres. These centres are assumed to be so close to the valence band (so far below the Fermi level E_F) that in the dark they are all occupied by an electron at the prevailing temperature. In both cases the recombination process is bimolecular. (All figures of this type indicate the state that exists after the transition has taken place, the charge carriers being represented by full circles. The situation of a charge carrier before the transition is indicated by a dashed circle.)

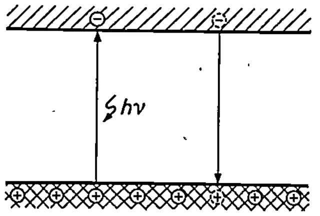


Fig. 4. Photo-excitation and recombination in a purely extrinsic P-type semiconductor with a large energy gap. (The impurity centres and the Fermi level have been omitted.) Under weak illumination the recombination process is pseudomonomolecular. Under very strong illumination the number of extra holes in the conduction band is no longer large compared with the number of charge carriers freed by the light, and the recombination is again a bimolecular process.

4) An insulator with impurity levels (concentration not too low) of which only a part, say half, is occupied (*fig. 5*). Further, as under (2), the photo-excitation is only from the impurity centres. If the excitation density is so low that the illumination causes very little change in the concentration of the empty impurity centres, the number of recombination partners available to the photo-electrons is nearly constant. The process is thus again monomolecular. Here too the process gradually changes with increasing illumination intensity from monomolecular to bimolecular.

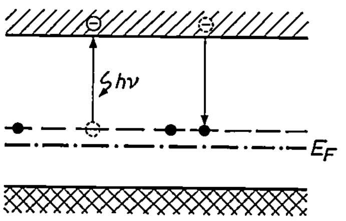


Fig. 5. Photo-excitation and recombination in an insulator with impurity centres that exchange electrons with the conduction band only and which in the dark are roughly half occupied (the impurity levels are approximately at the height of E_F). Here too the recombination process gradually changes with increasing illumination intensity from monomolecular to bimolecular.

Recombination via an impurity centre

As a final example we take a case, frequently encountered in practice, which cannot be so readily analysed. Here, as in the previous example, the substance is an insulator containing impurity centres, a substantial proportion of which are occupied. The impurity centres here, however, exchange electrons to a considerable extent with *both* bands, or, in other

[3] A treatment of reaction kinetics will be found in textbooks of theoretical chemistry, such as E. A. Moelwyn-Hughes, Physical chemistry, Pergamon Press, London 1957.

[4] Since the recombination cannot in reality be monomolecular, we shall leave out the prefix "pseudo" in the following.

words, they capture both free electrons and free holes. Centres of this kind are termed *recombination centres*: electrons that arrive in the conduction band by photoexcitation go, on their return journey, first to an unoccupied impurity level, and then on to the valence band (*fig. 6*). We shall now show that the recombination process taking place via this kind of impurity centre is *monomolecular* both under very weak and under very strong illumination. The *thermal* excitation of the electrons will again be disregarded because it does not essentially alter the situation in the present example.

Let N be the concentration of the recombination centres, n_r that of the filled recombination centres and p_r that of the empty ones, so that $N = n_r + p_r$. The number of photo-electrons passing per second from the conduction band to empty recombination levels is then $c_n\, n\, p_r$, where c_n is a proportionality factor referred to as the capture probability. The second jump which the electrons make — or, in other words, the capture of holes by filled centres — is governed by an analogous relation (with capture probability c_p), so that for the complete recombination process the following differential equations hold:

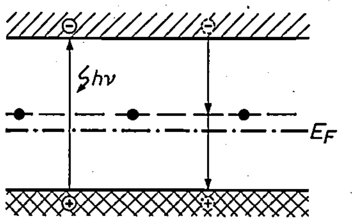


Fig. 6. As in fig. 5, but now with impurity centres that exchange electrons with *both* bands. The recombination process is now monomolecular both under weak and strong illumination.

$$\mathrm{d}n/\mathrm{d}t = G - c_n\, n\, p_r, \quad \text{(II,9a)}$$

$$\mathrm{d}p/\mathrm{d}t = G - c_p\, p\, n_r. \quad \text{(II,9b)}$$

Here G is again the excitation density.

We shall now consider two extreme cases and a transition situation in between them. In the first extreme case the illumination is so weak that the concentration n of the photo-electrons and the concentration p of the "photo-holes" are both small compared with the concentrations of the occupied as well as of the unoccupied impurity centres: $n,\ p \ll n_r,\ p_r$. In view of the condition of conservation of charge, the relation

$$\Delta n_r = n_r - n_r(0) = p - n \quad \text{(II,10)}$$

holds for the change Δn_r of n_r caused by the illumination. Here $n_r(0)$ is the concentration of the occupied centres before the illumination is switched on. Expressed in words: for every hole in the valence band there is a corresponding electron; if that electron is not in the conduction band it will be located in an impurity centre. Since p and n are both small in relation to n_r, it is evident from (II,10) that Δn_r will also be small with respect to n_r, so that n_r may be regarded as constant and equal to $n_r(0)$. Solving, on this assumption, equations (II,9) — putting the left-hand side of both expressions equal to zero, because we are concerned here only with the stationary state — and using the relation $\tau = n/G$ (I,5) to calculate the lifetime of electrons and holes, one then finds two dissimilar but constant values:

$$\tau_n = \frac{1}{c_n p_r(0)} = \frac{1}{c_n\,\{N - n_r(0)\}}, \quad \text{(II,11a)}$$

$$\tau_p = \frac{1}{c_p n_r(0)}. \quad \text{(II,11b)}$$

From the constancy of the lifetimes it follows that the recombination is monomolecular: from their inequality it follows that n and p must also be unequal (cf. I,5).

In the second extreme case the illumination is so strong that n and p are both large compared with N, and therefore the concentrations of the charge carriers are large in relation to the maximum possible change of concentration Δn_r in the centres. From the condition (II,10) it then follows that $p \approx n$. With these conditions the solution of equations (II,9) for the stationary state gives the relation:

$$\frac{n_r}{p_r} = \frac{c_n}{c_p}. \quad \text{(II,12)}$$

The concentrations of electrons and holes in the recombination centres are therefore in the same ratio to one another as the capture probabilities of these charge carriers. Since the sum of these concentrations must be equal to N, we find:

$$n_r = \frac{c_n}{c_n + c_p}\,N, \quad \text{(II,13a)}$$

and

$$p_r = \frac{c_p}{c_n + c_p}\,N. \quad \text{(II,13b)}$$

We see that in this second extreme case the occupation of the centres is independent of the intensity of illumination. The lifetime of the charge carriers will therefore again be constant, but this time will have the same value for electrons and holes:

$$\tau_n = \tau_p = \frac{1}{c_p n_r} = \frac{1}{c_n p_r} = \frac{1}{N}\left(\frac{1}{c_n} + \frac{1}{c_p}\right). \quad \text{(II,14)}$$

In spite of the different conditions, the recombination process in the example considered — which thus takes place via only one kind of recombination centre — is evidently *monomolecular* both under very weak and very strong illumination; this is in contrast with the situations in examples 3) and 4), where the recombination gradually changed from monomolecular to bimolecular with increasing illumination L. In the first extreme case, however, $n \neq p$, while in the second n and p are equal. In fact the expressions derived for both extreme cases are valid in wide ranges of values of L which we shall refer to as region *I* and region *III* (*fig. 7*). In these regions n and p are therefore proportional to L (and G).

Between these regions there lies a transition region *II* in which the electron and hole concentrations gradually become identical: p varies in this region sublinearly and n supralinearly with G. (If the substance is one in which, both in the dark and under illumination, only a small proportion of the centres is occupied by electrons, then in region *II* the electron concentration continues to change linearly and the hole concentration follows suit. The curve in such a case can be found to a good approximation using the method of calculation discussed in the final section of this article.)

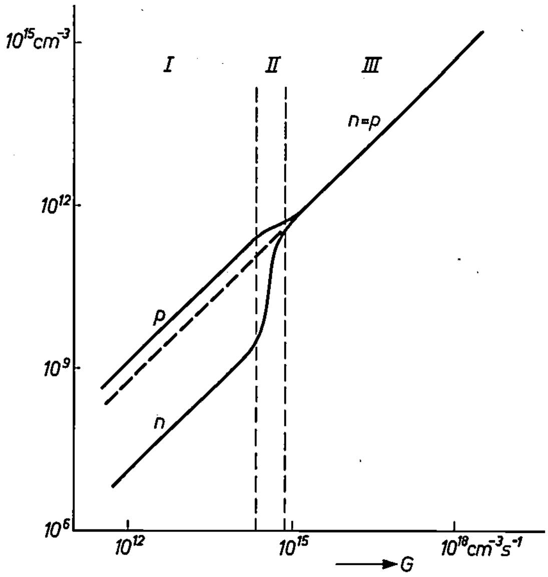


Fig. 7. The concentrations n and p as functions of the illumination intensity — here represented by the excitation density G — for the substance referred to in fig. 6. At very low and very high G both n and p are proportional to G (region *I* and region *III*). Between *I* and *III* there is a transition region *II* in which n increases supralinearly with G.

In the special case where the value $n_r(0)$ which n_r assumes in the dark happens to be exactly equal to $Nc_n/(c_n + c_p)$ — see (II,13a) — no change takes place in the occupation of the impurity levels when the illumination is switched on, and $n = p$ at every illumination intensity. The lifetimes as well are then independent of the intensity of illumination. At all illumination intensities the situation of region *III* then holds (see the dashed line in fig. 7).

Finally, two further remarks. We have just seen that upon recombination via a centre, even under strong illumination, the recombination process is monomolecular. There is no disputing this, but it does not imply that the γ of the photoconductor with recombination centres will always have the value 1 under strong illumination. The probability of direct recombination — also called band-band recombination — will never be exactly zero, and its contribution increases more strongly with the excitation density than that of the recombination via centres, the former being proportional to *both* concentrations, the latter only to one of them. With increasing illumination the band-band recombination will therefore in the long run make the greater contribution to the recombination and the value of γ will gradually move from 1 to $\frac{1}{2}$.

In the second place it should be noted that energy is liberated in every recombination process, this being the energy that was supplied upon excitation by the light quantum. This liberated energy is sometimes emitted again in the form of light (luminescence). The recombination can also be radiationless, the energy then being released in the form of lattice vibrations, so that the temperature of the crystal rises.

Activation

When a photoconductor is doped with a substance whose atoms (ions) can function as recombination centres, the result is that the lifetime of the charge carriers generally decreases. This comes about because of new recombination paths.

This is not always so, however. In some cases the lifetime of one of the two kinds of charge carrier, far from being shortened, is in fact enormously increased, so that doping has the effect of making the photoconductor much more sensitive. This effect is known as "activation", and an additive that produces it in a photoconductor is called an "activator". The activation mechanism can best be explained with an example.

Suppose that the substance at first contained N recombination centres/cm³ of type *1*, which have roughly the same capture probability c for electrons and for holes and which in the dark are half filled. Under weak illumination the lifetime of the charge carriers was thus equal to $2/Nc$. (See above.) Suppose further that another N recombination centres/cm³ are added by the doping, this time of type *2*. Characteristic of these type *2* recombination centres is that their capture probability is the same for holes, but is much smaller for electrons, for example $10^{-4}\,c$. Finally, we assume that in the dark the type *2* centres are nearly all filled. After the illumination has been switched on, the freed holes will now be trapped initially in roughly equal amounts by both types of centre. The photoelectrons, however, will be almost entirely trapped by the type *1* centres. As a result of this asymmetry a proportion of the type *2* centres will be emptied and nearly all the type *1* centres will be occupied by an electron. The illumination, which need only be quite weak, thus gives rise to a redistribution of the electrons and holes among the centres, which amounts to saying that half the electrons have transferred from type *2* centres to type *1* centres. See *fig. 8.*

If we now consider the manner in which the recombination takes place under these new circumstances, we see 1) that for an electron the possibility of recombination via type *1* centres has become very small because these centres are nearly all occupied, and 2) that although half the type *2* centres can still function, they nevertheless offer the electrons very little chance for recombination because their capture probability is only 10^{-4} *c*. As nevertheless in a stationary state the number of recombinations taking place per unit volume and time must be equal to the excitation density, the concentration of electrons in the conduction band is now very much higher than before: the doping has therefore had the effect of increasing the lifetime of the electrons.

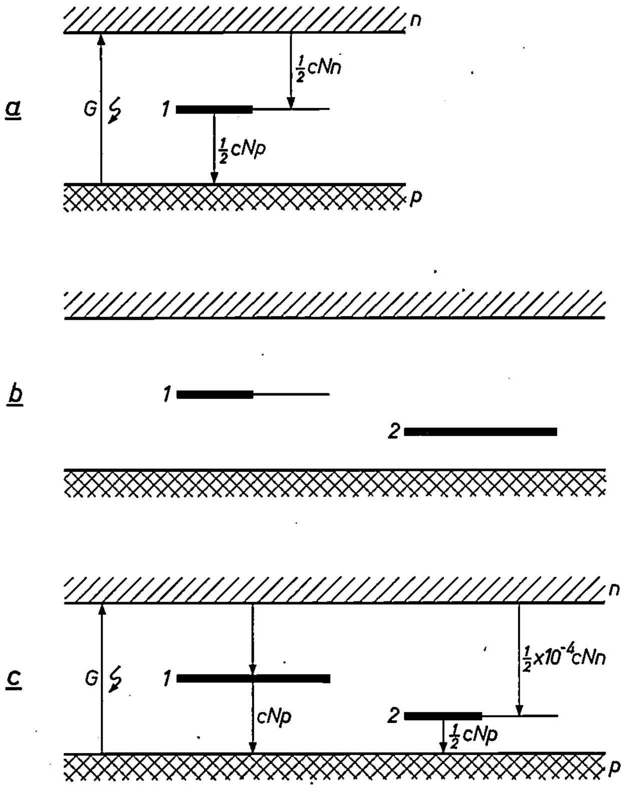


Fig. 8. Illustrating the fact that under certain circumstances the addition of impurity centres of a different type can considerably increase the sensitivity of a photoconductor (activation). The collective impurity centres of one type are represented here by a horizontal line. The thick part represents the occupied fraction of the centres and the thin part the unoccupied fraction. The expression written beside an arrow gives the number of recombinations occurring per cm³ and per second.
a) Energy band scheme of the original substance. There is one type of impurity centre (*1*), and half of these centres are occupied.
b) Band scheme of the substance after the addition of an equal number of type *2* centres; in the dark these are all occupied.
c) The same under illumination. Half of the electrons from *2* have gone to type *1* centres, so that the latter are nearly all occupied. Recombination via *1* is therefore virtually impossible, and recombination via centres *2* is also difficult because these centres have much less chance of trapping electrons from the conduction band. Since nevertheless the total number of recombinations per cm³ and per second is equal to *G*, *n* must now be much greater.

The change in the lifetime of the holes, on the other hand, is not so considerable. The number of holes recombining per cm³ and per second via type *1* centres is *cNp* and via type *2* centres the number is $\frac{1}{2}cNp$, giving a total of $1\frac{1}{2}cNp$. Since this number was $\frac{1}{2}cNp$ before doping, the lifetime of the holes has therefore become three times shorter. This is hardly a significant change compared with the enormous increase in the lifetime of the electrons: the doping has thus increased the sensitivity of the photoconductor by about 10^4 times.

To sum up, it can be said that activation of this nature depends on the fulfilment of three conditions. In the first place, the type *2* centres, the activator centres, must not have the same capture probability for electrons as for holes. Secondly, in the dark the activator centres must contain more electrons than there are unoccupied type *1* centres; upon illumination the latter can then be completely filled, so that the recombination path via *1* becomes "blocked". Thirdly, the concentrations *n* and *p* of free electrons and holes must be small compared with that of the centres.

The latter condition implies that the intensity of illumination is subject to an upper limit. If the illumination exceeds that limit, then in the long run *n* and *p* will become equal again, as in region *III* of fig. 7. The concentrations of the electrons trapped in the two types of centre adjust themselves to the new situation in a similar way to that described for region *III.* The lifetime then assumes the value that corresponds to the parallel operation of the two recombination paths, and the increased sensitivity is lost.

The influence of traps on the speed of response

In the foregoing we have confined our considerations to cases in which impurity centres were either absent (fig. 3*a*), did not interact with light-excited charge carriers (fig. 4), or had in fact taken over the role of the valence band (fig. 3*b* and 5), or were able to exchange charge carriers with both bands (fig. 6). As a first approximation the thermal excitation could be neglected.

In this section we shall consider the situations that can arise in substances containing impurity centres which can only exchange charge carriers with one band, i.e. centres that act as "traps". In this case the thermally excited transitions of charge carriers from these levels to the relevant band can no longer be neglected; on the contrary, the extent to which they take place determines the behaviour of the photoconductor. On the other hand, as in the last example in the previous section, we shall disregard *optically* excited transitions to or from the impurity centres. In this section we shall be concerned with materials containing traps of only one kind; see *fig. 9.* Furthermore, we shall confine our attention to the behaviour of the electrons.

A characteristic of the effect of traps on the behaviour of a photoconductor is that they have, as a rule, a very marked influence on the speed of response. This effect can easily be explained in qualitative terms. In a stationary state the concentration n obviously has a certain value that depends on the illumination intensity. Owing to the exchange of charge carriers between the traps and the conduction band, however, the same also applies to the concentration n_t of the charge carriers present in the traps; if the illumination intensity changes, then n and n_t must *both* change correspondingly, and in the same direction. If, for example, the illumination intensity is increased, then immediately afterwards the "current" G must bring about not only the necessary increase in n but also that in n_t. This of course takes longer than if no traps were present. Since in practice n_t is quite often much greater than n, the decrease in speed of response due to the traps may be very considerable. A complication — see below — is that the equilibrium between the conduction band and the impurity centres is not reached instantaneously. In terms of the model in fig. 1, one cannot in all cases consider the conduction band and all the traps together as one large reservoir.

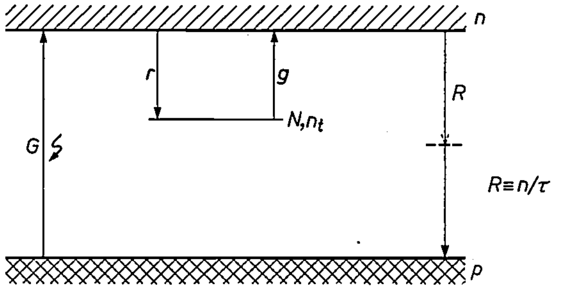


Fig. 9. Energy band scheme of substances containing traps of only one type. As in fig. 8, the traps are again collectively represented by a single horizontal line. Their concentration is N, the concentration of the occupied traps n_t. Once again, G is the density of the optical excitation, R the recombination density, g density of the thermal excitation from the traps, r that of the electron capture from the conduction band. The recombination may either take place directly, or via recombination centres (see small dashed line).

At the same time it can be shown qualitatively why, under *very strong* illumination, the traps do *not* make the response much slower. The traps are then nearly all filled with electrons, so that when the illumination intensity changes hardly any change in the number of trapped electrons is possible. This does not of course apply if the illumination intensity is reduced from such a high value to a very low one (or to zero). In this case, after an initial steep decline in the photocurrent, there is a small residual current ("tail") whose strength decreases much more slowly, just as in situations where the initial value of the illumination is not extremely high.

Before examining the influence of traps on the speed of response in more quantitative terms, we shall dwell for a moment on their effect on the *sensitivity* of a photoconductor. In the stationary state the concentration n of the electrons in the conduction band is governed by the condition that the excitation density G (cf. fig. 9) and the recombination density R are equal. This follows from the fact that n_t in a stationary state is also constant and therefore the "currents" g and r compensate one another exactly. So, in the first instance the presence of the traps does not alter the situation; the traps can, however, have some influence on the manner in which R depends on n, so that *indirectly* they may influence the sensitivity. To make this clearer, let us return for a moment to the general formula for the stationary state in an illuminated insulator: $\mathrm{d}n/\mathrm{d}t = 0 = G - bnp$ (cf. II,3a). Depending on the nature of the recombination mechanism, p in this expression is now the concentration of the free holes or the concentration of the holes bound to a recombination centre. It can be said that the traps do not alter the contribution of the electrons to the photocurrent if they do not affect the product bp. This is the case for example when the recombination process is monomolecular (see figs. 4, 5 and 6) and the illumination sufficiently weak.

In the more quantitative considerations about to follow we shall not consider extremely strong illumination and confine ourselves to situations in which the illumination is so weak that only a small fraction of the traps are occupied by electrons. We further assume, for simplicity, that the recombination takes place via recombination centres, and is thus monomolecular, so that τ may be regarded as constant. For the concentrations n and n_t (fig. 9) the following differential equations apply:

$$\frac{\mathrm{d}n}{\mathrm{d}t} = G - R + g - r, \quad . \quad . \quad . \quad . \quad \text{(II,15a)}$$

$$\frac{\mathrm{d}n_t}{\mathrm{d}t} = -g + r. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(II,15b)}$$

The "currents" g and r, which are governed by n and n_t and by the (constant) probabilities e_n and c_n

of the two transitions (cf. II,9), are given by:

$$g = e_n n_t, \quad \ldots \quad \text{(II,16a)}$$

$$r = c_n n (N - n_t). \quad \ldots \quad \text{(II,16b)}$$

The values of the transition probabilities e_n and c_n are characteristic of the type of impurity centres under consideration. There exists a relation between them that can be found by considering a state of thermal equilibrium: here $g = r$ or $e_n n_t = c_n n (N - n_t)$, whence:

$$e_n = c_n \frac{n(N - n_t)}{n_t}. \quad \ldots \quad \text{(II,17a)}$$

On the other hand, using the Fermi-Dirac distribution function (I,1) we can derive the following expression for the ratio of the energy level occupations n and n_t in thermal equilibrium:

$$\frac{n}{n_t} = \frac{N_c}{N - n_t} \exp\left[-(E_c - E_t)/kT\right]. \quad \ldots \quad \text{(II,18)}$$

(Here, as before, N_c is the effective density of states at the bottom of the conduction band, E_c is the energy of the bottom of the conduction band and E_t that of the traps.) With the aid of this equation we can reduce (II,17a) to:

$$e_n = c_n N_c \exp\left[-(E_c - E_t)/kT\right], \quad \ldots \quad \text{(II,17b)}$$

which gives the required relation between e_n and c_n.

It can be shown that (II,18) is also valid for a stationary state in which the substance is illuminated, and even for a non-stationary state provided n and n_t are in thermal equilibrium. We shall make use of this fact in the following considerations.

The transition probability (capture probability) c_n is often written as the product of the average thermal velocity of the electrons v_{th} and the capture cross-section S, which has the dimensions of area. The capture probability c_n depends only slightly on the temperature: when S is constant, $c_n \propto v_{th}$ hence $\propto T^{1/2}$. The probability e_n of the thermal excitation, on the other hand, is strongly temperature dependent. (See II,17b; as a rule $(E_c - E_t) \gg kT$.)

Using (II,16a), (II,17b) and where applicable (II,18), we shall now calculate the response for two extreme cases. Since we have assumed that the illumination is so weak that most traps are empty — i.e. $n_t \ll N$ — we may begin by simplifying (II,18) to:

$$\frac{n}{n_t} = \frac{N_c}{N} \exp\left[-(E_c - E_t)/kT\right]. \quad \ldots \quad \text{(II,19)}$$

The ratio of n to n_t under weak illumination is thus independent of the illumination intensity and equal to a constant; in the following the quotient n_t/n will be referred to as δ.

As the first extreme case we assume that g and r are both large with respect to the recombination density R; the electrons in the traps and in the conduction band will then be in mutual thermal equilibrium, even in the periods during which n and n_t are moving to a new equilibrium value. The decrease of n after the illumination is switched off — i.e. $G = 0$ — satisfies the following differential equation, which can be derived from (II,15), from (II,19) in the form $n/n_t = 1/\delta$ and from the defining equation $R = n/\tau$ — see (I,5):

$$(1 + \delta)\, \mathrm{d}n/\mathrm{d}t = -n/\tau. \quad \ldots \quad \text{(II,20)}$$

It follows from this that n decreases exponentially with a time constant τ_0 given by:

$$\tau_0 = (1 + \delta)\tau. \quad \ldots \quad \text{(II,21)}$$

To demonstrate the magnitude of the effect which traps have on the speed of response in the situation under discussion we shall calculate the magnitude of δ when reasonable values are chosen for the quantities in (II,19). Taking $N = 10^{16}$ cm^{-3}, $N_c = 10^{19}$ cm^{-3} and a value of 0.3 eV for $e(E_c - E_t)$, the "trap depth", we find $\delta \approx 100$. In a more strongly doped substance and at greater trap depth — situations frequently encountered in practice — δ can easily become 10^6 or 10^7. In the situation described here, where g and r are relatively large, the traps appear to have a very considerable influence on the speed of response.

As in most cases $\delta \gg 1$, then (II,21) can often be simplified to

$$\tau_0 = \delta\tau. \quad \ldots \quad \text{(II,22)}$$

This shows that the conduction band and the traps can then be regarded collectively as a single reservoir, whose magnitude is completely determined by the concentration of the traps.

As our second extreme case we assume that g and r are *small* compared with R. The traps and the conduction band now exchange charge carriers only with great difficulty, and when the illumination is switched off the thermal equilibrium between them is at first completely broken. The electron concentration n in the conduction band therefore falls at first relatively quickly, i.e. with a time constant equal to the lifetime τ. After this, however, the curve of electron concentration against time shows a long "tail". We shall now calculate the time constant of the latter.

As the flow r of electrons to the traps is proportional to the electron concentration n in the conduction band, then after the initial steep drop in n, the quantity r will also fall to a fraction of its original value. The thermal generation g has hardly changed, however, so that now $r \ll g$. Using this relation together with (II,16a) and (II,17b) we can reduce equation (II,15b) to:

$$\mathrm{d}n_t/\mathrm{d}t \approx -g = -c_n N_c n_t \exp\left[-(E_c - E_t)/kT\right]. \quad \text{(II,23)}$$

From this it follows that the concentration n_t of the trapped electrons decreases exponentially, with the time constant

$$\tau_t = \frac{\exp\,[(E_c - E_t)/kT]}{c_n N_c} \,. \quad . \quad . \quad (II,24)$$

This time constant is quite high, particularly at large $E_c - E_t$ and at low temperature. This slow "evaporation" of electrons from the traps maintains in the conduction band an electron concentration which is roughly equal to $g\tau$ and which therefore — since g is proportional to n_t — likewise decreases with the time constant τ_t. The response curves for the two situations discussed are given diagrammatically in *fig. 10.*

In practice a substance will of course often contain various types of trap. The response curves will then correspond to a superposition of the curves for each kind of trap. Such curves can differ considerably in character, for if there are traps of more than one kind present, the two situations just discussed may very well occur simultaneously, and moreover the deepest traps may even become entirely filled at the illumination intensity employed. All this makes it often very difficult, if not impossible, to derive, from the observed curves, unambiguous data on the concentration and on the nature of the traps.

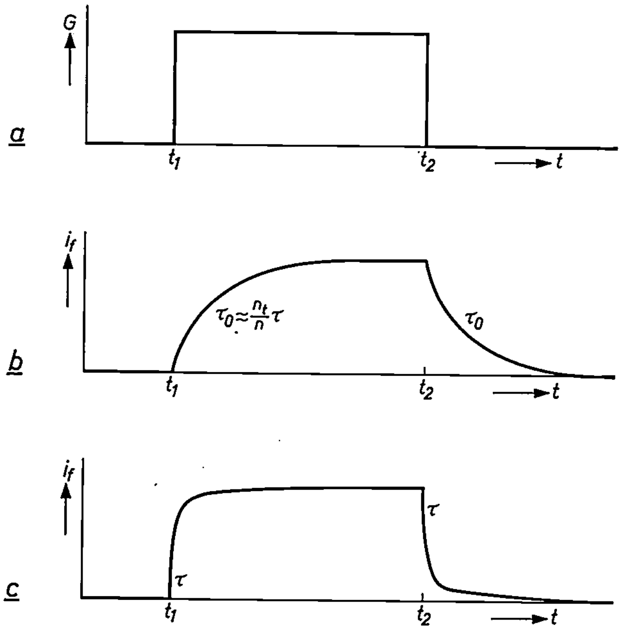


Fig. 10. Examples of the response of photoconductors containing traps on switching on or off the illumination. The illumination is so weak that n_t is always smaller than N (cf. fig. 9).
a) The variation of illumination with time; the strength of illumination is again represented by the excitation density G.
b) Response of a photoconductor in which the traps and the conduction band exchange electrons to such an extent that they are constantly in thermal equilibrium (g and $r \gg R$). The photocurrent varies exponentially with a time constant approximately n_t/n greater than the lifetime τ.
c) If g and r are small compared with R, then i_f changes very rapidly when the illumination is switched on or off (time constant $\approx \tau$), but there is a very long "tail".

In our considerations on the speed of response we have hitherto taken no account of the influence which the *contacts* can have on the response of a photoconducting device. If the contacts are of the injecting (replenishing) type [1], this effect is often negligible. At high field strengths and under strong illumination, however, it may happen that the space charges (band-bending) in the contact areas show a marked dependence on these quantities, and that the speed with which the band bending adjusts itself to the changed situation partly governs the speed of response. The influence of these effects cannot, however, be dealt with in a brief treatment.

If the contacts are of the blocking type the situation in the first instance is quite simple. At sufficiently high field strength all charge carriers freed by the light reach a contact before they recombine; the photocurrent is then saturated [2]. If there are no traps, then in this case the response time is identical with the average carrier transit time $T_{n,p}$, i.e. inversely proportional to the field strength. If traps *are* present, the response time is equal to the sum of the carrier transit time and the average time which a carrier spends in traps. The latter time is usually much longer than $T_{n,p}$ and therefore generally governs the speed of response. It also decreases with the field strength, not primarily in the sense that the carrier remains in the trap for a shorter time, but simply because an electron stops at fewer traps the faster it travels. Thus, in a photoconductor containing traps the speed of response also increases with the field strength when the current is saturated. A notable example of this is the photoconducting layer in a "Plumbicon" television camera tube [5].

The behaviour of photoconductors with two or more types of impurity centre

So far, except in dealing with activation, we have consistently referred to substances which contain at the most one kind of impurity centre. We have further simplified the situation in most cases by disregarding thermal excitation. In the previous section, where we did not do this, we assumed that the impurity centres present could only exchange electrons with the conduction band, i.e. they behaved as ideal traps; the other type of impurity centre was called a recombination centre.

In reality, impurity centres do of course exchange electrons with a band by thermal excitation, and moreover the probability that an electron in a trap will move to the valence band is not exactly zero. In this section, in which we shall now have something to say about the behaviour of photoconductors that contain impurity centres of two or more types, and in which we shall touch on problems which confront the research worker who wishes to devise a model of a substance from the results of his measurements, we shall no longer distinguish between impurity centres as pure traps and as pure recombination centres, but bear in mind that every impurity centre combines both aspects in one way or another.

[5] See E. F. de Haan, A. van der Drift and P. P. M. Schampers, The "Plumbicon", a new television camera tube, Philips tech. Rev. 25, 133-151, 1963/64.

Fig. 11 shows the energy band scheme of a semiconductor containing impurity centres of two types, the possible transitions being indicated by arrows. The letters G, E, B, R, I, C, L and F denote, as in the previous figures, the number of transitions per second and occurring per cm^3 (the transition densities). In the following we shall refer to the quantities in short as "currents". Provided the energy gap is not too small, the transitions denoted by dashed lines will occur only infrequently — at least, transitions due to thermal excitation — because of the magnitude of the energy jump. Since in the following considerations we shall also assume that no electrons can be optically excited from or to the impurity centres, and since we shall also neglect band-band recombination ($F = 0$), we can confine our attention to the transitions indicated by the full lines. It should be noted that these assumptions are made purely for the sake of simplicity and are in no way fundamental in nature.

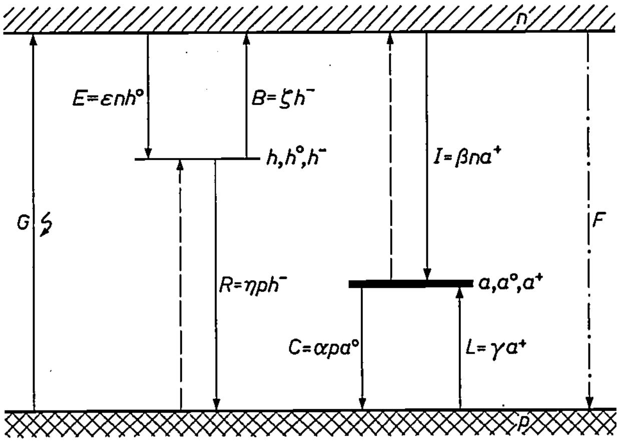


Fig. 11. Energy band scheme of a photoconductor with impurity centres of two types. The arrows represent the possible electron transitions. The capital letters indicate these transitions themselves as well as their density, (number per cm^3 and per second; the excitation and recombination "currents"). The concentration of the high centres is h, that of the low (deep) centres is a. For the first centres the concentration of the unoccupied levels is h^0, and that of the occupied ones h^-. For the deep centres the concentration of the unoccupied levels is a^+ and that of the occupied levels (occupied by one electron is a^0). The Greek letters represent the transition probabilities.

Let the concentration of the centres of the one type be h cm^{-3}, of which h^- are filled (with an electron) and h^0 are empty; let that of the other, deeper type be a cm^{-3} of which a^0 are filled and a^+ empty. We assume that in the dark the latter centres are all occupied by an electron (thick line) and that the other centres are all unoccupied (thin line). We further assume that in the dark the concentrations n and p are negligibly small (the substance is virtually an insulator). We shall now calculate how the concentration of the charge carriers in the two bands and in both kinds of impurity centre depends on the illumination intensity (again represented by the excitation density G). We shall consider only stationary states.

In every stationary state the following equations apply:

$$a = a^0 + a^+, \quad \ldots\ldots \quad \text{(II,25a)}$$
$$h = h^0 + h^-, \quad \ldots\ldots \quad \text{(II,25b)}$$
$$a^+ + p = h^- + n, \quad \ldots\ldots \quad \text{(II,25c)}$$
$$G = I + R, \quad \ldots\ldots \quad \text{(II,25d)}$$
$$C = L + I, \quad \ldots\ldots \quad \text{(II,25e)}$$
$$E = B + R. \quad \ldots\ldots \quad \text{(II,25f)}$$

The first two equations state that the sum of the number of occupied and the number of unoccupied centres must be equal to the total number (a and h respectively). Equation (II,25c) states that the total number of positive charge carriers is equal to the total number of negative ones. The last three equations state that in a stationary state the sum of the "currents" flowing towards one kind of centre must be equal to the sum of the "currents" that flow away from it (principle of detailed balancing). These "currents" depend on the concentrations in the manner indicated in fig. 11. At a certain value of G (the independent variable) the six concentrations n, p, h^0, h^-, a^0 and a^+ are thus completely determined by these six equations when the concentrations a and h and the material constants α, γ, β, η, ε and ζ are known.

The solution of the (non-linear) set of equations (II,25), even if it could be written in a closed form, would be very complicated and consequently not very helpful. If, for example, one tries to solve the equations by eliminating the unknowns, one obtains equations of high degree. It is therefore necessary to simplify the equations by neglecting terms that are relatively small. What simplifications are permissible, however, will depend on the energy level scheme and on the transition probabilities — quantities not yet known, quantities, indeed whose evaluation is the whole purpose of the measurements. As will be made clear below, a considerable degree of uncertainty is introduced into the analysis when one tries to explain a particular observation, for example the variation of the photocurrent i_f with the illumination intensity — i_f is proportional to $n\mu_n + p\mu_p$, see (I,7) — on the basis of a single hypothetical model with the appropriate simplifying assumptions. For, agreement between the theoretical and experimental curves gives no guarantee at all that the model chosen is the right one; usually various models are possible that lead to the same theoretical curve.

This uncertainty can be eliminated by means of a method of approximation given by Klasens. This method, which we shall describe here, offers a relatively simple means of obtaining a survey of *all* the possible solutions [6].

The method starts from the consideration that the values of the quantities in equations (II,25) normally cover a range of many orders of magnitude, so that, of two terms on the same side of the equals sign, as a rule one is so much larger than the other that the smaller one can be neglected [7]. Introducing these simplifications one then finds that for (II,25c) there are *four* possible forms, and for the five other equations *two*:

(1)	(2)	(3)	(4)	
$a^0 = a$	$a^+ = a$			(II,26a)
$h^0 = h$	$h^- = h$			(II,26b)
$a^+ = h^-$	$a^+ = n$	$h^- = p$	$n = p$	(II,26c)
$G = I$	$G = R$			(II,26d)
$C = L$	$C = I$			(II,26e)
$E = B$	$E = R$			(II,26f)

A particular situation prevailing in an illuminated photoconductor corresponds to a particular combination of these equations. It should be noted at this point that such a "situation" depends not only on the properties of the substance but also on the strength of illumination, i.e. on the value of G. A particular combination of equations (II,26) applies only in a limited interval of G values; outside that interval another has to be used.

The various combinations of equations can be made easier to handle by assigning to them a code number of six digits: the first digit denotes the column where the applicable form of eq. (II,26a) can be found, the second denotes the relevant column for (II,26b), and so on. For example, the code 111211 indicates that the following equations are applicable:

$$\left.\begin{array}{l} a^0 = a \\ h^0 = h \\ a^+ = h^- \\ G = R \\ C = L \\ E = B \end{array}\right\} \text{(code 111211).} \qquad \text{(II,27)}$$

As can be seen, a combination of equations simplified in the manner described is so simple that the solutions for the concentrations can at once be found. It can easily be verified that these read as follows (cf. fig. 11):

$$\left.\begin{array}{l} n = \dfrac{\alpha^{\frac{1}{2}}\zeta\, a^{\frac{1}{2}}}{\gamma^{\frac{1}{2}}\,\varepsilon\,\eta^{\frac{1}{2}}h}\, G^{\frac{1}{2}};\; h^- = a^+ = \dfrac{\alpha^{\frac{1}{2}}\, a^{\frac{1}{2}}}{\gamma^{\frac{1}{2}}\,\eta^{\frac{1}{2}}}\, G^{\frac{1}{2}}; \\ p = \dfrac{\gamma^{\frac{1}{2}}}{\alpha^{\frac{1}{2}}\,\eta^{\frac{1}{2}}\, a^{\frac{1}{2}}}\, G^{\frac{1}{2}};\quad h^0 = h;\quad a^0 = a. \end{array}\right\} . \qquad \text{(II,28)}$$

For the "currents" we then have:

$$\left.\begin{array}{l} I = \beta n a^+ = \dfrac{\alpha\,\beta\,\zeta\, a}{\gamma\,\varepsilon\,\eta\, h}\, G;\; C = L = \gamma a^+ = \dfrac{\alpha^{\frac{1}{2}}\gamma^{\frac{1}{2}}a^{\frac{1}{2}}}{\eta^{\frac{1}{2}}} G^{\frac{1}{2}}; \\ E = B = \zeta h^- = \dfrac{\alpha^{\frac{1}{2}}\,\zeta\, a^{\frac{1}{2}}}{\gamma^{\frac{1}{2}}\,\eta^{\frac{1}{2}}} G^{\frac{1}{2}};\;\; R = G\,. \end{array}\right\} \text{(II,29)}$$

For other combinations of simplified equations one again finds that the solutions are either simple power functions of G with low (integral or fractional) exponents, or constants.

Altogether there are 128 different possible combinations of simplified equations, whose solutions can all be calculated in the same simple manner. A survey of the solutions will be found in the reference given [6].

As already noted, each combination is valid only in a limited range of G values, and when the illumination is increased from zero it will be necessary from time to time to change the combination (the code) to give a correct description. The G values at which this should be done are those at which a neglected quantity has become equal in magnitude to its non-neglected companion. These can easily be found graphically by plotting the calculated variation of the six concentrations and "currents" against G: the end of the validity range of a combination is arrived at when, for a given pair of quantities, the curve for the neglected and that for the non-neglected quantity intersect. *Fig. 12* shows such a graph for equations (II,28) and (II,29), specific values having of course been chosen for the six material constants. As can be seen, a logarithmic scale has been adopted for both coordinates. This is necessary for the ordinate because the values of the concentrations can differ from one another by many decades. By making the abscissa G logarithmic as well one obtains straight lines for all curves; from the slope one can read directly the exponent by which the relevant quantity varies as a function of G.

In our example the lines $a^0 = a$ and $a^+ = (\alpha a/\gamma\eta)^{1/2}G^{1/2}$ are the first to intersect with increasing G. At the G value corresponding to the point of intersection ($G = G_1$) the contribution a^+ has reached the value a and the equation $a^0 = a$ (II,26a, column 1) is evidently no longer a good approximation for $a = a^0 + a^+$ (II,25a). In the neighbourhood of this G value we have to change from the approximation $a^0 = a$ to $a^+ = a$; the first digit of the code number thus changes from 1 to 2. Calculating the concentra-

[6] H. A. Klasens, The intensity-dependence of photoconduction and luminescence of photoconductors in the stationary state, Phys. Chem. Solids 7, 175-200, 1958.

[7] This approach is due to C. A. Duboc (Brit. J. appl. Phys. 6, Suppl. No. 4, p. S 107, 1955). The method described in this section is therefore sometimes referred to as the Duboc-Klasens method.

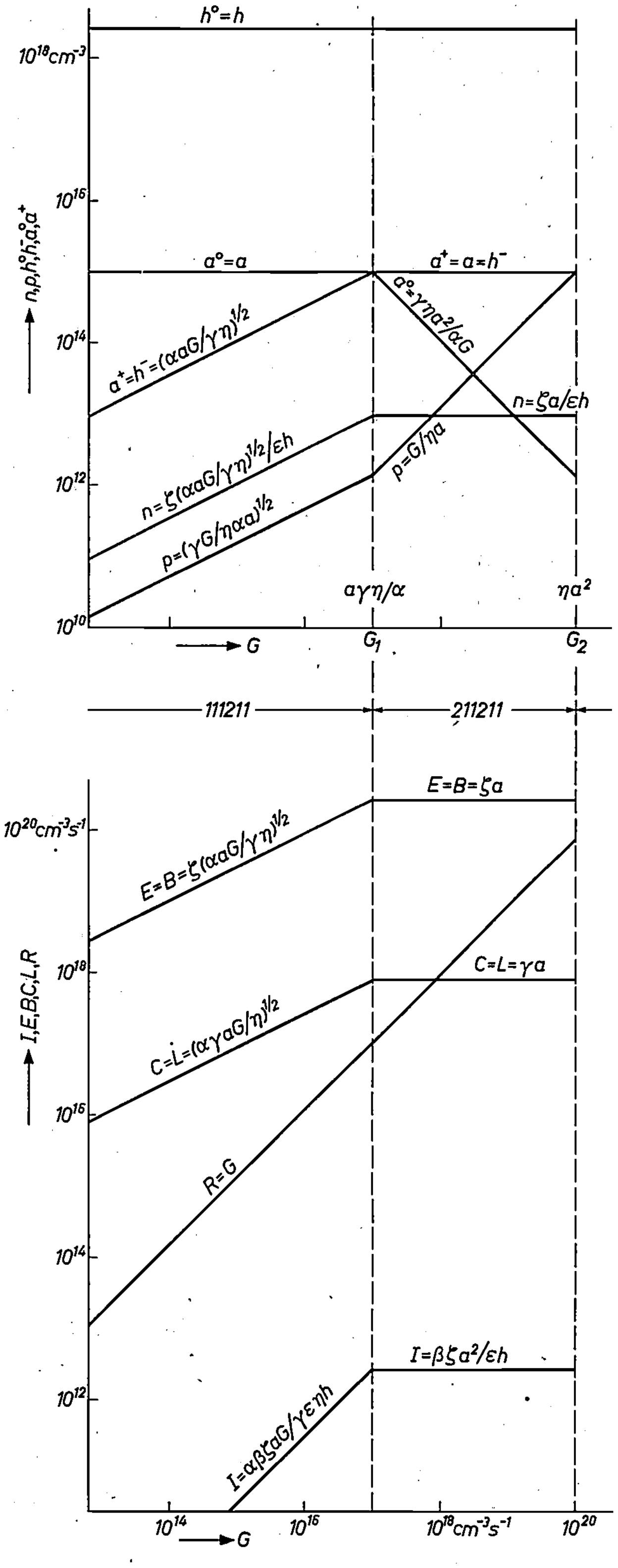


Fig. 12. Klasens diagram of the variation of the concentrations (upper) and "currents" (lower) with the excitation density G for a photoconductor as in fig. 11. In the log-log graphs all curves are approximated by straight lines. With increasing G it is necessary from time to time to change over to another approximation. In the present example this occurs at the values G_1 and G_2. Since the region traversed by G covers many decades, and since as a rule the approximation will only have to be changed a few times, the straight lines give a fairly true picture of the real shape of the curves.

tions for the new combination (211211) we then find the solutions indicated in the figure. The value of G_1 is easily found by for example equalizing the expressions on both sides for a^+, which gives $G_1 = a\gamma\eta/\alpha$. In reality, of course, the curves do not show a discontinuity but a rounded bend in the neighbourhood of G_1.

Following the behaviour of the curves further we also see that the combination 211211 eventually ceases to be valid: where the line for p intersects the line $a^+ = a$ the approximation $a^+ = h^-$ for (II,25c) has to change to $p = h^-$ (II,26c, column 3). The code number thus becomes 213211.

Having now sufficiently illustrated the principle of the method of calculation, we shall not examine what happens when the value of G is increased still further in the example given in fig. 12. Instead we shall consider a special case.

Extreme supralinearity

In research on photoconductors the situation is sometimes encountered in which for instance the photocurrent — that means n or p — increases in a small region of G values very much more rapidly than linearly with G, for example in proportion to G^4 or G^6. This effect is known as "extreme supralinearity". We shall now give an example to show that the explanation of this effect is to be found in a straightforward way from Klasens' system of approximations. We shall see that such an interval of G values is *not* identical with the validity interval of one approximation (cf. fig. 12), but that on the contrary "extreme supralinearity", when present only occurs around a G value where two such intervals adjoin one another.

Let us suppose we have a substance for which, in a certain interval of G values, the combination 111211 of our previous example is valid, but in which the concentration h of the higher situated centres is not, as in that case, greater than that of the other but in fact smaller; the horizontal line $h^0 = h$ is now, therefore, *below* the curve $a^0 = a$ (*fig. 13*). The end of region 111211 now appears when the line for h^- intersects the horizontal line $h^0 = h$. The approximation $h^0 = h$ for (II,25b) has evidently to be replaced there by $h^- = h$, so that the code number changes to 121211. We thus obtain the following set of equations (cf. fig. 11 and eq. II,26):

$$\left.\begin{array}{lll} \left.\begin{array}{l} a^0 = a, \\ h^- = h, \\ a^+ = h^-, \end{array}\right\} & & a^+ = h, \\ G = R, & \text{or:} & G = \eta p h^-, \\ C = L, & \text{or:} & a p a^0 = \gamma a^+, \\ E = B, & \text{or:} & h^0 n = \zeta h^-/\varepsilon. \end{array}\right\} \quad \text{(II,30)}$$

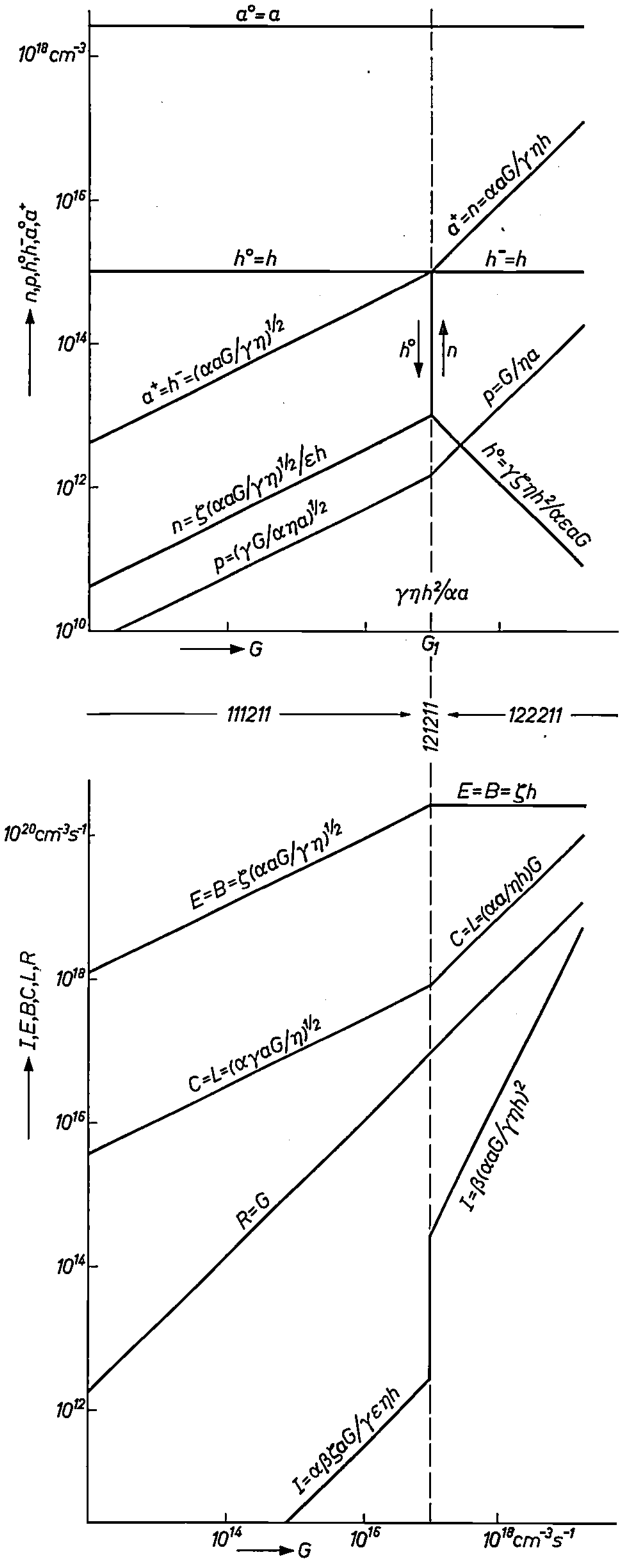


Fig. 13. Klasens diagram of a photoconductor as in fig. 11 in which the material constants (the concentrations of the impurity centres and the transition probabilities) have values such that the curves of n, I and h^0 have to be approximated by a jump at a certain value of G. In reality there is of course no jump at this value of G, but n and I increase very steeply in a small region of G values (extreme supralinearity); in this region the concentration h^0 decreases very rapidly.

Of these six equations the first five contain only four unknowns; these equations must therefore conflict. The two other unknowns — h^0 and n — both occur in the sixth equation. If we consider the two solutions for p that can be derived from the fourth and the fifth equation: $p = G/\eta h$ and $p = \gamma h/\alpha a$, we see that the conflict is removed when G has the value G_1 at which both these solutions are equal to one another: $G_1 = \gamma\eta h^2/\alpha a$. The approximation 121211 is thus not valid in an interval of values but only at this one value of G; for larger G values a successive approximation has to be used.

The extreme supralinearity becomes immediately apparent when we consider what happens to h^0 and n at $G = G_1$. All that we know about these from the sixth equation is that their product is constant. We therefore consider whether, and if so in what direction, n and h^0 change at $G = G_1$, and what form the variation takes in the interval adjoining the upper side of $G = G_1$. In this adjoining interval the approximation 122211 will apply. This can be derived from the following. In the interval 111211, h^- increased with G: at $G = G_1$ we therefore had to replace even the approximation $h^0 = h$ by $h^- = h$. This means that h^0 really decreases in the first interval, and that in our diagram (fig. 13) the curve for h^0 should drop vertically at $G = G_1$. In view of the constancy of nh^0 the concentration n at that point will therefore increase. In doing so n will ultimately pass the value h, which implies that the approximation $a^+ = h^-$ for (II,25c) must change to $a^+ = n$ (II,26c, column 2); the third digit of the code therefore does change, from 1 to 2. If we calculate with the new approximation the variation of the different quantities with G, we then find the solutions written beside the curves in fig. 13; by putting in $G = G_1$ we find the co-ordinate value at which the curves begin.

From these considerations it is apparent that the quantities h^0, I and n — and therefore the photocurrent as well — undergo a *jump* at the point $G = G_1$. The abrupt increase of n and I at this point is no other than the approximate representation, in our schematic approach, of the supralinearity mentioned at the beginning [8].

If we consider what happens around $G = G_1$ as a whole we see that with increasing G a marked change takes place there in the occupation of the energy levels: n becomes much larger and h^0 much smaller, which means that the higher levels nearly all become occupied.

[8] An investigation into the circumstances in which this effect can appear is described by F. N. Hooge and D. Polder, Conditions for superlinear intrinsic photoconductivity, J. Phys. Chem. Solids **25**, 977-984, 1964. See also: F. N. Hooge, Consequences of the conditions for superlinear intrinsic photoconductivity, Philips Res. Repts. **19**, 333-348, 1964.

The increase of n obviously results in greater photosensitivity. Here again we have a situation where the addition of impurity centres has an activating effect (see the first section). A schematic illustration of the change in occupation of the levels can be seen in *fig. 14*.

Finally, we shall examine the process of recombination in the three intervals of G, and calculate the respective lifetimes. From fig. 13*b* — and also from the fact that the fourth digit of the code number is always 2 — it may be inferred that for all three intervals the approximation $R = G$ is valid; the recombination in all three apparently takes place mainly via the higher centres. The lifetime τ is thus roughly equal to n/R, or:

$$\tau = n/\eta p h^{-}. \qquad \text{(II,31)}$$

From this it may be derived, for example by substituting the solution for n, p and h^{-}, applicable to each interval, that in the interval to the left of G_1 the lifetime is inversely proportional to $\sqrt{G}$ and that in the interval to the right of G_1 the lifetime is constant. For excitation densities smaller than G_1 the recombination process is therefore *bimolecular*, and for greater excitation densities *monomolecular* (cf. II,5 and II,6).

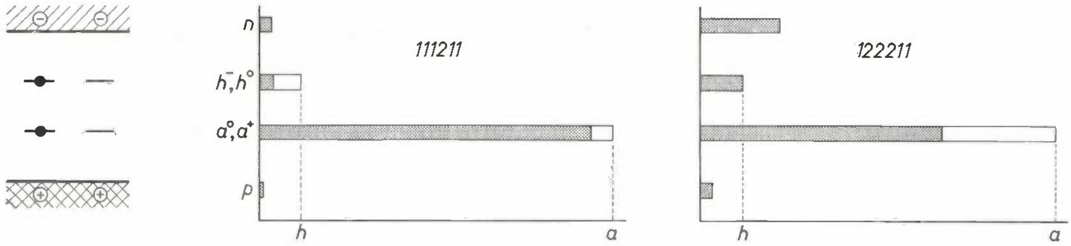


Fig. 14. Schematic representation of the very marked change in the energy level occupation which occurs in the photoconductor of fig. 13 around the value G_1. *a*) Situation in the interval 111211. *b*) Situation in the interval 122211. The occupation a^0 of the deep centres changes relatively little, because a is much greater than h.

Luminescence

We have already noted that when an electron falls from a higher to a lower energy level (recombination) the energy difference may emerge in the form of a light quantum. The study of this light emission — or "luminescence" — can provide valuable information on the processes taking place in an illuminated photoconductor.

Let us again take as an example a substance containing impurity levels of two types, as shown by the energy level diagram of fig. 11. In principle, all four transitions indicated in fig. 11 by downward pointing arrows can be accompanied in such a substance by the emission of light. Let us assume that in our example this is so for transition I; and that the three others are radiationless transitions. As the strength of the luminescence is proportional to the number of electrons changing levels per unit time, its variation with the illumination intensity produced by the incident radiation is identical, except for a constant, with that of the curves for I in a diagram of the type given in fig. 12 and fig. 13. Whereas the study of photoconductivity teaches us something about the concentrations n and p and the lifetime of the charge carriers, the behaviour of the luminescence can yield information on the "recombination currents"; both kinds of data supplement each other and together make it possible to obtain a picture of what takes place in a particular substance. In addition one can of course calculate the energy difference referred to at the beginning from the wavelength of the emitted light, and thus find the location of the relevant impurity level.

Just as the addition of impurity centres of a second kind can profoundly influence the photoconducting characteristics, it can also influence the luminescence, which is as mentioned before a "recombination current" made "visible". It will be clear that this effect can be investigated, just like the effect on the photoconductivity, with the aid of the method of approximation described above. All that this requires in the example referred to is to determine what happens with the curve for I.

Finally, we shall describe two striking examples of the influence of impurity centres of a second type. These will again demonstrate the very close relation existing between luminescence and photoconductivity.

If, for example, the higher centres (fig. 11) have a pronounced trapping character (R negligible) then the presence of these centres will give rise to afterglow (phosphorescence). This is simply the recombination (via the other centres) of the electrons which, after the illumination has been switched off, are still "evaporating" from the traps and which, as described, are the cause of the "tail" shown by the photocurrent.

If, on the other hand, the transitions R and E dominate — second example — then only a few of the recombinations take place via the deep centres and therefore I is much weaker than in the previous example. The same therefore applies for the luminescence. A substance which attenuates the luminescence in this way when used as a doping element in luminescent material is known as a "killer".

The extent to which a "killer" attenuates the luminescence depends on the temperature. If this is raised then the "current" L becomes stronger, and a^+, the number of unoccupied (deep) impurity centres, decreases. The same happens to I, and a greater fraction of the recombinations take place via the other centres, that is without the emission of radiation. If this effect is pronounced it is referred to as "thermal quenching".

Further, as can be inferred from fig. 12 and 13, the luminescence exhibits peculiarities similar to those shown by the photoconductivity. In fig. 12 we have a case where both the photoconductivity (n) and the luminescence (I) reach saturation with increasing G. In fig. 13 the peculiarities in the behaviour of n discussed in detail above — extreme supralinearity, activation — also occur in the behaviour of the luminescence.

To conclude this article we would like to say something about practical research work on photoconductors and luminescent substances, and in particular on the possibilities offered by the Klasens method mentioned in the last section.

In describing the method we took as our basis a simplified model of substances containing impurity centres of two types, and no account was taken of contact effects. As already noted, for such a situation there are as many as 128 possible combinations of approximated equations. For every substance to which this model applies we saw that the behaviour of different variables as functions of the excitation density G could be approximated by a series of a limited number of these 128 combinations.

In practice it will often be found, however, that an energy level diagram with only two types of impurity centre is not adequate. Some impurities still have a noticeable effect on the photoconductivity even though their concentration is so small that their presence can only just if at all be demonstrated by chemical means — the lowest concentration at which many impurities can be spectrochemically demonstrated is in the region of 10^{16} atoms/cm^3. For investigating such substances it would therefore be useful to have tables of approximations obtained by the Klasens method and based on energy level diagrams with more than two types of impurity centre.

In theory it is quite possible to produce such tables, but they would be so large as to be almost impractical. This by no means implies, however, that the method is not applicable in the more complicated situations. Indeed, even if the starting material is a photoconducting (luminescent) substance whose properties are well known, the method can be a valuable tool of research into the properties of new materials obtained from it by doping, that is by adding another type of impurity centre. Conversely, by judicious doping it is sometimes possible to give a well-known starting-material the properties required for a specific application. For this one must, of course, have information about the nature of the impurity centres produced by the various doping elements considered. The Klasens method can then be used to determine which of these elements can best be used and what is the best concentration to give the best approximation to the desired result. In some cases, the present state of the techniques of making extremely pure substances allows the natural impurities present to be disregarded, as their influence on a substance is negligible compared with that of the doping agent.

The difficulties are considerable when attempts are made to arrive at a good model of a substance of which hardly anything is known. It is often not even possible to answer directly the question: "how many types of impurity centre are there?". As we have seen, some impurities have a noticeable influence even in extremely small concentrations; vacancies and interstitial atoms also play a part, and finally there are impurities that give rise to impurity levels of more than one type. It is then more than ever necessary to investigate as many properties as possible — photoconductivity, luminescence, their temperature dependence, absorption spectrum, electron spin resonance etc. — and to try, by combining data and by trial and error, to arrive at a simple initial model. Analyses by the Klasens method will often indicate that more than one model is possible. One of the merits of the method is that the analysis indicates the lines along which the investigation should proceed to find out which model is best.

Summary. The equilibrium in an illuminated photoconductor is a *dynamic* equilibrium: in unit time each energy level is reached by as many charge carriers as leave it. Charge carriers can only form and disappear two by two: every electron has a corresponding hole. The recombination takes place either as a bimolecular reaction or as a pseudomonomolecular reaction. Recombination via an impurity centre is a pseudomonomolecular process both under strong and weak illumination. The addition of recombination centres of other types (doping) can sometimes substantially increase the photoconductivity (activation). Traps often make the response very much slower except under very intense illumination. Substances with impurity centres of two or more types show widely varying behaviour. The complete and exact theoretical description is complicated. The Klasens method of approximations provides a useful practical survey of all behaviour patterns.

Stereophonic radio broadcasting

II. Susceptibility to interference

N. van Hurck, F. L. H. M. Stumpers and M. Weeda

534.76:621.396:621.391.82

Part II of this article deals with an aspect of stereophonic signal transmission not covered in Part I, the fact that stereophonic radio receivers are more readily affected by interference than monaural ones. The results of some theoretical and experimental investigations into this are given.

In the stereophonic radio broadcasting systems described in part I [1], there are components present in the modulation impressed on the carrier that have frequencies above the audible range. Under the FCC system, to which most attention was given, these components take the form of a stereo sub-signal and a pilot signal. At the receiver, these are converted by the adaptor circuit into a difference signal which falls within the audible frequency range. The process is liable to give rise to interference not encountered in mono reception.

In the pages that follow we shall be taking a closer look at this aspect of stereophonic radio broadcasting, after a general introduction dealing with the various types of radio interference.

Types of radio interference

Interference to radio reception can have a variety of causes and its character, as perceived by the listener, depends upon its origin.

One type is *noise*. Cosmic noise is picked up by the aerial; but noise can also arise in the aerial or receiver [2]. The simplest form is *white* noise, which can be regarded as the superposition of a very large number of electrical oscillations whose combined energy, measured within a 1 c/s interval, is the same throughout the spectrum. In most practical cases only a limited frequency band is of interest, in which the noise can with sufficient approximation be considered as "white".

The type of interference commonly known as "man-made static", which is caused by a variety of electrical equipment — motors, the ignition systems of internal combustion engines, and thermostatically controlled switches such as those in electric irons — is of a quite different character. It often takes the form of a sequence of sharp pulses, and is naturally most troublesome in thickly populated areas, where it is likely to be more of a nuisance than noise.

A third type of interference is due to radio stations other than the one to which the receiver is tuned. Its effect is usually rather different from that of the first two types. Noise and pulse-like interference manifest themselves as a sizzling or crackling background; the transmissions from an interfering transmitter can be heard along with the desired one, though in a distorted form.

Another type of interference plays a bigger part in FM than in AM reception. It arises when the signal from the transmitter reaches the receiver along different paths of unequal length [3], and it can occur in mountainous areas as a result of reflection from hillsides. High buildings and similar structures may cause the same sort of interference in other places.

The various types of interference will now be considered in turn, and their undesirable effects on stereo and on mono reception compared. We shall also deal with the reception of stereo, as well as of normal mono transmissions, on a mono receiver.

The reproduction of a radio programme can be subject to interference that does not have an r.f. origin at all. The a.f. signal from the studio, prior to being fed to the transmitter, may contain noise that has originated in a microphone, record-player or tape-recorder. It is also possible for pulse-like interference to enter the receiver by a route other than the aerial. Moreover, an unwanted voltage of frequency 50 or 100 c/s, audible as hum, can arise in the receiver itself; this may be due to inadequate smoothing of supply voltages, or to induction in connections inside the set. Such types of unwanted signal may lead to confusion in the study of interference arising in the radio link itself. In experiments of the kind referred to below it is therefore always advisable to make sure that the observed interference does indeed originate in the radio link. This can usually be checked fairly easily.

Noise

For the listener with a *monaural set*, the noise level associated with reception of a stereo broadcast is rather higher than that associated with reception of a mono broadcast. The difference is however so small as to be hardly perceptible. The slightly higher noise

N. van Hurck, Dr. F. L. H. M. Stumpers and Ir. M. Weeda are with Philips Research Laboratories, Eindhoven. In 1962, 1963 and 1965 Dr. Stumpers took part in CCIR and EBU discussions relating to stereophonic broadcasting.

level on stereo is connected with the fact that, in FM reception, the signal-to-noise ratio is proportional to the square of the maximum frequency sweep [4]. Under the FCC system the frequency deviation of the sum signal must not exceed 90% of the allotted sweep. This means that a mono receiver tuned to a stereo broadcast is dealing with a signal whose frequency sweep is 10% less than that of the signal it would pick up from a corresponding mono transmission. As a result, the signal-to-noise ratio is 0.9 dB smaller.

Reception of a stereo programme on a *stereo receiver* may be accompanied by an appreciably higher noise level. The main reason for this was briefly mentioned in the introduction. In a mono receiver, no audible contribution to the noise level is made by interference components present at the output of the ratio detector, which are in the same frequency range as the stereo sub-signal, whereas in a stereo receiver they are converted into an audible form in the same way as the a.f. difference signal is obtained from the stereo sub-signal. The effect is aggravated by the fact that in FM reception, the interference caused becomes stronger as the frequency of the detected interfering voltage increases [4]. As a result of this interference occurring in the same frequency range as the stereo sub-signal will be stronger at the output than interference in the sum signal range; so that the adverse effect on the difference signal will be greater than that on the sum signal.

The rise in noise level accompanying the changeover from mono to stereo reception has been determined both mathematically and by experiment.

The signal-to-noise ratio associated with reception of a mono signal is given by:

$$\left(\frac{S}{N}\right)_{\mathrm{M}} = \tfrac{1}{2}\,\frac{v_{\mathrm{c}}^{2}}{n_{0}^{2}}\,\frac{\Delta f_{\mathrm{m}}^{2}}{f_{\mathrm{d}}^{3}\left(\dfrac{f_{\mathrm{a}}}{f_{\mathrm{d}}}-\arctan\dfrac{f_{\mathrm{a}}}{f_{\mathrm{d}}}\right)}, \quad . \; . \quad (1)$$

while the signal-to-noise ratio associated with reception of the right-hand and left-hand components of a stereo broadcast is given by [5]:

$$\left(\frac{S}{N}\right)_{\mathrm{S}} = \frac{0.81}{4}\,\frac{v_{\mathrm{c}}^{2}}{n_{0}^{2}}\,\frac{\Delta f_{\mathrm{m}}^{2}}{f_{\mathrm{d}}^{3}\left[\tfrac{3}{2}\dfrac{f_{\mathrm{a}}}{f_{\mathrm{d}}}+\left(\dfrac{f_{\mathrm{h}}^{2}}{f_{\mathrm{d}}^{2}}-\tfrac{3}{2}\right)\arctan\dfrac{f_{\mathrm{a}}}{f_{\mathrm{d}}}\right]}\,. \quad (2)$$

In both cases it is assumed that the received signal has the maximum allowed frequency sweep, denoted by Δf_{m}. The other quantities occurring in (1) and (2) are:

v_{c} = r.m.s. value of the i.f. signal voltage at the limiter input.

n_0 = r.m.s. value of the noise voltage within a 1 c/s interval, at the same point in the receiver circuit.

f_{d} = the frequency at which the a.f. signal is attenuaed by 3 dB as a result of de-emphasis. In Europe a value of 3.180 kc/s has been chosen for f_{d}: this is equivalent to an RC time constant of 50 μs.

f_{h} = the subcarrier frequency (38 kc/s).

f_{a} = the a.f. signal bandwidth (15 kc/s).

On inserting the stated values of f_{d}, f_{h} and f_{a} in equations (1) and (2) we find that $(S/N)_{\mathrm{S}}$ is about 150 times smaller than $(S/N)_{\mathrm{M}}$. This means that given the same signal and noise levels at the limiter input, the signal-to-noise ratio for stereo reception is 22 dB lower than that for mono reception of a mono broadcast. (Of course, the ear will not notice a difference of this magnitude unless the received signal is weak enough for noise to be audible for mono reception.)

Equations (1) and (2) only cover the electronic aspect of the problem, laying down the signal-to-noise ratio as a ratio of two powers and making no distinction between noise components of different frequencies within the audible range. In the estimation of noise, the human ear has however a part to play; two unwanted sounds of equal energy but of differing frequency are not experienced as having equal effect. To allow for this, a common practice is to place in front of the measuring instrument a "psophometric" filter with a frequency characteristic such as that shown in *fig. 1* [6]. If this curve is employed in the calculation of total noise power it will be found that the signal-to-noise ratio is 23 dB lower for stereo than for mono reception. This differs little from the result obtained by attaching equal weight to all frequencies in the audible range.

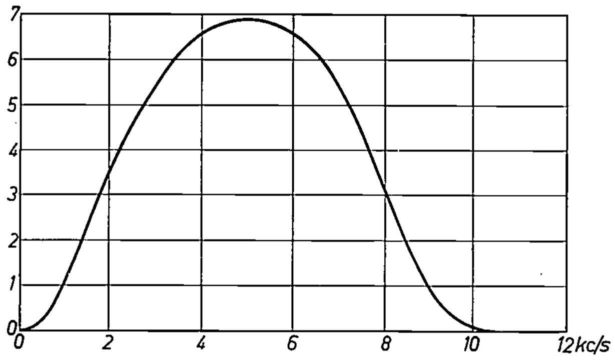


Fig. 1. Frequency characteristic of a "psophometric filter", as used in noise measurements to represent the differences in the subjective effect of audible interfering sounds having the same energy level but different frequencies.

[1] N. van Hurck, F. L. H. M. Stumpers and M. Weeda, Stereophonic radio broadcasting, I. Systems and circuits, Philips tech. Rev. **26**, 327-339, 1965 (No. 11/12).

[2] For a concise qualitative treatment, see K. S. Knol, Philips techn. Rev. **20**, 50-57, 1958/59.

[3] See J. Koster, Multipath transmission effects in FM reception and their simulation in the laboratory, Philips tech. Rev. 22, 393-402, 1960/61.

[4] See for example C. E. Tibbs and G. G. Johnstone, Frequency modulation engineering, Chapman and Hall, London 1956, or Stanford Goldman, Frequency analysis, modulation and noise, McGraw-Hill, New York 1948.

[5] A derivation of the equations may be found in W. P. Neidig, Stereo FM (1), R.G.T. Monitor (Philips) 4, 29-38, 1964.

[6] See CCITT. Avis P53B: Psophomètre utilisé sur les circuits pour transmissions radiophoniques, Livre rouge, Tome V (IIe Assemblée plénière, New Delhi, 1960), pp. 131-133.

As far as it relates to interference from the difference signal, equation (2) only takes account of noise components whose frequencies lie within the stereo sub-signal range (23 to 53 kc/s). Components with frequencies above 53 kc/s can also however give rise, in the AM detector that demodulates the stereo sub-signal, to unwanted frequencies in the audible range. For instance, in a normal *peak detector* handling a carrier with a frequency of f kc/s, audible interference arises not only from noise components whose frequencies lie between $f-15$ and $f+15$ kc/s, but also from components with frequencies between $2f-15$ and $2f+15$ kc/s, between $3f-15$ and $3f+15$ kc/s, and so on [7]. This effect is less in a *switch detector*: noise components centred on a frequency of $2f$ give no audible interference; while those centred on $3f$ give rise to only one-ninth (in power) of the interference that would arise in a peak detector.

The simplest way of avoiding such extra interference is to design the receiver in such a way that alternating voltages with frequencies over 53 kc/s do not reach the AM detector for the stereo sub-signal. In this connection the reader is referred to the adaptor circuit given in Part I of this article (p. 337, fig. 15), in which the stereo sub-signal reaches the detector by way of a tuned circuit K_2, and also to the filters marked Fi_2 in figs. 11 and 14 in Part I.

Equations (1) and (2) can be used to calculate the magnitude of the *aerial signal* necessary for a good signal-to-noise ratio. At maximum output power a 50 dB signal-to-noise ratio is generally considered to be good. If we assume that the receiver has a noise factor [8] of 3.5, that a dipole aerial with a noise temperature of 1200 °K is being used [9], we find that a field-strength of 5 μV/m is required for mono and one of 70 μV/m for stereo reception, for the specified signal-to-noise ratio of 50 dB.

In looking at these figures it must be borne in mind that they are much lower than the field-strength normally occurring in the area served by a radio station. In general broadcasting authorities try to achieve a field-strength of at least 1 mV/m throughout the service area. In these circumstances, with a normal receiving aerial the noise originating in the radio link is usually much weaker than the background noise that is inevitably present. Consequently the listener notices no rise in noise level when he switches from mono to stereo reception. More noise will however be heard on stereo if the transmitter is at a very great distance, for then its local field-strength will be very much lower.

Pulse-like interference

The nature of pulse-like interference depends to a very great extent on the type, and on the state of maintenance, of the electrical machines or appliances that give rise to it. This kind of interference cannot therefore be expressed in exact and universally valid figures; the experiments referred to below are based on the use of equal pulses, equidistant in time.

It has been found by experiment that the field-strength for stereo reception must be about 14 dB higher than that for mono reception if the effect of pulsed interference is to be the same at the output of the receiver.

The measurements in question were carried out in conformity with rules laid down by the European Broadcasting Union. The procedure is to connect the receiver to a signal generator whose output is modulated with monaural information, and which delivers to the receiver a voltage the same as that which would exist at the terminals of a dipole aerial in a 250 μV/m field. "Interference" pulses are then added whose level is 3 dB above that causing just perceptible interference under mono reception conditions. The system is then switched to stereo reception, the pulses being kept at the same level, and the strength of the input signal is raised until the interference effect, as subjectively assessed, is the same as before.

The figure of 14 dB quoted above implies that for the same degree of pulse-like interference a stereo transmitter must give, at the receiver, a field strength five times as great as that from a mono transmitter. Or, to put it another way, one can say that the range of a transmitter broadcasting a stereo programme is shorter than that of a mono transmitter of the same power. Let us assume, with particular regard to pulsed interference, that a field-strength of 250 μV/m is necessary for satisfactory reception of a mono transmitter. (The required field-strength for stereo reception will then be 1250 μV/m.) If we now refer to graphs published by the CCIR, we find that on this basis the service area of a transmitter with an effective radiated power of 100 kW and an aerial height of 200 m has a radius of about 95 km for mono and 50 km for stereo reception [10].

Interference from other stations

It is again very difficult to provide exact numerical information about the effect of interference from unwanted transmitters, for here too a large number of factors come into play which are difficult or even impossible to express quantitatively. Among these are the types of programme being broadcast by the wanted and the unwanted station, and the subjective effect of the resulting interference on the listener. Because of such factors the results obtained by laboratories that have investigated the subject show fairly wide disparities, even when the same programmes have been presented to listeners for appraisal. Some guidance is available from *fig. 2*, which displays the results of an experiment carried out with a receiver and adaptor built in our own laboratory. The wanted and unwanted signal took the form of programmes compiled by the "Institut für Rundfunktechnik" in Hamburg and recommended by the European Broadcasting Union for use in this kind of experiment; the field-strength of the wanted transmitter was adjusted to 1 mV/m. The strength of the interfering signal was gradually raised to the point where the resulting interference was just audible. In the graph, the ratio between the wanted and unwanted signal strengths has been plotted as a function of the separation between the two centre frequencies. Curve *M* relates to mono reception, curve *S* to stereo reception.

Fig. 2 shows that interference from a station on an adjoining frequency comes about more readily with stereo than with mono reception. If the two stations share the same carrier frequency, then with mono reception, an unwanted signal can be tolerated which is 11 dB stronger than with stereo reception. As there is an interval of 100 or 200 kc/s between the carrier frequencies, the ratio between the interference levels acceptable on mono and stereo becomes about 23 dB and 2 dB respectively.

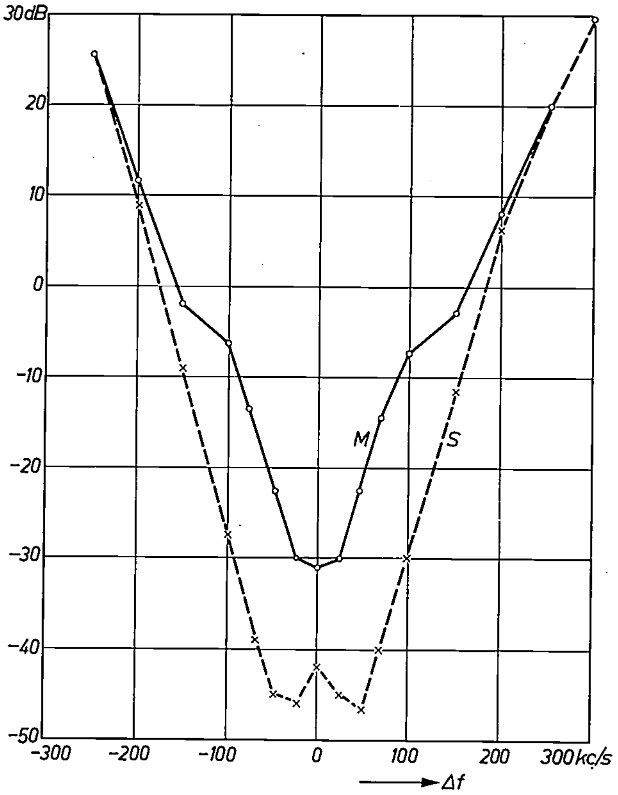


Fig. 2. Ratio between unwanted and wanted signal strengths as a function of the separation Δf between the carrier frequencies of the wanted und unwanted transmitters, when the unwanted signal level is just high enough to cause audible interference. Curve *M* relates to mono and curve *S* to stereo reception.

Two things should be borne in mind when the results plotted in fig. 2 are examined. Firstly, the standard test programmes used in the experiment have been designed to simulate the worst possible situation arising in practice (large frequency sweep in the interfering transmission and many quiet passages in the wanted programme). Secondly, subjects taking part in the experiment listened very critically before deciding whether interference was perceptible or not. In fact the average listener will usually not notice an interfering signal considerably stronger than that indicated in fig. 2.

The difference in the shape of the two curves in fig. 2 can be readily explained, in qualitative terms at least, The spectrum of a signal radiated by an FM broadcast transmitter occupies a band of about 180 kc/s.The whole band "fills up" with components when the carrier wave is modulated with speech or music, and this applies to both the wanted and the interfering station. Interference to mono reception is mainly due to components of the wanted and unwanted signals whose difference frequencies lie in the audible range, i.e. to components less than about 15 kc/s apart. Stereo reception is subject, in addition, to interference caused by those components in the spectra of the two transmitters whose difference frequencies lie within the frequency range of the stereo sub-signal (23 to 53 kc/s). As mentioned above when dealing with noise, these, because of their higher frequencies, produce even stronger interference than the first type. This explains why the maximum acceptable level of the interfering signal is lower on stereo than on mono. It will also be clear now why interference to stereo reception from another station is worst when the carrier waves of the wanted and unwanted transmitter are about 40 kc/s apart. At this frequency difference the carrier of the unwanted transmitter, whose amplitude is large during quiet passages in the modulation, gives rise to an interfering component with a frequency in the range of the stereo sub-signal.

Since a mono receiver does not utilize the stereo sub-signal the possessor of such a set hardly nötices any change in the interference from other stations, when the transmitter to which he is tuned switches from a mono to a stereo broadcast.

Interference due to reflections

The interference that arises from undesired reflection of radio signals leads to distortion of the modulation,

[7] This property of detectors is comparatively little known, probably because the i.f. stages of a normal radio receiver only pass a band centred on the intermediate frequency, and so signal or interference components with frequencies close to higher harmonics of the i.f. do not get through to the detector.

[8] The noise factor of a circuit is a measure of the contribution the circuit itself makes to the noise present at its own output terminals. For a more formal definition, see F. L. H. M. Stumpers and N. van Hurck, An automatic noise figure indicator, Philips tech. Rev. **18**, 141-144, 1956/57.

[9] This means that the aerial receives cosmic noise corresponding to the noise from a resistance equal to the radiation resistance of the aerial, and at a mean temperature of 1200 °K. This noise is not constant: according to measurements by H. V. Cottony and J. R. Johler (Proc. IRE 40, 1053-1060, 1952) it varies cyclically over a period of about 24 hours, the maxima and minima corresponding to aerial noise temperatures of 1700 °K and 700 °K respectively. For information about noise temperatures, see Reference data for radio engineers, published by the International Telephone and Telegraph Corporation, New York.

[10] It should be borne in mind that the siting of transmitters in an FM broadcasting network is generally such as to ensure a field-strength appreciably higher than 250 μV/m throughout the area served.

which can sometimes be very severe. Here again, a part is played by factors not susceptible to quantitative treatment. We shall therefore confine ourselves to reporting some of the results of an investigation in which reflection of the received signal was simulated in the laboratory. The experiments were arranged so that the "reflection" corresponded to path differences of up to about 15 km for the reflected and direct signals. In each experiment the amplitude of the reflected signal was adjusted to the point where its interfering effect just became perceptible. *Fig. 3* is a plot of the ratio between the reflected and direct signal strengths as a function of the difference *d* in their path lengths; curve *M* relates to mono, curve *S* to stereo reception. The graph shows that in most cases satisfactory stereo reception requires that the reflected wave must be about 10 dB weaker than is acceptable for mono reception. If the distortion due to spurious reflections is just noticeable on mono, it may in some situations become very severe on changeover to stereo reception.

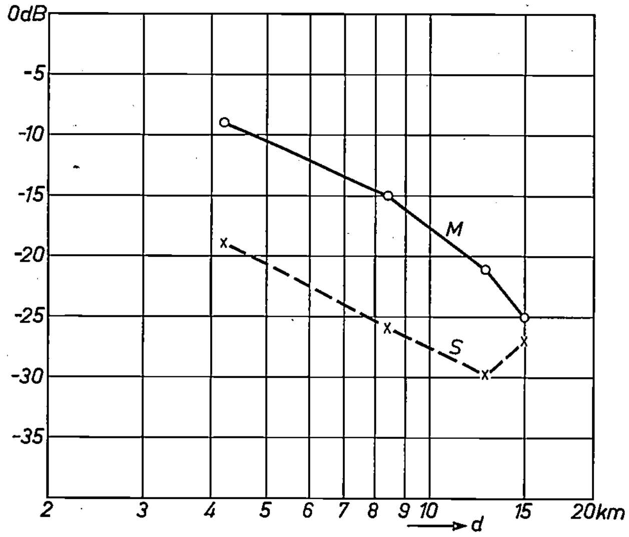


Fig. 3. The ratio between the amplitudes of a reflected and direct signal as a function of the difference *d* in path length, when the reflected signal is just large enough to give perceptible distortion. Curve *M* relates to mono and curve *S* to stereo reception.

Experiments with various systems of stereophonic broadcasting were performed in the Teutoburgerwald (in Germany), a region where reflections give considerable difficulties. These consistently showed, once again, that stereo reception is more susceptible than mono to interference from reflected waves. It was however considered that the FCC system gave better results than the other systems investigated.

It has also been found that the combination of a frequency discriminator of wide frequency characteristic, and an effective limiter, lessens the effect of interfering reflections. Another and more obvious remedy is to use a receiver with a directional aerial system.

Summary. A stereo receiver is more susceptible than a comparable mono set to each of the main types of interference — noise, pulse-like interference, transmitters on adjacent frequencies, and spurious reflection of the signal. One implication of this is that the local field-strength necessary for good quality reception from a given transmitter is higher for stereo than mono. The results of some relevant theoretical studies and experimental investigations are presented.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands	*E*
Mullard Research Laboratories, Redhill (Surrey), England	*M*
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France	*L*
Philips Zentrallaboratorium GmbH, Laboratory at Aachen, Weisshausstrasse, 51 Aachen, Germany	*A*
Philips Zentrallaboratorium GmbH, Laboratory at Hamburg, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany	*H*
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium.	*B*

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

G. Arlt: Resonance-antiresonance of conducting piezoelectric resonators.
J. Acoust. Soc. Amer. **37**, 151-157, 1965 (No. 1). *A*

G. Arlt and **G. R. Schodder**: Some elastic constants of silicon carbide.
J. Acoust. Soc. Amer. **37**, 384-386, 1965 (No. 2). *A*

A. C. Aten, J. H. Haanstra and **G. Diemer**: Peculiar spectral distribution of the light emission from ZnTe during d.c. pulse excitation.
Philips Res. Repts. **20**, 125-135, 1965 (No. 2). *E*

K. H. Beckmann and **B. Caspar**: On the energy balance for the passage of light through a thin absorbing film.
Philips Res. Repts. **20**, 190-205, 1965 (No. 2). *H*

G. Blasse: Crystal chemistry and some magnetic properties of mixed metal oxides with spinel structure.
Thesis Leiden, April 1964. *E*

G. Blasse: A new type of superexchange in the perovskite structure.
Proc. int. Conf. on Magnetism, Nottingham 1964, pp. 350-352, publ. Inst. Phys./Phys. Soc., London. *E*

G. Bosch and **A. G. van Vijfeijken**: On the specific heat of a system of quantum oscillators.
Physica **30**, 2317-2320, 1964 (No. 12). *E*

G.-A. Boutry, H. Dormont, R. Evrard and **R. Perrin**: Contribution à l'étude des propriétés photoélectriques du potassium pur, préparé et conservé dans l'ultravide.
C.R. Acad. Sci. Paris **261**, 383-386, 1965 (No. 2). *L*

H. Bremmer: Semi-geometric-optical approaches to scattering phenomena.
Proc. Symp. on Quasi-Optics, Polytech. Inst. Brooklyn 1964, pp. 415-435. *E*

J. C. Brice and **U. Pick**: The coercive force in permalloy thin films.
Brit. J. appl. Phys. **16**, 565-566, 1965 (No. 4). *M*

A. Bril and **W. L. Wanmaker**: Fluorescent properties of some europium-activated phosphors.
J. Electrochem. Soc. **111**, 1363-1368, 1964 (No. 12). *E*

A. J. Burggraaf and **J. Cornelissen**: The strengthening of glass by ion-exchange, Part 1. Stress formation by ion diffusion in alkali aluminosilicate glass.
Phys. Chem. Glasses **5**, 123-129, 1964 (No. 5).

J. Dieleman: Observation of Se^{77} superhyperfine structure on the electron-paramagnetic resonance of Fe^{3+}($3d^5$) in cubic ZnSe.
Philips Res. Repts. **20**, 206-212, 1965 (No. 2). *E*

J. A. W. van der Does de Bye: Cathodoluminescence of GaAs and GaP.
Comptes rendus 7e Congrès int. Recombinaison radiative dans les semiconducteurs, Paris 1964, pp. 243-250. *E*

W. F. Druyvesteyn: Magnetisation curves of superconducting indium lead alloys in a longitudinal and a transverse field.
Physics Letters **13**, 195-196, 1964 (No. 3). *E*

R. Evrard and **P. Beaufils**: Etalonnage des jauges à ionisation au moyen de la jauge absolue à suspension diamagnétique.
Le Vide **20**, 116-120, 1965 (No. 116). *L*

P. J. Flanders: A simple demonstration of magnetic interactions.
Amer. J. Phys. **33**, 346-347, 1965 (No. 4). *M*

H. G. Grimmeiss and **H. Scholz**: Optical and electrical properties of Cu-doped GaP. Part I: Photoconductivity.
Philips Res. Repts. **20**, 107-124, 1965 (No. 2). *A*

P. A. H. Hart: Partition effects in transverse electron-beam waves.
Thesis Eindhoven, June 1964. *E*

W. Hondius Boldingh: Grids to reduce scattered X-rays in medical radiography.
Thesis Eindhoven, Jan. 1964.

J. Kelly and **H. N. G. King:** The use and design of a high-voltage electron beam machine.
Microelectronics and Reliability **4**, 85-89, 1965 (No. 1). *M*

E. Kooi: Formation and composition of surface layers and solubility limits of phosphorus during diffusion in silicon.
J. Electrochem. Soc. **111**, 1383-1387, 1964 (No. 12). *E*

C. Kooy: Material transport in solid-state reactions.
Pure and appl. Chem. **9**, 441-452, 1964 (No. 3). *E*

J. Lemmrich: Der synchronisierte Induktionsmotor.
Elektrotechn. Z. **A 85**, 724-726, 1964 (No. 22). *H*

J. J. van Loef and **A. Broese van Groenou:** On the sublattice magnetization of $BaFe_{12}O_{19}$.
Proc. int. Conf. on Magnetism, Nottingham 1964, pp. 646-649, publ. Inst. Phys./Phys. Soc., London. *E*

F. K. Lotgering: Ferromagnetic interactions in sulphides, selenides and tellurides with spinel structure.
Proc. int. Conf. on Magnetism, Nottingham 1964, pp. 533-537, publ. Inst. Phys./Phys. Soc., London. *E*

J.-P. Mathieu: On damped vibration theory.
Int. J. mech. Sci. **7**, 173-182, 1965 (No. 3). *B*

J.-P. Mathieu: Contribution à l'étude des vibrations d'un système amorti.
Rev. univ. Mines **108**, 153-158 and 178-182, 1965 (Nos. 4 and 5). *B*

E. A. Muijderman: Spiral groove bearings.
Thesis Delft, March 1964. *E*

E. C. Munk: The equivalent electrical circuit for radial modes of a piezoelectric ceramic disc with concentric electrodes.
Philips Res. Repts. **20**, 170-189, 1965 (No. 2). *E*

D. J. van Ooijen and **A. S. van der Goot:** Critical currents of superconducting niobium-oxygen alloys.
Philips Res. Repts. **20**, 162-169, 1965 (No. 2). *E*

G. W. van Oosterhout: The structure of goethite.
Proc. int. Conf. on Magnetism, Nottingham 1964, pp. 529-532, publ. Inst. Phys./Phys. Soc., London. *E*

O. Reifenschweiler and **K. R. Fröhner:** A new principle of ion extraction from a gas discharge plasma.
Nucl. Instr. Meth. **30**, 298-302, 1964 (No. 2). *E*

R. J. Ritsma and **F. L. Engel:** Pitch of frequency-modulated signals.
J. Acoust. Soc. Amer. **36**, 1637-1644, 1964 (No. 9). *E*

F. C. de Ronde: Improvement of the wall-current detector.
IEEE Trans. on microwave theory and techniques **MTT-12**, 616, 1964 (No. 6). *E*

P. Schagen: Alternatives to thermionic emission.
Brit. J. appl. Phys. **16**, 293-303, 1965 (No. 3). *M*

A. J. Schrijner and **A. Middelhoek:** The determination of the density of Ta, Nb, and anodically formed Ta_2O_5 and Nb_2O_5.
J. Electrochem. Soc. **111**, 1167-1169, 1964 (No. 10).

H. Severin: Ferrite bei hohen Mikrowellenleistungen.
Nachrichtentechn. Z. **18**, 7-17, 1965 (No. 1). *H*

P. J. Severin: The quenching of light by microwaves incident on a negative glow plasma in a cold cathode glow discharge in helium, neon and argon.
J. quant. Spectrosc. radiat. Transfer **4**, 763-774, 1964 (No. 6). *E*

M. J. Sparnaay, A. H. Boonstra and **J. van Ruler:** The influence of chemisorption upon the electrical properties of germanium surfaces.
Surface Sci. **2**, 56-63, 1964. *E*

F. A. Staas, A. K. Niessen, W. F. Druyvesteyn and **J. van Suchtelen:** Guided motion of vortices in type II superconductors.
Physics Letters **13**, 293-295, 1964 (No. 4). *E*

J. M. Stevels: Glass-ceramics.
Science of Ceramics **2** (Proc. Conf. Noordwijk aan Zee 1963), pp. 425-431, Academic Press, London 1965. *E*

A. L. Stuijts: Low-porosity ferrites.
Proc. Brit. Ceramic Soc. **2**, 73-81, Dec. 1964. *E*

I. Sunagawa and **P. J. Flanders:** Structural and magnetic studies in hematite single crystals.
Phil. Mag. **11**, 747-761, 1965 (No. 112). *M*

M. L. Verheijke: A note on the non-proportional response of NaI(Tl) crystals to gamma rays.
Nucl. Instr. Meth. **30**, 357-358, 1964 (No. 2). *E*

J. H. N. van Vucht, D. J. van Ooijen and **H. A. C. M. Bruning:** Some investigations on the niobium-tin phase diagram.
Philips Res. Repts. **20**, 136-161, 1965 (No. 2). *E*

W. L. Wanmaker: Het 5e Internationale Symposium over de reactiviteit van vaste stoffen, 2-8 augustus 1964 te München.
Chem. Weekblad **60**, 705-712, 1964 (No. 52).

P. A. C. Whiffin and **J. W. Orton:** Rhodium as an impurity in single crystals of zinc tungstate.
Brit. J. appl. Phys. **16**, 567-569, 1965 (No. 4). *M*

J. S. van Wieringen and **J. G. Rensen:** Paramagnetic resonance at high temperatures.
Electronic magnetic resonance and solid dielectrics, Proc. XIIth Coll. Ampère, Bordeaux 1963, pp. 229-231, North-Holland Publ. Co., Amsterdam 1964. *E*

W. J. Witteman and **R. Bleekrode:** Pulsed and continuous molecular far infra-red gas laser.
Physics Letters **13**, 126-127, 1964 (No. 2). *E*

Volume 27, 1966, No. 2 pages 33-68 Published 18th August 1966

VOLUME 27, 1966, No. 3/4

Osmium dispenser cathodes

P. Zalm and A. J. A. van Stratum

621.3.032.213.2:621.385.6

It might reasonably be expected that metals of low work function would be most suitable in the development of dispenser cathodes, when high emission current density or low operating temperature are required. This turns out, perhaps unexpectedly, not to be the case. Very promising results have now been achieved with less common metals such as rhenium or ruthenium, and, in particular, with osmium.

A characteristic of dispenser cathodes, such as the "L" cathode [1] and the impregnated cathode [2], is that a functional distinction exists between the emissive surface and a reserve supply of the material that serves to keep the work function of the emissive surface sufficiently low. One possible arrangement for an L cathode is illustrated in *fig. 1.* The emission from a cathode of this type takes place from the surface of a porous tungsten body (the substrate), whose work function is lowered by adsorbed Ba and BaO (the adsorbate). Behind the tungsten body the L cathode has a storage chamber which contains a mixture of tungsten powder and barium-calcium-aluminate of composition $5BaO.3CaO.2Al_2O_3$. The structure of an impregnated cathode is somewhat different: there is no storage chamber and the barium-calcium-aluminate is adsorbed in the pores of the porous tungsten body. In the barium-calcium-aluminate, which is found to consist of three phases [3], i.e. $Ba_3Al_2O_6$, $Ba_2CaAl_2O_6$ and CaO, a chemical reaction takes place at the operating temperature of the dispenser cathodes, probably in accordance with the equation:

$$W + 3Ba_3Al_2O_6 + 6CaO \rightleftarrows 3Ba_2CaAl_2O_6 + Ca_3WO_6 + 3Ba, \quad . \; . \quad (1)$$

and this ensures an adequate supply of barium to the emissive surface.

The metallic character of these cathodes and the presence of a sufficiently large reserve of activating material make them particularly useful for applications which require a combination of long life and heavy current densities — for continuous as well as pulsed operation — and good mechanical properties.

The development of microwave tubes, such as disc-seal triodes, klystrons and magnetrons for higher and higher frequencies and output powers has benefited from these characteristics of dispenser cathodes, and has in its turn stimulated further investigations into these cathodes. Dispenser cathodes are also being used with success in many other types of thermionic valve. In all cases, however, the valves have been for professional applications.

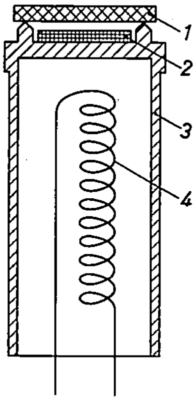


Fig. 1. Basic construction of an L cathode. *1* porous tungsten body. *2* chamber containing a reserve supply of barium-calcium-aluminate. *3* molybdenum cylinder. *4* filament.

Dispenser cathodes have not so far been used in non-professional valves for various reasons. Most important among these are the relatively high price, the high operating temperature (1110 °C) and, for some applications, the excessive evaporation of the activating material (Ba and BaO).

Dr. P. Zalm and A. J. A. van Stratum are with Philips Research Laboratories, Eindhoven.

[1] H. J. Lemmens, M. J. Jansen and R. Loosjes, A new thermionic cathode for heavy loads, Philips tech. Rev. **11**, 341-350, 1949/50.

[2] R. Levi, The impregnated cathode, Philips tech. Rev. **19**, 186-190, 1957/58.

[3] This has been demonstrated by C.A.M. van den Broek and A. Venema in this laboratory. Having regard to recent investigations, we feel justified in assuming the validity of equation (1) for the reaction.

These factors are to some extent interrelated. The high operating temperature of 1110 °C inevitably entails expensive methods of mounting the filament. In order to achieve good heat transfer and at the same time sufficient electrical insulation it is necessary to cement the filament of an L cathode into the molybdenum cylinder (*3* in fig. 1). This is a rather costly operation. The high operating temperature required is also responsible for a rapid evaporation of the activating material. Again, the development of a cheaper dispenser cathode (the pressed cathode [4]) was hampered by this evaporation. In this type the cathode body is made of compressed tungsten powder and barium-calcium-aluminate. Efforts were made to reduce excessive evaporation in these cathodes by using mixtures of tungsten and molybdenum powders, at the expense, however, of the permissible current density, as the work function of the cathodes was increased by the addition of molybdenum.

In these considerations it is seen that attempts to widen the useful applications of dispenser cathodes must be concentrated on lowering the operating temperature.

The dependence of the current density j of a thermionic cathode on the temperature T is given by Richardson's equation:

$$j = AT^2 \exp(-e\Phi/kT), \quad \dots\dots \quad (2)$$

where Φ is the work function of the cathode surface, e the electron charge, and k is Boltzmann's constant. The factor A, which is in theory 120 A/cm^{2}deg^2, is for various reasons much smaller in practice, but little can be done to improve it. If, therefore, one wishes to achieve a specified thermionic current density at a lower operating temperature, the work function has to be reduced.

If this can be done, it is then also possible to increase the current density if required, while maintaining the operating temperature at the same value. This again may be useful for certain applications in professional valves.

Possible ways of reducing the work function

The work function of the surface of a conductor is governed by two parameters: the chemical potential μ_c/e of the electrons in the conductor, and the electrostatic potential barrier φ_s at the surface [5]:

$$\Phi = -\mu_c/e - \varphi_s. \quad \dots\dots \quad (3)$$

The first of these two parameters, μ_c, depends only on the chemical composition of the conductor. The second parameter, φ_s, is determined in the first place by the surface structure, i.e. for a given metal, by the choice of crystal plane, but it is also dependent on external influences, in our case on the presence of foreign ions on the surface.

Modification of the potential barrier φ_s by means of adsorbed ions is a traditional method of reducing the work function of cathodes. The adsorbate used may for example be caesium, barium, lanthanum or thorium. For normal oxide cathodes Ba is used — here not adsorbed on a metal but on an oxide of an alkaline earth metal; dispenser cathodes also use Ba(BaO), adsorbed on a substrate of polycrystalline tungsten. For every combination of adsorbate and substrate there is an optimum degree of surface coverage at which the work function reaches a minimum value.

In order to lower further the work function of our dispenser cathodes, one might choose a different adsorbate or a different substrate. It has been found that in practice a different adsorbate does not lead to better results. Although Cs(Cs$_2$O) lowers the work function more than Ba(BaO), it is not as a rule suitable for use in thermionic valves because of its high volatility: this would so reduce the maximum permissible temperature of the cathode that the emission density would drop to a lower value than found with a Ba(BaO) layer.

However, unexpected results have been obtained at Philips Research Laboratories in Eindhoven through the use of a different *substrate*. These results depend on the paradoxical fact that *the effective work function* [6], *given an optimally covered surface, is smaller for a higher work function of the non-covered substrate*. This was predicted on theoretical grounds [7] and was later confirmed by extensive theoretical and experimental investigations with caesium as adsorbate [8].

We shall here explain this unexpected result by means of a highly simplified model. We shall then deal in more detail with the results obtained with dispenser cathodes in which the substrate is osmium instead of tungsten.

Work function theory based on a simple model

Fig. 2 shows the potential curve at the surface of the emissive body, the surface being assumed to be homogeneous (i.e. without atomic structure). We assume

[4] R. C. Hughes and P. P. Coppola, The pressed cathode, Philips tech. Rev. 19, 179-185, 1957/58.

[5] See C. Herring and M. H. Nichols, Rev. mod. Phys. 21, 185, 1949.

[6] A different choice of materials generally changes the factor A in the Richardson equation (2). To make it possible to compare cathodes *at a given temperature*, it is usual to calculate from the measured j the value of Φ for A = 120 A/cm^{2}deg^2. This value is called the effective work function. See e.g. E. B. Hensley, J. appl. Phys. 32, 301, 1961.

[7] P. Zalm, Rep. 21st Ann. Conf. phys. Electronics M.I.T., page 62, 1961.

[8] A. J. Kennedy, Adv. Energy Conversion 3, 207, 1963.
N. S. Rasor and C. Warner, J. appl. Phys. 35, 2589, 1964.

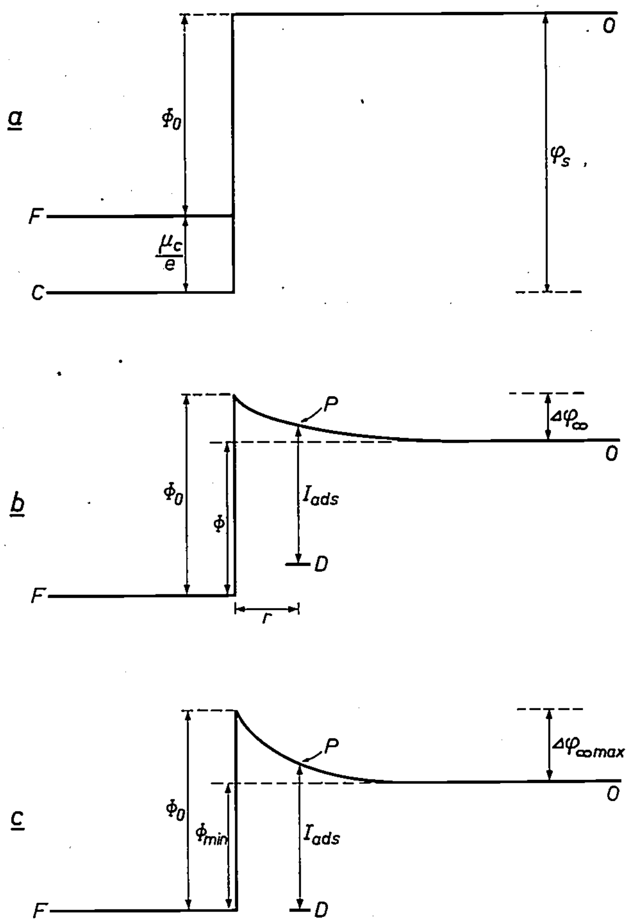


Fig. 2. Potential curve at the surface of an emissive body. As is customary, the graph shows the positive direction of the potential downwards. F Fermi level. O potential level in vacuum. Φ_0 work function.

a) Without adsorbate. C is the lower limit of the conduction band in the metal. For μ_c/e and φ_s see eq. (3).

b) With a certain quantity of adsorbate. The adsorbed atoms give rise to donor levels D which are at a potential difference I_{ads} below the local potential P. The work function has been reduced from Φ_0 to Φ.

c) With optimum coverage. The level D coincides with the Fermi level. The work function has the minimum value, Φ_{min}.

for simplicity that all the electrons in the metal are at the Fermi level F; at a considerable distance away they are at the vacuum level O. The potential difference which the electrons have to overcome with the aid of thermal energy in oider to leave the metal — the work function — is denoted by Φ_0 for the case where there are no adsorbed ions (fig. 2*a*). The values of μ_c and φ_s are here taken into account (see eq. 3). The value of φ_s is now altered by the adsorption of atoms able to give up electrons to the metal — thus acting as a kind of surface donor. In fact the ions formed at the surface, together with their mirror image charge, effectively form a charged capacitor, so that the total potential curve has the form shown in fig. 2*b*. This lowers the work function by an amount $\Delta\varphi$.

How far can this process of adsorption and lowering of Φ continue? If only one atom is adsorbed it gives rise to a donor level which is lower than the local potential (as yet unchanged) by an amount which we shall call I_{ads}: If this donor level is higher than the Fermi level, an electron will as a rule be given up to the metal. As the number N of adsorbed ions per unit area increases (all at a distance r from the ideal surface) the local potential decreases — causing a decrease of Φ — and consequently the height of the donor level above the Fermi level also decreases (fig. 2*b*). There will no longer be any ionization, and thus the process of adsorption with ionization and lowering of Φ cannot continue, as soon as the level of the donors *coincides* with the Fermi level. In this situation — the optimally covered surface — where N_{opt} ions are adsorbed and the local potential at r is lowered to a height I_{ads} above the Fermi level, we have the lowest work function, Φ_{min}; see fig. 2*c*.

In reality the nature of the process gradually changes before we reach this stage, because the surface coating contains an increasing fraction of *non*-ionized donors, and moreover the effect of the adsorption on the potential curve gradually decreases — or even changes sign — because to a certain extent the adsorbed ions can form dipoles. All these factors can however be disregarded in our model.

Consider now a different substrate, which has a *higher* work function Φ_0 in the uncoated state. The same considerations can be applied in this case. Although I_{ads} and r depend slightly on the substrate, their influence can be neglected, Here also, for an optimally covered surface, the local potential at the location r will have dropped to a height I_{ads} above the Fermi level. If we try to draw a continuous potential curve that passes through this fixed point and through the starting point, located at the metal surface, which is at a height Φ_0 above the Fermi level (*fig. 3*), we see that the final level, at a considerable distance from the surface, will naturally be smaller if the starting point is made higher. A substrate with a *higher* Φ_0 therefore gives a *lower* Φ_{min}. This is the paradoxical effect already mentioned. Its explanation, using the picture we have sketched, is that the number of ionized donors providing the optimum coverage, N_{opt}, rises with increasing Φ_0 to such an extent that the increase of Φ_0 is more than compensated.

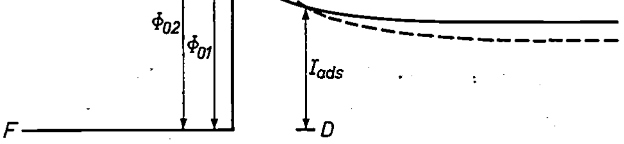


Fig. 3. Starting from a *higher* Φ_0 ($\Phi_{02} > \Phi_{01}$), a *lower* Φ_{min} (dashed curve) may be expected when the surface is optimally covered.

A rough calculation based on the model yields expressions for N_{opt} and for the final work function Φ_{min}, which lead to the same conclusion.

The calculation is as follows.

If there are N ionized atoms present per unit area, then the potential at the surface with respect to that at a considerable distance is increased by:

$$\Delta\varphi_\infty = \frac{1}{\varepsilon_0} N e r . \quad \ldots\ldots\ldots \quad (4)$$

In fact, if we have a parallel-plate capacitor with vacuum dielectric and a spacing between the plates of $2r$ — i.e. the distance between the ions and their mirror image — then the voltage with respect to the plane of symmetry is given by eq. (4) if the charge per unit area is Ne. The local potential at the location of an ion can be calculated by assuming that the ion has a circular area πd^2 available to it (by definition of N, $\pi d^2 N = 1$), while all other adsorbed ions are distributed homogeneously over the rest of the surface. This is a plausible simplification in view of the migration of the adsorbed ions. The homogeneously distributed ions, together with their mirror image, again form a parallel-plate capacitor with a spacing between the plates of $2r$, but in which there is now a circular hole of diameter $2d$. It can be shown by a simple calculation that, at the location of the ion, in the centre of this hole, there must be a potential difference with respect to the metal surface, of magnitude:

$$\Delta\varphi_r = N\left(\sqrt{d^2 + 4r^2} - d\right)e/2\varepsilon_0 ,$$

or, as $\pi d^2 N = 1$:

$$\Delta\varphi_r = \sqrt{\pi N}\left(\sqrt{1 + 4\pi r^2 N} - 1\right)e/2\pi\varepsilon_0 . \quad \ldots \quad (5)$$

We have seen that the optimum ion coverage, N_{opt}, is reached as soon as the local potential has dropped to the height I_{ads} above the Fermi level. Therefore:

$$\sqrt{\pi N_{\text{opt}}}\ \left(\sqrt{1 + 4\pi r^2 N_{\text{opt}}} - 1\right)e/2\pi\varepsilon_0 = \Phi_0 - I_{\text{ads}} . \quad \ldots \quad (6)$$

On the other hand we then have according to eq. (4):

$$\Delta\varphi_{\infty\,\text{max}} = \Phi_0 - \Phi_{\text{min}} = N_{\text{opt}}\, re/\varepsilon_0 . \quad \ldots \quad (7)$$

By eliminating N_{opt} from (6) and (7) we find the relation of interest between Φ_{min} and Φ_0:

$$\Phi_0 - I_{\text{ads}} = \sqrt{\frac{e(\Phi_0 - \Phi_{\text{min}})}{4\pi\varepsilon_0 r}} \left\{\sqrt{1 + \frac{4\pi\varepsilon_0 r}{e}(\Phi_0 - \Phi_{\text{min}})} - 1\right\} . \quad \ldots \quad (8)$$

It is easily verified that, provided the condition $\Phi_0 > I_{\text{ads}}$ is fullfilled, $\mathrm{d}\Phi_{\text{min}}/\mathrm{d}\Phi_0$ is in fact negative — and this is the paradoxical effect mentioned above.

Still keeping to our model, Φ_{min} can easily be calculated for the case where N_{opt} is still fairly small, so that the ions cover only a small part of the metal surface. In this case the expression between brackets in eq. (6) is approximately $2\pi r^2 N_{\text{opt}}$, and eq. (8) simplifies to:

$$\Phi_{\text{min}} = \Phi_0 - \sqrt[3]{(\Phi_0 - I_{\text{ads}})^2 e/\pi\varepsilon_0 r} . \quad \ldots \quad (9)$$

For a caesium adsorbate we put $r = 1.65 \times 10^{-10}$ m and $I_{\text{ads}} =$ 3.87 volt, which are the values for the free caesium ion and the ionization potential of the free caesium atom. We may then expect that, because of the condition referred to, a negative $\mathrm{d}\Phi_{\text{min}}/\mathrm{d}\Phi_0$ will arise in metals for which $\Phi_0 > 3.87$ V. On the other hand, equation (9) can only be applied to metals for which $\Phi_0 < 5.1$ V: at greater Φ_0 the value of N_{opt} is so high that the approximation used is no longer sufficiently accurate.

As an example we can use the equations to calculate Φ_{min} for the adsorption of caesium on two different tungsten surfaces: a (100) plane and a (112) plane. For the (100) plane we have $\Phi_0 = 4.65$ V [9]. With the values of I_{ads} and r for caesium the equations yield: $N_{\text{opt}} = 9.25 \times 10^{13}/\text{cm}^2$ and $\Phi_{\text{min}} = 1.88$ V. For the tungsten (112) plane a *higher* value of Φ_0 applies: 4.85 V, and the equations give: $N_{\text{opt}} = 1.08 \times 10^{14}/\text{cm}^2$ and $\Phi_{\text{min}} = 1.63$ V, that is to say a substantially *lower* work function.

Although, in view of the nature of the model, the calculations only have a qualitative significance, these results do not, remarkably enough, differ very much from the values found by experiment. If they are to be compared, it should be noted that only the *total* quantity of adsorbate can be determined experimentally, i.e. the sum of ionized and non-ionized donors, whereas N_{opt} relates only to the ionized ones. The ratio between the two quantities when the surface is optimally covered can be calculated on a statistical basis; for caesium it is 1 : 2, so that for the two cases we find from the calculated values of N_{opt} a total occupation of $2.8 \times 10^{14}/\text{cm}^2$ for the (100) plane of tungsten and $3.2 \times 10^{14}/\text{cm}^2$ for the (112) plane. In recent measurements [10] on a polycrystalline tungsten surface, which had been thermally etched before the adsorption of caesium the minimum work function found was $\Phi_{\text{min}} = 1.78 \pm 0.01$ V, the adsorption amounting to 2.6 to 2.8×10^{14} atoms/cm^2.

Measurements on various substrates using a barium adsorbate

Much less is known experimentally about the way in which the work function of the substrate affects the resultant work function when Ba(BaO) is used as adsorbate instead of caesium, nor is it possible to calculate Φ_{min} for barium from the purely ionogenic model used, as the value to be inserted for I_{ads} is not known. However, investigations by Rittner [11] on tungsten and molybdenum, and by Sackinger and Brodie [12] on rhenium, gave analogous results: here also Φ_{min} is found to decrease with increasing Φ_0.

On the basis of the foregoing, we have investigated various metals to determine their usefulness as a substrate for dispenser cathodes. The choice of these metals was dictated by the requirements that the work function in the uncoated state should be higher than 4.54 V, the work function of tungsten, and that the metals should not alloy with Ba, nor react with the barium-calcium-aluminate mixture used as the filling in dispenser cathodes.

Rhenium, ruthenium, iridium and osmium are metals that meet these requirements. The work function of polycrystalline rhenium is 5.1 V. The work functions of polycrystalline iridium and osmium have been found in our laboratory [13] to have values of 5.50 ± 0.05 V and 5.93 ± 0.05 V, respectively. The work function of ruthenium is not known exactly but is in any case higher than that of tungsten. These four metals were applied in layers between 0.1 to 1 μm thick to the porous tungsten body of a dispenser cathode, which in other respects was made in the normal way, either as an L cathode or as in impregnated cathode. The pores of the tungsten have dimensions of the order of 10 μm, so that the thinly applied metal coating in no way inhibits the replenishment process in the dispenser cathode.

As expected, after adsorption of Ba(BaO) the effective work function decreased as the work function of the substrate increased. The differences found between dispenser cathodes using Re, Ru, Ir and Os were not very considerable. The work function of the cathodes coated with osmium, however, was plainly the lowest. We therefore constructed a number of osmium-coated cathodes and subjected them to various extensive tests.

Osmium-coated cathodes

Current density

Fig. 4 shows the current density as a function of temperature for an L cathode with 0.5 μm thick osmium coating and for a normal L cathode. The measurements were made with a simple diode with an anode-cathode spacing of 0.4 mm, at an anode voltage of 1000 V and under pulsed loading, the pulse length being 50 μs and the repetition frequency 50 pulses per second. It can be seen from the figure that at 800 °C the current density of the Os cathode is about 10 times higher than that of an uncoated cathode. At 1050 °C the current densities still differ by a factor of 2.5. It can also be seen that the osmium cathode needs a temperature 100 °C lower than that of a normal L cathode for the same current density.

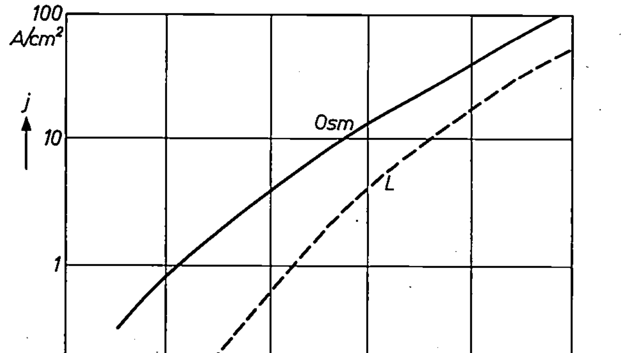


Fig. 4. Current density j as a function of cathode temperature T, for a normal L cathode (dashed curve) and for an L cathode with an osmium coating 0.5 μm thick (full curve).

Calculation of the work function from the curves gives $\Phi = 1.60$ V and a value of $A = 10$ A/cm²deg² for the Os cathodes, compared with $\Phi = 1.95$ V and $A = 50$ A/cm²deg² found for the uncoated cathodes.

More detailed information on the emission characteristics of an Os cathode and a normal L cathode is given in *fig. 5*. In this figure the abscissa represents the effective work function at 800 °C (i.e. the Φ value calculated for A assumed at 120 A/cm²deg² [6]), and the ordinate gives for each value of Φ_{eff} the fraction of the total cathode surface at which that value is found. It is seen that the average Φ_{eff} for the Os cathode is 1.68 V and that for a normal L cathode 1.90 V. An-

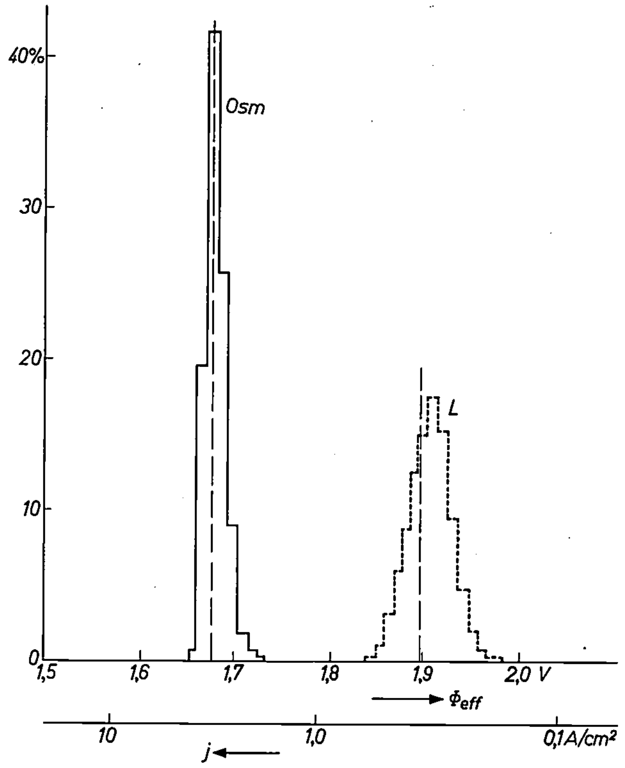


Fig. 5. Emission from the same two cathodes as in fig. 4, at 800 °C. The abscissa represents the effective work function Φ_{eff} (and also the corresponding current density), and the ordinate gives, for each value of Φ_{eff}, the fraction of the total cathode surface at which that value is found. (The measurements were made by a method described some time ago in this journal; see C. G. J. Jansen, A. Venema and Th. H. Weekers, Philips tech. Rev. 24, 402, 1962/63.)

[9] This value for the work function can be derived from the experimental data given by A.G.J. van Oostrom, Validity of the Fowler-Nordheim model for field electron emission, thesis, Amsterdam, June 1965 (Philips Res. Repts. Suppl. 1966, No. 1).

[10] B. H. Blott, B. J. Hopkins and T. J. Lee, Surface Sci. 3, 493, 1965 (No. 5).

[11] E. S. Rittner, R. H. Ahlert and W. C. Rutledge, J. appl. Phys. 28, 156, 1957; E. S. Rittner and R. H. Ahlert, J. appl. Phys. 29, 61, 1958.

[12] J. P. Sackinger and I. Brodie, Rep. 21st Ann. Conf. phys. Electronics M.I.T., page 40, 1961.

[13] The values were derived from photoelectric measurements made by G. J. M. van der Velden.

other feature of note in the graph is the reduced spread in the work function of the Os cathode. This feature can be important in valves in which the emitted electrons should as far as possible have the same energy.

Life

The life of an L cathode is proportional to the quantity of barium-calcium-aluminate in the storage chamber. For a given porosity of the tungsten body, the quantity of Ba (and BaO) lost by evaporation per unit time and per unit cathode surface depends only on the cathode temperature (*fig. 6*). The life of a normal L cathode operating at 1110 °C is in the region of 15 000 hours. An osmium-coated cathode with the same reserve supply of barium-calcium-aluminate may be expected to have the same life if operated at 1110 °C; if, however, the Os cathode is operated at a temperature of 1010 °C, which is sufficient for the same emission, the cathode should have a 6 to 10 times longer useful life.

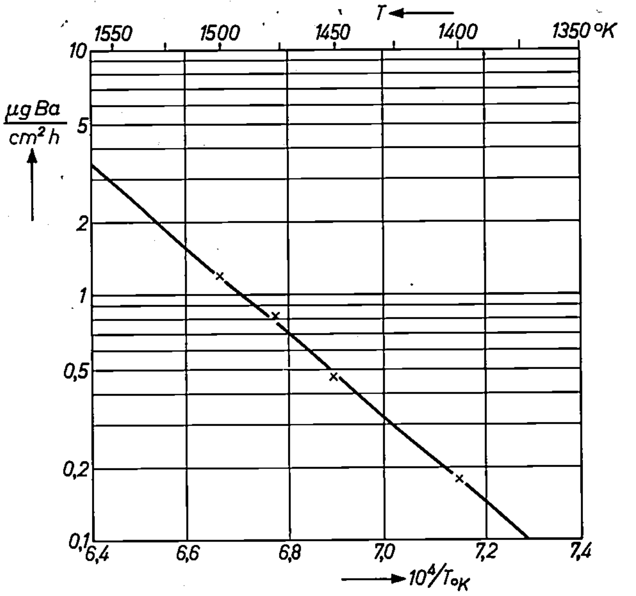


Fig. 6. Evaporation of Ba as a function of cathode temperature T at a given porosity of the tungsten body (after unpublished measurements performed by C. A. M. van den Broek and A. Venema).

Two series of cathode life-test experiments have been started to check this prediction, using Os cathodes of 8 mm^2 surface area, with 0.5 μm osmium, mounted in diodes with molybdenum anodes and operated at 100 mA, i.e. 1.3 A/cm^2, at a cathode temperature of 1110 °C in one series and at 1010 °C in the other. These experiments have now been in progress for 14 000 hours — there are unfortunately only 9000 hours in the year. During this time the saturation emission has remained virtually constant. This means that the advantage of the lower operating temperature has been brought into effect; whether the extremely long life predicted will be achieved in these experimental valves is not yet known. In fact, it is conceivable that the osmium coating will eventually disappear as a result of evaporation, diffusion or migration. Nothing of this has yet been noticed at temperatures lower than 1050 °C, even with the thinnest coatings of 0.1 μm. At temperatures higher than 1150 °C, however, there are some indications of coating loss. A separate series of tests has been started to investigate this effect as a function of cathode temperature and coating thickness, but these tests will take a considerable time to complete because of the long lifetimes involved.

Osmium-coated impregnated cathodes are more subject to coating loss. At a cathode temperature of 1110 °C the "osmium properties" have disappeared after only about 5500 hours of operation, and the cathode then behaves in the same way as an uncoated impregnated cathode. At a cathode temperature of 1060 °C this happens after about 7500 hours, but at a temperature of 1010 °C this type of cathode also shows no signs of osmium loss, even after 14 000 hours.

The exhaustion of the reserve of barium-calcium-aluminate, which we have mentioned as the primary cause of limitation of life, is of course not the only possible cause. The filament may break, the porous tungsten body may become deformed, the valve characteristics may be affected by the evaporated Ba and BaO (insulation faults, grid emission). At the lower operating temperature of the osmium cathodes, all these factors as well will only make their effect felt after a much longer running time than with normal L cathodes.

Secondary electron emission

In some applications, for example in magnetrons, the contribution of the secondary electron emission to the total current density is very important. Although no particular attention was paid to this point in the development described here, it is nevertheless a welcome incidental feature that the secondary emission coefficient of the osmium-coated cathodes is appreciably higher than that of normal dispenser cathodes. It is worth noting that in this respect as well, osmium is by far the most suitable of the metals investigated.

In *fig. 7* the results of measurements of the secondary emission coefficient are presented as a function of the energy of the primary electrons, for an osmium-coated L cathode, for a normal L cathode and for a normal impregnated cathode. Although the differences in the secondary emission coefficient are evidently not very large, they are nevertheless important. In a 4 mm magnetron type DX 164, for example, the maximum current density is 190 A/cm^2 with an uncoated impregnated cathode, but 320 A/cm^2 with an Os cathode.

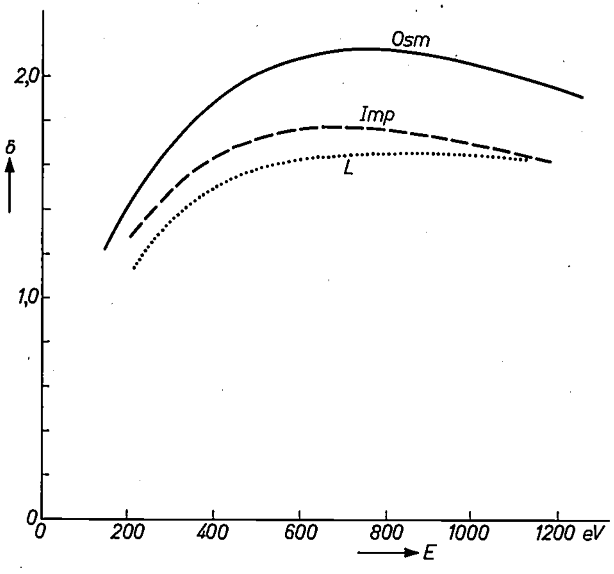


Fig. 7. Secondary emission coefficient δ as a function of the energy E of the primary electrons (in eV) for a normal L cathode (*L*), an impregnated cathode (*Imp*) and an osmium-coated L cathode (*Osm*).

Other practical data

As shown in the foregoing, the reduction in the temperature required for operation increases the possibility of using dispenser cathodes in non-professional valves, where the expense of initial purchase and replacements is such an important consideration. The development of such applications is still in the initial stage. As indicated, however, the new cathodes with their higher performance may also be useful in various professional valves, in particular in microwave valves. We have already mentioned magnetrons. The use of Os cathodes is also being investigated in disc-seal triodes [14] and in various reflex klystrons.

In the *disc-seal triodes* osmium-coated cathodes are under test both at 1150 °C (the usual temperature for uncoated cathodes in these tubes) and at 1020 °C. In both cases the cathode loading is 0.75 A/cm^2 and up to now the useful life is longer than 10 000 hours. Because of the low barium evaporation and the imperceptibly small deformation of the tungsten body, the valve characteristics, particularly for cathodes operated at 1020 °C, are remarkably constant compared with those of valves with uncoated cathodes. From these data the expected life of Os cathodes at an operating temperature of 1020 °C (970 °C black-body temperature) may be estimated at several times 10 000 hours — more than sufficient for valves to be used in interplanetary missions.

The use of Os cathodes makes it possible to make *reflex klystrons* for even shorter wavelengths than have up to now been possible. Normal uncoated L cathodes have a reasonably long life in continuous operation at loadings up to 8 A/cm^2. This is just sufficient, as B.B. van Iperen has shown, for a 2 mm reflex klystron to be feasible [15], albeit with marginal characteristics. Using an Os cathode, G. H. Plantinga of this laboratory has succeeded in producing a practical 2 mm reflex klystron (type DX 247) in which the current density is 20 A/cm^2. With the same type of cathode he was also able to develop a 1.5 mm reflex klystron. In this case continuous loading at about 25 A/cm^2 is required.

[14] On disc-seal triodes see J. P. M. Gieles, Philips tech. Rev. **19**, 145, 1957/58.

[15] B. B. van Iperen, Philips tech. Rev. **21**, 221, 1959/60.

Summary. To widen the useful scope of dispenser cathodes, the work function of the emissive tungsten surface should be reduced to a value lower than that already achieved with the coating of adsorbed barium. This has been made possible by the paradoxical effect that the work function of a metal surface optimally covered with barium is lower for a *higher* work function of the *uncoated* metal. This effect is explained with the aid of a simple model, and by means of an approximate calculation equations are derived for estimating the work function to be expected. The predictions are found to agree very reasonably with experiments. With a view to the practical application of the effect, dispenser cathodes were made with the porous tungsten body variously coated with a thin layer (0.1 to 1 μm) of rhenium, ruthenium, iridium or osmium. The best results, i.e. a lowering of the work function from 1.95 V to 1.60 V, were obtained with osmium. This means that an osmium-coated L cathode operated at 800 °C has a current density 10 times higher than a normal L cathode. This improvement can be utilized either for operating the cathode at a lower temperature, thus making more economical designs and longer lifetimes possible (e.g. in disc-seal triodes several times 10 000 hours) or for achieving even higher current densities (e.g. 25 A/cm^2 in a 1.5 mm reflex klystron).

Physiological aspects of the lighting of tunnel entrances

D. A. Schreuder

628.971.6:624.19:612.84

In the last few years the Philips Lighting Laboratory in Eindhoven has made an extensive investigation to find how road tunnels should best be illuminated so that traffic can drive through safely without having to travel unduly slowly. Some of the experiments are described in the present article. They relate in particular to the effects on the vision of a driver as he approaches and enters a long and therefore relatively dark tunnel. The practical conclusions drawn from these experiments have had an important bearing on the design of the lighting system for the Coen tunnel, now being built under the North Sea Canal to the west of Amsterdam.

Introduction

The growing volume of traffic, both on land and water, has the result that it is increasingly necessary to make use of tunnels for the intersections of traffic routes. For a simple road under-pass the tunnel is usually short enough for a driver to see from one end to the other, and no artificial lighting may be needed.

The situation is quite different in tunnels built under rivers and canals, or through mountains. A driver passing through such tunnels is unable to see obstacles silhouetted against the light of the exit. Here, effective lighting during the daytime (and at night as well) is essential [1].

The problems in tunnel lighting design which we want to examine in this article, arise in the last instance because, without installing exorbitantly expensive and elaborate systems, it is not possible to illuminate the inside of the tunnel in such a way that the brightness equals or even approaches that of a sunlit road. Even in well-lit tunnels, drivers will experience a decrease in the luminance level: this may often fall to a value no higher than 1/500 of the luminance outside. A driver's eyes can only cope with such a change of luminance if the change does not take place too quickly. If the eye is not given enough time, its state of adaptation remains for some time far behind that appropriate to the luminance level encountered. The driver then sees virtually nothing.

The relatively slow adaptation of the eye is not, however, the only visual difficulty involved. Difficulties can already begin to arise while the motorist is still approaching the tunnel entrance in full daylight. These are a result of the *induction* effect, in which the sensitivity (state of adaptation) of a particular part of the retina is influenced by that of the neighbouring parts. The strong illumination of a fairly large area of the retina prevents the sensitivity of the part of the retina on which the tunnel entrance is imaged from adjusting itself to the relatively low level of luminance at the entrance: it adjusts itself instead to an appreciably higher level. If this effect is very pronounced, the motorist sees the tunnel entrance as a "black hole" in which no details can be discerned. In other words, he cannot see into the tunnel at all [2]; see *fig. 1.*

The lighting engineer has to design transitional lighting for a tunnel in such a way that the state of adaptation of the eyes of a driver approaching the tunnel entrance does not lag too far behind, and also in such a way that no inconvenient „black-hole effect" occurs. In the Philips Lighting Laboratory at Eindhoven, our investigations have been directed towards a better understanding of these effects, so that rules and recommendations can be drawn up for the design of tunnel lighting systems. The principal results of these investigations are firstly the conclusion that the black-hole effect is not troublesome if the luminance in the first part of the tunnel is at least 10% of the level outside, and secondly that adaptation presents no problems when 10 to 15 seconds are available for the transition from daylight to the lowest level of luminance inside the tunnel. For a motorist travelling at a speed of 72 km/h (45 m.p.h.) these times correspond to distances of not less than 200 to 300 metres.

In this article we shall describe the investigations and set out the main conclusions from the results. We shall then briefly discuss some experiments made with models at Eindhoven, in which these conclusions were put to the test. Finally, we shall give a concise

Dr. Ir. D. A. Schreuder is with the Lighting Laboratory of the Philips Lighting Division, Eindhoven.

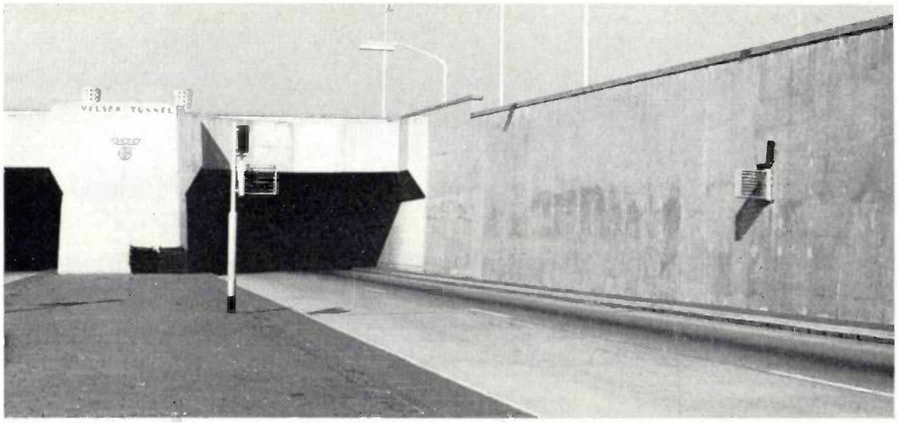


Fig. 1. An insufficiently lit tunnel entrance may look like a black hole. An approaching driver is unable to see into the tunnel. (Since the photograph was taken the entrance lighting for this tunnel has been very much improved.)

survey of the recommended rules for the design of a tunnel lighting system [3], and a short description of the entrance lighting, designed along these lines, for the Coen tunnel now being completed under the North Sea Canal to the west of Amsterdam. Before dealing with the experiments concerning the relevant adaptation phenomena, we shall go into the nature of these effects in somewhat more detail, and show that adaptation and induction, although related, can in our case be studied separately.

Distinction between adaptation and induction

In this article we shall take induction to be the effect in which, for a non-uniform distribution of luminance over the object, the sensitivity of an element of the retina does not correspond exactly to the luminance of the part of the object imaged on it, but depends on the luminance of other parts as well. In some measure this is due to the scattering of light in the eye.

The sensitivity of a given element of the retina is called the *state of adaptation* of that element. The same state of adaptation of a certain part of the retina can result from many different luminance distrubutions at the object. One of these distributions, of course, is a uniform one, and the luminance appropriate to this distribution is termed the *adaptation level*.

If we do not have steady-state conditions, but the luminance of the visual field or part of it changes, then one must expect a change in induction effects, and therefore a change in the state of adaptation as well. Some time after the change, each part of the retina reaches a new state of adaptation. The change is the process of *adaptation* and the time necessary for the adjustment is the *adaptation time*.

Assuming that a state of adaptation can be defined at any given instant during the adaptation process, we can describe adaptation in terms of a succession of adaptation levels, that is to say by a curve representing luminance values of a hypothetical visual field. During adaptation the hypothetical luminance value usually differs from the actual luminance in the visual field, and the difference between them is called the *adaptation defect*. If the luminance of the visual field changes slowly, the adaptation defect is small; if, however, the luminance changes abruptly, the adaptation defect is considerable. We shall confine ourselves in the following to what happens when the luminance in a part of the visual field decreases abruptly, or at least rapidly, which is the usual situation on entering a tunnel. The adaptation effects at the tunnel exit are much less inconvenient, and will not be considered.

The relatively low speed with which the state of adaptation of a part of the retina adjusts itself to a change in the luminance of the corresponding part of the visual field is responsible for the familiar after-images on the retina. Suppose, for example, that a person is observing a screen which is uniformly illuminated except for one small patch, which is brighter. The luminance of this small patch is then reduced to that of the rest of the screen. The state of adaptation of the part of the retina which received the image of the path does not follow the change of luminance instantaneously. The sensitivity at first remains appropriate to a higher luminance. The observer now underevaluates the luminance of the area of the screen where a moment ago the luminance was still high, i.e. he sees a dark area of the same shape. For obvious reasons, this is referred to as a negative after-image.

A change in the distribution of luminance in the

[1] Tunnels that can be seen through, but which also require artificial lighting in the daytime, will not be considered in this article.
In this connection see: D. A. Schreuder, Short tunnels, Int. Lighting Rev. **16**, 95-99, 1965 (No. 3).

[2] The fact that there are two distinct disturbing effects was first noted by the late Dr. A. M. Kruithof (Philips tech. Rev. **10**, 299-305, 1948/49).

[3] A more extensive treatment of this topic will be found in: D. A. Schreuder, The lighting of vehicular traffic tunnels, Philips Technical Library, Eindhoven 1964. The conclusions drawn from the experiments have also been reported in: Aanbevelingen voor tunnel verlichting, Electro-techniek **41**, 23-32 and 46-53, 1963, and have been issued separately by the Nederlandse Stichting voor Verlichtingskunde (Netherlands Illuminating Engineering Society).

visual field generally results in a change of induction. Since induction changes involve an adjustment of retinal sensitivity, it is reasonable to consider this as an adaptation effect. As here the sensitivity of the whole retina (or a considerable part of it) changes, the effect may be called *general adaptation*.

The pupillary reflex, i.e. the change in the diameter of the pupil as a result of a change in the luminance of the visual field, is often also considered as an adaptation effect. However, as under practical conditions the variations of pupil diameter are relativily small, the influence of the pupillary reflex can be neglected in discussing the lighting of tunnel entrances [4].

What happens when a driver approaches a tunnel entrance can now be described somewhat schematically as follows. When the driver is still at some considerable distance from the tunnel, the state of adaptation of his visual system is not yet influenced by the dark entrance. As he comes nearer a moment is reached when, owing to the changes in the distribution of luminance in the visual field, the state of adaptation of the driver's eyes starts to be influenced: the adaptation begins. The point on the road where this happens is termed the *adaptation point*. Until this point is reached, the state of adaptation of the whole retina remains unchanged and any limitation of observation is due solely to induction. Up to this moment the eye is adapted to the mean value of luminance in the open field outside the tunnel. After the adaptation point has been passed, the limitation imposed on observation is due solely to adaptation effects, (the *change* of induction being also an adaptation effect). The induction and adaptation effects do not therefore limit observation simultaneously, but consecutively, and this means that they can be studied separately.

The location of the adaptation point and the luminance of the visual field in front of the tunnel entrance are two important factors. The adaptation point is not at exactly the same place for every driver, nor is it the same for every journey; it may often be no more than 30 or 40 metres in front of the tunnel entrance. Since a driver travelling at a speed between 70 and 100 km/h must be able to see at least 100 metres of the road ahead, this means that in most cases he must be able to see into the tunnel before he has reached the adaptation point.

The outside luminance determines to a considerable extent the inconvenience experienced by the driver upon approaching a tunnel; this applies both to induction and adaptation effects. It is therefore important to know the maximum value of the outside luminance that will be encountered. Daylight recordings have shown that the illumination level measured in the horizontal plane in the summer is quite often higher than 100 000 lux. Under conditions of total reflection, this would correspond to a luminance of about 30 000 cd/m^2. The reflection factor of the surfaces near the tunnel entrance is not usually more than 0.25. Values of luminance higher than 8000 cd/m^2 need not be taken into account, and in this article we shall take this value as the maximum for practical purposes.

Induction

We shall now describe the laboratory experiments carried out at Eindhoven to determine the requirements for the lighting of a tunnel to avoid the black-hole effect. These experiments were concerned first and foremost with the visibility of objects: a study was made to determine the luminance required at the part of the visual field corresponding to the tunnel entrance to enable objects silhouetted against it to be distinguished. Particular attention was paid to the dependence of this luminance level on the contrast between the object and its direct background. In the experiments, the observer's state of visual adaptation was kept constant to ensure that the effects involved would be purely induction effects, or, in other words, to ensure a fair imitation of the conditions experienced by a driver who has not yet passed the adaptation point. Later in this article it will be shown that the conclusions on tunnel lighting which have been drawn from these experiments are not significantly affected by the fact that, in other respects, there are differences between the experimental conditions and those actually found on the road.

As a yardstick of visibility we introduced a "critical" object. We chose a square object, 20 × 20 cm, situated 100 m in front of the vehicle — at 100 m a length of 20 cm corresponds to 7 minutes of arc — and having a reflection factor giving a contrast of 20% with the direct background [5]. We considered the visibility of the critical object as sufficient when there was a 75% probability of seeing it "at a glance", that is to say in about 0.1 s. Contrasts of 20% between vehicles and the background against which they are seen are fairly frequent, smaller contrasts are seldom found. The choice of this critical object and of the time of observation are of course to some extent arbitrary. It will be shown, however, that the results of the experiments do not differ very widely, even when the various parameters are altered considerably.

The arrangement for the induction experiments is represented schematically in *fig. 2*. Visual adaptation was determined by a screen S of luminance L_1 illuminated by lamps V_1 (adaptation level = L_1). A shutter *B* in this screen could be opened for a short time to display an inner field of luminance L_2 containing an

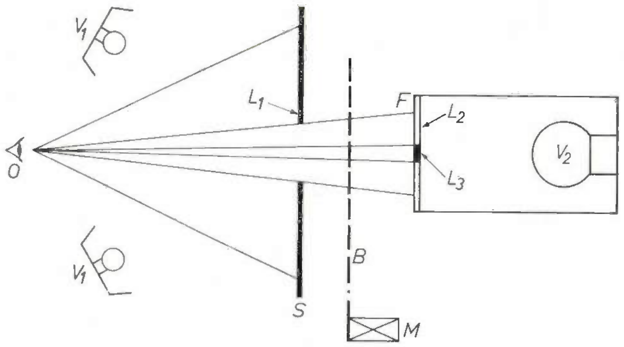


Fig. 2. Arrangement for induction experiments. S large screen illuminated by lamps V_1. The luminance L_1 of the screen determines the adaptation level of the observer's eye O. B shutter driven by motor M. Every six seconds the shutter opens and displays for 0.1 s an inner field of luminance L_2, in which there is an object of luminance L_3 (see fig. 3). The inner field is produced by illumination of the neutral filter F by the lamp V_2.

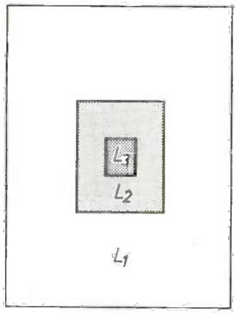


Fig. 3. Schematic representation of the observer's visual field when the shutter is open. He sees an object with luminance L_3 against a background with luminance L_2. The luminance L_1 of the screen governs the adaptation.

object of luminance L_3 (*fig. 3*). When the shutter was closed the luminance at the location of the inner field was also L_1. The inner field with the object was produced by means of a glass plate F on which there was a vacuum-evaporated neutral filter. The transmission at the object location was lower than elsewhere, and the filter was illuminated from behind by a lamp V_2. By means of a rotating shutter, driven by a synchronous motor M, the inner field was made visible every six seconds for the quoted 0.1 second. This time is short enough to meet the requirement that no adaptation should occur. The coarse adjustment of luminances L_1 and L_2 was made by means of neutral filters and the fine adjustment by varying the voltage of the lamps. (If large variations are made by changing the lamp voltage alone there is an excessive change in the colour of the light.) Since the ratio L_2/L_3 depends solely on the filter, a variation of the illumination of the filter F caused no change of contrast. To find out how far the results were governed by the choice of the critical object, various filters F were used with different L_2/L_3 ratios, that is to say measurements were carried out with different contrasts between object and background. A photograph of the experimental arrangement can be seen in *fig. 4*.

At the beginning of each series of experiments the observer looked at the screen for some time (about 5 minutes) with the shutter closed. After his eyes had become sufficiently adapted, the shutter was opened a number of times. The same glass plate F was used throughout in one series of experiments — constant contrast between object and background — but the luminance L_2 was varied; each value of L_2 occurred five times in a single series. The observer now had to indicate, every time the inner field was exposed, whether he had seen the object or not. A difference from the situation on the road was of course that the observer knew where the object was and what it looked like, so that he did not have to look for it. With the aid of statistical methods, a curve was drawn through the 25 to 35 points obtained from each series of measurements. These curves (regression curves) indicated the dependence of the percentage of positive answers on L_2. The values of L_2 appropriate to observation probabilities p of 50% and 75% were derived from these curves. Similar series of experiments were carried out for various other values of L_1 and various other contrasts, with ten observers in each series.

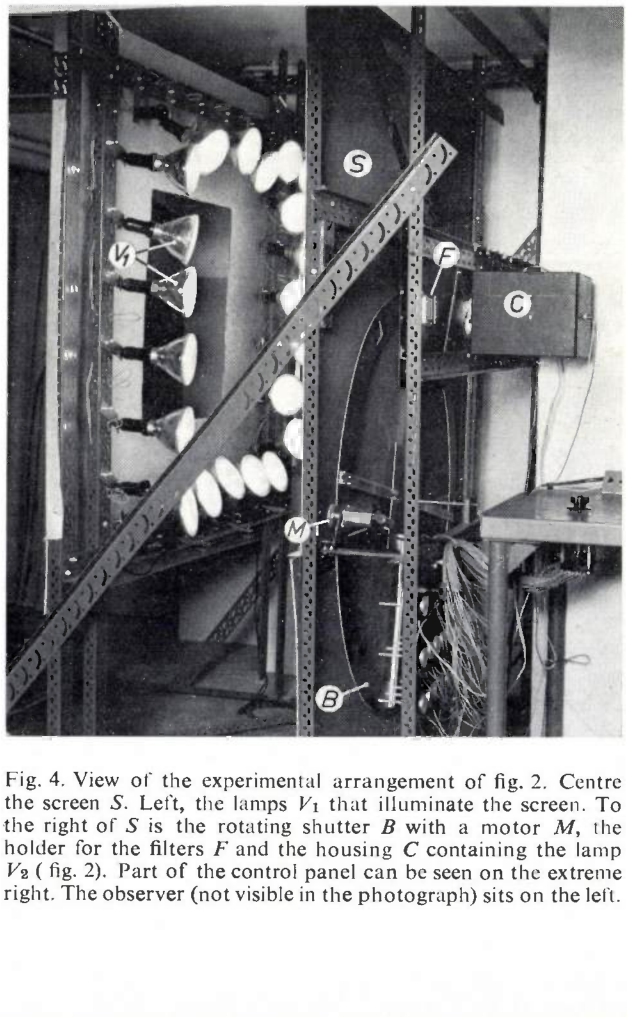


Fig. 4. View of the experimental arrangement of fig. 2. Centre the screen S. Left, the lamps V_1 that illuminate the screen. To the right of S is the rotating shutter B with a motor M, the holder for the filters F and the housing C containing the lamp V_2 (fig. 2). Part of the control panel can be seen on the extreme right. The observer (not visible in the photograph) sits on the left.

[4] See: H. Bouma, Receptive systems — mediating certain light reactions of the pupil of the human eye, Thesis, Eindhoven 1965; also published in Philips Res. Repts. Suppl. 1965, No. 5.

[5] We take contrast to be the difference between the luminances of object and background, divided by the background luminance. The ratio is expressed as a percentage.

The results [6] for $p = 75\%$ are summarized in *fig. 5*, which gives the minimum value of L_2 needed to be able to observe an object in 75% of the cases at an adaptation level L_1. The curves relate to different values of contrast C, as indicated beside the graph.

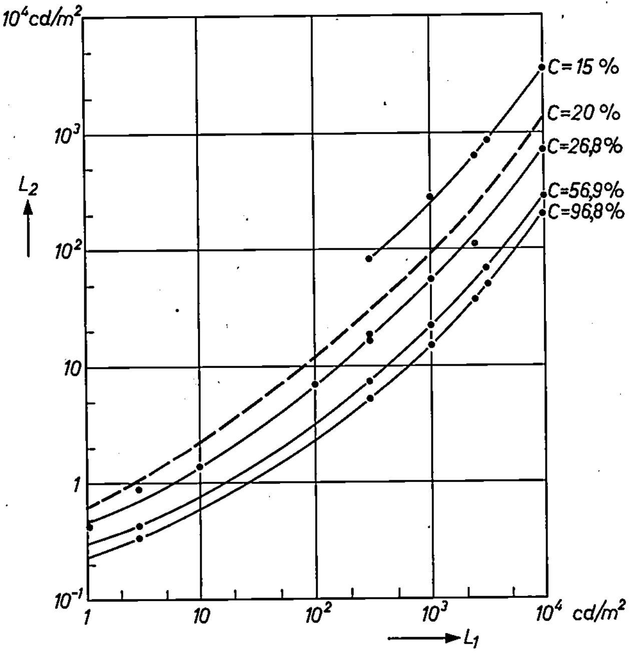


Fig. 5. Results of induction experiments. The luminance L_2 which was necessary in 75% of the cases to see the object within a time of 0.1 sec is plotted as a function of the screen luminance L_1, with the contrast C as parameter. The values for the critical object ($C = 20\%$, angle of vision 7′) were found by interpolation.

Two important conclusions may be drawn:

1) To see an object that has a contrast of 20% against its background, where $L_1 = 8000$ cd/m², the luminance L_2 at the beginning of the tunnel must be about 1000 cd/m².
2) The minimum value of L_2 is roughly proportional to L_1 for a considerable range of L_1 values; as an approximation: $L_2 = 0.1\,L_1$. This is particularly important because it shows that there is no reason why the gradual luminance transition required by the slowness of adaptation should not be realized with the aid of daylight-screening grids fitted at the entrance of the tunnel. As the transmission of such grids is constant, their use means that L_2 is always the same fraction of L_1.

It is of importance to know how much these results change when a different critical object is chosen or when the other conditions of the experiment are changed. Fortunately it is found that even a fairly considerable variation of most parameters has little effect. It may therefore be concluded that this figure can indeed be used to describe the black-hole effect. We shall now briefly discuss the influence of the various parameters.

If one accepts an observation probability of 50% instead of 75%, L_2 may permissibly be 20% lower. Where the object is *lighter* than its background (negative contrast), then if the contrast is not too great the minimum value of L_2 remains roughly the same. If the *angle* at which the object is seen is four times greater (28′ instead of 7′), then L_2 may be on an average about 30% lower. If the *presentation time* is made three times longer (0.3 s instead of 0.1 s), the average value of L_2 may be about 40% lower. (If the object is presented for an even longer time, the condition of constant adaptation is no longer fulfilled, and the results found will therefore not be directly comparable with ours [7].) Finally, it can be seen from fig. 5 that the influence of the *contrast* on the minimum value of L_2 is only significant when the contrasts are very small. At contrasts of 20% or more the influence of C is also relatively slight. If, for example, we take 25% instead of 20% for the contrast of the critical object — a substantial change considering the frequency at which the various C values occur — L_2 decreases by about 40%.

It is clear from the foregoing that, as we have assumed, the conclusions concerning the minimum requirements which the lighting of a tunnel entrance must fulfil in order to avoid the black-hole effect are not greatly dependent on the choice of the critical object. Our choice therefore seems to have been justified. The practical consequences of the results described will be dealt with at the end of this article.

Some investigators have assumed that, where glare occurs, the decrease of the perceptibility of contrasts is due solely to the scattering of light in the eye [8]. This theory, however, does not entirely explain the shape of the curves in fig. 5. The light scattering will be proportional to L_1, and the same must apply to the decrease of the perceptibility. One would therefore expect to find straight lines, with a slope of 45°, in fig. 5. The lines in fig. 5 are however by no means straight. This means that the induction phenomena described here must involve adaptation effects as well as scattered light. The nature of these effects cannot of course be deduced from our experiments.

No experiments with which the above-mentioned experiments might directly be compared have been reported in the literature. A certain analogy does exist, however, between the results shown in fig. 5 and those of Schumacher [9] and Lythgoe [10].

Adaptation

When a driver has passed the adaptation point, the sensitivity of his visual system is influenced by the presence of the dark entrance to the tunnel. In the introduction we have shown that the resultant phenomena are to be divided into 1) a change in the state of adaptation of particular parts of the retina (local adaptation) and 2) a change of induction (general adaptation). For a light-coloured façade in particular, the adaptation point is not very far from the entrance, so that these processes do not begin until the driver is almost about to enter the tunnel.

The induction phenomena do not have a very pronounced after-effect: general adaptation takes place very rapidly. As far as this is concerned, a sudden lowering of the level of luminance does not have too adverse an effect on the visibility of objects. Local adaptation, on the other hand, takes place slowly, as we have seen, so that rapid local changes of luminance leave after-images on the retina [11]. As these disappear slowly, and generally cover large parts of the visual field, they can seriously impair the visibility of the road ahead.

Our adaptation experiments were designed in such a way as to allow special study of the second effect, the disappearance of after-images. They were based on the principle of the subjective appraisal of luminance transitions. In all these experiments the luminance of the visual field was varied, so that full account could be taken of the influence of the variation of general adaptation and that of the pupillary reflex. This means that in fact the adaptation defect was measured. The experiments were also arranged so that quantitative requirements for the lighting of a tunnel entrance could be derived directly from the results. Two methods were adopted. In the first method a study was made of the time taken — after a sudden sharp decrease in the luminance of the visual field — for the adaptation delay to decrease to an acceptable value, or, in other words, how long it takes for the after-images to disappear. The second method was designed to study how quickly the luminance may decrease, and what law it should follow, to avoid excessive adaptation delay. In both cases the starting level for the luminance was at least 8000 cd/m^2, which is the maximum value of luminance of practical importance in the open.

In the experiments by the first method a screen of about 6 square metres was uniformly illuminated to produce a luminance of 8000 cd/m^2. The screen was observed from a distance of about 5 m. After a pre-adaptation time of about two minutes, at an instant known to the observer, the luminance was reduced within 10 ms to 13 cd/m^2. The time was recorded that elapsed from that instant to the instant at which the observer reported that the after-images had decreased to an acceptable degree. These after-images of course covered a considerable part of the visual field. The experiment was repeated ten times by each of ten observers; the results are presented in *fig. 6*.

For the experiments by the second method the same basic arrangement was used. Now, however, the observer himself was able to vary the luminance of the screen. The observer was instructed to reduce the luminance as quickly as possible from 8000 cd/m^2 to about 4 cd/m^2 without experiencing inconvenience from after-images. This experiment was also repeated ten times by each of ten observers, and in each experiment the variation of screen luminance was recorded. The luminance transition found to be acceptable in 75% of the observations is shown in *fig. 7*.

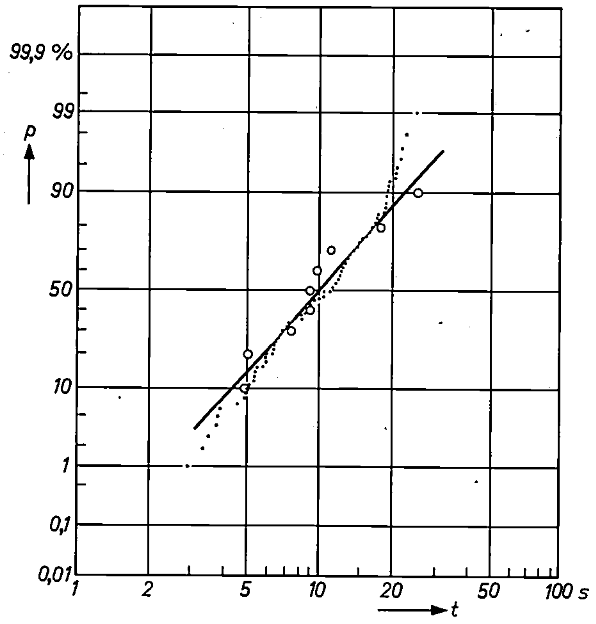


Fig. 6. Results of adaptation experiments. The time t taken for the after-images to decrease sufficiently is plotted against the percentage p of cases in which t was equal to or greater than the ordinate value (cumulative frequency distribution). The full line curve relates to the experiments in which the luminance was abruptly reduced from 8000 to 13 cd/m^2 (small circles represent the experimental points). The dots are derived from fig. 7 and relate to experiments in which the luminance was gradually reduced from 8000 to 13 cd/m^2.

This figure also gives the result of the experiments by the first method (point *1*). It can be seen that this point lies close to the curve, even though in the one case the luminance transition was abrupt and in the second case gıadual. Conversely, fig. 6 shows the

[6] See: J. B. de Boer, Untersuchungen der Sehverhältnisse bei Tunneleinfahrten, Lichttechnik **15**, 124-127, 1963, and also: D. A. Schreuder, Contrast sensitivity in test field with bright surround, J. Opt. Soc. Amer. **55**, 729-731, 1965 (No. 6).

[7] See: R. O. Schumacher, Die Unterschiedsempfindlichkeit des helladaptierten menschlichen Auges, Thesis, Berlin Technische Hochschule 1942, and E. Kern, Der Bereich der Unterschiedsempfindlichkeit des Auges bei festgehaltenem Adaptationszustand, Z. Biol. **105**, 237-245, 1952.

[8] See: J. J. Vos, On the mechanism of glare, Thesis, Utrecht 1963.

[9] See the Thesis in reference [7].

[10] R. J. Lythgoe, The measurement of visual acuity, Sp. Rep. Ser. Med. Res. Comm. **173**, London 1932.

[11] A division of this nature into two processes, one fast and one slow, was first made by J. F. Schouten (Thesis, Utrecht 1937, p. 16). See also: R. M. Boynton and N. D. Miller, Visual performance under conditions of transient adaptation, Illum. Engng. **58**, 541-550, 1963.

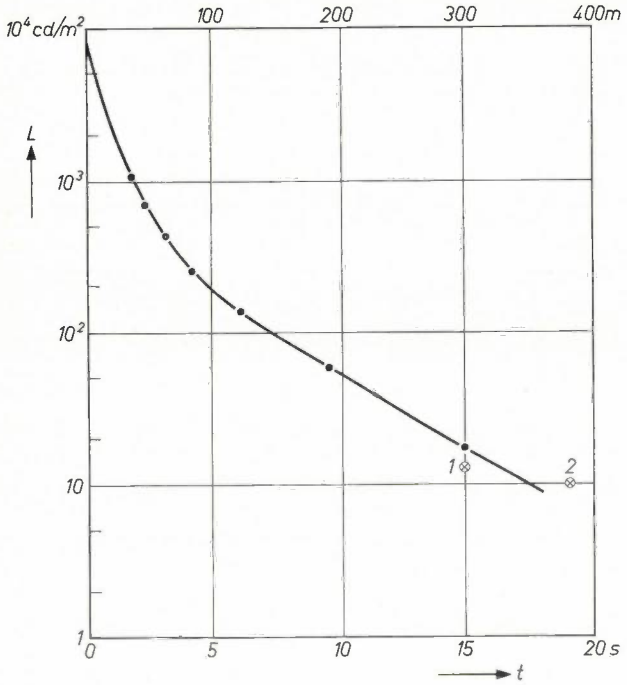


Fig. 7. Results of the adaptation experiments by the second method. The curve indicates how the luminance should decrease so as to avoid inconvenience from after-images. Point *1* is derived from fig. 6, and point *2* gives the result of experiments with models. The distance scale relates to a speed of 72 km/h (i.e. 20 m per second).

results of the second series of experiments, this time in the form of a cumulative frequency-distribution of the adaptation times for a transition from 8000 to 13 cd/m². Here too the agreement is striking. The results of the adaptation experiment may be summarized as follows.

1) About 15 seconds is needed for a transition from 8000 cd/m² to 13 cd/m²; these are the normal values for the open road and for a well-lit tunnel interior respectively. (In 15 seconds a driver travelling at 72 km/h covers a distance of 300 m.)
2) The time ultimately required for adaptation depends only on the initial and final levels of luminance, and not on the variation of the luminance during the transition.

Practical value of the experimental results: experiments with models

The arrangement used for the adaptation experiments just described, like that for the induction experiments, differed considerably from a real tunnel. It is therefore reasonable to ask whether the unmodified results are applicable to tunnel entrances. To clarify this question we repeated some of the experiments, using a model (scale 1/20) of the entrance zone of a long tunnel; it is of course very difficult to carry out such experiments on a true scale [12]. The model is shown in *fig. 8*, and the visual field of the observer in *fig. 9*. The model was fitted with various lighting systems for simulating the transition from the open road to the tunnel. These transition lighting arrangements could be presented to the observer in a random sequence. In all experiments the initial luminance outside the tunnel was 5000 cd/m² and the final level in the interior of the tunnel was 10 cd/m². The time available for the transition could be varied between 3 and 21 seconds. On each occasion the observer had

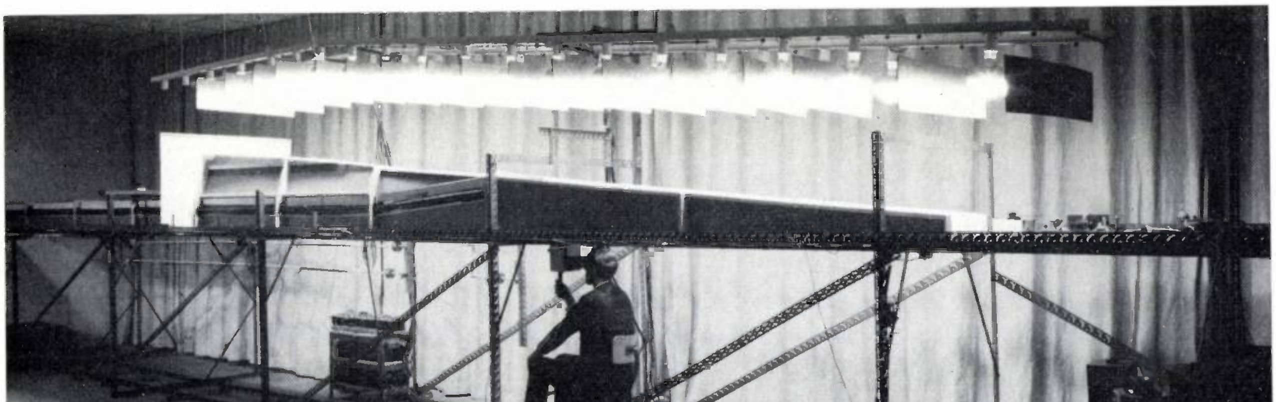


Fig. 8. Arrangement for experiments with the model (scale 1/20). The observer rides on rails below the road surface of the model. He can see the road by means of a periscope: the view obtained is like that from the driving seat of a car (see fig. 9). The model of the tunnel entrance is placed on the road. The lamps above it simulate daylight. The tunnel is on the extreme left.

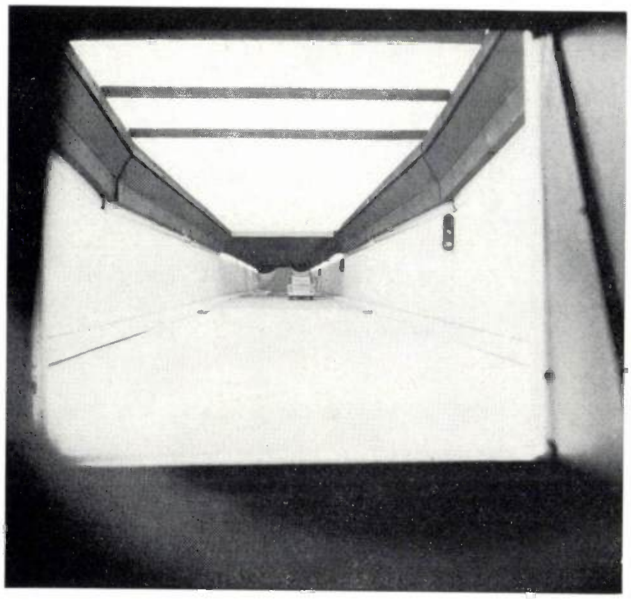


Fig. 9. Observer's field of vision in experiments with the model.

to report whether or not he had been inconvenienced by after-images. The results are summarized in *fig. 10*, where the percentage of favourable answers is plotted versus the time *t* taken for the transition. The value of *t* at which the percentage was 75% is also the abscissa of point *2* in fig. 7; the ordinate of this point is the final luminance level (10 cd/m^2) chosen for these experiments. A comparison of the curve in fig. *7* with points *1* and *2* shows there is reasonable agreement between all results. Further confirmation of the validity of our results is to be found in experiments carried out by Kabayama, which are discussed below.

Another question that needs to be considered is what the adaptation behaviour will be when the initial luminance is not 8000 cd/m^2, but some other level. This question can be answered from the results of Kabayama's experiments [13]. He used an arrangement which broadly resembled the one we used for our adaptation experiments by method 2 (see fig. 7). The criterion was that the object to be observed should remain partly visible. His results are presented by curves *1* to *6* in *fig. 11*. These curves relate to the pre-adaptation levels mentioned in the table and are all based on the results of observations made by three subjects. Curves *7* and *8* in this figure represent our results. In *fig. 12* the same results are repeated, but now with the *relative* luminance as the ordinate; all the curves begin at the same point. As can be seen,

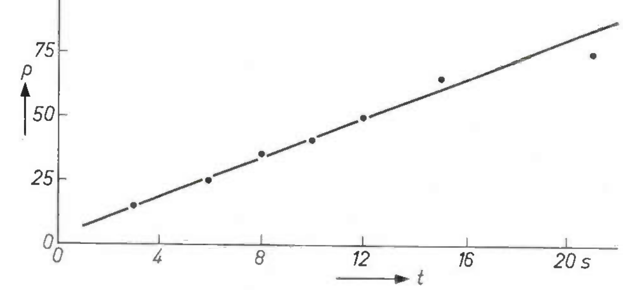
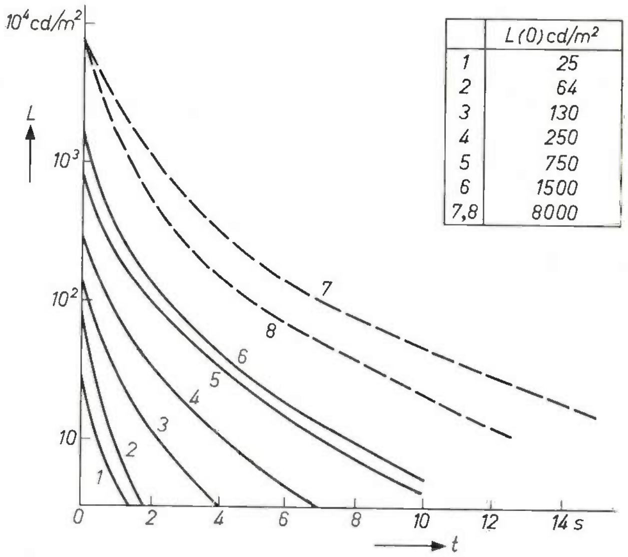


Fig. 10. Results of adaptation experiments with the model. The number of times *p* (in %) that a transition from 5000 to 10 cd/m^2 was judged to be acceptable is plotted against the time *t* available for the transition. The value of *t* corresponding to *p* = 75% is the abscissa of point 2 of fig. 7.



	$L(0)$ cd/m^2
1	25
2	64
3	130
4	250
5	750
6	1500
7,8	8000

Fig. 11. Permissible luminance curves at different pre-adaptation levels (see table). Curves *1* to *6* give the results of experiments by Kabayama [13] where the criterion was that an object should remain partly visible. Curve *7* is the one from fig. 7 and curve *8* is the one found by us for transition times judged to be acceptable in 50% of the cases.

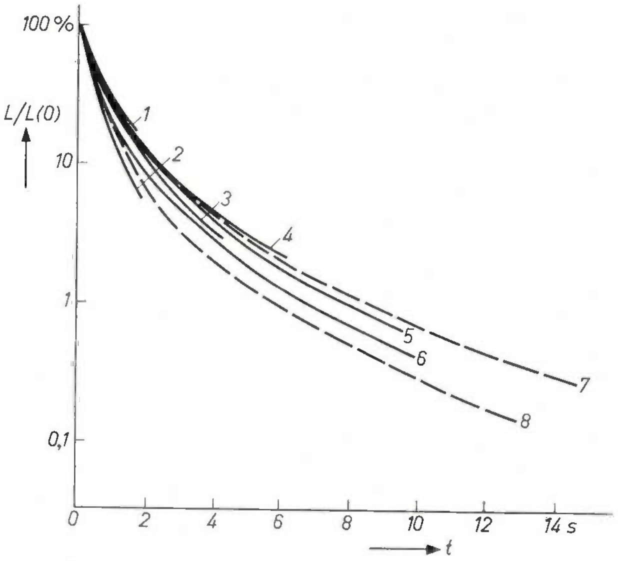


Fig. 12. If the ordinate values of the curves in fig. 11 are reduced to the same initial value, the curves obtained almost coincide.

[12] This arrangement is described in reference [3] and also by J. J. Balder and D. A. Schreuder, Problems in tunnel lighting, Int. Lighting Rev. **10**, 24-26, 1959.

[13] H. Kabayama, Study on adaptive illumination for sudden change of brightness (in Japanese with English summary), J. Illum. Engng. Inst. Japan **47**, 488-496, 1963.

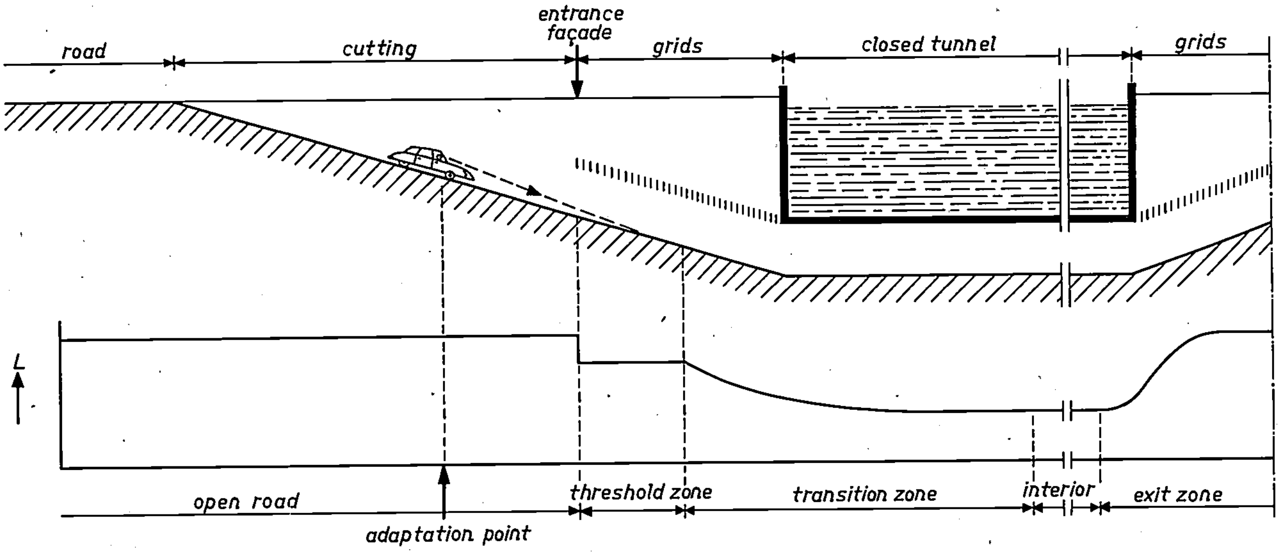


Fig. 13. A road approaching and passing through a tunnel with the zones appropriate to highway engineering (above) and lighting engineering (below). The luminance curve is represented more accurately in fig. 14.

Kabayama's curves and ours are all of the same form. This indicates that the relative luminance curve does not depend on the starting level of luminance, and demonstrates the close agreement between our results and those of Kabayama.

Practical considerations; the Coen tunnel

On the basis of the results given above, we shall now examine how a lighting system for a tunnel entrance should be arranged. In the first place, to avoid the black-hole effect the luminance in the first part of the tunnel should not be lower than about 10% of the level outside the tunnel. We shall call this part of the tunnel the *threshold zone;* see *fig. 13.* Since objects must be visible at a distance of 60 to 100 m, the threshold zone must extend 80 to 120 m beyond the adaptation point, as there must also be about 20 m of illuminated road surface to give a contrasting light background for the objects. As the adaptation point is usually 30 or 40 m in front of the tunnel entrance, the threshold zone need not as a rule be longer than about 50 m. (To avoid misunderstanding, it should be noted that from now on we take the tunnel entrance to be at the point where the luminance begins to decrease; in a tunnel with daylight-screening grids this is the point where the grids begin.)

The threshold zone is followed by the *transition zone*. This is the zone where the luminance has to decrease to the level in the interior of the tunnel. If the driver can see about 100 m in front of him, he sees the beginning of the transition zone as he passes the adaptation point. i.e. at the moment from which induction effects can be disregarded. The luminance curve in the transition zone should therefore follow that shown in fig. 7; and its starting point must correspond to the outside luminance. *Fig. 14* shows the complete luminance curve for the four zones of interest — the open road, the threshold zone, the transition zone and the tunnel interior. At a speed of 72 km/h, the use of such a luminance curve avoids the difficulties of the black-hole effect and of adaptation. As can be seen, the threshold and transition zones taken together are a few hundred metres long, although the luminance level chosen for the interior is by no means low (a good 10 cd/m²).

From the lighting engineer's point of view, it is not important whether the desired luminance is obtained with artificial light or in some other way. In most cases the economically optimum solution proves to be one in which the threshold zone and the first part of the transition zone are lit by subdued daylight obtained

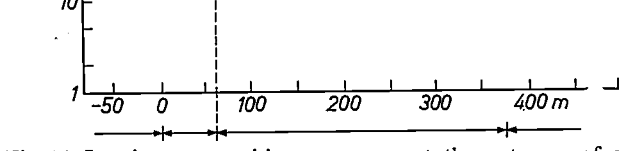


Fig. 14. Luminance transition necessary at the entrance of a tunnel so that a driver travelling at 72 km/h does not experience the black-hole effect or after-images.

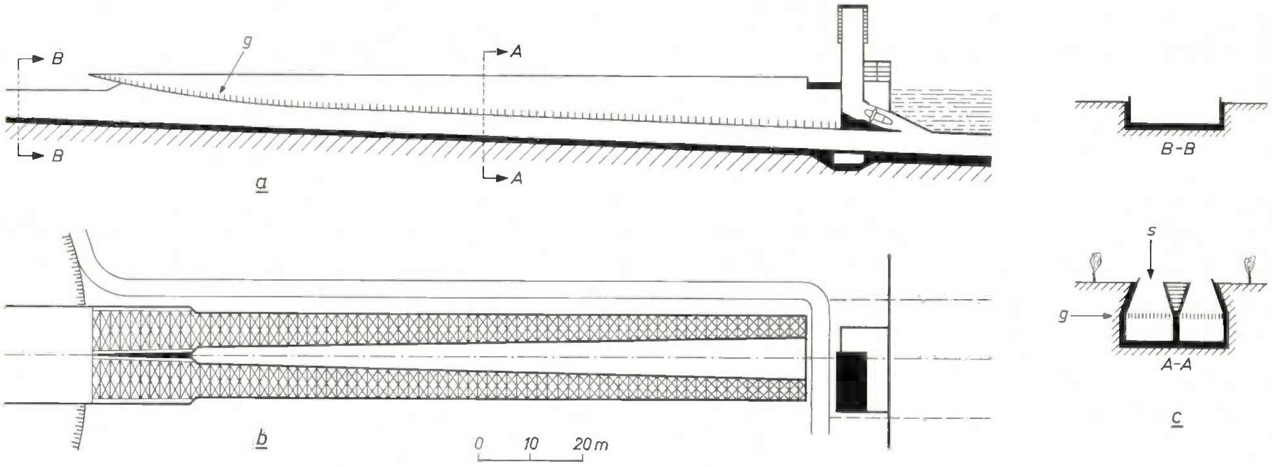


Fig. 15. Entrance area of the Coen tunnel under the North Sea Canal (after Griffioen [15]). *a*) Longitudinal section. *b*) Plan view. *c*) Two transverse cross-sections. A grid *g* is fitted at approximately constant height above the road surface. The grid receives its light by way of a gradually narrowing opening *s*.

through a diffusing screen or grid. The use of daylight-screening has the further advantage that changes in the illumination level outside the tunnel do not alter the relative luminance variation. As we have seen optimum visual conditions are then maintained.

Grids of this kind have to meet various requirements: we shall mention the two most important ones. In the first place they must be open, to let snow and rain through. In the second place, their elements must be so designed that under no circumstances can direct sunlight illuminate the road surface. To meet this requirement, and still achieve a light transmission of 10%, it is necessary to use a highly reflecting material; this must not, of course, give specular reflections [14].

The Coen tunnel

We shall conclude this article with a short description of the entrance lighting of the Coen tunnel at present under construction [15]. In broad lines, the luminance transition in this tunnel will meet the requirements discussed above. The construction adopted is illustrated in *fig. 15*. An aluminium grid begins about 150 m in front of the closed part of the tunnel (*fig. 16*). This grid, whose height above the roadway is approximately constant, receives its light by way of a gradually narrowing opening at surface level. The transmission of the grid itself is roughly 30%. This construction gives rise to a gradual decrease of road surface luminance. At the beginning of the closed part of the tunnel extra lighting has been installed over a length of about 70 m, producing a luminance of about 30 cd/m^2. In the interior of the tunnel the luminance will be about 15 cd/m^2. The artificial lighting comes from tubular fluorescent lamps whose luminous flux is automatically adapted to the level outside the tunnel by a thyristor circuit controlled by photoconductors.

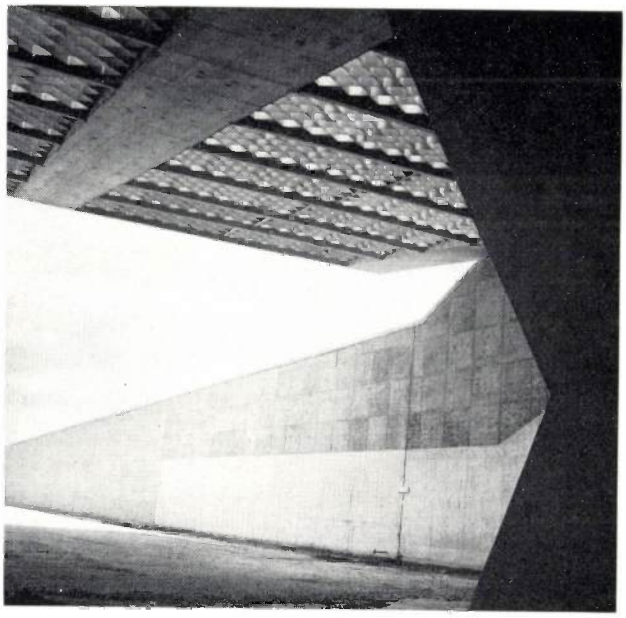


Fig. 16. North entrance to the Coen tunnel, showing the overhead grids.

Fig. 17 shows the proposed luminance curve (curve *1*). For comparison the figure includes the curve from fig. 14, giving the luminance curve which our results show to be desirable (curve *2*) and also the results of measurements in the Velsen tunnel (under the North Sea Canal, opened in 1956; curve *3*) and in the Rotter-

[14] For further particulars see: D. A. Schreuder, Über die Beleuchtung von Verkehrstunneln, Lichttechnik **17**, 145A-149A, 1965 (No. 12).

[15] See: A. Griffioen, De Coentunnel, III. De ventilatie en verlichting, Ingenieur **75**, B213-B224, 1963.

dam tunnel (under the river Maas; opened in 1942; curve *4*). The measured curves have been shifted horizontally with respect to curve *2* so as to produce the best fit. As can be seen, the luminance curve proposed for the Coen tunnel approximates closely to that of fig. 14. A comparison of the situations for the three tunnels shows clearly how traffic requirements have resulted in lighting of a higher quality.

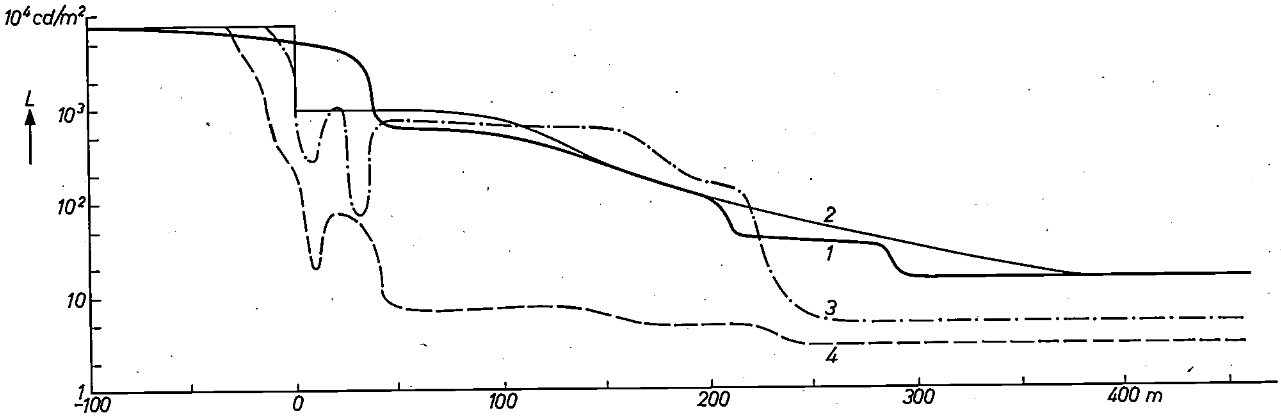


Fig. 17. Luminance curve calculated for the Coen tunnel entrance (curve *1*). For comparison the curve from fig. 14 (curve *2*) is also shown, together with the measured luminance curves for the Velsen and Rotterdam tunnels (curves *3* and *4* respectively).

Summary: A driver approaching an inadequately lit tunnel entrance is at first inconvenienced solely by induction effects (he sees the entrance as a black hole). After he has passed the point where visual adaptation begins (the adaptation point) his vision is impaired purely by adaptation effects, in particular by after-images. Induction experiments in the laboratory have shown that the black-hole effect is avoided when the luminance L_2 at the tunnel entrance is greater than 10% of the outside luminance L_1. This proportionality allows the use of overhead grids for lighting the tunnel entrance by subdued daylight. The inconvenience caused by after-images was investigated by abruptly decreasing the luminance of a large screen and studying the time taken for the effect of the after-images to decrease to an acceptable level, and also by allowing observers to vary the luminance themselves. The optimum luminance curve was derived from these experiments. The time required for adaptation with an initial luminance level of 8000 cd/m² and a final level of 13 cd/m² proves to be roughly 15 sec (corresponding to 300 m at a driving speed of 72 km/h (45 m.p.h.). Experiments with models have confirmed the results. If the adaptation point is 50 m in front of the entrance and L_1 = 8000 cd/m², there must be a 50 m zone in the tunnel where $L_2 \approx$ 1000 cd/m². This should be followed by a transition zone where L_2 decreases in accordance with the experimental curve, with L_1 as initial value (by extrapolation). These requirements are largely met by the lighting planned for the Coen tunnel near Amsterdam.

Microfractography of thin films

J. M. Nieuwenhuizen and H. B. Haanstra

620.187.5

The special structure of evaporated thin films, which are of particular interest to the electronics industry at the present time, has given rise to the concept of a fourth state of aggregation of the material. A surprisingly direct insight into this state has now been obtained from "fractographs" — photographs of fracture surfaces — which have been made, by means of the electron microscope, for aluminium films about 1 micron thick.

Thin films of materials of various kinds show effects and properties which do not appear, or are much less pronounced, in the bulk material. Thin films are therefore the subject of intensive investigation, and their properties have led to important applications. For several decades they have been used for example as getters in thermionic valves, and for coating glass surfaces to give them particular reflecting (or non-reflecting) properties. More recently, thin films have found application in electronic circuit components, especially for digital techniques [1], and the possibilities offered by thin film components in integrated circuits are of considerable practical importance.

The thin films concerned are generally produced by *evaporation* on to a substrate or carrier in a vacuum, and their thickness may be anywhere between a few to many thousands of atoms. In the latter case the thickness of the film may be of the order of 1 micron. These relatively thick films can sometimes be given considerably different characteristics if in the evaporation process the molecular beam is directed not vertically upon the substrate but more or less *obliquely*. Obliquely deposited films of semiconductors such as silicon, gallium arsenide or tellurium, show a very marked photovoltaic effect, which is virtually absent in vertically deposited films. Obliquely deposited films of some substances are dichroic, vertically deposited films of the same substances are not. In obliquely deposited magnetic layers, for example of permalloy (80Ni20Fe), a strong magnetic anisotropy may occur, in which — at least if the angle of incidence is not made too great — the preferred direction of magnetization is perpendicular to the plane of incidence of the atomic beam. The magnitude of the mechanical stresses (if any) may also depend on the angle from which the film is evaporated [2].

All these differences can only be explained on the assumption that for different evaporation angles films of different structure are formed. It has now proved possible to make these differences in structure directly visible with the electron microscope. In the following we shall briefly describe the procedure and show some results.

For "thin" thin films, where the processes of interest are nucleation and the growing together of islands of the evaporated material to form an unbroken coating, the electron microscope has previously been employed with success as a research tool [3]. In this case the films themselves are placed as specimens in the electron beam of the microscope. The "thick" thin films with which we are concerned here, however, are not sufficiently transparent to electrons. As with bulk metals, only the surface can be made accessible to electron-microscopic investigation, viz. by making a transparent *replica*. Conclusions about the internal structure of the film can then only be drawn in so far as this structure appears in relief on the surface.

This information is very limited and inadequate,

J. M. Nieuwenhuizen and H. B. Haanstra are with Philips Research Laboratories, Eindhoven.

[1] See W. Nijenhuis and H. van de Weg, Developments in the field of electronic computers during the last decade, Philips tech. Rev. **26**, 67-80, 1965.

[2] On the photovoltaic effect, see: B. Goldstein and L. Pensak, J. appl. Phys. **30**, 155, 1959; E. I. Adirovich, V. M. Rubinov and Yu. M. Yuabov, Sov. Phys. Solid State **6**, 2540, 1965 (No. 10).
On dichroism: A. Kundt, Ann. Physik u. Chemie **27**, 59, 1886.
For magnetic anisotropy: D. O. Smith, M. S. Cohen and G. P. Weiss, J. appl. Phys. **31**, 1755, 1960.
For mechanical stresses: J. D. Finegan and R. W. Hoffman, J. appl. Phys. **30**, 597, 1959.

[3] G. A. Bassett, J. W. Menter and D. W. Pashley, in: Structure and properties of thin films, Proc. int. Conf. Bolton Landing, New York 1959, page 11; J. van de Waterbeemd, Physics Letters **16**, 97, 1965 (No. 2); J. van de Waterbeemd, Philips Res. Repts. **21**, 27, 1966 (No. 1).

and we therefore looked for a technique in which a *cross-section* of the film could also be reproduced in the replica.

It is not sufficient just to break off a piece of substrate with its film, for if a replica of such a piece is made that extends over the surface and the fracture face, the part in which we are interested is to be found at the edges of the replica, which usually curl up and are not so suitable for good observations. This can be avoided by applying an intermediate film which is afterwards dissolved [4], but the finest details are then likely to be less distinct, and there is the possibility that spurious details may be introduced by the structure of the intermediate layer.

Etching away parts of the film before making the replica did not produce useful results either, the transition from the original surface to the substrate was then too gradual; cross-sections of the layer providing clear information about the internal structure are not obtained in this way.

A method that did produce useful results was to make a scratch in the thin film with the point of a razor blade. On each side of the scratch the film crumbles away here and there; fragments of the film become detached from the substrate, and the piece of film that remains shows a well-defined fracture surface, extending from the surface of the film to the substrate.

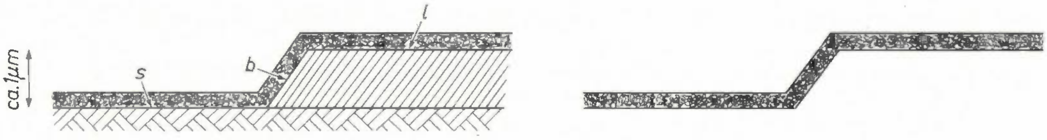


Fig. 1. When the surface of a thin film is scratched with the point of a razor blade, a fairly clean fracture surface *b* is obtained at some places. A continuous replica can now be made of the film surface *l*, the fracture surface and the exposed surface *s* of the substrate. The step-like replica can be detached without damaging it.

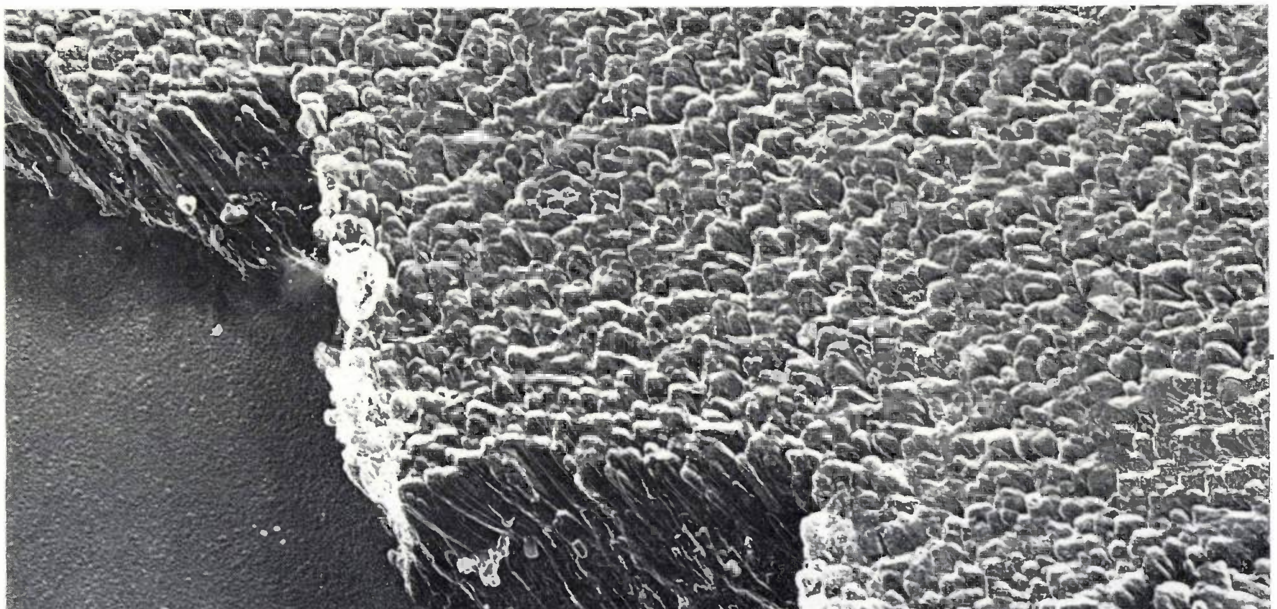


Fig. 2. Electron-micrograph of a replica of an aluminium film about 1 micron thick, obtained by the method described. The picture of the fracture surface shows that the aluminium film is built up from parallel crystal columns, inclined at a certain angle.

It was verified from stereomicrographs made of all the specimens that the replica had retained the form illustrated in fig. 1.

At such places carbon can be deposited in the usual way to make a replica. A continuous replica is then obtained of the surface of the film, the fracture surface and the exposed surface of the substrate; see *fig. 1.* A difficulty with such a replica, which contains a kind of step with two sharp bends, is that the sharp edges easily break, and to prevent this from happening the carbon layer is removed extremely slowly. In this investigation we worked mainly with thin films of aluminium evaporated on glass at room temperature. The aluminium film under the replica can be removed by dissolving it in very dilute hydrofluoric acid (0.05% HF in water); the replica then comes away from the glass at the same time. The dissolving process may in some cases last a whole week, but the result is that the replica remains intact [5].

Electron-photomicrographs of such replicas now give clear pictures of the fracture surface of a thin film. A "microfractograph" of this type is shown in *fig. 2.* The step in the replica is clearly visible (although the edge of the step was not straight here, as in normal steps, but a zigzag line). The micrograph shows that the film is built up from columnar crystals, with parallel longitudinal axes inclined at a certain angle to the normal to the substrate. To learn more about this angle, which we shall call φ_k and which is obviously of importance in the structure of the film, we made use of the following procedure.

It is usual to make the relief structure of electron-microscopic specimens more visible by shadowing, that is to say by evaporating a heavy metal such as platinum on to the specimen from an oblique angle: at places where no platinum atoms have been deposited, the electrons in the beam pass through the specimen with relatively little scattering, and at such places the micrograph shows a high density. We carry out this shadowing procedure for our evaporated thin films in the same vacuum vessel in which the films were produced, and we evaporate the Pt atoms on to the scratched aluminium film from exactly the same angle (in fact from the same source location) as used for the Al atoms of the substrate, moreover making sure that when the film is scratched [6] the position of the substrate in the vacuum vessel is not changed. The length of the resultant Pt shadow thus as it were marks the angle φ_a from which the aluminium is evaporated on to the substrate.

It follows immediately from the fact that the Pt shadow in the photograph in fig. 2 does not have zero length that the columnar crystals of the layer have *not* grown exactly in the direction of the incident aluminium atoms (see *fig. 3*): they are more upright ($\varphi_k < \varphi_a$). The columns are however always in the plane of incidence of the Al atoms; we have been able to establish this in experiments with five different angles of incidence, up to about 80°. The angle φ_k is easily calculated from the known φ_a and the lengths k and a indicated in fig. 3, which can be measured on the micrograph. The values thus found are plotted in *fig. 4* against φ_a. The relationship can be represented very well by the equation [7]:

$$\tan \varphi_a = 2 \tan \varphi_k.$$

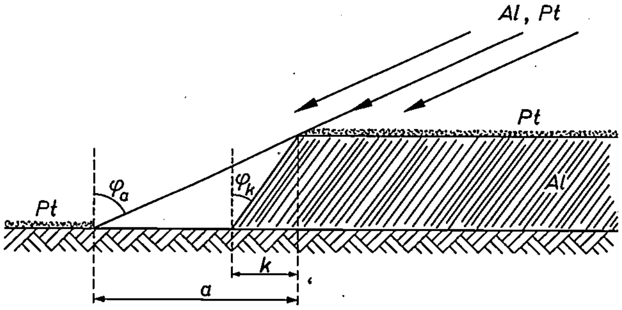


Fig. 3. The columns of the film are at an angle φ_k to the normal to the substrate. This angle proves to be smaller than the angle φ_a at which the beam of aluminium atoms (and the beam of platinum atoms in the shadowing process) is incident on the film. The columns are however in the plane of incidence of the atoms. The length of the Pt shadow is a minus k.

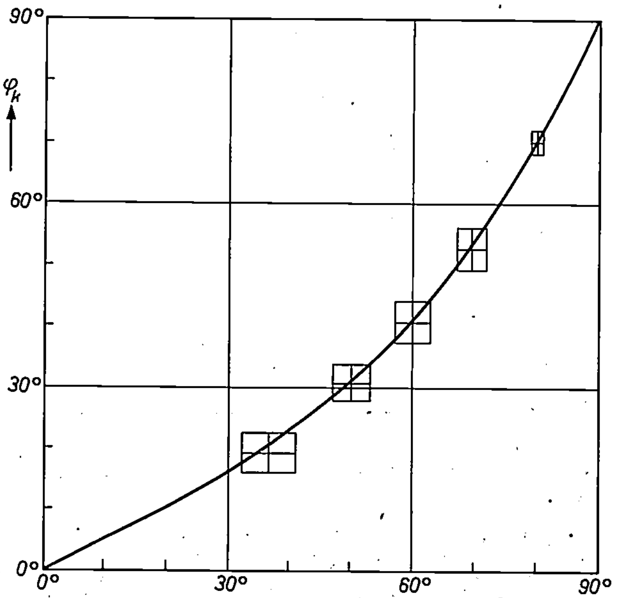


Fig. 4. Relationship found experimentally between the angles φ_a and φ_k of the direction of evaporation and the orientation of the columns. The relation is given to a very good approximation by the curve shown which represents the equation: $\tan \varphi_a = 2 \tan \varphi_k$.

[4] R. Ya. Berlaga and M. I. Rudenok, Sov. Phys. Solid State 3, 458, 1961.

[5] Mrs. J. Andreas-Bosdijk did a large part of the experimental work for the development of this technique.

[6] The reader will notice that the shadowing is done before the carbon replica is made. When the carbon layer is detached, the deposited Pt atoms adhere to it.

[7] Our attention was drawn to this by Drs. G. W. van Oosterhout.

Fig. 5 shows another clear micrograph of a fracture surface. These and similar micrographs show details in the columns that provide some foundation for a probable explanation of the mechanism underlying the growth of the columns; we shall not, however, go into this subject here [8]. Finally, *fig. 6* shows the fracture surface of a composite film, built up by successive evaporation of aluminium from *two* directions. In the second part of the film the columns are accordingly inclined in a different direction, giving the cross-section a kind of herringbone structure. (This experiment, incidentally, provides conclusive proof that the columnar structure in micrographs such as those in figs. 2 and 5 cannot be a result of the method of preparation itself — a possibility that could not *a priori* be excluded.) To obtain a pronounced picture of the herringbone structure, the second layer must be given the chance to nucleate independently, e.g. on an amorphous oxide layer, produced by admitting air for a moment into the vacuum vessel between the two evaporation processes. If this is not done, the columns do not abruptly change their direction of growth, and

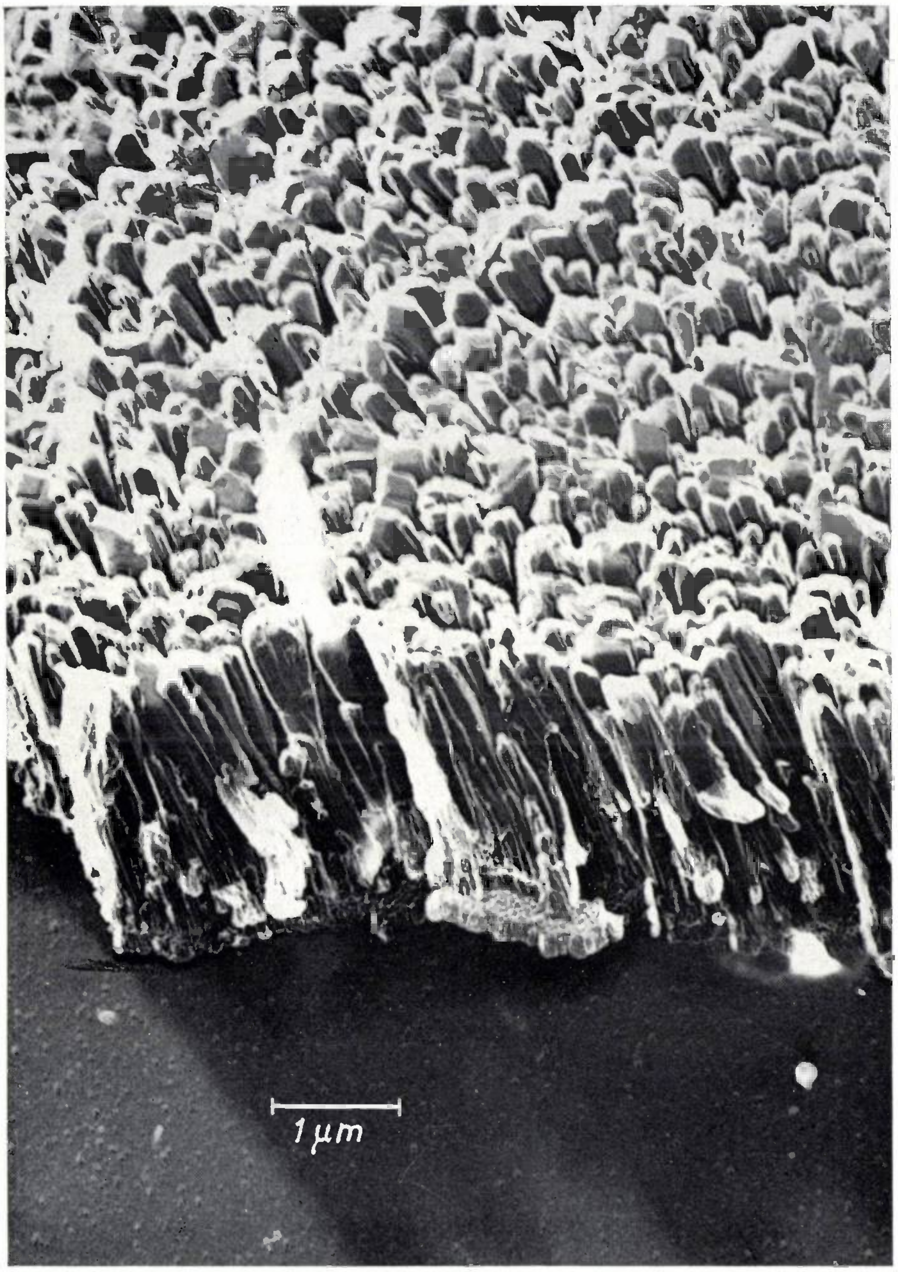


Fig. 5. A particularly detailed fractograph obtained by the method described.

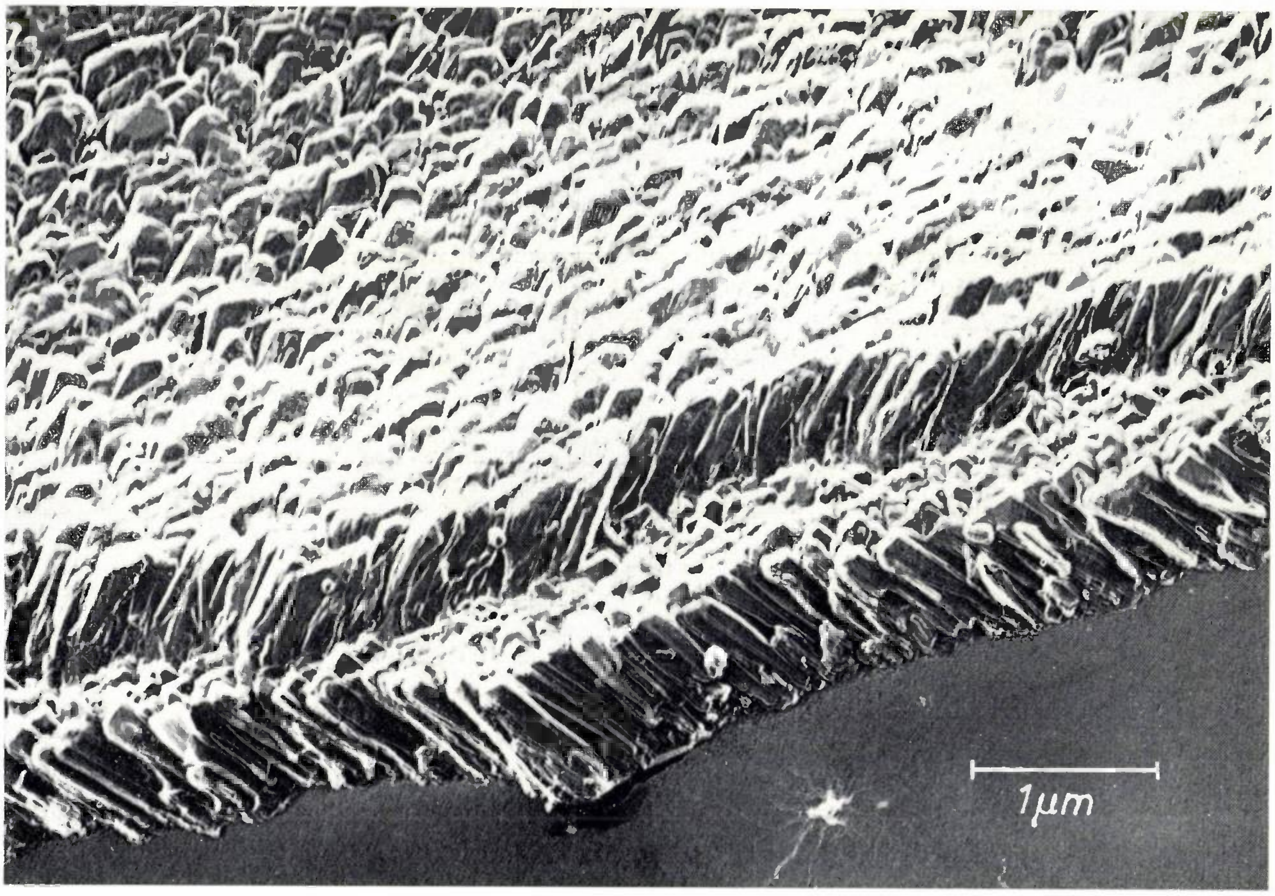


Fig. 6. Fractograph of a film grown by the successive evaporation of aluminium from *two* directions. Air was admitted into the vacuum vessel between the two evaporation processes.

the change in the fracture surface is brought out much less distinctly.

The information supplied by the structure patterns obtained in the fractographs takes us a stage nearer to an explanation of the relationship, which we mentioned at the beginning of this article, between various properties of thin films and the direction from which the films are evaporated. This information can moreover contribute towards a better technological control of evaporated films.

Summary. The characteristics of some evaporated thin films depend to a marked extent on the angle at which the atoms arrive at the substrate. Replicas of fracture surfaces of relatively thick films of aluminium (thickness approx. 1 micron) have been made successfully by means of a simple technique. Photographs of these taken with the electron microscope show clear pictures of the cross-section of such a film. From these "fractographs" conclusions can be drawn concerning the structure of the films and the way in which the structure is influenced by the evaporation angle.

[8] See: C. Kooy and J. M. Nieuwenhuizen, Structural effects in thin films observed by electron microscopy of the film cross-section, shortly to be published in Proc. Coll. on basic problems of the physics of thin films, Clausthal 1965.

An automatic exposure control system for X-ray diagnostics

H. Elgström and E. Zieler

621.386.8

X-ray technology has always been a field of application for the results of many other branches of knowledge — medicine, physics, optics and even mechanical engineering. Nowadays, electronic engineering as well makes many contributions to developments in the X-ray field. The automatic exposure control system described here illustrates the interplay between physical considerations and electronic circuit techniques.

The reasons for automatic exposure control

In nearly all branches of engineering increasing use is being made of automation. This is not always dictated by purely economic considerations; the elimination of human errors, liberation from dull routine tasks and the greater constancy of results are often put forward as arguments for automation. The same arguments applied in the development of an automatic exposure control system for X-ray diagnostics. Automatic exposure control, then, is required: a) to improve the average quality of the radiographs, and b) to simplify the operation of the X-ray equipment.

The ideas underlying automatic exposure control are not new. They were put forward at Messrs. C. H. F. Müller as long ago as 1929 (by H. Franke), and patents were taken out[1], but it was twenty years before components became available that were sufficiently advanced for use in reliable automatic exposure timers. For the last ten years or so, instruments of various proprietary makes have been marketed[2], and in recent years, owing to certain refinements, and in particular to the achievement of very short switching times, the automatic exposure timer has found fairly wide application. In the following we shall describe the "Amplimat" system, developed by Messrs. C. H. F. Müller[3].

First, it will be useful to examine the conditions to be fulfilled to obtain a good radiograph. The film, as in ordinary black-and-white photography, has to be exposed to a certain quantity of radiation (the product of intensity and time). The choice of this exposure, however, is much more critical than in ordinary photography, firstly because there is here no printing process that can compensate for errors in the original exposure, and secondly because the film used for radiography (it is coated with emulsion on both sides), in combination with intensifying screens, has a much steeper tonal gradation than standard photographic films. *Fig. 1* shows the density (blackening) of an X-ray film and that of an amateur film as a function of the logarithm of the exposure. For an amateur film the tonal gradation γ is less than 1, while for an X-ray film γ is about 3. The higher γ of the X-ray film is needed because the anatomical details to be made visible in a radiograph often produce only very slight differences of intensity in the radiation pattern. As a result of this steeper tonal gradation the exposure is no longer within the useful part of the characteristic curve (between the dashed lines in fig. 1) if it deviates as little as 40% from the average value.

In radiography the exposure is adjusted at the radiation source. There are three variables available for this adjustment; these are a) the voltage on the X-ray tube, b) the current through the X-ray tube, and c) the exposure time; and these have to be selected for each exposure. The voltage governs the contrast of the X-ray image and is therefore primarily chosen so that the contrast is suitable for the clinical subject under investigation. This leaves the tube current and the exposure time; the quantity of radiation is proportional to each of these and their choice should result in the desired film density. For every type of radiograph and the appropriate tube voltage it is necessary to determine from experience, and by estimating the body thickness of the subject, what the milliampere-second product should be. As the density of the object can vary considerably, incorrect exposures can occur. How serious the errors can be is illustrated in *fig. 2*. This shows, as a function of body thickness, the mAs product that was required to obtain a chest radiograph of the same average density at the film for a large number of patients. The dashed line gives the values from an exposure table, based on experience with average cases. It can be seen from the points that for a thickness of

H. Elgström and Dr. E. Zieler are with C. H. F. Müller GmbH, Hamburg.

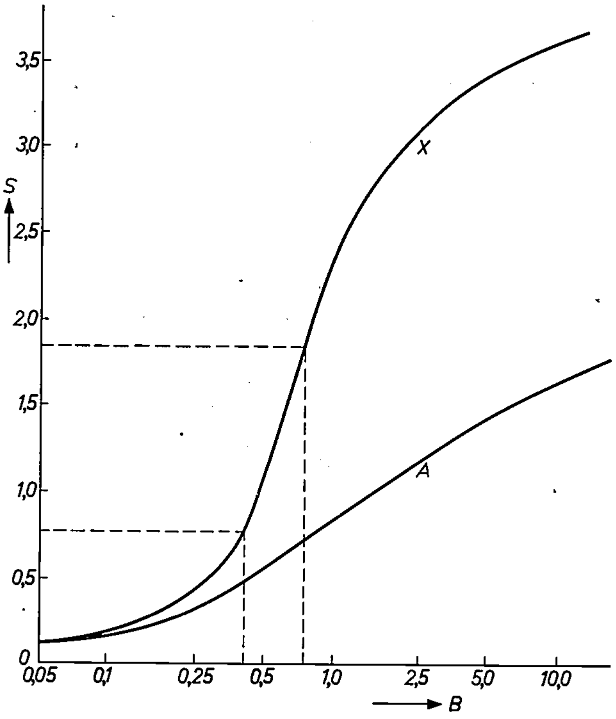


Fig. 1. Film density S as a function of exposure B (intensity × time). The exposure latitude of an X-ray film (X) is very much smaller than that of a normal photographic film (A).

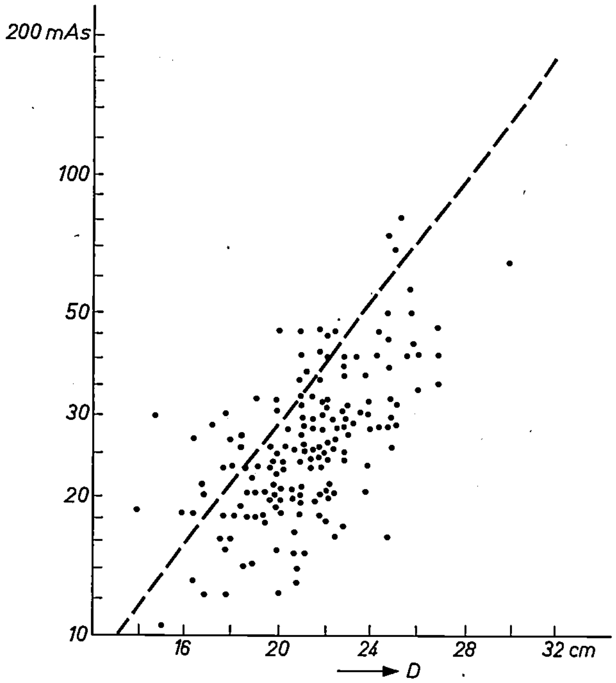


Fig. 2. Spread in values of mAs product that were required to obtain identical density on a large number of chest exposures on patients of different body thickness D. After Stieve [4].

20 cm the correct exposure value may lie between 12 and 45 mAs, or, for a thickness of 25 cm, between 16 and 70 mAs. Thus, if the radiographs are taken using the data from the exposuie table, the exposure for many of the radiographs will be quite wrong. This demonstrates the advantage of automatic exposure control, which replaces the uncertainty of an estimate by an objective measurement.

General arrangement of the Amplimat

The general arrangement of our automatic exposure control system is illustrated in *fig. 3.* It comprises an ionization chamber, which acts as radiation detector, an electronic section, consisting of amplifiers and power-supply, and, on the left, a mains switch for the high-tension transformer in the X-ray generator. It is now current practice to incorporate the circuits for operating the Amplimat, which will be discussed presently, in the X-ray generator control panel.

The detector in the Amplimat is located between the patient and the film cassette. In principle the detector could also be located *behind* the cassette; this arrangement is used in some exposure timers. Our choice was largely dictated by the following considerations:

1) As the X-rays are considerably attenuated by the cassette with film and intensifying screens, a much more sensitive amplifier is needed when the detector is behind the cassette.
2) The attenuation of the X-rays by the intensifying screens depends to a great extent on the radiation quality, and therefore, if the detector is behind the cassette, the sensitivity needs to be corrected to suit the selected tube voltage. Such a correction is fairly difficult to achieve.
3) Our arrangement is not affected by details of cassette construction (the material of the cover, whether or not the cover contains lead, or by the presence of any of the vaiious types of spring clip).

The detector placed in the radiation in front of the cassette (and, of course, so designed that it gives no

[1] DRP 574 441 of 10.3.1929. See also: H. Franke, Ionimetrische Bestimmung optimaler Belichtungszeiten, Congress issue of Fortschr. Röntgenstr. 40, 99-100, 1929; Der Belichtungsautomat, 1. Congress issue Fortschr. Röntgenstr. 42, 153-154, 1930.

[2] K. Bischoff, Der Iontomat, ein neuer Belichtungsautomat für Röntgenaufnahmen, Fortschr. Röntgenstr. 71, 994-1002, 1949.
M. Fourlon, Les appareils photo-électriques pour la détermination des temps de pose, J. Radiol. Electrol. 33, 39-42, 1952.
P. C. Hodges, Photoelectric timing in general radiography, Acta radiol. Suppl. No. 116, 605-612, 1954.
Th. Lohmann, Ein neuer Belichtungsautomat für die Lungen-Diagnostik, Röntgenblätter 7, 259-267, 1954.
F. E. Stieve, Untersuchungen über das Wirkungsprinzip von Belichtungsautomaten für Röntgenaufnahmen, Fortschr. Röntgenstr. 85, 491-510, 1956.

[3] E. Zieler, Der Amplimat, ein Belichtungsautomat für allgemeine Röntgenographie, Fortschr. Röntgenstr. 86, 382-393, 1957.

[4] F. E. Stieve, Belichtungsautomaten in der Praxis, Röntgenblätter 9, 325-333 and 363-372, 1956.

shadow in the X-ray image, see below) measures the radiation "dose"[5] that emerges from the patient during the exposure. This dose governs the photographic density of the film. By means of a density selector the apparatus can be set to the dose that will give the film the appropriate density: as soon as the dose reaches this pre-set value, the automatic exposure control disconnects the high-tension generator from the mains. Control of the exposure time is thus achieved, and this quantity no longer has to be pre-set on the X-ray unit. The tube voltage and tube current, however, still have to be chosen. The consequences of introducing automatic exposure control for the selection of these variables will be discussed under the next heading.

the maximum permissible value.

Now, with automatic exposure control, the exposure time is not known beforehand, and the optimum load on the X-ray tube cannot be preset. If a very high current is chosen, as required for a short exposure, it may well happen that during the time which the load characteristic of fig. 4 shows to be permissible, the dose required to produce the correct film density is not reached; the exposure will then be terminated by the operation of the safety device, which switches off the tube, and the result will be an under-exposed film. If a much lower current, permissible for a longer time, is selected, radiographs of the correct density are obtained, but the exposure time will be longer, and the movement blur in the X-ray image will be greater than is strictly necessary.

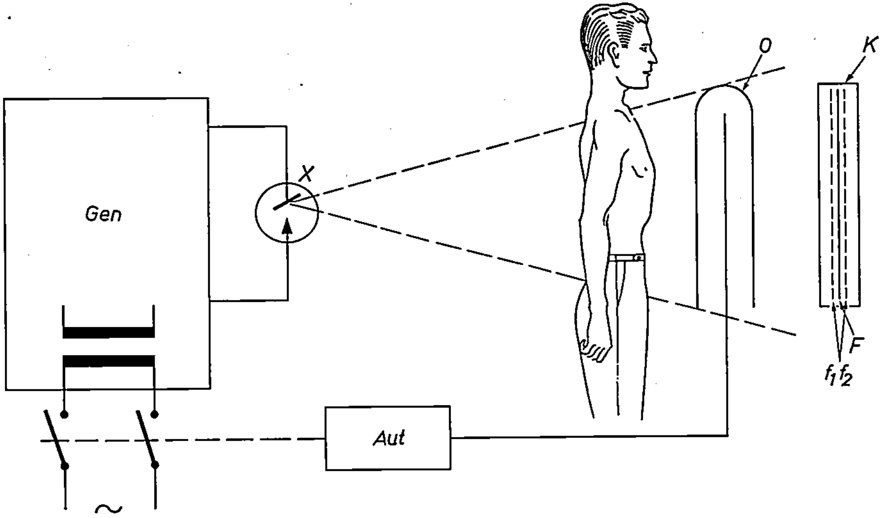


Fig. 3. Arrangement for radiography using the "Amplimat" automatic exposure control system. *Gen* X-ray generator. *X* X-ray tube. *O* detector (ionization chamber). *Aut* electronic section of automatic exposure control system. *K* cassette with film *F* and intensifying fluorescent screens $f_1 f_2$.

Loading the X-ray tube

The living subjects with which diagnostic radiography is concerned have to be exposed to the X-rays for as short a time as possible if movement blur on the film is to be kept to a minimum. The beating of the heart can give rise to movements of the organs in the thorax: according to Berger[6], these can have velocities as high as 400 mm/s. As short an exposure as possible necessitates the highest possible loading of the X-ray tube. In principle the maximum permissible load is limited by the temperature at the focus. This temperature rises during the exposure, at a rate that increases with the loading. This results in a relation between the permissible load on the tube and the exposure time, the load characteristic, which is shown in *fig. 4*. For any given exposure time the characteristic indicates the maximum power that may be used without the final temperature of the focal spot exceeding

These disadvantages are avoided by a method proposed by Bouwers at the Philips Research Laboratories, Eindhoven, as long ago as 1933[7]. The principle of this method is a gradual decrease of the load during the exposure. The exposure begins with the maximum available power (i.e. the highest current that can be used at the pre-selected voltage), so that the focus very quickly reaches its maximum permissible temperature (*fig. 5*). The current is then gradually reduced by an automatic device in such a way that the temperature at the focus remains approximately constant. The tube is now fully utilized all the time, no matter how long it takes before the exposure timer cuts out, and the exposure time is minimized in each case. This does away with the need to set the current, and all that need be done is to pre-select the tube voltage (single-knob operation).

Philips and C. H. F. Müller have already applied such a method of load control for their X-ray diagnostic units in the thirties. It was first used, in a somewhat imperfect form, in the SUPER D, and later a technically more satisfactory version was used in the MAXIMUS D. Since it involves the use of more equipment, however, this loading method failed to gain a foothold in conventional radiography with pre-selected mAs settings. The merits of the method only appear to full advantage when used with automatic exposure control.

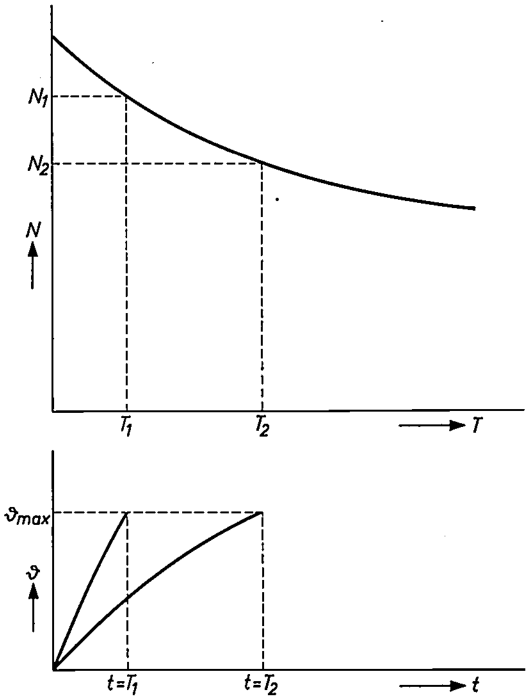


Fig. 4. Load characteristic of a diagnostic X-ray tube: the permissible (constant) load N is plotted against exposure time T. The lower graph shows the temperature ϑ at the focus, for two loads N_1 and N_2, as a function of time t. The maximum permissible temperature $\vartheta_{\max}$ is reached exactly at the end of the respective exposure times T_1 and T_2.

Physical processes in the ionization chamber

The measurement of the X-radiation may be based on the ionizing action of the rays or on the generation of fluorescent light. In the latter case a fluorescent screen is used in conjunction with a photomultiplier tube. The photocurrent charges a capacitor, and the accumulated charge is a measure of the X-ray dose. This method is efficient in those cases where visible light is produced, and is therefore widely used, for example, in fluorography.

For general X-ray diagnostic work, however, the method based on ionization has considerable advantages. This method allows the use of relatively simple detectors, which can be placed in front of the cassette (we have already explained why we think this to be desirable). Moreover, the detector can be adapted simply and inexpensively to the medical requirements. Another noteworthy feature is that the ion charge in the chamber always contributes in equal measure to the output signal irrespective of the part of the image where it was generated. In detectors for fluorescent radiation this condition is never exactly fulfilled, since the conduction of the fluorescent light to the photocathode always entails losses, depending on the path of the light. In the development of the Amplimat these considerations led us to adopt the ionization method.

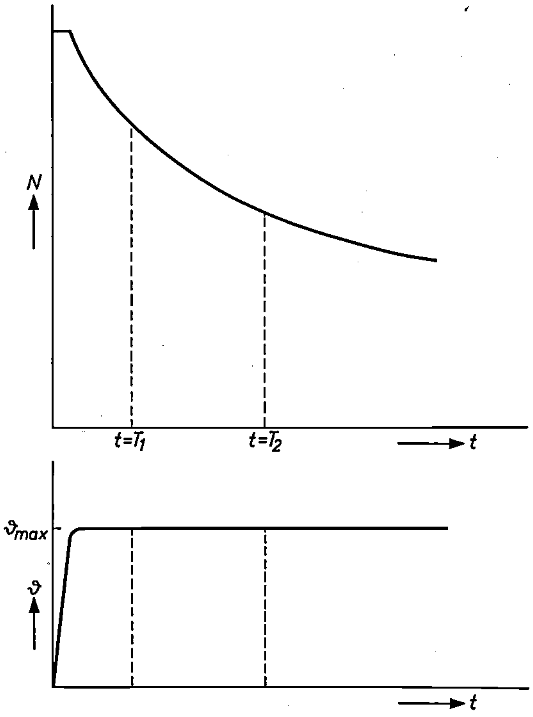


Fig. 5. Exposure with diminishing load. For all exposures the load N on the X-ray tube begins at the maximum available current and then decreases with the same function of time t. Because of the very high initial load the temperature at the focus very quickly reaches the permissible value $\vartheta_{\max}$, and then remains roughly constant.

In its simplest form the ionization chamber consists of two parallel plate electrodes with a gas-space between them. A d.c. voltage, the chamber voltage, is applied across the electrodes. When X-radiation is incident on the chamber, some of the X-ray quanta are absorbed in the electrode material and in the gas. Each absorbed quantum releases an electron from its atomic bond and imparts considerable energy to it; this

[5] For brevity we shall use the term "dose" in this article to refer to the quantity of radiation, although it is not a dose that can be expressed in röntgens (see page 96).

[6] A. Berger, Zum Problem der Bewegungsunschärfe im Röntgenbild der Lunge und des Herzens, Röntgenblätter **14**, 369-380, 1961.

[7] A. Bouwers, Verkürzung der Aufnahmezeit durch eine neue Belastungsmethode, Fortschr. Röntgenstr. **47**, 703-706, 1933.

electron in its turn produces a large number of ion pairs in the gas. If the chamber is filled with air at normal pressure and temperature, and if the electrodes consist of a material equivalent to air (e.g. graphite), then one röntgen produces, by definition [8], 2.1×10^9 ion pairs per cm^3. The chamber voltage sets up an electric field in the ionization space, and under the influence of this field the ions travel to the electrodes. The voltage is made high enough for all the ions to reach the electrode (saturation current, see page 97). It follows from the number of ion pairs mentioned that in a chamber as described above, and with an ionization volume of 1 cm^3, a saturation current of 3.3×10^{-10} A corresponds to a dose rate of 1 R per second.

Since the doses required for radiographs made with intensifying screens range from about 0.5 to 1 mR, the ionization currents are extremely small and therefore a highly sensitive d.c. amplifier has to be used. To keep the amplifier as simple as possible, every attempt is made to influence the ionization processes in such a way that the highest possible ion current is obtained. The design of the chamber offers various ways of achieving this.

To begin with, there is the electrode material: for this we use a metal, that is to say a material of relatively high atomic number. The electrons released from the metal by the absorption of X-ray quanta give rise to considerably stronger ionization of the gas than that caused by electrons from air-equivalent walls. It is true that the amplification factor for the ionization current now depends on the energy of the X-ray quanta absorbed (and therefore on the X-ray tube voltage), so that an ionization chamber of this kind is not suitable for measuring radiation in röntgens. In our case, however, this is an advantage rather than a disadvantage. Since the usual material for fluorescent intensifying screens is $CaWO_4$, i.e. a material of high atomic number, the energy-dependence given to the ionization current by the use of metal chamber walls is rather like the energy-dependence of the quantity of light generated in the intensifying screens. This means that, for a given dose, the signal delivered by the detector, while dependent on the X-ray tube voltage, depends on it in much the same way as the quantity of light generated at the film.

To increase the ionization current, it is also desirable to use a large chamber volume, and a gas of high atomic number, at a high pressure. For simplicity, however, the gas we use is air at atmospheric pressuie. As regards the volume, it will be seen later (page 97) that the surface area of the chamber is determined by the situation during the exposure, so that a change of volume is possible only by changing the distance between the electrodes. *Fig. 6* shows the effect which this has on the ionization current. With air-equivalent electrodes the ionization current varies linearly with the spacing, as shown by *a*; the curves *b* represent the ionization current variation for three copper electrodes of different thicknesses. The shape of these curves can be explained as follows. For the left hand part of the curves the electrode spacing is smaller than the range of the photoelectrons and Compton electrons emitted by the copper electrodes. The electrons have not yet entirely used up their kinetic energy for ionization when they strike the opposite wall of the chamber. The bend in the curves represents the situation where the electrons have just reached this wall. If the electrode spacing is further increased, the electrons released from the wall can no longer increase the ionization current, so that any further rise of current is due entirely to electrons liberated in the air itself. The electrode spacing at which the bend in curves *b* occurs depends on the energy of the electrons released in the wall, and therefore also on the energy of the X-ray quanta which these electrons liberate.

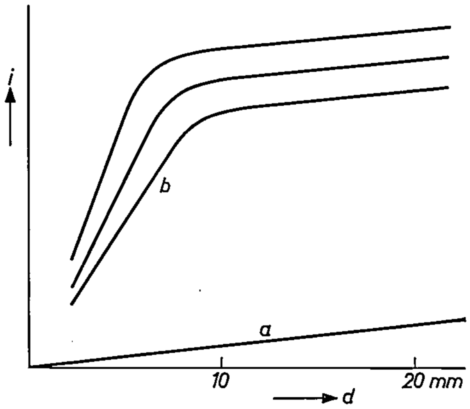


Fig. 6. Effect of inter-electrode distance *d* on the ionization current *i* of an ionization chamber (schematic). The straight line *a* relates to air-equivalent electrodes, the three curves *b* to copper electrodes of different thickness.

This effect of the electrode spacing provides a means of matching the variation of chamber sensitivity with tube voltage more exactly to the voltage-dependence of the intensifying screens. Another means to this end is the choice of the type and thickness of the electrode material. *Fig. 7* shows a few examples of how this affects the relation between the ionization current and the voltage on the X-ray tube, at constant dose rate. At low voltages the absorption of the electrons in the metal predominates, since the energy of the electrons is low. When the voltage is raised, more and more electrons can be released from deeper layers into the gas space, but the absorption of the X-rays in the metal layer decreases. There is therefore a maximum in the medium voltage range, and it is evident that the height

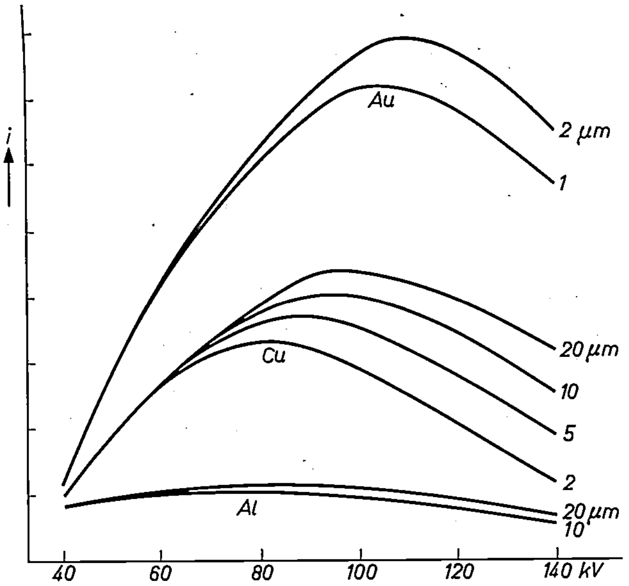


Fig. 7. Ionization current i as a function of X-ray tube voltage, for electrodes of different metals and different thickness. (The extremely thin metal layers are applied to sheet plastic.)

and situation of the maximum must depend on the material and its thickness.

The foregoing shows how the voltage-dependence of the ionization chamber can be matched to that of the intensifying screen in such a way that for any given combination of film and screens, the automatic exposure control always produces uniformly dense radiographs. The results obtained with a fine-grained calcium-tungstate screen are shown in *fig. 8* [9]. It can be seen that the constancy of the photographic density is sufficient for practical purposes.

If they remain for a longer time in the gas space, ions of opposite sign will neutralize each other. The probability that an ion will reach an electrode will be greater if it has a higher velocity in the gas space, i.e. if the field strength in the ionization chamber is greater. It follows from this that the ionization current depends on the chamber voltage. If the voltage is sufficiently high, all the ions reach the electrodes and the current reaches its saturation value. For automatic exposure control, however, it is not enough that practically all the ions formed should reach the electrodes; it is also important that their transit time should be small compared with the time available for the detection process. It will be shown below that, for the shortest exposure which the Amplimat is required to control, the switching off signal has to be given not more than about 2 ms after the beginning of the exposure. The maximum transit time of the ions must therefore be a great deal shorter if there is to be no perceptible error of measurement at these exposure times. To meet this requirement the chamber must be operated well above the voltage required for saturation, so that from now on we can neglect the influence of the chamber voltage on the ionization current.

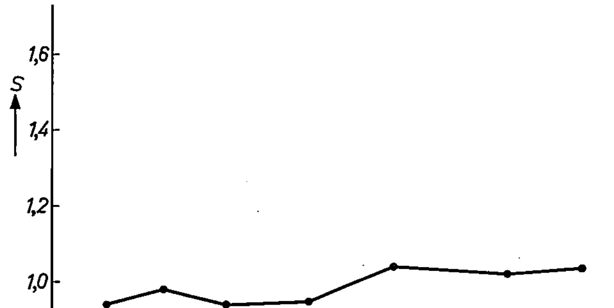


Fig. 8. Density S plotted against tube voltage for radiographs of a given object, using a fine-grained calcium-tungstate intensifying screen and the Amplimat. The fact that S is fairly constant indicates that the wavelength-dependence of the ionization chamber has been well matched to that of the intensifying screen used.

Design of the ionization chamber

The location of the ionization chamber (see fig. 3) imposes the following requirements on the design:

a) The absorbed part of the X-radiation should be as small as possible, as this can now make no contribution at the film and the dose received by the patient must be increased correspondingly.

b) The chamber should contain no components likely to cause undesired shadows in the X-ray image. The chamber should therefore be larger than the largest size of film used, otherwise the edge of the ionization chamber will show up on the radiograph.

c) The chamber should be as thin as possible, otherwise, because of the finite dimensions of the X-ray focus, the increased distance required between patient and film leads to poorer definition on the radiograph.

These requirements indicated the use of a thin chamber of large area. Reduction of the chamber thickness (requirement c) is limited by the effect, described above, of the thickness on the voltage-dependence (through the range of the electrons). We chose a thickness of 16 mm; see the photograph of the ionization chamber in *fig. 9*. This value is a good practical compromise, as at a tube voltage of about 150 kV (the maximum voltage for present-day diagnostic-type generators), the electrons released from the electrodes can have a range of 14 to 15 mm in air.

It is not self-evident that the entire chamber thickness will be available for the electron paths. The signal electrode, or collector, which accumulates the charge to

[8] See J. Hesselink and K. Reinsma, Dosemeters for X-radiation, Philips tech. Rev. **23**, 55-66, 1961/62, where various important concepts of dosimetry are defined.

[9] E. Zieler, Untersuchungen zur Belichtungsautomatik unter besonderer Berücksichtigung des Amplimat, Fortschr. Röntgenstr. **88**, 718-731, 1958.
O. Schott, Erfahrungen mit dem Belichtungsautomaten, Röntgen- und Laboratoriumspraxis **10**, 131-139, 1957.

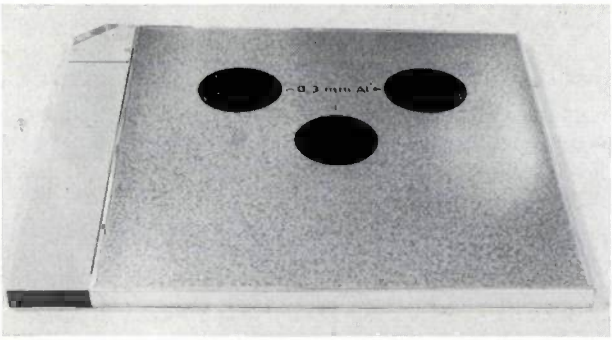


Fig. 9. Amplimat ionization chamber, with three measurement fields.

be applied to the amplifier input, cannot be placed close to the chamber wall: in such a position the screening which is always necessary would make the electrode capacitance to earth so high that the collected charge would no longer be sufficient to produce readily measurable potential differences. The collector electrode is therefore situated in the *centre plane* of the chamber, and the chamber voltage is applied between the collector electrode and two interconnected electrodes placed symmetrically against the walls. The capacitance of these electrodes to earth does not affect the measurements. In order to enable electrons to travel distances equal to the complete chamber thickness, notwithstanding this electrode arrangement, a number of holes are provided in the collector electrode. The photo-electrons and Compton electrons can thus ionize the gas in both halves of the chamber and not only in the half where they were liberated. The fact that the electrons now have to travel part of the time against the direction of the field that attracts the ions is immaterial in view of their high energy.

From requirement (b) it follows that the X-radiation should not be measured over the entire surface of the ionization chamber. Obviously, for relatively small objects, such as the neck and the skull for example, unattenuated radiation that has passed by the patient would be incident on the chamber, and the automatic exposure control would switch off the tube voltage before the film had received the necessary dose at the location corresponding to the object. The radiation is therefore measured only at a specific part of the chamber surface. This measurement field, as it is called, must not on the other hand be too small, since the quantity measured should correspond to a certain average density of the part of interest of the film, and should not be influenced by the fortuitous distribution of small details. In our opinion a measurement field of about 50 cm^2 is a useful compromise.

The location of the measurement field is of considerable practical importance. The Amplimat was the first instrument that made use of the three-field principle (see fig. 9), which was soon taken up by other manufacturers. It is evident that for most applications a centrally situated field alone is sufficient, since the image of the radiographed object is arranged to be at the centre of the film. But there are also dual organs which are situated symmetrically in the body, such as the kidneys and the lungs. For normal chest radiographs a central field is unsuitable, since it would measure the dose at the location of the heart shadow, where the film density is of no interest. Two symmetrically disposed fields give the ideal arrangement for this type of radiograph, each field covering the centre of one lung, as shown in *fig. 10*. The measurement fields can be switched in separately or in arbitrary combinations.

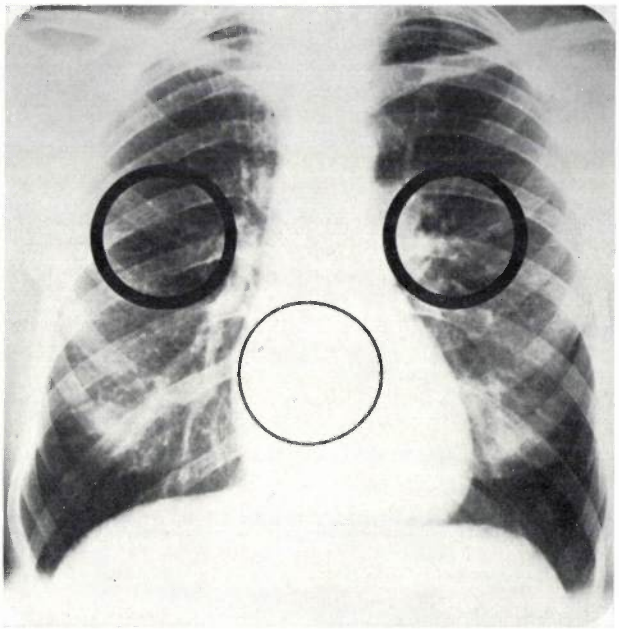


Fig. 10. Location of the measurement fields for an X-ray photograph of the chest. The centre field is not in use.

This facility is particularly useful, for example, when taking a chest radiograph of a patient for whom one lung only exhibits strong pathological absorptions, or for chest exposures taken from the side.

Fig. 11 shows a partly cut-open view of an ionization chamber, and *fig. 12* gives a cross-section at the location of a measurement field (for clarity not drawn to scale). At the central plane of the chamber there is a highly insulating plastic sheet, which is coated on both sides with an electrically conducting lacquer at the location of the measurement fields. This double coating constitutes the collector electrode, and is not visible in the radiograph. The plastic sheet with the three-part collector electrode is illustrated in *fig. 13*. The perforations mentioned above can be seen clearly. The electrical supply leads also consist of strips of conductive lacquer, since even very thin copper wires

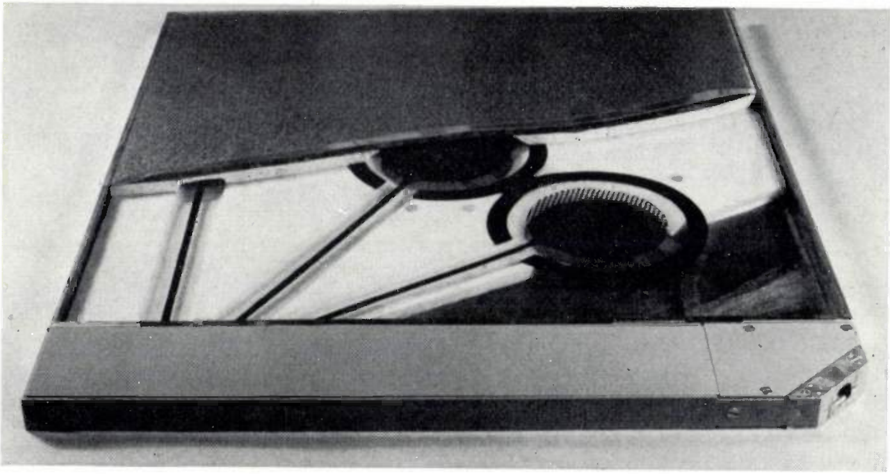


Fig. 11. Ionization chamber, partly cut open, showing the electrode systems for two of the measurement fields.

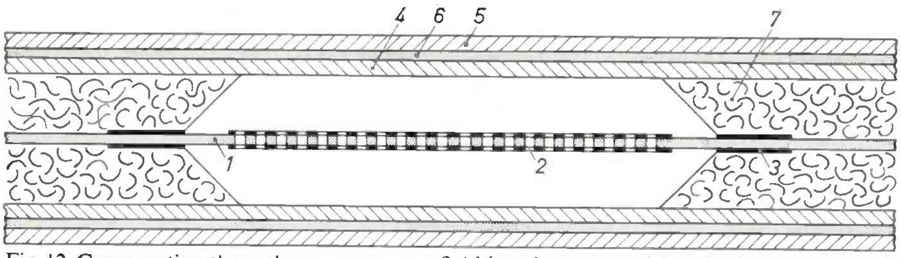


Fig. 12. Cross-section through a measurement field (not drawn to scale). *1* plastic sheet on which the centre electrode (collector) *2* and a guard ring *3* have been formed with conducting lacquer. The centre electrode and the foil it covers are perforated to allow the ionizing electrons to pass through it. *4* electrode consisting of two pieces of copper foil. *5* chamber casing (aluminium sheet 0.3 mm thick). *6* insulating sheet. *7* foamed plastic.

would show up in the radiograph. The other pair of electrodes are very thin sheets of copper of the same size as the whole ionization chamber, so that these as well are invisible in the radiograph. Except at the measurement field locations, the entire space inside the ionization chamber is filled with foamed plastic of very low density. The air space remaining constitutes the effective sensitive volume. Only in this space do the ions which are formed travel to the collector electrode under the influence of the electric field.

The density of the foamed plastic must be low enough to prevent the radiation scattered in it from making any significant contribution to the ionization in the sensitive volume, and from causing any disturbing fog on the film. The edges of the foamed plastic are made sloping so that the edges of the free air space do not show in the radiograph (fig. 12); this measure is effective even for fairly high density. The copper electrodes to which the chamber voltage is applied are insulated from the thin sheet-aluminium casing by plastic sheet. This type of construction gives the ionization chamber reasonable mechanical stability, and the X-rays available for the radiograph are not unduly attenuated. In view of the very low ionization current the insulation must be particularly good, and with modern plastics this is indeed possible. However, the insulation resistance of the foamed plastic is not sufficient here, and channels are therefore introduced in it under the supply strips to the measurement fields. The field electrodes are surrounded by guard rings, which again are applied to the plastic sheet in the form of conducting lacquer (see figure 12 and 13). The rings are given a potential equal to the average potential of the collector electrodes. By means of these measures stray currents are kept to a minimum.

The pre-amplifier

The ionization current i delivered by the ion chamber is between 10^{-8} and 10^{-11} A. If this current charges a capacitor C, the voltage V produced across the capacitor is proportional to the X-ray dose applied, since:

$$V = \frac{1}{C}\int_0^T i\,\mathrm{d}t, \quad \ldots\ldots\quad (1)$$

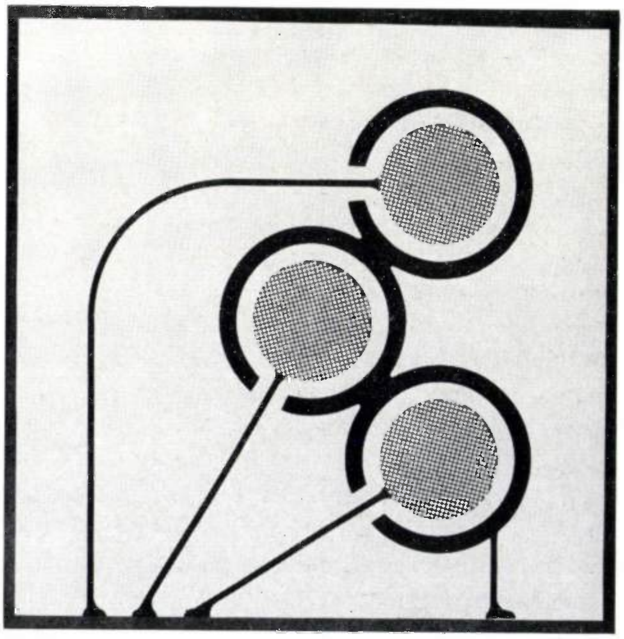


Fig. 13. Plastic sheet at the central plane of the ionization chamber, with the collector electrode consisting of three parts for the three measurement fields, and the guard rings.

to where T is the total charging time. The voltage V is therefore the quantity to be measured. The longest exposure for which X-ray generators are generally rated is 5 s. At this exposure the ionization current is at its lowest, namely 10^{-11} A. (The change which the ionization current undergoes during the exposure, due to the decreasing of the load, can be disregarded for our present purposes.) If, for example, we use a 100 pF capacitor, the final voltage across the capacitor is 0.5 V. The available "driving power" is then of the order of 10^{-12} W. Since this low power must be sufficient to switch off an X-ray generator of 50 kW reliably, it is evident that well-designed screening is required. It is best if the first amplifier stage can be accommodated in the casing of the ionization chamber itself. This method has been adopted in the Amplimat, as there is then no longer any particular difficulty in feeding the signal to the switching circuits without interference, even over fairly long cables. The pre-amplifier thus combined with the ionization chamber should preferably be contained in a volume no thicker than the ionization chamber itself. This has been accomplished by using subminiature components (fig. 17).

Up till now the most difficult problem was the choice of the first valve, whose grid current is restricted to 10^{-12} A, if a 10% error is allowed in the most unfavourable case. Commercial electrometer valves that meet such requirements, such as the Valvo 4065 and the Telefunken DF 703, cannot be used here because they have a directly heated cathode and therefore do not have sufficient mechanical stability. In cooperation with the manufacturers, however, it proved possible to develop a special valve possessing sufficient stability. A second valve must now be added to keep the signal from the pre-amplifier sufficiently free from interference.

As the longest exposure is 5 seconds, the signal to be amplified contains very low frequencies. A d.c. amplifier is therefore required. We employ a directly coupled d.c. amplifier, as this is the simplest system if only two stages are to be used. In very sensitive d.c. amplifiers the most intractable interfering effect is the drift in the quiescent (zero signal) current. We therefore designed a circuit in which the quiescent current is kept automatically constant, taking as our design basis the well-known circuit for electronically stabilized power supplies.

In a stabilized power supply (*fig. 14*) part of the output voltage V_{st} is compared with the reference voltage V_n. Any deviation produces an amplified voltage change across the anode resistance R_a, and this voltage change appears at the control grid of the next valve B_2 and has the effect of considerably reducing the original variation. Thus, any fluctuation of V_0 is only very slightly perceptible in V_{st}. If there is no load connected to V_{st}, a constant output voltage V_{st} therefore means a constant current i or i_a.

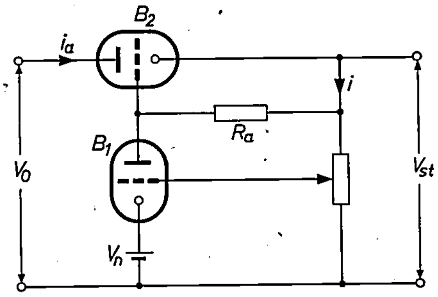


Fig. 14. Circuit showing the principle of a stabilized power supply.

Our circuit (*fig. 15*) employs the same principle. V_2 is the anode voltage source for the electrometer valve B_1. The current i flowing in the quiescent state is kept constant in the same way as in fig. 14, the voltage iR_k being compared here with the reference voltage V_n. A constant i also means a constant anode current i_a in the next valve B_2 (there is a constant ratio between the screen grid current of B_2 and the anode current). The ionization chamber with its chamber voltage source is indicated in fig. 15 by dashed lines. The capacitor C which is charged by the ion current is incorporated in the grid circuit. Its (negative) voltage V_c is added to iR_k and the sum is compared with V_n. The circuit ensures that this sum remains constant:

$$V_c + iR_k = \text{constant.} \quad . \quad . \quad . \quad . \quad . \quad (2)$$

If V_c varies, then i must also vary in such a way that eq. (2) is always satisfied. The change of i_a accompanying the change of i is the quantity measured: during the exposure i_a increases in proportion to the dose.

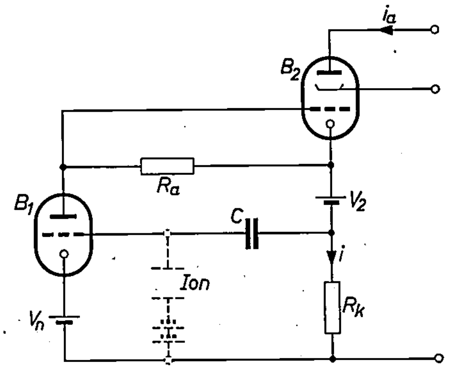


Fig. 15. Circuit of the pre-amplifier incorporated with the ionization chamber. B_1 special electrometer valve with grid current $< 10^{-12}$ A and indirectly heated cathode (used for mechanical stability). The quiescent current i is kept constant by means of the reference-voltage source V_n. The quantity to be measured is the change in i_a that occurs as soon as the ionization chamber *Ion.* supplies a charging current to the capacitor C. This change is proportional to the dose reached.

We should more accurately write, instead of eq. (2) the expression:

$$\Sigma V = 0, \text{ hence } V_n - V_g = V_c + iR_k. \quad (3)$$

Because of the amplification in the valves, however, the variation of V_g remains so small that eq. (2) is sufficiently accurate. At these small variations of V_g the valve characteristic is linear to a very close approximation so that for the amplifier as a whole there is a linear relation between the output and input quantities, as *fig. 16* shows. The linear region extends up to 13 mA, which means that a Δi_a of 12 mA can be used. The quiescent current is 0.7 mA.

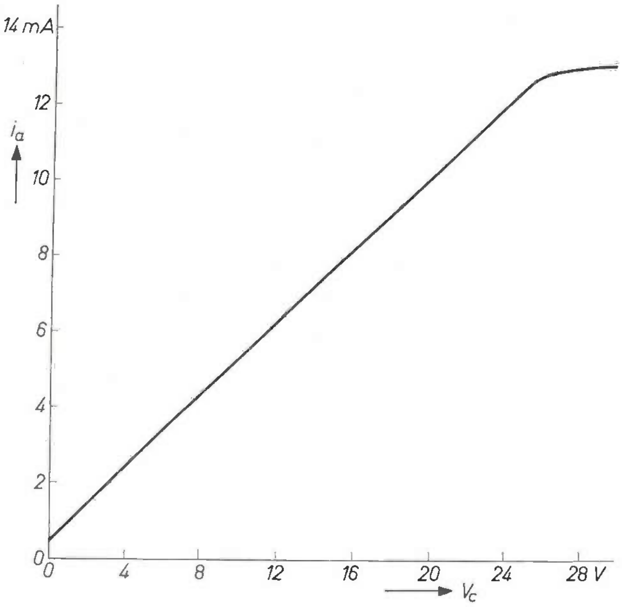


Fig. 16. Pre-amplifier characteristic: i_a is shown as a function of capacitor voltage V_c. The characteristic is linear over a current range Δi_a = 12 mA.

The constancy over long periods depends primarily, of course, on the constancy of the reference-voltage source (a Zener diode). We operated the circuit continuously for several months and found that the fluctuations in the quiescent value of i_a never exceeded ± 25 μA, although all supply voltages were taken direct from the mains with no other stabilization.

Fig. 17 shows the pre-amplifier with part of the casing removed. In addition to the two valves and the other components mentioned it contains four small relays, three of which are responsible for connecting the selected measurement fields to the amplifier input in any required combination. There is a separate capacitor for each field. As, with uniform irradiation, the ion current is proportional to the surface area of the measurement field, the sensitivity of the chamber is not affected when the various measurement fields are combined, as the ratio between field surface area and capacitance remains constant.

The complete circuit

Fig. 18 shows the circuit of the complete system. The mains lead to the high voltage transformer can be seen at the top. This circuit includes the contacts of two heavy-duty relays. Relay S_1 is controlled by the exposure control circuit and is switched on before the exposure begins. When the exposure signal is given, the contacts of relay S_2 close and this starts the exposure. In parallel with relay S_2 a third relay S_8 is energized: S_8 is contained in the pre-amplifier and in its non-energized state it short-circuits the measurement capacitor C_1. This short-circuit is removed by the energizing of S_8 for as long as the exposure lasts. The exposure is terminated by relay S_1 opening its contacts.

In *fig. 19* the various signals are shown schematically as a function of time. During the exposure the capacitor C_1 is charged by the ionization current from the ion chamber. If we assume as a first approximation that the radiation is constant (fig. 19*a*) then the (negative) voltage V_c rises linearly with time (fig. 19*b*). As we have seen, the anode current i_a is directly proportional to V_c, so that i_a also increases linearly with time (shown by the dashed line in fig. 19*c*; the solid line will be explained later). The pre-amplifier is connected by a cable to the main unit. The anode current i_a flows through one of the conductors in this cable and through the adjustable anode-resistance R_s of the valve B_2. The voltage V_1 across R_s increases with i_a and when it reaches a fixed value (V_{1s}), the mains relay S_1 receives the switching-off signal. By adjusting R_s one can pre-set

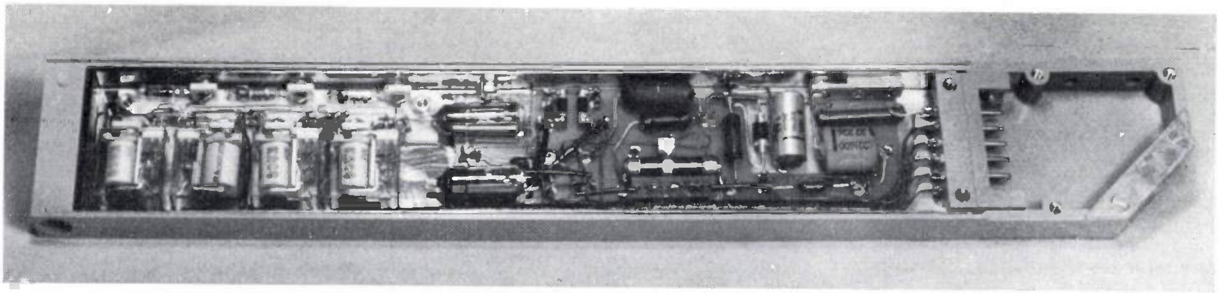


Fig. 17. The pre-amplifier opened up. It is located in the lower part of the ionization chamber casing shown in fig. 11.

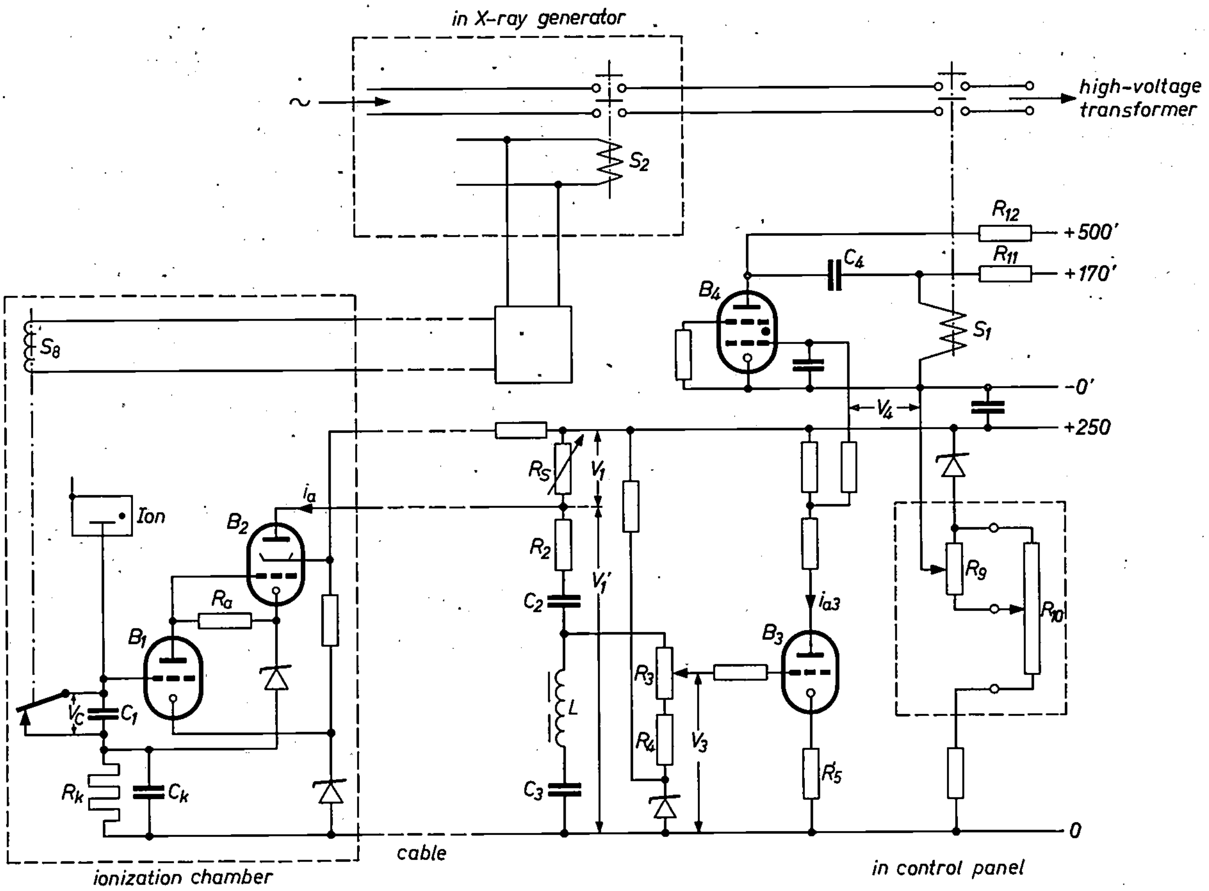


Fig. 18. Complete circuit diagram of the Amplimat, with the connections to the X-ray generator. On the left is the ionization chamber with pre-amplifier, which is connected by a cable to the rest of the circuit. B_4 is the thyratron, which is ignited at a given moment by the increasing signal voltage, causing relay S_1 to release and terminating the exposure. For the meaning of the other letters see fig. 15 and text.

the i_a value required for reaching the value V_{1s}, and so pre-select the dose (i.e. the film density) at which the radiation is switched off. The change in the voltage V_1, or rather in the complementary voltage V_1' (figures 18 and 19) appears by way of R_2 and C_2 as voltage V_3 on the grid of the next amplifier valve B_3. R_5 applies strong negative feedback to this valve, and the working point is at the upper end of the characteristic. Due to these measures the anode current i_{a3} of this valve follows the grid voltage V_3 over a wide range (fig. 19*f*), while the voltage V_3 on the grid of thyratron B_4 follows the current i_{a3}. When V_4 has reached a certain value the thyratron ignites (fig. 19*g*) and the relay S_1 releases contact.

Some details of the circuit

We shall now briefly describe some special features of the circuit.

R-C coupling

The capacitor C_2 in fig. 18 not only serves as a potential divider but also for the coupling of two d.c. amplifier stages. The *R-C* circuit that determines the signal transmission consists of C_2 and the two resistors R_3 and R_4 that follow it. The *R-C* circuit here is designed on principles different from those commonly adopted in telecommunication circuits. For an input voltage varying linearly with time t the output voltage from the *R-C* circuit varies in accordance with RC $[1 - \exp(-t/RC)]$. After a certain time ($t \gg RC$) this voltage becomes constant; however, we need only the first part of the rise ($t < RC$) in order to transmit as faithfully as possible the (theoretically linear) *increase* of the voltage. We therefore have to choose the time constant RC large enough to ensure that even at the longest exposure of 5 seconds, the output voltage will not deviate appreciably from a linear rise. With $RC = 20$ s the deviation is 11%; the difference in density which this produces on the X-ray film is imperceptible.

Compensation of the relay switching-time

Since the Amplimat is required to work with any conventional X-ray generator, it contains its own relay for disconnecting the high-voltage transformer from the mains. With a relay of this kind there is a finite delay between the instant when the energizing current is switched off and the instant when its contacts open, breaking the circuit. If the dose required for correct film density is reached with an exposure time not much larger than the release time of the relay, the film will be considerably over-exposed, because the film continues to be exposed between the instant at which the switching-off signal is given and the instant at which the tube voltage is actually switched off. To avoid such over-exposures the Amplimat circuit contains a device which compensates the influence of the relay release time by giving the signal an "early start". This is accomplished by means of the capacitor C_k in parallel with R_k in the pre-amplifier (fig. 18). As long as a constant anode current flows, C_k is charged to the voltage iR_k. The charge on the capacitor follows any charge in the anode current with a slight delay given by the time constant R_kC_k. Because of this the negative feedback through the cathode circuit is also delayed and, as the current begins to rise, the valve at first still gives full amplification; the cathode resistance is in effect short-circuited. After the charge on the capacitor has followed the anode-current variation, however, the slope only has the lower value determined by the negative feedback. This gives rise to the signal waveforms that were illustrated in fig. 19. The switching-off signal is now given earlier, by a time equal to R_kC_k.

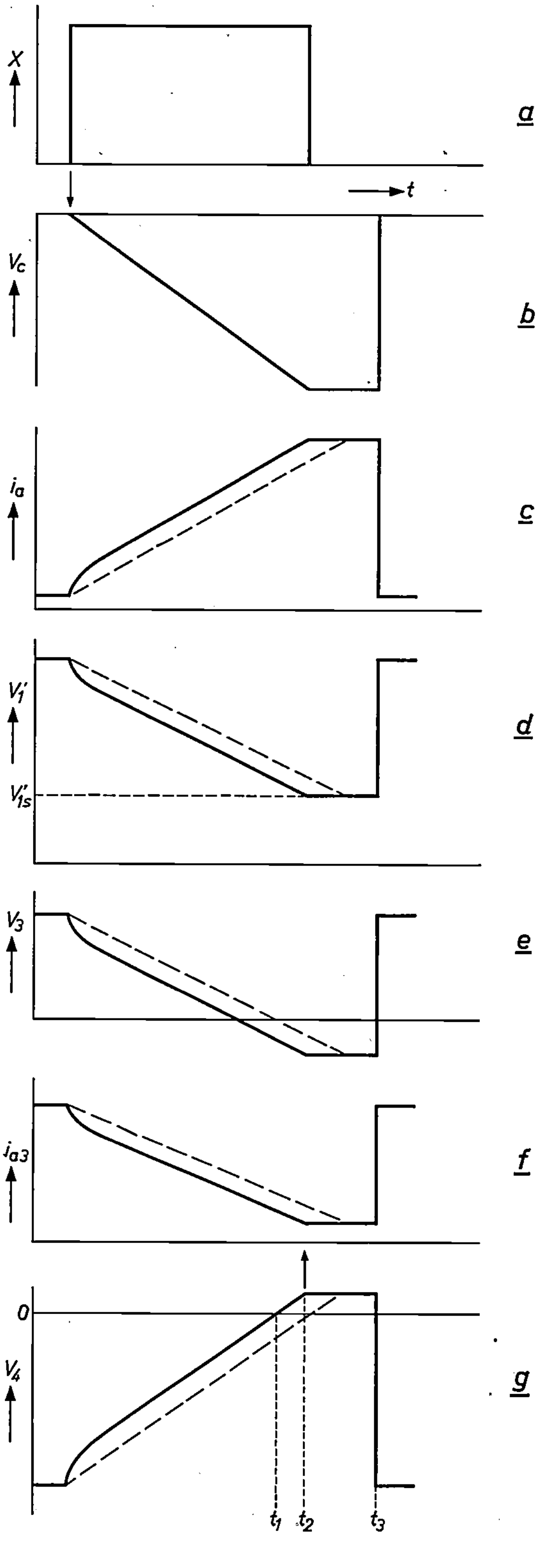


Fig. 19. Waveforms of the various signals. The X-radiation in this case is assumed to be constant. At t_1 the thyratron ignites, at t_2 the relay S_1 releases and at t_3 the relay S_8 releases.

The switching time of the relay is 3 ms, and the time constant R_kC_k is chosen to correspond with this.

Reducing the relay release time

The compensation of the release time, as described above, is of course no longer helpful if the correct exposure time approaches the relay release time. Exposures shorter than the release time are completely ruled out. To give the Amplimat the widest possible useful scope it was therefore necessary to minimize the release time of the relay. By means of a special design, using moving parts of very low mass and of much greater resilience than normal, we were able to make the time which the contact mechanism needs for its movement exceptionally short.

The electrical circuit of the coil in the relay also has its effect on the switching time. If the holding current of the relay coil is simply interrupted, it decays exponentially with the time constant L/R (see curve I in *fig. 20*), L being the inductance of the coil and R the total ohmic resistance in the coil circuit. In the example illustrated, which corresponds to normal designs the current takes between 1.5 and 2 ms to drop to zero. This time was substantially reduced by using the circuit arrangement

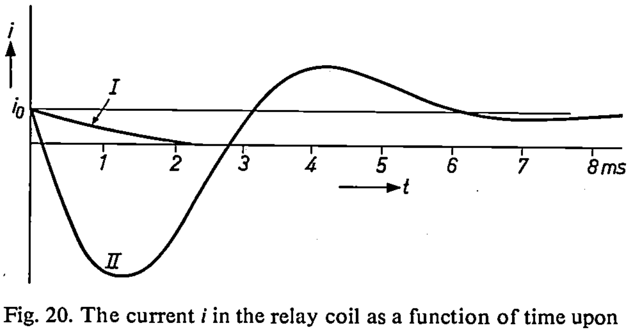


Fig. 20. The current i in the relay coil as a function of time upon switching off, using the normal circuit (I) and the circuit illustrated in fig. 18 (II). In II the magnetic field of the coil drops to zero much faster than in I. i_0 is the holding current.

already shown in fig. 18. The holding current for the relay (S_1) flows through the resistor R_{11}; capacitor C_4 is charged to a high voltage through R_{12}. In the quiescent state the thyratron B_4 is not conducting; its ignition would bring about the release of the relay. After ignition the capacitor is discharged through the thyratron and the relay coil. This current flows in the opposite sense to that of the holding current and decays as a damped oscillation. The total current follows curve *II* in fig. 20. After only 0.1 ms it has dropped to zero, so that the electromagnet is de-energized much faster than for curve *I*. The subsequent magnetization in the opposite sense is of no significance, since the current required to close the relay again is not reached.

As a result of these measures the relay contacts open 3 ms after the switching-off signal.

L-C filter

In the foregoing we have disregarded the fact that the X-radiation is not constant, as represented schematically in fig. 19*a*. In fact it contains a ripple, caused by the high voltage rectifier. The fundamental frequency of the ripple depends on the rectifier circuit. X-ray generators of average power are usually supplied from the single-phase a.c. mains, and the ripple then has a fundamental frequency of 100 c/s (two-peak generator). Generators of higher power are supplied from a three-phase supply; with the rectifier used in this case the fundamental frequency is 300 c/s (six-peak generator). The intensity of the radiation, and the signal i_a derived from it, are modulated by this voltage ripple.

We have seen that to compensate for the relay release time, the feedback in the pre-amplifier is made frequency-dependent (capacitor C_k in fig. 18). This has the result that the alternating voltage of 100 c/s or 300 c/s present in the signal is amplified more than the desiied signal, which increases linearly with time. The voltage at the grid of the thyratron does not therefore follow the same function of time as the dose, which follows the same function of time as the charge on the measuring capacitor. Although the increase of the thyratron voltage corresponds *on average* to the increase in the amount of radiation that governs the film density, a deviation can occur at any instant, depending on the point reached in the ripple cycle. This could cause considerable disparities in film density, particularly when the exposure is very short.

To obtain reproducible results, the alternating voltage, which is selectively amplified as a result of the frequency-dependent negative feedback, has to be attenuated again, but without losing the early start we want to give to the useful signal corresponding to the average voltage increase across the measuring capacitor. For this, a tuned circuit L-C_3 is used as a filter (see fig. 18), which, depending on the setting of R_s, reduces the amplitude of a 100 c/s component to between 1/25 and 1/50, and that of a 300 c/s component to between 1/40 and 1/80.

More effective filtering, which could be achieved with a higher Q for the *L-C* circuit, is not necessary and is even undesirable since the duration of the transient oscillations of a tuned circuit increases with the Q. The filter transients have their greatest effect at short exposures — the very situation in which the superimposed alternating voltage can cause the worst errors. If we again assume a linearly increasing input voltage (with no a.c. component) then at a given radiation intensity and setting of R_s the straight line *a* in *fig. 21* represents the nominal value of the switching voltage V_4 as a function of time (disregarding the constant bias). The action of the capacitor C_k in parallel with R_k in the amplifier, has shifted the straight line upwards by an amount which corresponds to a shift to the left over a time interval C_kR_k = 3 ms (line *b*). In reality, however, V_4 follows curve *c*, owing to the transient oscillation. Depending on the magnitude of the switching-off signal in the figure (strictly speaking this voltage is fixed and the scale of the ordinate in the graph is changed to correspond with the radiation intensity and the adjusted value of R_s) the deviation of curve *c* from *b* gives rise to a more or less marked deviation of the switched dose from its nominal value. The design of the tuned circuit is optimum when curve *c* cuts the straight line *b* as early as possible without excessive overshoot. The curve

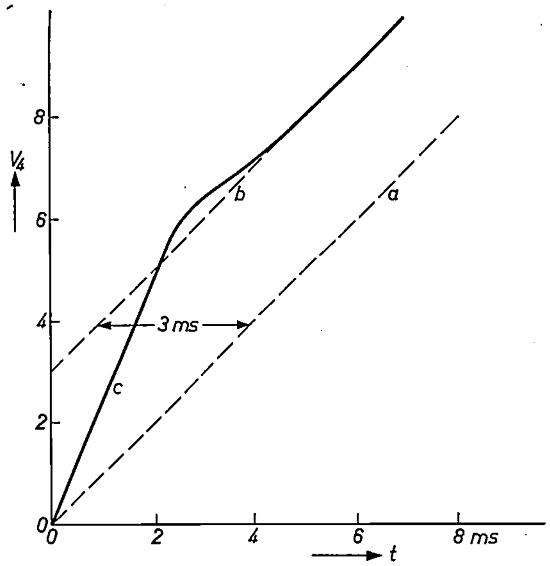


Fig. 21. Thyratron grid voltage V_4 as a function of time. At a given (constant) radiation intensity and a given setting of the density selector R_s, the variation of V_4 should follow the dashed line *b*. Due to the action of capacitor C_k in fig. 18, this line is shifted parallel to line *a*, thus compensating for the relay release time of 3 ms. In practice a curve such as *c* is obtained, because of the transient in the *L-C* filter.

obtained with the Amplimat is represented quantitatively in fig. 21. The nominal value is reached after only 2 ms, while the overshoot causes a maximum dose deviation of no more than about 7%.

Compensation of voltage effects

We have explained above (page 96) that the voltage-dependence of the ionization current is matched to that of the film density for a given combination of film and intensifying screens. In medical X-ray practice, however, various combinations of film and screens are in use. To make the Amplimat as widely useful as possible, a facility is provided for applying a correction if it is found that the use of intensifying screens of different voltage-dependence causes the density of the film to vary with the anode voltage of the X-ray tube. The correction is made with two potentiometers, R_9 and R_{10} in fig. 18, which affect the switching voltage V_4 and so control the density of the radiograph. The potentiometer R_{10} is mounted in the control panel so that its slider can be linked with the X-ray tube voltage control. The effect of this on V_4 can be approximately adjusted for the appropriate type of intensifying screen by means of R_9. This adjustment is made when the complete system is set up.

The shortest switching time in practice

Investigations by Stieve [10] have shown that radiologists often find it desirable to be able to reduce the times of direct exposures to 6 ms. It is in fact quite possible to obtain the correct film density with such short exposure times, if a high power X-ray generator is used and a high tube voltage is chosen. The shortest switching time for the exposure control system must now be no longer than this, or the use of the automatic control sets a limitation on the use of the system. The first automatic exposure timers, introduced some ten years ago, were unable to meet this requirement because the device was switched off from the mains with the standard switchgear used for X-ray generators. Automatic exposure control only became really successful after the intıoduction of a separate switching relay which, as described above, was able to give a considerable reduction in the switching time. Apart from this relay the only other element in the Amplimat that can cause any significant delay time is the *L-C* filter with its transient. The shortest switching time is the sum of the transient time of the filter and the release time of the relay, and from fig. 21 it is seen to be $2 + 3 = 5$ ms.

It may well be asked to what extent it is possible to ensure a *reproducible* dose with repeated exposures of such short duration. Apart from the spread in the release time of the switching relay, which may be some tenths of a millisecond, it is necessary to take account of the following two causes of deviation.

a) The initial switching transient of the high-voltage transformer may differ from one exposure to another. This is because, at the end of an exposure, the magnitude of the residual magnetization of the iron core of the high-voltage transformer is determined by the instant of switching off. For single-phase operation, this instant may coincide exactly with that at which the voltage passes through zero, and may also occur at any other point in the positive or negative half-cycle. When the voltage is switched on again, the magnitude of the primary current is dependent, during the first half-cycle, on the instant of switching-off. This in turn affects the magnitude of the high voltage during the first few milliseconds. Obviously, with short exposures, such fluctuations during the first few milliseconds will also produce deviations in the dose controlled by the Amplimat, since the (average) intensity during the measurement is not identical with the intensity during the release time of the relay.
b) The X-radiation contains the ripple originating in the rectifier (see page 104). The dose on the film is therefore built up more or less intermittently, while due to the operation of the *L-C* filter, the signal voltage, after the switching transient, shows an approximately *linear* increase with time, as if the dose were produced by a d.c. voltage on the X-ray tube. Depending on the instant at which the signal voltage reaches the value required for terminating the exposure, the dose that is actually delivered departs to a certain extent from this linear law.

The spread of the dose values will be relatively greater as the ripple in the tube voltage increases and the exposure time decreases. If the deviation from the nominal value is not to be greater than 20% (the corresponding deviation in the average film density is entirely acceptable as a practical limit) the switching time of the Amplimat should not fall short of 25 to 30 ms when using a two-peak X-ray generator. If a six-peak generator is used the limit is not reached until about 5 ms, because of the smaller ripple. In this case the ionization chamber has to give its switching signal within only 2 ms of the beginning of the exposure.

We have investigated the resultant spread experimentally for the six-peak generator. *Fig. 22* shows the delivered dose as a function of exposure time, for a large number of exposures. The dose was measured

[10] F. E. Stieve, The automatic exposure timer as a basis for automation in roentgen work, Acta radiol. **53**, 459-480, 1960.

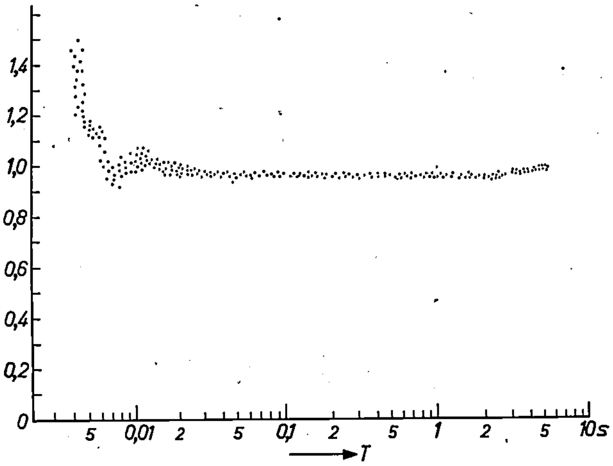


Fig. 22. Delivered dose for a large number of radiographs at different exposure times, made with the Amplimat controlling a six-peak generator. The dose actually delivered was measured with a dosemeter at the normal film location. The deviations from the nominal value (here put at 1) only exceed 20% for switching times below 5 ms.

with a dosemeter whose ionization chamber was placed immediately behind the measuring chamber of the Amplimat, i.e. at the normal location for the film. At long exposures the spread is found to be only a few per cent; it increases at short exposures. The dose also increases as the time limit of 3 ms is approached. This is easily understood since exposure times under this limit are not possible, however high the intensity. At the shortest exposure of 5 ms, however, the deviation from the desired value is never greater than 20%.

Summary. In the "Amplimat" automatic exposure control system a flat ionization chamber is placed between the patient and the X-ray film. The ionization chamber contains three measurement fields which can be selected to suit the anatomy of the subject of the radiograph. The ionization current charges a capacitor, and the increasing voltage across the capacitor supplies a signal that switches off the X-ray generator as soon as it reaches a pre-set value. The physical factors that determine the design of the ionization chamber are discussed in some detail, in particular the measures taken to ensure that the dependence of the chamber sensitivity on the X-ray tube voltage is approximately the same as the voltage-dependence of the film density, when used with a given intensifying screen. During prolonged exposures the ionization currents may be very small (e.g. 10^{-11} A in a 5 s exposure). To ensure reliable switching at such low currents, a pre-amplifier which uses subminiature circuit techniques and a specially designed electrometer valve is combined with the ionization chamber to form a single unit. The release time of the switching relay sets a lower limit to the exposure time. This limit has been reduced by mechanical and electrical means to 3 ms. The effect of the relay release time on the delivered dose is compensated by a special device in the amplifier. Measures are also taken to compensate for the ripple in the X-radiation produced by standard two- or six-peak generators, and for the effect of the voltage influence when different intensifying screens are used. Experiments have shown that with a two-peak generator a minimum exposure of 25 to 30 ms is possible without excessive deviations from the pre-set dose. With a six-peak generator the exposure can be as short as 5 ms without the spread in the dose exceeding the permissible 20% deviation.

Grease-lubricated spiral groove bearing for a straight-through shaft

621.822.5:621.892.9

The spiral groove bearing is a new and very promising form of "self-acting bearing", i.e. one in which the running surfaces are completely sepaiated from each other by a thin layer of lubricant and in which the rotation itself forces the lubricant between the running surfaces. This latter action is achieved in the present case by the pumping action of the spiral grooves Oil may be used as a lubricant, but grease is also suitable, especially for small bearings (shaft diameters of a few millimetres). The great advantage of grease is that it does not leak away from the bearing so readily when the shaft is stationary.

These points were explained in detail some time ago in an article in this journal [1]. This article included an illustration (fig. 26) of a small test motor with its shaft supported at each end by a grease-lubricated conical spiral groove bearing, capable of taking both thrust and radial loads. In the course of life tests, this motor has meanwhile been run for 15 000 hours and, during this period, has been stopped and restarted about 150 000 times without being re-lubricated.

The above relates to a "blind" bearing, i.e. one in which the shaft terminates. The very good results achieved with the test motor encouraged us to look for an anlogous design for use with a *straight-through shaft*, i.e. a spiral groove bearing, capable of taking both thrust and radial loads, that could be lubricated with grease and would require no further re-lubrication during its life.

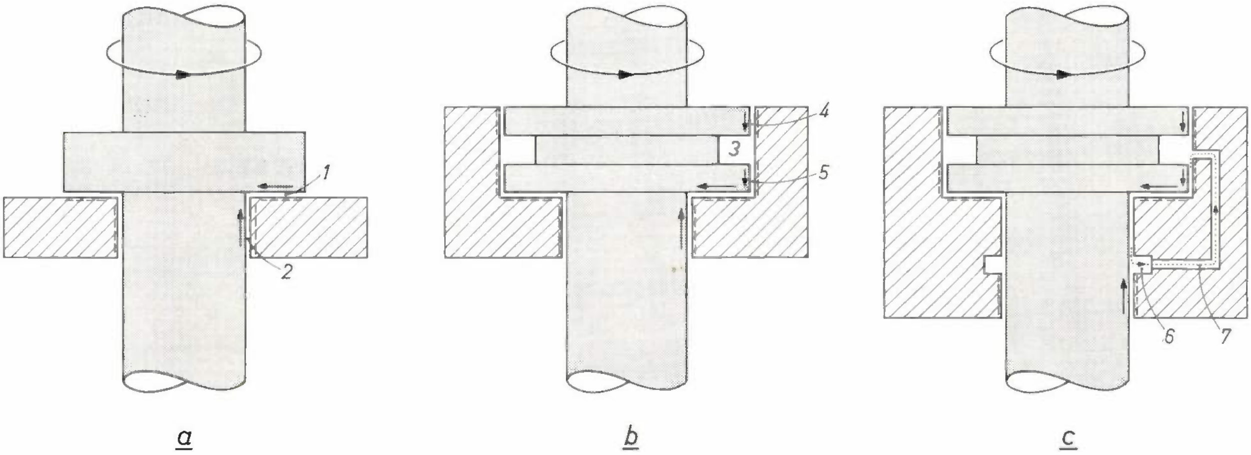


Fig. 1. Grease-lubricated spiral groove bearing for straight-through shaft.
a) The simplest possible form, with thrust bearing *1* and cylindrical bearing *2*.
b) An extension of this design, with arrangements to return the grease forced out of bearing *1* when the shaft is stationary.
c) The final design. While the shaft is rotating, there is a small continuous flow of grease to the annular space *6*, from which it is returned to chamber *3* through channels *7*.

Our solution to this problem is illustrated in *fig. 1c*. By way of explanation, let us first consider the simplest imaginable configuration for the required bearing, as shown in fig. 1*a*. The thrust bearing *1* has to take up the axial load, and cylindrical bearing *2* the radial load. Bearing *1*, designed for a bearing gap of, say, 5 to 10 μm, has spiral grooves, forcing the grease in the direction of the arrow. The pressure thus set up increases strongly towards the centre and provides the required thrust load carrying capacity for bearing *1*; at the same time it ensures that journal bearing *2*, which has a typical clearance of 20 to 30 μm, is always filled with grease. Bearing *2* is, therefore, itself capable of providing the necessary radial bearing capacity. To prevent the grease forced towards the centre by spiral grooves *1* from leaking away from the bottom of bearing *2*, the latter must be provided with a series of helical grooves to pump the lubricant in the opposite direction.

In order to build up a pressure equal to that provided by bearing *1*, the grooves in bearing *2*, owing to the much wider gap and other causes, would have to be of quite considerable length. It would be better, therefore, to entrust part of this counter-pumping effect to a section of the thrust beating *1*: its bearing surface can be divided into two rings, only the outer one being provided with the inwardly-directed pattern of spiral grooves, the inner one carrying a pattern giving the counter-pumping effect (a herringbone pattern, as shown in fig. 24 of reference [1]).

[1] E. A. Muijderman, New forms of bearing: the gas and the spiral groove bearing, Philips tech. Rev. **25**, 253-274, 1963/64.

There are several reasons why the simple bearing in fig. 1*a* is not entirely satisfactory. Firstly, the most obvious one: when the shaft is *stationary*, the thrust load will gradually force some of the grease in gap *1* outwards, and this grease will accumulate around the outer circumference of the bearing. When the motor is started this grease will be thrown off by the centrifugal force, and eventually there will be insufficient grease to maintain the lubricating film.

This can be counteracted by the design shown in fig. 1*b*. Here, the shaft has a flange with an annular space *3* containing a reserve of grease and which will collect the lubricant forced out of the gap *1* when the shaft is stationary. As soon as the shaft begins to rotate, this grease is forced outwards by centrifugal force and pressed against the inner wall of the bearing housing. Two sets of helical grooves *4* and *5* are cut in this wall, which both have a downward pumping action. The sole purpose of grooves *4* is to prevent some of the grease from escaping upwards, while grooves *5* pump the grease towards the thrust bearing *1*, so that both this bearing and journal bearing *2* are always full of grease.

This does indeed correct the deficiency inherent in the design shown in fig. 1*a*, but there is yet another and rather more subtle difficulty. It is difficult to balance accurately the opposed impelling forces that the grooves in bearings *1* and *2* exert on the grease, particularly because these forces will be determined by the thrust and radial loads, which can vary independently of each other. It is therefore impossible in practice to prevent a slight flow of grease. In order to overcome this difficulty, we have included a return channel for the grease that is pumped away. The assembly is dimensioned in such a way that the pumping action of bearing *1*, which is reinforced by grooves *5*, is always predominant. As may be seen in fig. 1*c*, the journal bearing has an annular space *6* to catch the grease. Sealing from below is ensured by the grooves in bearing *2*, and there are a number of return passages *7* from the cavity *6* to the annular space *3* at the input side of the thrust bearing *1*.

The grease circulation on which the final design is based has the further advantage that any air-bubbles that might be left when the bearing is packed with grease are very rapidly removed from the bearing gap. It is, therefore, unnecessary to pack the bearing in vacuo.

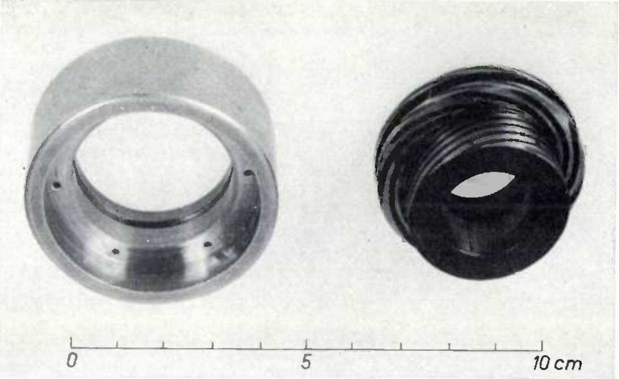


Fig. 2. Parts of the new spiral groove bearing. The cylindrical bearing housing is shown on the left, and the rotating part, made of hard material that is attached to the straight-through shaft is shown on the right. Here the groove patterns (*1*, *2*, *4* and *5* in fig. 1) are not cut in the wall of the bearing housing, but in the hardened moving part; this makes no difference to the operation of the bearing. The moving part is shown the opposite way round from the housing so that the groove patterns may be seen. The ends of the grease channels, which, in this design, run in a slightly different way from that shown in fig. 1*c* (*7*), may be seen in the bearing housing.

Fig. 2 is a photograph of one design of the bearing described above.

This extension, given here in outline, of the application of this type of bearing to straight-through shafts opens up wide new prospects. It may be expected to find many uses as a bearing for relatively high loads and speeds or long life (more than 20 000 hours). For such conditions cheap sintered bearings (impregnated bearings) are unsuitable, and the type of bearing described here will render the use of compratively expensive and bulky ball bearings is unnecessary. Further advantages, such as the fact that grease-lubricated bearings, like oil-lubricated bearings, are much quieter than sintered or ball bearings, and the fact that they withstand heavy thrust loads well, will, in our opinion greatly encourage their use.

G. Remmers

G. Remmers is with Philips Research Laboratories, Eindhoven.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands	*E*
Mullard Research Laboratories, Redhill (Surrey), England	*M*
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France	*L*
Philips Zentrallaboratorium GmbH, Laboratory at Aachen, Weisshausstrasse, 51 Aachen, Germany	*A*
Philips Zentrallaboratorium GmbH, Laboratory at Hamburg, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany	*H*
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium.	*B*

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

W. Albers and **H. J. Vink**: De atomaire bouw van vrijwel stechiometrische verbindingen.
Chem. Weekblad **61**, 61-69, 1965 (No. 6). *E*

P. Beekenkamp: The influence of the coordination number of boron on the properties of alkali borate glasses.
Physics of non-crystalline solids, Proc. int. Conf. Delft 1964, pp. 512-524, North-Holland Publ. Co., Amsterdam 1965. *E*

H. J. van den Berg: A new technique for obtaining thin lipid films separating two aqueous media.
J. mol. Biol. **12**, 290-291, 1965 (No. 1). *E*

G. Blasse: On the octahedral radius of the trivalent cobalt ion.
J. inorg. nucl. Chem. **27**, 748-750, 1965 (No. 3). *E*

G. Blasse: Magnetic properties of mixed metal oxides containing trivalent cobalt.
J. appl. Phys. **36**, 879-883, 1965 (No. 3^{II}). *E*

G. Blasse: Ferromagnetism and a new type of ferrimagnetism in oxygen spinels.
Solid State Comm. **3**, 67-69, 1965 (No. 4). *E*

G. Blasse: Influence of covalent bonding on some magnetic properties of transition metal oxides.
Bull. Soc. Chim. France, April 1965, 1212-1214. *E*

G. Blasse: New compounds with perovskite-like structures.
J. inorg. nucl. Chem. **27**, 993-1003, 1965 (No. 5). *E*

G. Blasse: Some magnetic properties of mixed metal oxides with ordered perovskite structure.
Philips Res. Repts. **20**, 327-336, 1965 (No. 3). *E*

R. Bleekrode and **W. C. Nieuwpoort**: Chemical lasers, II.
Z. angew. Math. Phys. **16**, 107-110, 1965 (No. 1). *E*

R. Bleekrode and **W. C. Nieuwpoort**: Flame laser: model and some preliminary experimental results.
Appl. Optics, Suppl. 2: Chemical lasers, pp. 179-180, 1965. *E*

M. Borot: Luminescence par bombardement cathodique dans l'arséniure de gallium.
Onde électr. **45**, 1204-1215, 1965 (No. 463). *L*

C. J. Bouwkamp: An infinite product.
Proc. Kon. Ned. Akad. Wet. **A 68**, 40-46, 1965 (No. 1). *E*

J. C. Brice: The angular dispersion of anisotropy in thin ferromagnetic films.
Brit. J. appl. Phys. **16**, 965-968, 1965 (No. 7). *M*

A. Bril, **W. L. Wanmaker** and **C. D. J. C. de Laat**: Fluorescent properties of red-emitting europium-activated phosphors with cathode ray excitation.
J. Electrochem. Soc. **112**, 111-112, 1965 (No. 1). *E*

P. H. Broerse: Camerabuizen.
Ned. T. Natuurk. **30**, 409-419, 1964 (No. 12). *E*

J. Burmeister: Kristallzüchtung inkongruent schmelzender Verbindungen.
Phys. Stat. sol. **10**, 223-226, 1965 (No. 1). *A*

J. Burmeister: Schmelzen von Zinkoxyd durch Hochfrequenzerhitzung.
Phys. Stat. sol. **10**, K 1 - K 2, 1965 (No. 1). *A*

K. H. J. Buschow: Rare earth-aluminium intermetallic compounds of the form RAl and R_3Al_2.
J. less-common Met. **8**, 209-212, 1965 (No. 3). *E*

K. H. J. Buschow: The lanthanum-aluminium system.
Philips Res. Repts. **20**, 337-348, 1965 (No. 3). *E*

K. H. J. Buschow and **J. H. N. van Vucht:** Das System Erbium-Aluminium und ein Vergleich mit dem System Yttrium-Aluminium.
Z. Metallk. **56**, 9-13, 1965 (No. 1). *E*

G. Cayman: Laser action in an alloyed GaAs junction.
Solid-State Electronics **8**, 455-456, 1965 (No. 4). *E*

J. R. Chamberlain: The use of "fast" transitions in quantum counter applications.
Physics Letters **16**, 261-262, 1965 (No. 3). *M*

A. Claassen and **L. Bastings:** Potentiometric determination of vanadium in iron and steel containing chromium and tungsten.
Z. anal. Chemie **202**, 241-250, 1964 (No. 4). *E*

L. J. Collins and **J. Smith:** Mirror mounts for experimental optical masers.
J. sci. Instr. **42**, 499-500, 1965 (No. 7). *M*

H. J. van Daal: Mobility of charge carriers in silicon carbide.
Thesis Amsterdam, Dec. 1964. *E*

K. Deneke: A remark on the semiconducting compound "$AgFeTe_2$".
J. appl. Phys. **36**, 653, 1965 (No. 2). *A*

C. J. Dippel and **H. Jonker:** Ein neues Reproduktionsverfahren mittels Metallkeim-Introduktion.
Reprographie, Ber. I. int. Kongress für Reprographie, Cologne 1963, pp. 187-190, publ. Helwich, Darmstadt 1964. *E*

J. A. W. van der Does de Bye: Temperature regulation by controlled injection of liquid nitrogen.
Rev. sci. Instr. **36**, 104-105, 1965 (No. 1). *E*

W. F. Druyvesteyn: Influence of a transport current on the magnetization of superconducting indium-lead alloys.
Physics Letters **14**, 275-276, 1965 (No. 4). *E*

W. Elenbaas: The thermal-conduction loss in the high-pressure mercury-vapour discharge.
Philips Res. Repts. **20**, 213-225, 1965 (No. 3).

G. Engelsma and **G. Meijer:** The influence of light of different spectral regions on the synthesis of phenolic compounds in gherkin seedlings in relation to photomorphogenesis, I. Biosynthesis of phenolic compounds, II. Indoleacetic acid oxidase activity and growth.
Acta bot. neerl. **14**, 54-72 and 73-92, 1965 (No. 1). *E*

G. Frank: Das Gasgleichgewicht im System Germanium-Brom.
Ber. Bunsenges. phys. Chemie **69**, 119-124, 1965 (No. 2). *A*

P. Gerthsen and **F. Kettel:** Messung der Wärmeleitfähigkeit von Gasen in rotierender Zylinderanordnung.
Z. angew. Phys. **19**, 378-380, 1965 (No. 4). *A*

J. A. Geurst: Calculation of high-frequency characteristics of thin-film transistors.
Solid-State Electronics **8**, 88-90, 1965 (No. 1). *E*

J. A. Geurst: Calculation of high-frequency characteristics of field-effect transistors.
Solid-State Electronics **8**, 563-566, 1965 (No. 6). *E*

R. W. Gibson: Solid ultrasonic delay lines.
Ultrasonics **3**, 49-61, 1965 (April/June). *M*

A. H. Gomes de Mesquita: The structure of a silicon carbide polytype 24*R*.
Acta crystallogr. **18**, 128, 1965 (No. 1). *E*

A. H. Gomes de Mesquita, C. H. MacGillavry and **K. Eriks:** The structure of triphenylmethyl perchlorate at 85 °C.
Acta crystall. **18**, 437-443, 1965 (No. 3). *E*

C. Grenier and **B. Elschner:** Alternating-current losses in Nb/Zr superconductive coils.
Philips Res. Repts. **20**, 235-252, 1965 (No. 3).

H. L. Günther and **G. Hötzl:** Dehnungsmeßstreifen aus Silizium.
Z. Instrumentenk. **73**, 13-14, 1965 (No. 1). *H*

W. van Haeringen: Perturbation approach to the polaron self-energy in the intermediate coupling range.
Phys. Rev. **137**, A 1902 - A 1909, 1965 (No. 6A). *E*

J. Haisma: Gas lasers.
Z. angew. Math. Phys. **16**, 74-84, 1965 (No. 1). *E*

P. A. H. Hart: The internal noise of a transverse wave electron beam.
Tubes pour hyperfréquences, Trav. 5e Congrès int., Paris 1964, pp. 29-31. *E*

H. U. Harten: Die Grenzfläche Halbleiter-Elektrolyt.
Festkörperprobleme III, 81-123, Vieweg, Brunswick 1964. *H*

Y. Haven and **B. Verkerk:** Diffusion and electrical conductivity of sodium ions in sodium silicate glasses.
Phys. Chem. Glasses **6**, 38-45, 1965 (No. 2). *E*

H. Henkes: Some remarks on a paper of G. Kohlstrung.
Phys. Stat. sol. **9**, K 105 - K 107, 1965 (No. 2).

F. N. Hooge: Injection lasers.
Z. angew. Math. Phys. **16**, 89-97, 1965 (No. 1). *E*

F. N. Hooge and **T. Vrijheid-Lammers:** The influence of counterdoping on the distribution of mangenese over substitutional and interstitial sites in germanium.
Philips Res. Repts. **20**, 292-305, 1965 (No. 3). *E*

J. Hornstra: Dynamic properties of grain boundaries.
Science of Ceramics **2** (Proc. Conf. Noordwijk aan Zee 1963), 191-202, Academic Press, London 1965. *E*

H. Hoven and **A. Stegherr:** Über das Ätzen von Schliffen aus Bi_2Te_3-Bi_2Se_3- und Bi_2Te_3-Sb_2Te_3-Mischkristallen.
Prakt. Metallogr. **2**, 114-118, 1965 (No. 3). *A*

H. J. Hubers: Thermal stresses at glass-to-metal seals.
Symp. on the contact of hot glass with metal, Scheveningen 1964, pp. 785-806.

H. J. Hubers: Contraintes dans les objets en verre et leur influence sur la résistance à la rupture.
Silicates industr. **30**, 25-35, 1965 (No. 1).

B. B. van Iperen and **H. J. C. A. Nunnink:** The velocity-modulated electron beam as a harmonics generator for millimeter and submillimeter waves.
Tubes pour hyperfréquences, Trav. 5e Congrès int., Paris 1964, pp. 124-129. *E*

G. H. Jonker, J. T. Klomp and **Th. P. J. Botden:** High-temperature seals on pure dense alumina.
Science of Ceramics **2** (Proc. Conf. Noordwijk aan Zee 1963), 295-302, Academic Press, London 1965. *E*

L. Kapel and **J. H. Kroon:** Glass on metals and metallic oxides: wettability phenomena.
Symp. on the contact of hot glass with metal, Scheveningen 1964, pp. 739-762.

A. van Katwijk: A grammar of Dutch number names.
Found. Language **1**, 51-58, 1965 (No. 1). *E*

K. Klamer: Technologie van elektrokeramiek.
Klei en Keramiek **15**, 30-42, 1965 (No. 2).

S. R. de Kloet: Accumulation of RNA with a DNA-like base composition in Saccharomyces carlsbergensis in the presence of cycloheximide.
Biochem. biophys. Res. Comm. **19**, 582-586, 1965 (No. 5). *E*

E. Kooi: Influence of X-ray irradiations on the charge distributions in metal-oxide-silicon structures.
Philips Res. Repts. **20**, 306-314, 1965 (No. 3). *E*

E. Kooi and **M. M. J. Schuurmans:** Temperature-gradient effects during heat treatments of oxidized silicon.
Philips Res. Repts. **20**, 315-319, 1965 (No. 3). *E*

C. Kramer: A low-frequency pseudo-random noise generator.
Electronic Engng. **37**, 465-467, 1965 (No. 449). *E*

J. Krugers and **O. Reifenschweiler:** Activation analysis.
Practical instrumental analysis, pp. 117-129, Elsevier, Amsterdam 1965. *E*

W. Kuypers and **M. T. Vlaardingerbroek:** Measurement of electron-beam noise.
Philips Res. Repts. **20**, 349-356, 1965 (No. 3). *E*

W. Kwestroo and **J. Visser:** The ultrapurification of hydrofluoric acid.
Analyst (J. Soc. Anal. Chem.) **90**, 297-298, 1965 (No. 1070). *E*

J. van Laar and **J. J. Scheer:** Fermi level stabilization at semiconductor surfaces.
Surface Sci. **3**, 189-201, 1965 (No. 2). *E*

H. de Lang: Optics in laser research.
Z. angew. Math. Phys. **16**, 7-14, 1965 (No. 1). *E*

J. Liebertz and **C. J. M. Rooymans:** Die Ilmenit/Perowskit-Phasenumwandlung von $CdTiO_3$ unter hohem Druck.
Z. phys. Chemie Neue Folge **44**, 242-249, 1965 (No. 3/4). *A, E*

B. Lopes Cardozo: Adjusting the method of adjustment: SD vs DL.
J. Acoust. Soc. Amer. **37**, 786-792, 1965 (No. 5). *E*

F. K. Lotgering: On the spontaneous magnetization of $MnFe_2O_4$.
Philips Res. Repts. **20**, 320-326, 1965 (No. 3). *E*

G. Meijer and **G. Engelsma:** The synergistic influence of a pre-irradiation on the photoinhibition of gherkin seedlings.
Photochem. Photobiol. **4**, 251-258, 1965 (No. 3). *E*

O. W. Memelink: De werking van halfgeleiderdioden en transistoren.
Ingenieur **77**, E 15 - E 21, 1965 (No. 7). *E*

O. W. Memelink: De werking en eigenschappen van de silicium gestuurde gelijkrichter.
Ingenieur **77**, E 39 - E 42, 1965 (No. 11). *E*

B. J. Mulder: Photoconductivity spectra of stable and metastable single-crystals of perylene.
Rec. Trav. chim. Pays-Bas **84**, 713-728, 1965 (No. 6). *E*

E. A. Muyderman: Construeren met spiraallagers.
Polytechn. T. A **20**, 154 A - 166 A, 1965 (No. 4). *E*

J. Neirynck and **Ph. van Bastelaer:** Catalogue of step-responses of narrow-band FM systems.
Rev. MBLE **8**, 98-109, 1965 (No. 2). *B*

A. G. van Nie: Grundlagen der Mikrowellenbündel-Technik.
Nachrichtentechn. Z. **18**, 17-21, 1965 (No. 1). *E*

A. K. Niessen and **F. A. Staas:** Hall effect measurements on type II superconductors.
Physics Letters **15**, 26-28, 1965 (No. 1). *E*

A. K. Niessen, J. van Suchtelen, F. A. Staas and **W. F. Druyvesteyn:** Guided motion of vortices and anisotropic resistivity in type-II superconductors.
Philips Res. Repts. **20**, 226-234, 1965 (No. 3). *E*

W. C. Nieuwpoort and **R. Bleekrode:** Chemical lasers, I.
Z. angew. Math. Phys. **16**, 101-106, 1965 (No. 1). *E*

J. Ober: The large-signal behaviour of a travelling-wave tube with an attenuating central helix section.
Philips Res. Repts. **20**, 357-376, 1965 (No. 3). *E*

D. J. van Ooijen: Flux quantization noise on hollow superconducting cylinders.
Physics Letters **14**, 95-97, 1965 (No. 2). *E*

W. P. Osmond: The magnetic structure of spinels containing paramagnetic octahedral cations but diamagnetic tetrahedral cations.
Proc. Phys. Soc. **85**, 1191-1196, 1965 (No. 6). *M*

G. J. Oudemans: Compaction of dry ceramic powders.
Science of Ceramics **2** (Proc. Conf. Noordwijk aan Zee 1963), 133-146, Academic Press, London 1965. *E*

F. E. Roberts: High temperature Hall measurements on GaAs.
Physics Letters **17**, 21-22, 1965 (No. 1). *M*

J. G. van Santen and **A. J. F. de Beer:** A digital voltmeter with a photoconductive potentiometer.
Solid-State Electronics **8**, 7-12, 1965 (No. 1). *E*

D. A. Schreuder: Contrast sensitivity in test field with bright surround.
J. Opt. Soc. Amer. **55**, 729-731, 1965 (No. 6).

M. S. Seltzer: Diffusion of manganese into gallium arsenide.
J. Phys. Chem. Solids **26**, 243-250, 1965 (No. 2). *E*

P. J. W. Severin: The interaction of microwaves with the cathode fall and negative glow in a glow discharge.
Thesis Utrecht, Oct. 1964. *E*

A. L. Stuijts and **C. Kooy:** Influence of technological factors on the sintering behaviour of a ferrite.
Science of Ceramics **2** (Proc. Conf. Noordwijk aan Zee 1963), 231-242, Academic Press, London 1965. *E*

A. L. Stuijts, J. Verweel and **H. P. Peloschek:** Dense ferrites and their applications.
IEEE Trans. on communication and electronics **83**, 726-736, 1964 (No. 75). *E*

K. Teer: On the optimum configuration for a condenser microphone.
Acustica **15**, 256-263, 1965 (No. 5). *E*

G. W. Tichelaar: Brosse breuk van ongelegeerd lasmetaal bij lage temperaturen.
Lassymposium 1964, pp. 31-41, and Discussies van het Lassymposium 1964, pp. 23-28, publ. Wyt, Rotterdam. *E*

D. R. Tilley: Critical fields and flux penetration pattern in anisotropic type II superconductors.
Proc. Phys. Soc. **85**, 1177-1184, 1965 (No. 6). *M*

H. J. L. Trap: Electric properties of some aluminoborate glasses.
Physics of non-crystalline solids, Proc. int. Conf. Delft 1964, pp. 635-645, North-Holland Publ. Co., Amsterdam 1965. *E*

J. van der Veen: A simple method for the preparation of some methylphenylsulfonium nitrates.
Rec. Trav. chim. Pays-Bas **84**, 540-544, 1965 (No. 4). *E*

M. L. Verheijke: Efficiencies of a coincidence spectrometer for positron annihilation radiation.
Nucl. Instr. Meth. **34**, 132-136, 1965 (No. 1). *E*

A. G. van Vijfeijken and **A. K. Niessen:** Longitudinal and transverse voltages in superconductors.
Physics Letters **16**, 23-24, 1965 (No. 1). *E*

H. J. Vink: Bouw en opbouw van vaste anorganische stoffen.
Chem. Weekblad **61**, 1-11, 1965 (No. 1). *E*

J. Volger: The generation of heavy currents in superconducting circuits.
International Advances in Cryogenic Engineering (part 2 of Adv. Cryog. Engng. **10**), pp. 98-104, 1965. *E*

K. Walther: Dispersion of a step-modulated carrier wave in a waveguide.
Proc. IEEE **53**, 410, 1965 (No. 4). *H*

W. L. Wanmaker and **M. G. A. Tak:** The retention of copper and bromide in electroluminescent ZnS:Cu,Br phosphors.
Philips Res. Repts. **20**, 278-291, 1965 (No. 3).

J. van de Waterbeemd: Kinetics of growth and structure of thin films of tin.
Physics Letters **16**, 97-98, 1965 (No. 2). *E*

C. Weber: Calculation of electron guns taking into account space charge and thermal velocities.
Tubes pour hyperfréquences, Trav. 5e Congrès int., Paris 1964, pp. 47-49. *E*

K. R. U. Weimer, H. Bodt and **M. T. Vlaardingerbroek:** Interaction of an electron beam plasma system with slow-wave structures.
Tubes pour hyperfréquences, Trav. 5e Congrès int., Paris 1964, pp. 465-468. *E*

G. A. Wesselink and **A. Bril:** Fluorescent properties of trivalent neodymium in lanthanum and yttrium orthoniobates and tantalates.
Philips Res. Repts. **20**, 269-277, 1965 (No. 3). *E*

P. C. van der Willigen: De ontwikkeling van het CO_2-lassen van staal.
Metalen e.a. Constr.mat. **20**, 62-71, 1965 (No. 3). *E*

G. Zanmarchi: Absorption of light near the band edge in α-SiC 24R.
Philips Res. Repts. **20**, 253-268, 1965 (No. 3). *E*

H. Zimmer: Field emission at 9000 Mc/s in a superconducting cavity.
Electronics Letters **1**, 24, 1965 (No. 1). *H*

Volume 27, 1966, No. 3/4 pages 69-112 Published 26th August 1966

PHILIPS TECHNICAL REVIEW

VOLUME 27, 1966, No. 5

Peltier cooling

W. Lechner

537.322.15:536.48

The investigations carried out at the Philips Aachen Laboratory into the technology and applications of Peltier cooling which were concluded some years ago, have led to the production of Peltier batteries of various types. In view of the interest in thermoelectric refrigeration, it appeared that a summary of the principles, techniques, and fields of application of this method of cooling would be of value.

Although physicists have carried out a great deal of research on the thermoelectric effects since their discovery more than a century ago (Seebeck 1822, Peltier 1834, Thomson 1854), it is only in fairly recent years (at least for the Peltier effect), that materials have been found and developed that make their technical application a possibility. These are the semiconducting materials. Thermoelectric refrigeration systems based on the Peltier effect (Peltier batteries) have been so greatly improved in recent years that they are now in use in various fields. These are of course those fields where thermoelectric systems have special advantages over traditional refrigerators — their small size, the fact that they can be adapted to the object to be refrigerated and can be easily though accurately regulated — and where their relatively low efficiency, which is still a disadvantage, is of minor importance.

To present a survey of this method of refrigeration and its applications, we shall first of all briefly describe the thermoelectric effects and the principles underlying the design of Peltier elements and batteries. The thermoelectric effects will then be treated in terms of atomic theory, after which we shall examine the principal characteristics of the Peltier element as a heat pump (the refrigerating engineer prefers to call this a cold pump). In doing so we shall place emphasis on the relationship between these characteristics (e.g. the efficiency and attainable temperature reduction) and the physical properties of the semiconductors employed; this relation is expressed in a "figure of merit". The rest of the article will be concerned with the choice and preparation of the semiconducting material (compounds derived from bismuth telluride), with the operating conditions of Peltier batteries, and with some typical applications in measuring techniques and in laboratory use. It is hoped that the examples chosen will give some idea of the measures needed to solve various types of refrigeration problem with moderate to very strict requirements, for example for constancy of temperature. It is also hoped that the article will act as something of a stimulus and suggest ideas for new applications.

The thermoelectric effects

The first of the thermoelectric effects to be discovered was the *Seebeck effect*. This effect is based on the fact that in many substances a temperature gradient $\mathrm{d}T/\mathrm{d}x$ is associated with an electric field E:

$$E = \alpha \frac{\mathrm{d}T}{\mathrm{d}x}, \qquad (1)$$

α being the Seebeck coefficient of the material. If we make a closed circuit of two conductors A and B which have different Seebeck coefficients α_A and α_B (see *fig. 1*) and we keep the junctions at different temperatures, then a voltmeter inserted anywhere in the circuit will indicate a voltage: this is the Seebeck volt-

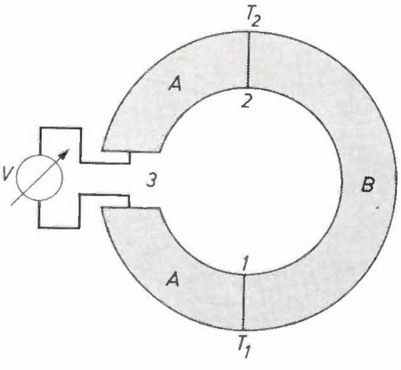


Fig. 1. The Seebeck effect. When the junctions *1* and *2* between conductors A and B are kept at two different temperatures T_1 and T_2, an e.m.f. V is generated in the circuit in accordance with eq. (2).

W. Lechner is with the Aachen laboratory, Philips Zentrallaboratorium GmbH.

age. In other words, the Seebeck voltage is the net e.m.f. of the thermocouple thus constructed. It is given by:

$$V = \int_{3(A)}^{2} E\mathrm{d}x + \int_{2(B)}^{1} E\mathrm{d}x + \int_{1(A)}^{3} E\mathrm{d}x$$

$$= \int_{1(A)}^{2} E\mathrm{d}x - \int_{1(B)}^{2} E\mathrm{d}x = \int_{T_1}^{T_2} (\alpha_A - \alpha_B)\mathrm{d}T. \quad . \ (2)$$

Thus, what is measured in practice is only the difference $\alpha_{AB} = \alpha_A - \alpha_B$ between the "absolute" Seebeck coefficients α_A and α_B. We call α_{AB} the "differential" Seebeck coefficient of the combination AB. The absolute Seebeck coefficient can be found by a measurement using as a reference a good metal conductor (e.g. copper) which, as a rule, has a relatively small thermo-e.m.f.

The *Peltier effect* is based on the occurrence of an energy-flow as a result of an electric current. The density of the Peltier energy-flow is proportional to the density of the electric current; the proportionality factor is the (absolute) Peltier coefficient Π. Other energy-flows, for example thermal conduction and the removal of Joule heat, are not taken into account here. If we substitute a d.c. source for the voltmeter in fig. 1 (see *fig. 2*), an electric current I then flows through A and B. If there is a difference between the absolute Peltier coefficients Π_A and Π_B of A and B, and hence between the energy-flows Q_A and Q_B through A and B, then the difference between the two energy-flows is released as heat at one junction and heat is taken up from the surroundings at the other. The element thus works as a heat pump. The heat absorbed or produced per second at a junction is given by:

$$Q = \Pi_{AB} I, \quad . \ . \ . \ . \ . \ . \ . \ . \ (3)$$

where

$$\Pi_{AB} = \Pi_A - \Pi_B. \quad . \ . \ . \ . \ . \ (4)$$

Π_{AB} is the "differential" Peltier coefficient of the combination AB. (From now on we shall usually omit the subscripts for the sake of brevity. The context will make it clear whether an absolute or a differential coefficient is intended.)

In general, α and Π depend to some extent on temperature. This temperature dependence may differ considerably from one material to another; for thermodynamic reasons, however (and this applies both to the absolute and the differential coefficient), the Thomson relation is always valid:

$$\Pi = \alpha T, \quad . \ . \ . \ . \ . \ . \ . \ . \ (5)$$

where T is the absolute temperature [1].

To make the following discussion clearer, we shall first of all describe the construction of Peltier elements and batteries in somewhat more detail. *Fig. 3* shows how the scheme of fig. 2 is arranged in practice. The arrangement used is determined by the method of

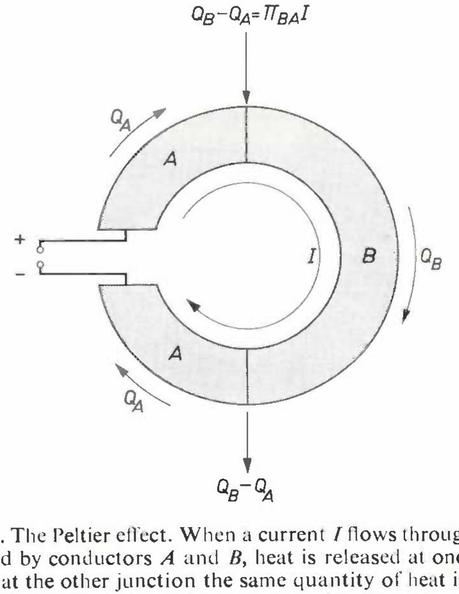


Fig. 2. The Peltier effect. When a current I flows through a circuit formed by conductors A and B, heat is released at one junction, while at the other junction the same quantity of heat is taken up from the environment. This quantity of heat is given by equations (3) and (4). (The d.c. source, like the voltmeter in the case of the Seebeck effect, can be introduced in the circuit at any arbitrary point, e.g. at one of the junctions, as in practical Peltier elements.)

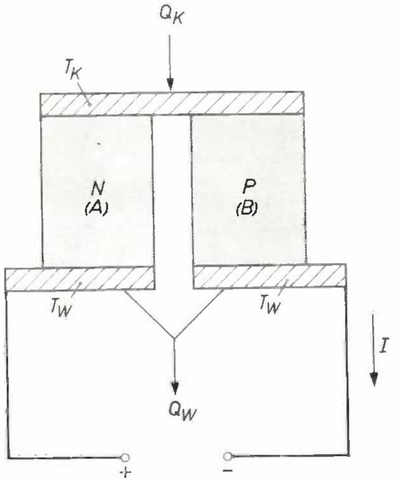


Fig. 3. The Peltier element as a heat pump. The element consists of a block of N-type material (A), a block of P-type material (B) and metal contact plates. At the cold side, at temperature T_K, an amount of heat Q_K is taken up per second from the environment; at the hot side, at temperature T_W, an amount of heat Q_W is released per second (this can be dissipated e.g. by water-cooling). I is the current. In general Q_W and Q_K differ from the Peltier heat $\Pi_{BA} I$ because they are also affected by Joule heat and thermal conduction.

[1] There is a second Thomson relation, in which a third thermoelectric effect occurs, called the Thomson effect. If there is an electric current in the material and at the same time a temperature gradient, an amount of heat, referred to as Thomson heat, is generated in addition to the Joule heat. The Thomson heat, which may be positive or negative, usually has little practical significance, and it will not be taken into consideration in the present article.

preparing the semiconductor materials, and the need for rapid transfer of the heat produced (or taken up) at the junctions. The junction contacts are in the form of broad flat pieces of metal, which are good conductors of electricity. One conductor (*A*) consists of an *N*-type semiconductor, and the other (*B*) of a *P*-type semiconductor. The uppermost contact in fig. 3 absorbs the heat withdrawn from the cooled object (a quantity of Q_K per second), and a temperature T_K is established here. The other contact, which is kept at a temperature T_W by means of appropriate heat exchangers (e.g. by liquid cooling), gives up a quantity of heat Q_W per second.

To increase the refrigerating capacity, and for easier adaptation to the object to be cooled, a number of Peltier elements of this type can be connected together to form a Peltier battery (see *fig. 4*), in such a way that they are electrically in series, but in parallel for heat transmission. The chains of elements thus produced are sandwiched between metal plates. Between the plates and the elements there is a very thin intermediate layer, which is an electrical insulator but as good a conductor of heat as possible. An alternative way of obtaining an intermediate layer is to use metal plates with anodized surfaces. This gives mechanical strength to the arrangement while at the same time solving the insulation problem, so that the user can attach both the object to be cooled and the heat exchangers straight on to the battery.

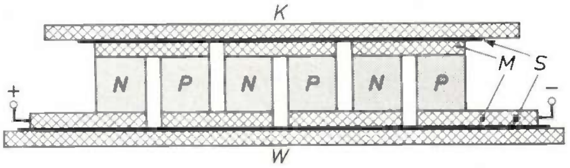


Fig. 4. Diagram illustrating the construction of a Peltier battery from a series of blocks of *N*-type and *P*-type semiconductors. *M* metal junctions. *S* electrically insulating layer. *K* cold side. *W* hot side.

Variations of this arrangement, designed for special applications, are discussed in the last section of this article.

The thermoelectric effects in terms of atomic theory

The conventional model for electrical conductivity in semiconductors and metals readily provides a qualitative explanation of the thermoelectric effects and of the fact that these effects can be considerably larger in semiconductors than in metals.

In this model the electric current is transported by free charge-carriers — electrons in metals and *N*-type semiconductors, holes in *P*-type semiconductors. The resistance arises because the charge-carriers are scattered by lattice vibrations (phonons) or by physical or chemical imperfections in the lattice.

The essential difference between metals and semiconductors is well expressed in the energy band scheme. The electrons are in quantum states or energy levels which are grouped together in bands; the electrons are distributed among these states in accordance with the Fermi-Dirac statistics. The distribution function *f* (the average number of electrons per state, $0 \leqslant f \leqslant 1$) is equal to unity (all energy levels occupied) for $E \ll E_F$, and *f* is equal to zero (all energy levels empty) for $E \gg E_F$ where E_F is the Fermi energy; *f* has values different from 1 or 0 only in an energy region round E_F, of the order of magnitude of kT. Completely empty or completely filled bands do not contribute to the conductivity.

In semiconductors E_F lies between two energy bands (see *fig. 5*). In *N*-type material conduction takes place by means of electrons in the upper, otherwise completely empty band, called the conduction band; in *P*-type material conduction takes place by means of holes in

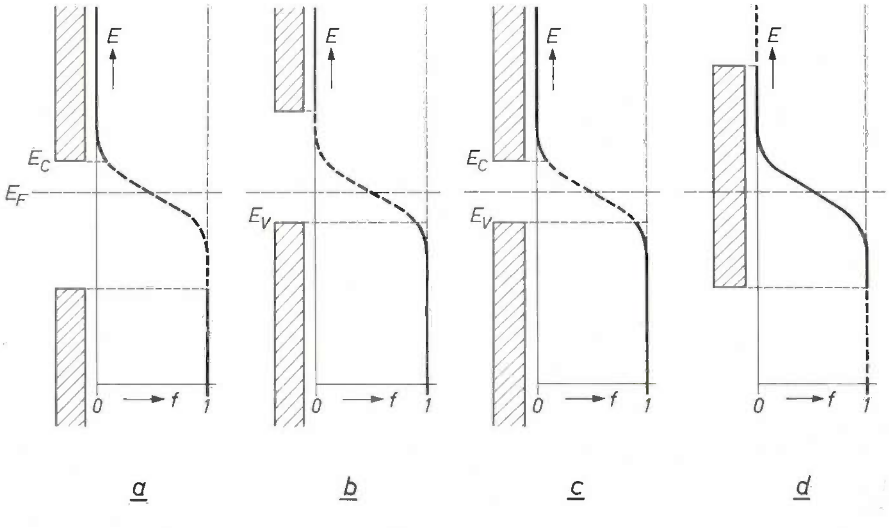


Fig. 5. Fermi-Dirac distribution in semiconductors and metals. The horizontal axis corresponds to the distribution function *f* of the states (the average number of electrons per state), the vertical axis shows the energy. The shaded strips on the left of each graph represent the permitted energy bands.
a) *N*-type semiconductor. The electric current is carried by a small number of electrons in the conduction band (bottom edge E_C).
b) *P*-type semiconductor. The charge-carriers are holes in the valence band (top edge E_V).
c) Intrinsic semiconductor. Charge-carriers are present in both bands.
d) Metal. The conduction is due to electrons with an energy around E_F.

the lower, otherwise completely filled band, called the valence band. These electrons and holes originate from donors and acceptors respectively. In intrinsic semiconductors both bands contain charge-carriers, because of the excitation of electrons from the valence band to the conduction band. In metals E_F lies somewhere in the middle of an allowed energy band; the conduction is then due to electrons with energies around E_F.

When the Fermi level in *N*-type or *P*-type semiconductors is very near the edge of the band in which the conduction takes place, or is even *in* that band, the semiconductor is said to be "degenerate"; it is then no longer true that $f \ll 1$ for *N*-type material or that $1 - f \ll 1$ for *P*-type material. For simplicity we shall assume in the following that the semiconductors are not degenerate. Although in practice degenerate semiconductors are in fact very important in Peltier cooling, the theory of non-degenerate semiconductors offers a better basis for an understanding of the thermoelectric effects.

The Fermi level is important as a reference energy level, since the condition for thermal equilibrium in an "electron gas", e.g. at a junction between two conductors or semiconductors, is that the Fermi level should everywhere be at the same height. Thus, if the energy of an electron is always reckoned from the Fermi level, the energies of electrons in various parts of a combination of conductors can immediately be compared with one another. The energy of a *hole* at a level *below* the Fermi level (as found, for example, in *P*-type semiconductors) is then invariably *positive*.

The Peltier effect

1) The Peltier effect in extrinsic semiconductors.

We consider a circuit formed from *N*-type and *P*-type semiconductors in series (*fig. 6*). In the *P*-type section the charge-carriers (holes) flow with the current, in the *N*-type section the charge-carriers (electrons) flow against it. In both sections, charge-carriers flow from the junction where they are generated (the upper one in fig. 6) to the other junction (the lower one) where they recombine. The relative location of the energy bands (*fig. 7a*) at the junctions is determined by the rule that the Fermi level shall everywhere be at the same height. At one junction (the cold one) the energy required for generating charge-carriers is taken from the environment as thermal energy. The energy produced upon recombination at the other junction is given up to the environment in the form of heat. The result is therefore a transfer of heat, and this heat transfer resulting from an electric current is the Peltier effect.

The above description of the Peltier effect is applicable to the combination of an *N*-type with a *P*-type semiconductor. The generation or recombination of a hole-electron pair is nothing else, however, than the transition of an electron between two energy levels. Generally speaking, a Peltier effect will occur in a circuit of two semiconducting materials if the energies of the states in which the conduction takes place are different in the two semiconductors (fig. 7*b*).

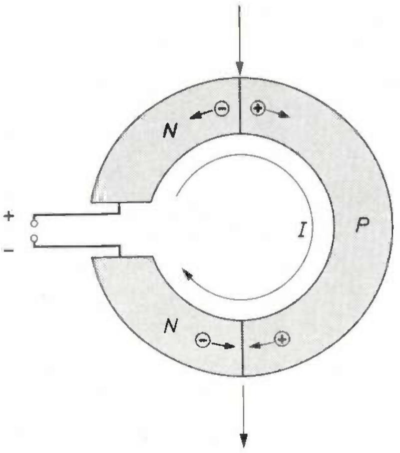


Fig. 6. When a current flows in a circuit formed from an *N*-type and a *P*-type semiconductor, charge-carriers are transported from one junction to the other in both semiconductors.

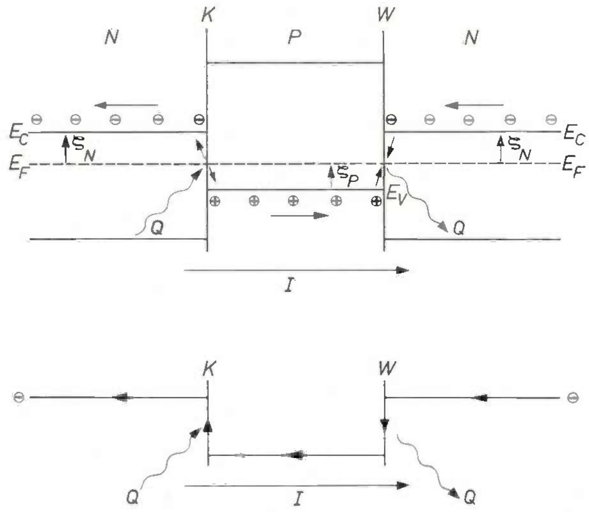


Fig. 7. *a*) The energy bands in a combination of an *N*-type and a *P*-type semiconductor. E_F Fermi level. E_C lower edge of conduction band in the *N*-type semiconductor. E_V upper edge of the valence band in the *P*-type semiconductor. *I* electric current. At the cold junction *K* hole-electron pairs are generated, and thermal energy *Q* is taken up. At the hot junction *W* an amount of heat *Q* is given up upon recombination. $\zeta_N = E_C - E_F$, $\zeta_P = E_F - E_V$.
b) The energies of the conduction states in a circuit of two semiconductors. At the cold junction *K* the conduction electrons are excited to a higher energy state by thermal energy taken up from the environment. At the hot junction *W* they return to the lower state, and heat *Q* is given up.

The giving out or taking up of heat occurs at the junctions where the electron makes an energy jump. The Peltier effect is thus a differential effect for two conductors; only the *difference* in the energies of the conduction states is relevant. It is useful, however, for formal purposes to introduce an *absolute* Peltier coefficient for a given material, and this is done in the following way. Taking the Fermi level as a fixed reference level, a particular energy can be assigned to each electron. This energy is carried along with the electron, so that an electric current gives rise to an energy-flow. The ratio of the energy-flow to the electric current is the absolute Peltier coefficient.

The sign of the absolute Peltier coefficient for N-type material is opposite to that for P-type material. This is because, although in both cases the energy flows with the charge-carriers, the electric current in a P-type conductor flows in the same direction as the carriers, but in an N-type conductor it flows in the opposite direction to the carriers. The greatest Peltier effects are therefore to be expected from combinations of N- and P-type material.

An expression for the absolute Peltier coefficient of an N-type semiconductor can be found as follows. The mean transported energy per electron is the sum of its potential and kinetic energies. Its potential energy (see fig. 7*a*) is $\zeta_N = E_C - E_F$; this is the energy of an electron without kinetic energy at the edge of the conduction band. The kinetic energy is $A_N kT$. If the mean *transported* kinetic energy per electron were the same as the mean kinetic energy, one would expect A_N to have a value of 3/2. The mean transported kinetic energy is, however, greater, because the higher-energy electrons not only transport *more* energy; they do so *faster*. Moreover it often happens, for example in scattering due to ionized lattice imperfections,. that the higher-energy electrons undergo relatively less scattering. Because of this kind of effect a value of 2 or 4 is found for A_N, depending on whether the electrons are mainly scattered by acoustic lattice vibrations or by ionized lattice imperfections.

The mean transported energy per electron is therefore $A_N kT + \zeta_N$. The transported charge per electron is —e, so that the absolute Peltier coefficient (the ratio of transported energy to transported charge) is therefore:

$$\Pi_N = -\frac{1}{e}(A_N kT + \zeta_N). \quad . \quad . \quad . \quad . \quad (6)$$

Similarly, for P-type semiconductors one finds:

$$\Pi_P = +\frac{1}{e}(A_P kT + \zeta_P), \quad . \quad . \quad . \quad . \quad (7)$$

where $\zeta_P = E_F - E_V$ (see fig. 7*a*).

Substituting (6) and (7) in (3) and (4) we find:

$$Q = \frac{I}{e}\Big[(A_N + A_P)kT + \zeta_N + \zeta_P\Big]. \quad . \quad (8)$$

At the hot junction, I/e is the number of pairs of charge-carriers recombining per second, and the equation shows that the total energy of these pairs is converted into heat. At the cold junction (8) has the opposite sign: the heat taken up from the environment is converted into the energy of the pairs of charge-carriers created.

In the theoretical treatment of Peltier cooling the Seebeck coefficient α is often used instead of the Peltier coefficient Π. The Seebeck coefficient is directly linked with Π by the Thomson relation (5): $\Pi = \alpha T$. We shall therefore give the expression for the Seebeck coefficient of an N-type or P-type material as derived from equation (6) or (7) and (5):

$$\alpha_N = -\frac{k}{e}\left[A_N + \frac{\zeta_N}{kT}\right], \quad . \quad . \quad . \quad . \quad (9)$$

$$\alpha_P = +\frac{k}{e}\left[A_P + \frac{\zeta_P}{kT}\right]. \quad . \quad . \quad . \quad . \quad (10)$$

It should be noted that the temperature dependence of α is not only given by the T appearing explicitly in (9) and (10). In general, ζ also is temperature-dependent. In semiconductors of interest in thermoelectric applications it is often found that the temperature dependence of ζ compensates the explicit factor T to some extent. As the first term can usually be regarded as temperature-independent, α is frequently taken in practice to be independent of temperature. In this rough approximation, Π is then proportional to T.

2) The Peltier effect in metals.

The Peltier effect in *metals* is substantially less than in extrinsic semiconductors. In the light of what has been written above this can be understood as follows. In semiconductors the thermoelectric effects are of opposite sign for electrons and for holes. If these are present at the same time, the effects will partly (or, in special cases, wholly) compensate one another. This is the situation in intrinsic semiconductors, and in these the thermoelectric effects are indeed relatively small. In several respects the situation in metals is very much the same as that in intrinsic semiconductors in which the energy gap goes to zero (cf. fig. 5).

The difference between metals and extrinsic semiconductors which is of significance in thermoelectric effects can be summarized as follows: in an extrinsic semiconductor, charge and energy are transported by charge-carriers on only one side of the Fermi level, whereas in a metal they are transported by carriers on both sides of the Fermi level.

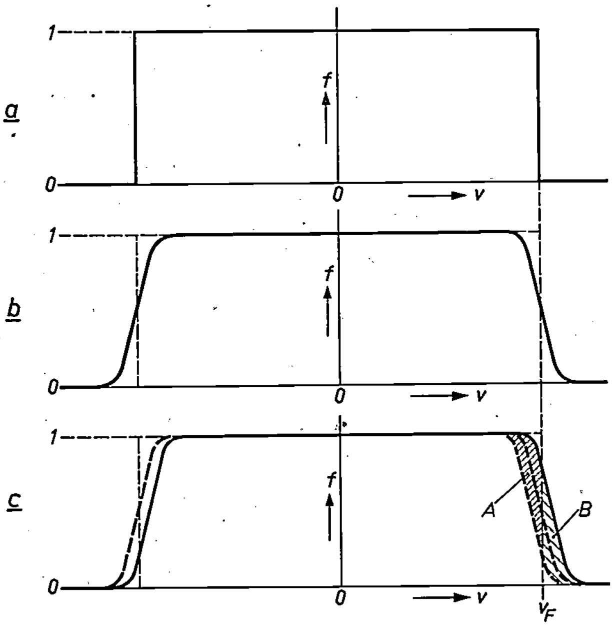


Fig. 8. The Peltier effect in metals. The distribution function f of the states in which the electrons have a velocity v is shown as a function of v in a one-dimensional representation. v_F Fermi velocity. *a)* $T = 0$ °K, $I = 0$. *b)* $T > 0$ °K, $I = 0$. *c)* $T > 0$ °K, $I \neq 0$. The electrons, represented by the regions A and B, transport negative and positive energy, respectively, towards the right. The absolute Peltier coefficient is determined by the difference between the absolute values of these quantities of energy, and is zero to a first approximation.

It will be useful to consider the Peltier effect in metals in somewhat more detail. For simplicity we shall confine ourselves to a one-dimensional model. In *fig. 8* the distribution function f (this represents the occupation of the energy states) is shown as a function of the velocity of the electrons in these states. At $T = 0$ °K all states below the Fermi level (i.e. with velocities lower than the Fermi velocity v_F) are full and those above it are empty. Above absolute zero this sharp boundary in the distribution becomes blurred. In both cases the velocities of the electrons to the left and to the right compensate each other for zero current exactly. If a current flows through the conductor the distribution is displaced slightly. The electrons with uncompensated velocities, represented by the regions A and B, are responsible for the net charge and energy transport. If we reckon the energy from the Fermi level, we see that the electrons in B transport positive energy to the right, whereas the electrons in A transport negative energy to the right. These energy quantities are about equal in absolute value and compensate one another to a first approximation. At first sight there is thus no Peltier effect in metals. There is nevertheless a small effect: this arises because the two quantities of energy transported do not completely compensate one another. Two reasons why compensation is incomplete may be mentioned. Firstly, the averages over A and B of the energy transport per electron are not exactly equal in absolute value. Secondly, there is generally a slight difference between the densities of states above and below the Fermi level, so that the number of electrons in A is not exactly equal to the number in B. The relatively large Peltier effect in semi-metals and in some transition metals can be explained as being due to a steep gradient in the density of states at the Fermi level.

The Seebeck effect

1) The Seebeck effect in semiconductors.

In view of the close relation that exists between the Peltier effect and the Seebeck effect, we shall also, for completeness, give a qualitative explanation of the Seebeck effect. Let us first consider a semiconductor. Assume that a temperature gradient exists perpendicular to a given cross-section through a piece of semiconducting material. The charge-carriers from the hot side which pass through this cross-section are somewhat "hotter" than the charge-carriers from the cold side; i.e. their mean (thermal) velocity is somewhat greater. If the concentration at both sides were the same, there would therefore be a net current of charge-carriers from the hot to the cold side ("thermal diffusion"). In an isolated piece of semiconductor this thermal diffusion current is compensated in the steady state by two currents in the opposite direction: in the first place the transported charge-carriers will build up a concentration gradient that results in a diffusion current, and in the second place they will generate a charge, and hence an electric field which in turn gives rise to a current. This electric field, produced as a result of a temperature gradient, is the Seebeck effect.

We shall now give a simple and very rough estimate of the magnitude of the effect, confining ourselves to the case where the diffusion current due to the concentration gradient is insignificant, and assuming that the energy of the charge-carriers consists only of thermal kinetic energy. Let l be the mean free path of the charge-carriers. The charge-carriers arriving at a given cross-section from the hot (or cold) side have a mean velocity appropriate to the cross-section where they last collided and acquired the lattice temperature, i.e. to a cross-section at a distance $\frac{1}{2}l$ away from the hot (or cold) side of the given one. This gives a thermal diffusion current:

$$\tfrac{1}{2}(nev_{x+\frac{1}{2}l} - nev_{x-\frac{1}{2}l}) = \tfrac{1}{2}\, ne \frac{\mathrm{d}v}{\mathrm{d}x}\, l\,.$$

Let us put

$$\tfrac{1}{2}mv^2 = \tfrac{1}{2}kT,$$

so that

$$\frac{\mathrm{d}v}{\mathrm{d}x} = \tfrac{1}{2}\frac{k}{mv}\frac{\mathrm{d}T}{\mathrm{d}x}\,.$$

If we also set the thermal diffusion current equal to the current σE due to the field E (the conductivity being given by

$$\sigma = ne\mu = \frac{ne^2l}{mv}\,,$$

where $\mu = el/mv$ is the mobility), we obtain $E = \alpha \mathrm{d}T/\mathrm{d}x$, where $\alpha = \frac{1}{4}k/e$. In our approximation α contains only the quantities k and e, and in agreement with the Thomson relation, in the same way as they are contained in the "kinetic term" in Π divided by (see eq. 6).

In a more rigorous treatment, leading to (9) and (10), one has to bear in mind that the temperature, the Fermi level (with respect to the edges of the bands) and the concentration of the charge-carriers are all closely related, and that these quantities are also dependent on location.

2) The Seebeck effect in metals.

In metals the Seebeck effect is again very small. *Fig. 9* shows the number of electrons f per energy state as a function of the absolute value of the velocity corresponding to that state at two temperatures (one-dimensional). If there is a temperature gradient the two curves give the velocity distributions of the electrons moving to the right and to the left respectively through a given cross-section. The area C represents a surplus Δn_C of electrons with a mean velocity v_C, say to the right; D represents a surplus Δn_D of electrons with a mean velocity v_D to the left. If $\Delta n_C = \Delta n_D$ the thermal diffusion (and hence the Seebeck effect) is determined by Δn_C and $v_D - v_C$; the effect is then very small. The effect can be appreciably greater if $\Delta n_C \neq \Delta n_D$ because of a steep gradient in the density of states.

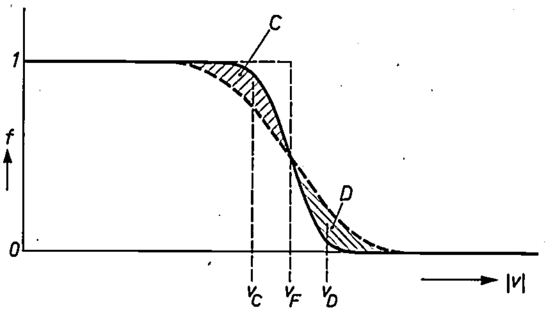


Fig. 9. The Seebeck effect in metals. The number of electrons per state (the distribution function) f is shown for two temperatures as a function of the absolute value of the velocity corresponding to that state. v_F Fermi velocity. If a temperature gradient exists, area C represents a surplus Δn_C of electrons moving to the right, and area D represents a surplus Δn_D moving to the left. v_C and v_D are the mean velocities of the electrons represented by C and D. If $\Delta n_C = \Delta n_D$, the thermal diffusion, and hence the Seebeck effect, is given by Δn_C and $(v_D - v_C)$.

Cooling power, temperature reduction, efficiency and figure of merit

It is obvious that for a good Peltier element the Peltier coefficient, and therefore in view of (5), the Seebeck coefficient, should be as high as possible. The electrical conductivity should also be as high as possible, the thermal conductivity on the other hand should be as low as possible. It will be shown that the quality of an element can be characterized by a single quantity, called the *figure of merit* $Z = a^2/RL$. Here L is the (parallel) thermal conductivity of the two arms together, and R is the electrical (series) resistance. The figure of merit Z has the dimension °C^{-1}; the value of Z in the best elements is between 2 and 3×10^{-3} °C^{-1}. Z is determined mainly by the physical properties of the semiconductors used, i.e. the Seebeck coefficient α, the electrical conductivity σ and the thermal conductivity λ; broadly speaking, Z is independent of the dimensions, since the effects on R and L compensate one another in the product RL. The precise influence of the dimensions will be discussed later.

The efficiency η is the ratio of the cooling power Q_K (the heat removed per second from the object cooled) to the electrical power consumed N: $\eta = Q_K/N$. In refrigeration engineering η is always regarded as an important characteristic. With the Z values at present attainable the Peltier element cannot yet compete with the conventional compressor-type refrigerator, for example in domestic refrigeration; to make this possible the figure of merit would have to be improved by a factor of 2 to 3. The Peltier element nevertheless has a wide field of potential applications where a high efficiency is of secondary importance. In these applications the maximum attainable reduction of temperature is often a more important characteristic.

A close relation exists between the cooling power Q_K and the temperature reduction $\Delta T = T_W - T_K$. The quantity of heat which has to be removed from the cold junction, by means of the Peltier effect, consists of the heat Q_K taken up from the cooled object and the heat $L\Delta T$ conducted from the hot junction, as well as a fraction of the Joule heat evolved. At a given temperature T_W at the hot junction and a current I, the magnitude of $Q_K + L\Delta T$ is thus established for any given Peltier element. In the following it will be shown that for every element there is a certain current at which $Q_K + L\Delta T$ is a maximum; as a rule the element will be operated at that current value. The cooling power Q_K therefore becomes smaller as the temperature difference ΔT is increased; the temperature reduction is greatest when $Q_K = 0$ (this may be obtained by thermal insulation between the refrigerated object and the environment). Conversely at smaller ΔT values a greater Q_K is available.

The dependence on the current follows from the energy balance at the cold junction. Taking into account that half of the Joule heat is removed at the cold junction (the other half is removed at the hot junction), the energy balance is given by:

$$Q_K = \Pi_K I - \tfrac{1}{2} I^2 R - L\Delta T, \quad . \; . \; . \quad (11)$$

or, with $\Pi_K = \alpha T_K$, by:

$$Q_K + L\Delta T = \alpha T_K I - \tfrac{1}{2} I^2 R. \quad . \; . \quad (12)$$

The voltage across the element (see also eq. 2) is $V = IR + \alpha\Delta T$; the electrical power is thus:

$$N = I^2 R + \alpha I \Delta T. \quad . \; . \; . \; . \; . \; . \quad (13)$$

We note that in this derivation α, σ and λ are considered to be temperature-independent between T_K and T_W. From (12) it is at once seen that there is a value of I giving a maximum for Q_K at given temperatures; for small I, the first term on the right-hand side is dominant and Q_K increases with I; at greater values of I, Q_K decreases again, because the second term now becomes dominant. The maximum value Q_M for Q_K at given ΔT, the value I_M of the current I at which it is reached, and the corresponding electrical power N_M are found from (12) and (13), with $\partial Q_K/\partial I = 0$:

$$Q_M + L\Delta T = \tfrac{1}{2} Z L T_K^2, \quad . \; . \; . \; . \quad (14)$$

$$I_M = \frac{\alpha T_K}{R}, \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad (15)$$

$$N_M = Z L T_K T_W. \quad . \; . \; . \; . \; . \; . \; . \quad (16)$$

Here Z is the figure of merit mentioned above:

$$Z = \alpha^2/LR. \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad (17)$$

Equation (14) thus gives, for a given Peltier element, the maximum cooling power at a given temperature difference, and vice versa. A graph of (14) giving a plot of Q_K versus ΔT — the cooling power diagram — gives an outline of the performance of a Peltier element as a "cold pump"; we shall deal with this presently. First, however, we shall derive from (14) an expression for the *maximum attainable temperature reduction* ΔT_0, another important characteristic of a Peltier element. With $Q_M = 0$, it follows from (14) that:

$$\Delta T_0 = T_W - T_0 = \tfrac{1}{2}ZT_0^2, \quad . \quad . \quad . \quad . \quad (18)$$

where T_0 is the lowest attainable temperature on the cold side. The value ΔT_0 of ΔT is reached at $I = I_0 = \alpha T_0/R$, when $Q_K = 0$. At a given T_W the value of ΔT_0 is evidently determined entirely by Z. ΔT_0, like Z, is therefore on the whole independent of the dimensions; the associated current value I_0 at given values of T_W, α and σ is however determined by the dimensions. If ΔT_0 is neglected in comparison with T_W, then ΔT_0 is proportional to Z. *Fig. 10* gives curves of ΔT_0 against T_W, after (18), for various practical values of Z.

The following comment should be made in connection with (18) and fig. 10. It can be seen in fig. 10 how ΔT_0 decreases as T_W decreases. Because of this Peltier cooling is not advantageous with low starting temperatures, e.g. after pre-cooling by conventional means or with the aid of a cascade arrangement. It follows from (18) that not only ΔT_0 but the relative temperature reduction $\Delta T_0/T_W$ as well becomes smaller as T_W decreases. This means that there is little prospect of being able to employ Peltier cooling in low-temperature physics. If, for example, a value of $Z = 3 \times 10^{-3}$ °C^{-1} had been achieved (assumed temperature-independent), then at $T_W = 260$ °K a temperature reduction of $\Delta T_0 = 60$ °K would be attainable, which is of technical significance. At $T_W = 10$ °K however, the temperature reduction would be merely an insignificant 0.15 °K.

The *cooling power diagram* is greatly simplified if a linear approximation of (14) is used, in which ΔT is neglected in comparison with T_W. In this approximation (see *fig. 11*), it follows from (14), using (16) and (18) that:

$$Q_M \approx \tfrac{1}{2}N_M\left(1 - \frac{\Delta T}{\Delta T_0}\right). \quad . \quad . \quad . \quad (19)$$

If we take for N_M in (19) the value for $\Delta T = 0$: $N_M = ZLT_W^2$, then (19) is a fairly good approximation even if ΔT is not particularly small compared with T_W. The error in (19) is zero for $\Delta T = 0$ and $\Delta T = \Delta T_0$; the maximum error is 8.5 to 10 % for $Z = 2$ to 3×10^{-3} °C^{-1}. This accuracy is quite sufficient for practical purposes.

From the foregoing it may be inferred that a Peltier element, at a given temperature T_W for the hot side, is fully characterized by the maximum temperature reduction ΔT_0, the current I_0 at which it is reached, and the electrical power required N_M (or the voltage V_M across the element). These quantities can be determined by means of a simple experiment in which the current is varied until the maximum temperature reduction is established, the temperature of the hot side being kept constant. Once these quantities are known, the maximum cooling power Q_M at any ΔT can be read from the diagram. The amount of heat Q_W to be removed from the hot junction, $Q_W = Q_M + N_M$, which is an important quantity in heat exchanger design, is then also known (as is also the efficiency $\eta_M = Q_M/N_M$). The relation between the quantities which is expressed in the cooling power diagram therefore makes a calorimetric measurement of Q_M and Q_W unnecessary.

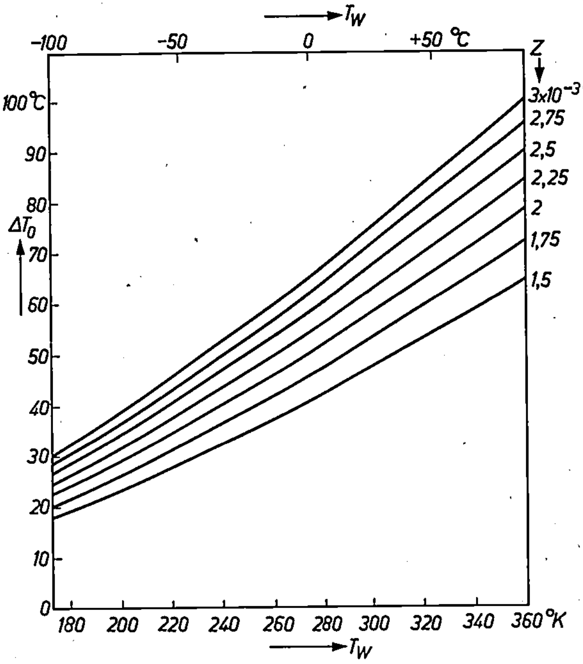


Fig. 10. Maximum temperature reduction ΔT_0 as a function of the hot junction temperature T_W, after equation (18), for various values of the figure of merit Z.

The voltage required to establish the optimum current I_M, and hence the maximum cooling power Q_M, is:

$$V_M = I_M R + \alpha\Delta T = \alpha T_K + \alpha\Delta T = \alpha T_W.$$

This voltage, for a given semiconductor combination

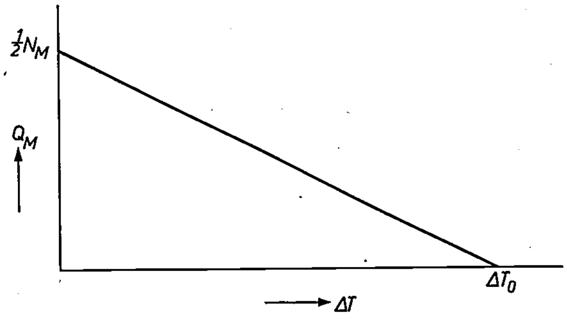


Fig. 11. Cooling power diagram after equation (19), giving the maximum cooling power Q_M as a function of the temperature reduction ΔT. ΔT_0 is the maximum temperature reduction. N_M is the electrical power consumed at $\Delta T = 0$.

(i.e. for a given α) depends only on the temperature of the hot junction T_W, and not on ΔT or on other factors, such as the figure of merit or the dimensions. Thus, for compounds derived from bismuth telluride, for example, $\alpha \approx 3.5 \times 10^{-4}$ V/°C; for $T_W = 300$ °K, V_M is then ≈ 0.105 volt.

In practice it is usual to connect a number of Peltier elements together to form a chain. A particular cooling power under particular temperature conditions may in principle be distributed over any number of elements. The optimum current is then inversely proportional to the number of elements, since a) the cooling power and therefore the electrical power per element are inversely proportional to this number, and b) the voltage over an element is however independent of the number of elements. It follows from (15) that the appropriate resistance per element, and therefore the ratio of length to cross-section of each arm, is proportional to the number of elements. The voltage over the whole chain is also proportional to this number. The choice made will depend on various technical considerations, such as the space available for the elements, the area of the hot junctions and the nature of the current source employed.

It is evident that the cooling power diagram of fig. 11 and equation (19) are also valid for chains of elements and Peltier batteries if Q_M and N_M are here taken to mean the *total* maximum cooling power and the corresponding *total* electric power.

Fig. 12 shows by way of example the cooling power diagrams for three types of Peltier battery. The curves are drawn for $T_W = 300$ °K. They can however also be used for other temperatures T_W' at the hot junction. One then only has to multiply the values given on the ordinate by $(T_W'/T_W)^2$ as (14) and (16) show with $T_K = T_W$; the value of ΔT_0 associated with T_W' can then be read from fig. 10.

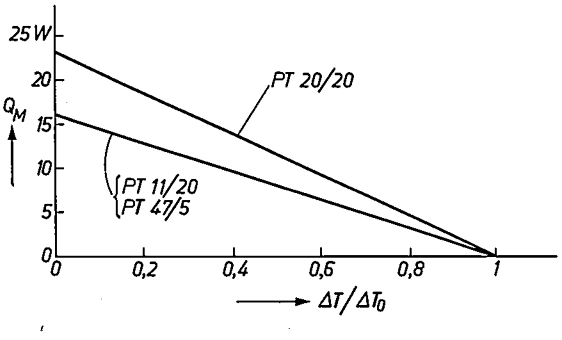


Fig. 12. Cooling power diagram for three types of Peltier battery. The diagram shows the maximum cooling power Q_M as a function of the temperature reduction ΔT divided by the maximum temperature reduction ΔT_0, at $T_W = 20$ °C and $\Delta T_0 = 50$ °C, for the types:
PT 20/20 at $I_M = 20$ A, $V_M \approx 2.0$ V,
PT 11/20 at $I_M = 20$ A, $V_M \approx 1.1$ V,
PT 47/5 at $I_M = 5.5$A, $V_M \approx 4.7$ V.

As there is relatively little demand for a high efficiency in present applications, we have so far left it out of account. Nevertheless, it is interesting to know what efficiency can be achieved, and indeed the question would become very important if, for example, materials with a better Z could be developed.

The maximum efficiency at given values of T_W and ΔT (or Q_K) is found from $\partial\eta/\partial I = 0$, where $\eta = Q_K/N$ and using (12) and (13):

$$\eta_0 = \frac{MT_K - T_W}{\Delta T(M+1)}, \quad \ldots\ldots \quad (20)$$

where

$$M = \sqrt{1 + \frac{Z}{2}(T_W + T_K)}. \quad \ldots. \quad (21)$$

At given T_W and T_K the maximum efficiency η_0 is therefore determined solely by Z, η_0 being a monotonic increasing function of Z. *Fig. 13* gives curves of η_0 as a function of ΔT for $T_W = 300$ °K and for various values of Z. It should be noted that η is not limited to values between 0 and 1, but can theoretically assume all values between 0 and ∞. If η is to be used as a measure of the "quality" of a Peltier element, in particular for comparison with other cold pumps, care should be taken for what temperature difference the comparison is made. If, for example, one takes $\Delta T = \Delta T_0$, then $Q_K = 0$ and hence $\eta = 0$; this applies however to any cold pump. Another point of fundamental importance in such comparisons is the cooling power for which the cold pump was designed. The efficiency of a Peltier element does not depend on this whereas the efficiency of a conventional compressor-type refrigerator drops steeply if the rated cooling power is reduced. An economic comparison can therefore only be made on the basis of detailed technical data for the cold pumps in question.

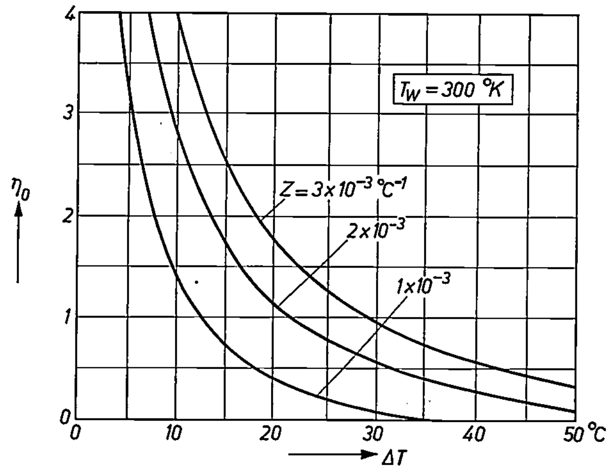


Fig. 13. Maximum efficiency η_0 as a function of the temperature reduction ΔT for various values of the figure of merit Z. The hot junction temperature here is $T_W = 300$ °K.

Choice and development of thermoelectric materials

Figure of merit of the material

Although the figure of merit $Z = \alpha^2/RL$ (eq. 17) is chiefly determined by the thermoelectric properties of the materials used, it still depends to some extent on the dimensions of the Peltier element. By appropriate choice of these dimensions, which affect the value of Z through the denominator RL, a particular material can be given a maximum Z. As the two arms of the element are electrically in series, but in parallel for heat conduction we have:

$$RL = \left(\frac{l_N}{\sigma_N q_N} + \frac{l_P}{\sigma_P q_P}\right)\left(\frac{\lambda_N q_N}{l_N} + \frac{\lambda_P q_P}{l_P}\right),$$

where l_N, l_P are the lengths, q_N, q_P the cross-sections, σ_N, σ_P the electrical conductivities and λ_N, λ_P the thermal conductivities of the N-type and P-type materials. The geometric quantities now appear only in the combination:

$$g = \frac{l_P}{l_N}\,\frac{q_N}{q_P},$$

so that

$$RL = \frac{\lambda_N}{\sigma_N} + \frac{\lambda_P}{\sigma_P} + g\frac{\lambda_N}{\sigma_P} + \frac{1}{g}\,\frac{\lambda_P}{\sigma_N}. \quad . \; . \; (22)$$

This expression has a minimum at:

$$g = \sqrt{\frac{\sigma_P}{\sigma_N}\,\frac{\lambda_P}{\lambda_N}},$$

and the corresponding maximum value of Z is:

$$z = \left[\frac{\alpha_P - \alpha_N}{\sqrt{\lambda_N/\sigma_N} + \sqrt{\lambda_P/\sigma_P}}\right]^2. \quad . \; . \; . \; (23)$$

We have therefore obtained, as a criterion for the performance to be obtained from a semiconductor combination, a figure of merit determined solely by the physical properties of the semiconductors. This figure of merit, which had already been found in principle by Altenkirch [2], is the criterion in most general use, although others are also employed.

As a measure of the thermoelectric quality of a single semiconductor another figure of merit can be used, viz:

$$z_N = \alpha_N{}^2\sigma_N/\lambda_N \text{ or } z_P = \alpha_P{}^2\sigma_P/\lambda_P \quad . \; . \; (24)$$

for N-type and P-type materials respectively. Although an unambiguous figure of merit for a complete element cannot be calculated from (24) and (23), the use of (24) is justified in practice since almost invariably the absolute values of α and σ/λ in N-type materials are approximately the same as those in P-type materials, while α_N and α_P are of opposite sign. It then follows from (23) that: $z \approx z_N \approx z_P$.

In equation (22) the ohmic contact resistances of the element should really be taken into account as well. In practice, however, it is found that in quantity-produced Peltier elements, the contact resistance has a negligible effect. The fact that for production reasons the dimensions of the element are not entirely optimum, is also of no significance in practice. In all practical cases, then, equation (23) gives the effective figure of merit for the complete Peltier element and this is determined solely by the thermoelectric properties of the material.

Semiconductor properties and figure of merit

When looking for suitable materials for thermoelectric applications the initial requirement will be a high figure of merit $z = \alpha^2\sigma/\lambda$. It is not possible to make a reliable choice on purely theoretical grounds, but the following considerations may serve as a guide. In practice the choice is still to a great extent an empirical one.

The charge-carrier concentration n has an important bearing on the estimated figure of merit for a given material. The situation is represented rather schematically in *fig. 14*, in which α, σ, $\alpha^2\sigma$ and λ are shown as functions of n. As n increases, α decreases (because the material becomes increasingly degenerate, see page 116), σ increases ($\sigma = ne\mu$, where μ is the mobility of the charge-carriers) and $\alpha^2\sigma$ has a maximum, which is in the region of $n = 10^{19}$ cm^{-3} for all semiconductors. The thermal conductivity is the sum of the lattice and the electron thermal conductivities: $\lambda = \lambda_g + \lambda_e$. Almost always in semiconductors $\lambda_e \ll \lambda_g$, so that as a first approximation one may neglect λ_e in comparison with λ_g. Since λ_g is independent of n, the value of z also has a maximum at $n \approx 10^{19}$ cm^{-3}.

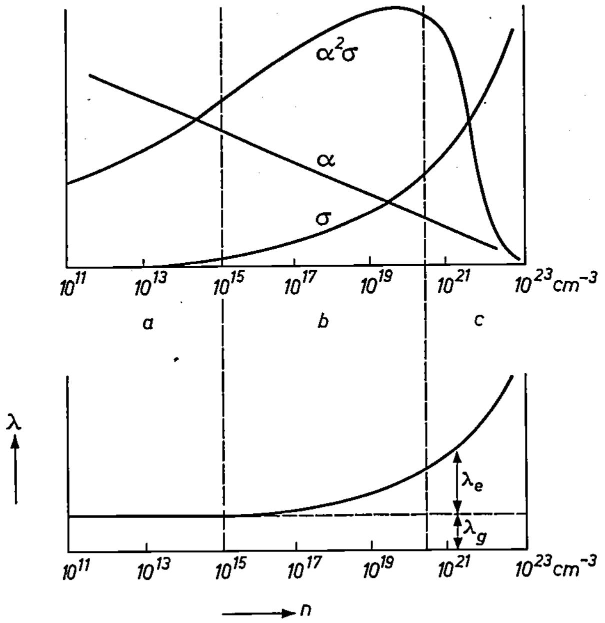


Fig. 14. The quantities α, σ, λ and $\alpha^2\sigma$ as a function of the charge-carrier concentration n, (this is a very schematic representation). a insulators; b semiconductors; c metals. λ_g and λ_e are the respective contributions to λ due to lattice and electronic thermal conduction.

In metals λ_e is dominant. Because of this σ/λ is constant (Wiedemann-Franz law), and therefore z is proportional to α^2. Because of the relatively low values of α in metals, the values of z that can be attained are also very low. The failure to find materials suitable for thermoelectric applications, until a few years ago, was due to the fact that research up till then had been confined almost exclusively to metals.

The reasoning that leads to a maximum of z for semiconductors at $n = 10^{19}$ cm^{-3} is not entirely correct. While it is true that the optimum concentration of 10^{19} cm^{-3} is reached in the semiconductor region, this concentration is so high that the electron gas in the semiconductor becomes strongly degenerate. Equations (9) and (10) are therefore no longer valid, while at the same time the mean mobility μ of the charge-carriers becomes dependent on their concentration. When this is taken into account, it is even found that in some cases $a^2\sigma$ no longer has a maximum. By what appears, on the face of it, to be a coincidence the electronic contribution to the thermal conduction λ_e then ensures that there is still a maximum for $a^2\sigma/\lambda$ [3] and once again the maximum is at $n \approx 10^{19}$ cm^{-3}.

It is in two respects a fortunate circumstance that the optimum value of n lies in the region of 10^{19} cm^{-3}. Firstly this value lies in the semiconductor region — the very region in which the concentration of charge-carriers can be adjusted by doping. The exact doping required for optimum results has to be determined by experiment. Secondly, the doping of about 10^{19} cm^{-3} which is required is already so high that there is no point in specifying a chemical purity greater than about $10^{19}/10^{23} = 10^{-2}\,\%$ for the starting materials (assuming approximately 10^{23} atoms or molecules per cm^3). Commercial grades of starting material usually meet this requirement, so that it is not necessary — as with germanium and silicon for transistors, where the concentration has to be about 10^{16} cm^{-3} — to reduce the chemical impurities to about $10^{-5}\,\%$ by zone refining.

There still remains the question as to which semiconductors have the highest figure of merit at optimum doping. Let a_{opt} and σ_{opt} be the values of a and σ at the optimum charge-carrier concentration n_{opt}. A somewhat simplified theory shows that a_{opt} is independent of the particular semiconductor properties: $a_{\text{opt}} \approx 2k/e$ ($= 172$ µV/°C); in practice $a_{\text{opt}} \approx 200$ to 300 µV/°C. For σ_{opt} we have $\sigma_{\text{opt}} = n_{\text{opt}}e\mu$. The optimum concentration n_{opt} is itself dependent on the structure of the relevant energy band, which is related to the effective mass of the charge-carriers m^*; theoretically, $n_{\text{opt}} \propto m^{*3/2}$. From (24) it now follows that the maximum figure of merit is proportional to $\mu m^{*3/2}/\lambda_g$, and this is therefore also a kind of figure of merit. The various quantities involved here are interrelated in a complicated way. For example, μ/λ_g shows some tendency to increase with increasing atomic weight, which means that semiconductors of high average atomic weight are generally more suitable than lighter compounds. The compounds derived from bismuth telluride, which have proved so far to be the best thermoelectric materials, do indeed have a high average atomic weight. On the other hand, it is possible for instance to lower the lattice thermal conductivity λ_g by making *mixed crystals*, the lattice waves being strongly scattered if the crystal lattice consists of irregularly distributed atoms of widely varying mass. When this is done the electronic structure should change as little as possible, and in particular the mobility of the charge-carriers should not be reduced. The mixed crystals must therefore be alloys of chemically closely related compounds with analogous valencies, whose constituent elements have atomic weights which are as different as possible. The lattice thermal conductivity is not the only factor determining the optimum ratio of the mixed crystal. The effective mass m^* is also affected by this ratio; in bismuth telluride and its associated isomorphous compounds, it has a maximum at a different ratio from the one at which the lattice thermal conductivity is a minimum. The best ratio has to be found by experiment.

Preparation and properties of semiconductors derived from bismuth telluride

The semiconductor materials which have given the highest figures of merit and which are usually used for Peltier devices, are bismuth telluride Bi_2Te_3 and the mixed crystals of this and the isomorphous compounds bismuth selenide Bi_2Se_3 and antimony telluride Sb_2Te_3. These crystals are rhombohedral. The mixed crystals with Bi_2Se_3, which are doped with say I or Br, are used as *N*-type material, while the mixed crystals with Sb_2Te_3 (doped for example with Pb) constitute the *P*-type material. The crystal structure can be regarded as a hexagonal laminated structure, in which the layers, each consisting of one type of atom, are arranged in the sequence ... Te$^{[I]}$ — |Te$^{[I]}$ — Bi — Te$^{[II]}$ — Bi — Te$^{[I]}$ — |Te$^{[I]}$ ... These crystals are easily cleaved, probably owing to the weakness of the bond between the Te$^{[I]}$ layers. The plane of cleavage is perpendicular to the c axis.

The semiconducting material is generally prepared by one of the following two methods. a) The melting method, in which the compounds are melted together and the melt is frozen in a particular direction by the "normal freezing" or Bridgman method, to avoid cracks and blowholes and to produce a good single crystal. b) The sintering method, in which the previously melted compounds are finely ground, sieved, compressed and finally sintered. In the following, for brevity we shall refer to *drawn* or to *sintered* material.

Because the crystals are easily cleaved, the drawn material is brittle and therefore difficult to handle. Sintered material which has greater mechanical strength, is easier to handle but its figure of merit is somewhat lower. This is connected with the anisotropy of the thermoelectric properties, as a result of the

[2] E. Altenkirch, Phys. Z. **10**, 560, 1909 and **12**, 920, 1911.

[3] J. D. Wasscher, W. Albers and C. Haas, Solid-State Electronics **6**, 261, 1963.

laminated structure. While there is scarcely any difference between the thermo-e.m.f.'s parallel (||) to the cleavage plane and perpendicular to it (⊥), the electrical conductivity and the thermal conductivity parallel to the cleavage plane are greater than those perpendicular to it. It is found experimentally that:

$$(\sigma_{||}/\sigma_{\perp})_{N,P} > (\lambda_{||}/\lambda_{\perp})_{N,P}.$$

The ratio σ/λ in the figure of merit thus gives:

$$(\sigma_{||}/\lambda_{||})_{N,P} > (\sigma_{\perp}/\lambda_{\perp})_{N,P}, \text{ so that } (z_{||})_{N,P} > (z_{\perp})_{N,P}.$$

Peltier elements of drawn material in which the cleavage planes are oriented parallel to the direction of the current therefore have a better figure of merit than elements of sintered material in which the crystallites are randomly oriented but evenly distributed. *Fig. 15* gives the measured maximum temperature reduction ΔT_0 as a function of the hot junction temperature T_W for Peltier elements of both drawn and sintered material. By means of the curves for constant figure of merit (derived from eq. 18) one can read from fig. 15 the figure of merit that occurs in practice for a particular starting temperature. The temperature-dependence of the thermoelectric properties is then also taken into account, but in any case this has relatively little influence on the figure of merit in the temperature interval measured. In connection with these measurements it should be noted that extra heat conduction losses have been avoided by means of evacuation, and heat radiation losses have been avoided as well, by cooling the measuring equipment down to a few °C above the temperature of the cold junction of the Peltier element.

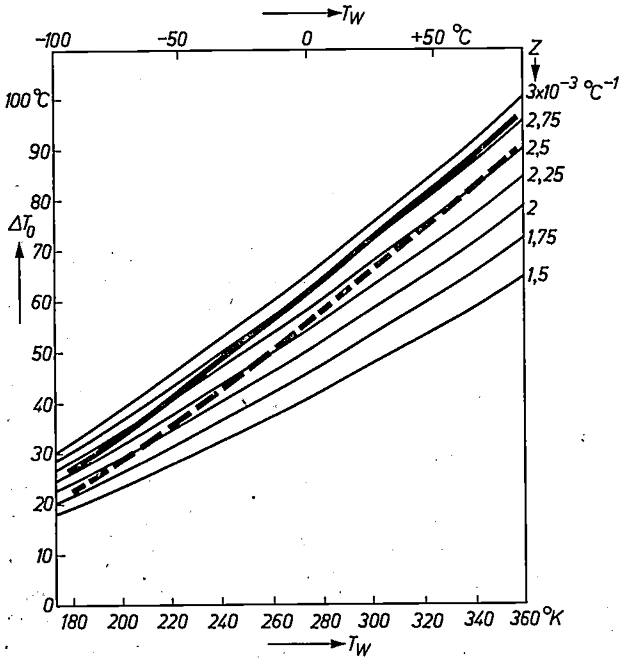


Fig. 15. Maximum temperature reduction ΔT_0 as a function of the hot junction temperature T_W, for Peltier elements made from semiconducting materials derived from bismuth telluride. The solid line relates to drawn material, the broken line to sintered material. The curves for constant figure of merit, from fig. 10, are given for comparison.

For semiconductors derived from bismuth telluride, neither the optimum mixed-crystal ratio nor the optimum doping is particularly critical. This probably accounts for the somewhat surprising fact that the figures of merit of Peltier elements made in *production quantities* by various manufacturers show very few differences (about 1%) at all values of T_W between about —90 and +100 °C, in spite of the undoubtedly different doping and mixed-crystal ratios that were employed. The manufacturers concerned have obviously managed to obtain about the maximum figure of merit from their materials. These remarks apply to elements made of drawn material. The same can be said of elements made from sintered material, but — as mentioned — the figures of merit are slightly lower. The curves in fig. 15 are typical for these figures of merit, which therefore could be taken as standard figures of merit for Peltier devices derived from bismuth telluride. At $T_W \approx 300$ °K (room temperature) the figure is 2.8×10^{-3} °C^{-1} for elements of drawn material, and 2.4×10^{-3} °C^{-1} for sintered material. The difference in the figures of merit of sintered and drawn material is of little practical significance. We have been discussing the figure of merit of the complete Peltier element and not that of the individual materials derived from bismuth telluride. For these materials peak values are often quoted (e.g. 2.8×10^{-3} °C^{-1} for *N*-type material and 3.5×10^{-3} °C^{-1} for *P*-type material) which, in principle should make possible higher figures of merit for the complete element. The curves in fig. 15 should not be extrapolated, for it is quite possible that the figures of merit may drop considerably just outside the region indicated.

If Peltier elements or chains of elements are made into complete Peltier batteries with cover plates and plastic cases etc. (or if the elements are encapsulated in a plastic material) additional heat conduction losses will occur which can reduce the "effective figure of merit" of the battery by as much as 20%.

Power supply

The d.c. current supplied to a Peltier battery often contains a ripple which can lower the performance of the battery. With increasing ripple the ratio of the mean value of the current to the r.m.s. value is smaller; the mean value determines the Peltier effect, the r.m.s. value determines the Joule heat. The drop in performance due to this is often overestimated, and it is assumed that the ripple current must be kept to a very low level. Calculations on the effect of the ripple (which can easily be verified by measurement) show however that a ripple in the supply current causes only a slight decline in the maximum refrigerating capacity or the maximum temperature reduction. When there is ripple in the supply current the figure of merit is *apparently* reduced by a factor $1/(1+w)^2$, where w is the ripple factor (the r.m.s. value of the ripple divided by the mean total current). The effect of the ripple on, for example, the maximum attainable temperature reduction ΔT_0 can be calculated with the aid of (18); *fig. 16* shows the relative decrease of ΔT_0 as a function of w. Since the decrease is almost independent of Z, it is given only for $Z = 2.5\times10^{-3}$ °C^{-1}.

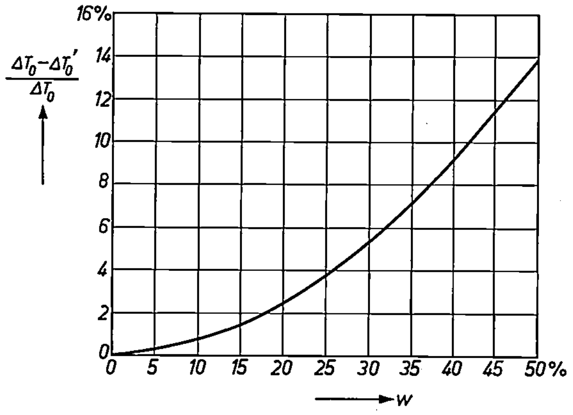


Fig. 16. Relative decrease $(\Delta T_0-\Delta T_0')/\Delta T_0$ of the maximum temperature reduction as a function of the ripple factor w of the supply current for $Z = 2.5\times10^{-3}$ °C^{-1}. $\Delta T_0'$ and ΔT_0 are the maximum temperature reductions for ripple factors of w and 0.

As an example let us take a d.c. current with a superimposed sine wave (see *fig. 17a*). The ripple factor is $w = I_{\sim}/I_{\mathrm{m}}\sqrt{2}$. If the ratio of $I_{\sim}$ to I_{m} is about $\frac{1}{4}$, i.e. w is in the region of 18%, then the decrease in the maximum temperature reduction is a mere 2%. As a further example let us consider a rectified alternating current as supplied by an ordinary rectifier circuit without smoothing filters. The ripple factor of this current (see fig. 17*b*), which may consist of any number of sine pulses, is expressed by the number of current pulses p per cycle:

$$w = \frac{\pi}{p\sin\pi/p}\sqrt{\tfrac{1}{2}+\frac{p}{4\pi}\sin\frac{2\pi}{p}-\left(\frac{p}{\pi}\right)^2\sin^2\frac{\pi}{p}}\,. \qquad (25)$$

The values of w calculated from this expression are listed below for some commonly used rectifier circuits.

p	w in %	type of rectifier circuit
2	48.2	single-phase full wave, bridge circuit
3	18.3	star circuit
4	9.8	—
6	4.2	three-phase full wave
8	2.4	—
12	1.0	—

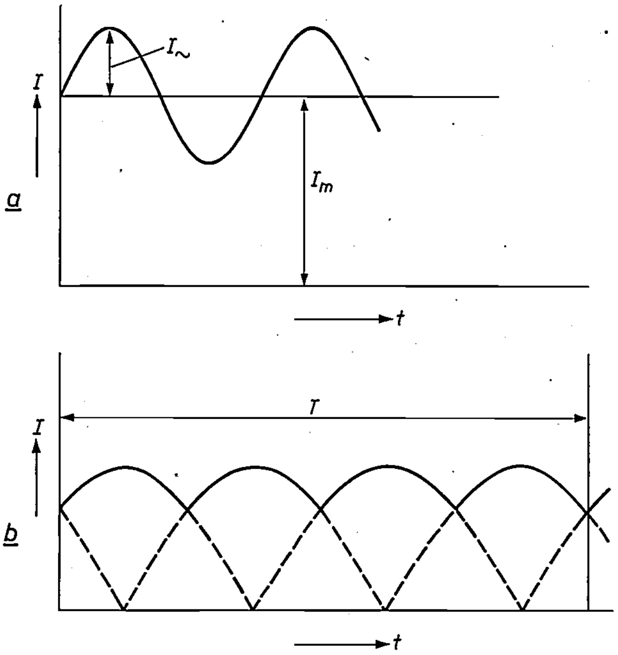


Fig. 17. Examples of a d.c. current with superimposed ripple. *a*) Sinusoidally modulated d.c. current. I_{m} average current. $I_{\sim}$ amplitude of sine wave. *b*) Rectified alternating current consisting of four sine pulses (p = 4).

From these values and fig. 16, it can be seen that, even with a simple rectifying circuit without elaborate smoothing filters, a good approximation to the performance achieved with a pure d.c. current can be obtained.

Practical Peltier batteries: design and applications

Peltier elements are well suited for cooling small objects which do not require a high cooling power, and for stabilizing the temperature of such objects. Typical examples are specimens or samples in physical chemistry or medical laboratories, where the quantities of material used are generally small or the measuring equipment contains temperature-sensitive components. *Fig. 18* illustrates some simple applications, intended as practical aids in laboratory work.

For this kind of application any number of Peltier elements can be joined together to form a battery with any required cooling power. The shape of the battery is also readily adaptable to the object to be cooled or to a particular arrangement (e.g. as a plug-in unit).

A great advantage of this purely electrical method of refrigeration compared with the use of say, liquid baths or compressor-type refrigerators is the facility of continuous and accurate control by means of the supply current. This enables exceptional temperature constancy to be achieved. The object can also be made to undergo rapid temperature variations or passed through a prescribed temperature cycle, since the ele-

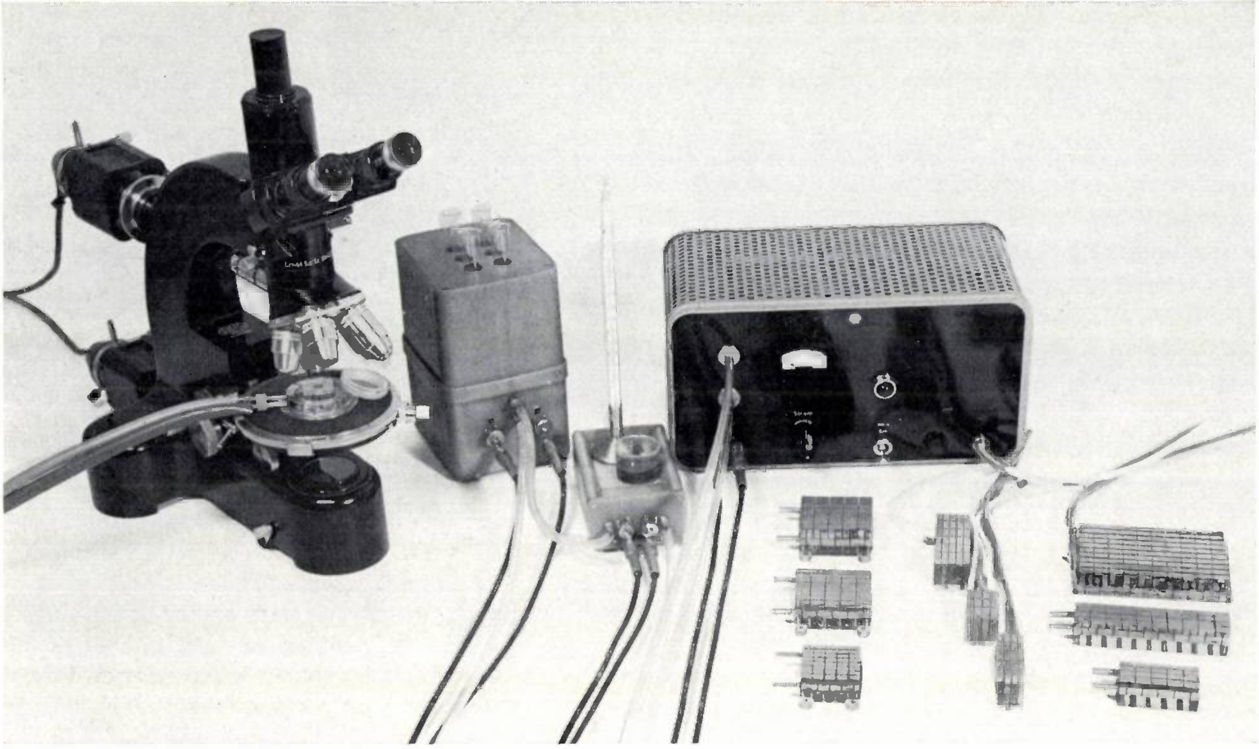


Fig. 18. Peltier cooling devices for laboratory use. The photograph shows a specimen table with thermometer, a test-tube cooler and a cooling and heating device mounted in a microscope, and an adjustable d.c. source. The partly liquid-cooled Peltier batteries in the right foreground are laboratory models of various sizes, consisting of 12, 20, 30, 40 or 100 elements.

ment supplies heat to the object when the current is reversed.

Fig. 19 shows the Peltier batteries types PT 20/20 and PT 11/20, which operate at a current of 20 A and a voltage of about 2 or 1.1 V respectively (the first figure in the type number gives the number of elements). To avoid the difficulties associated with the supply of such large currents, another type was developed, PT 47/5, which is designed for 5.5 A, 4.7 V, and has the same cooling power as the PT 11/20.

The heat generated in Peltier batteries can best be removed by liquid cooling, and the types contained in plastic casings are provided with a cooling jacket. Tapped holes are provided for fixing the battery to the object to be cooled.

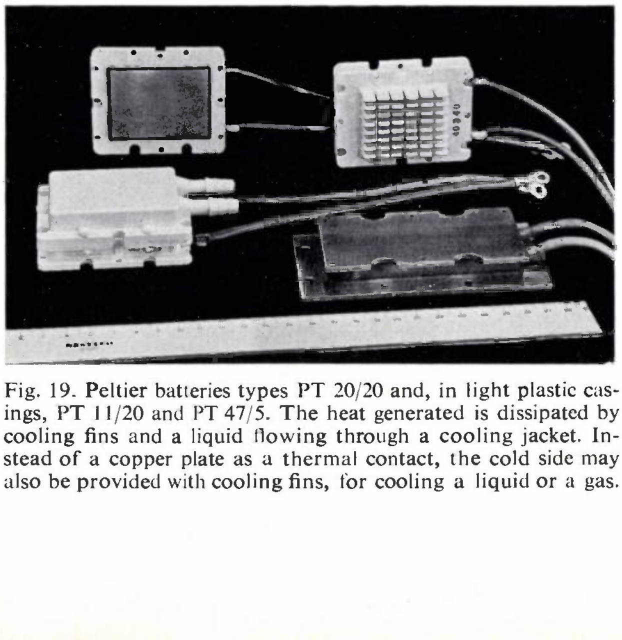


Fig. 19. Peltier batteries types PT 20/20 and, in light plastic casings, PT 11/20 and PT 47/5. The heat generated is dissipated by cooling fins and a liquid flowing through a cooling jacket. Instead of a copper plate as a thermal contact, the cold side may also be provided with cooling fins, for cooling a liquid or a gas.

Fig. 20 shows an example of an *air-cooled* Peltier

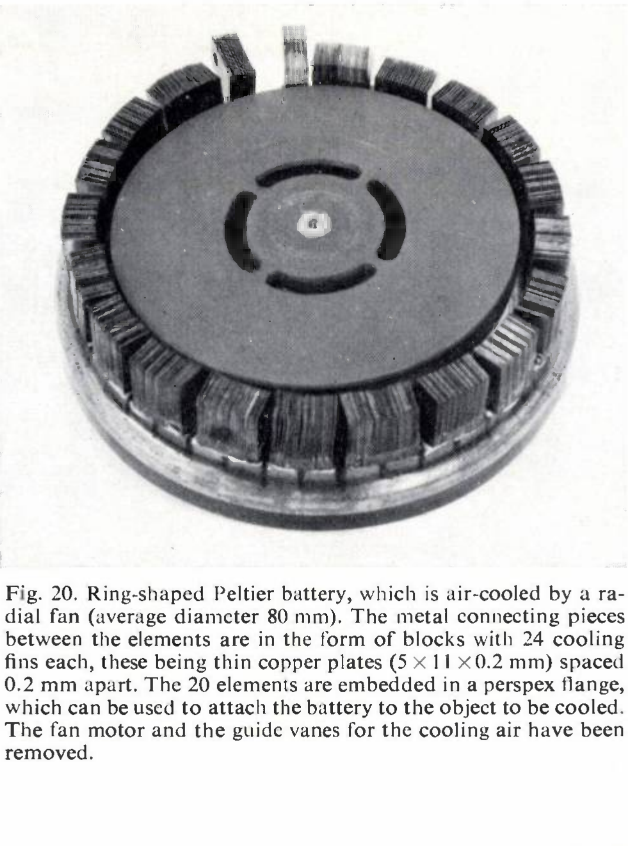


Fig. 20. Ring-shaped Peltier battery, which is air-cooled by a radial fan (average diameter 80 mm). The metal connecting pieces between the elements are in the form of blocks with 24 cooling fins each, these being thin copper plates (5 × 11 × 0.2 mm) spaced 0.2 mm apart. The 20 elements are embedded in a perspex flange, which can be used to attach the battery to the object to be cooled. The fan motor and the guide vanes for the cooling air have been removed.

battery: the unusual shape is chosen to give more effective cooling. Tominimize the heatconduction path each metal connectio n piece at the hot junction is in the form of a block with 24 cooling fins consisting of thin copper plates of 5×11×0.2 mm, spaced 0.2 mm apart. A single radial fan can be used to cool a fairly large number of elements arranged round it in a circle; the cooling air passes through one element at a time so that no element receives air that has already been heated. In this design the heat generated in a 20-element ring-shaped battery (this may be as much as 45 to 60 watts, depending on the temperature difference between the hot and cold junctions), can be dissipated without the temperature of the hot junction rising more than 3 to 6 °C relative to the ambient air temperature.

A few simple small thermostats can be seen in fig. 18, these are satisfactory if there is no need for a very constant temperature. One of the thermostats is a test-tube cooler, consisting of an aluminium block with holes for the test-tubes; the block has several Peltier batteries attached to it. The other thermostat is a cooling table for beakers and other laboratory glass-ware. With these devices it is possible to maintain temperatures of —40 °C and above (depending on the temperature of the cooling water) constant to within 0.1 °C; control is effected with the aid of a thermometer in one of the holes of the test-tube cooler or in a separate metal foot on the cooling table.

Technical requirements sometimes make it necessary to cool the object under investigation with a current of air or gas; it may be necessary to prevent contact between the object and the heat-transferring components. For instance, in the measurement of the temperature coefficient of the capacitance of miniature ceramic capacitors there would be errors due to stray capacitances if the capacitors were directly fixed to the metal at the hot junction. The cooling wind tunnel shown in *fig. 21* was designed for this special purpose (dimensions 240×150×67 mm). It is equipped with four water-cooled Peltier batteries (laboratory models) of 40 elements each; in this device the metal connections at the *cold junction* have cooling fins for heat exchange with the air-stream. This apparatus, which has a cooling power corresponding to that of about 7 batteries of the PT 20/20 type, can refrigerate an object to —35 °C at an ambient temperature of 20 °C, and with cooling water at 12 °C.

An example of an application with stricter requirements for temperature constancy is a cooling device developed in this laboratory by H. Polnitzky [4]. This device (shown in *fig. 22*) stabilizes the reference temperature of thermocouples at freezing point (0°C). With the aid of several Peltier elements, a layer of water about 1 mm thick is cooled until it freezes. The layer has a much lower resistance in the liquid state than in the frozen state. The change in resistance at 0 °C controls the Peltier current in such a way that a metal plate is kept at 0 °C with an accuracy of ±0.01 °C. Holes can be drilled in this plate to take a number of thermojunctions.

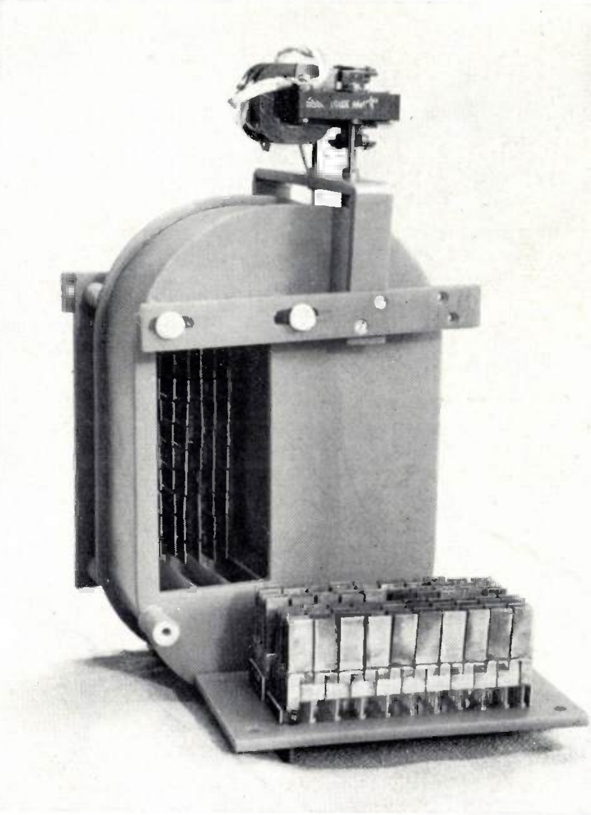


Fig. 21. Cooling wind-tunnel. Dimensions 240×150×67 mm. Circulating air is cooled (left) by four Peltier batteries (laboratory models) and flows to the right along the objects to be cooled. The four batteries are distributed over two plug-in units. One unit has been detached and its thermal insulation removed. The fan motor for the air circulation can be seen at the top. The objects to be cooled are introduced through an opening on the right, which is not visible here.

A thermostat of very high precision has been built and tested by F. Becker [5]. In this device the liquid in a bath to be kept at constant temperature is stirred with a centrifugal pump. A Peltier battery of the PT 20/20 type, whose capacity is continuously controlled by an optical control thermometer [6], is responsible for the supply and removal of heat. With this technique the temperature in a 10 litre thermostat can be kept constant to within ±0.001 °C at 25 °C.

[4] H. Polnitzky, Eispunkt-Thermostat, Z. Instrumentenkunde **73**, 11-13, 1965 (No. 1).

[5] F. Becker and W. Walisch, Präzisionsthermostat mit optischem Regelthermometer und gesteuerter Peltier-Kühlung, Z. phys. Chem. NF **34**, 369-378, 1962.

[6] W. Walisch and F. Becker, Kontinuierliche Temperaturregelung mit Hilfe eines optisch gesteuerten Regelthermometers, Z. phys. Chem. NF **29**, 371-376, 1961.

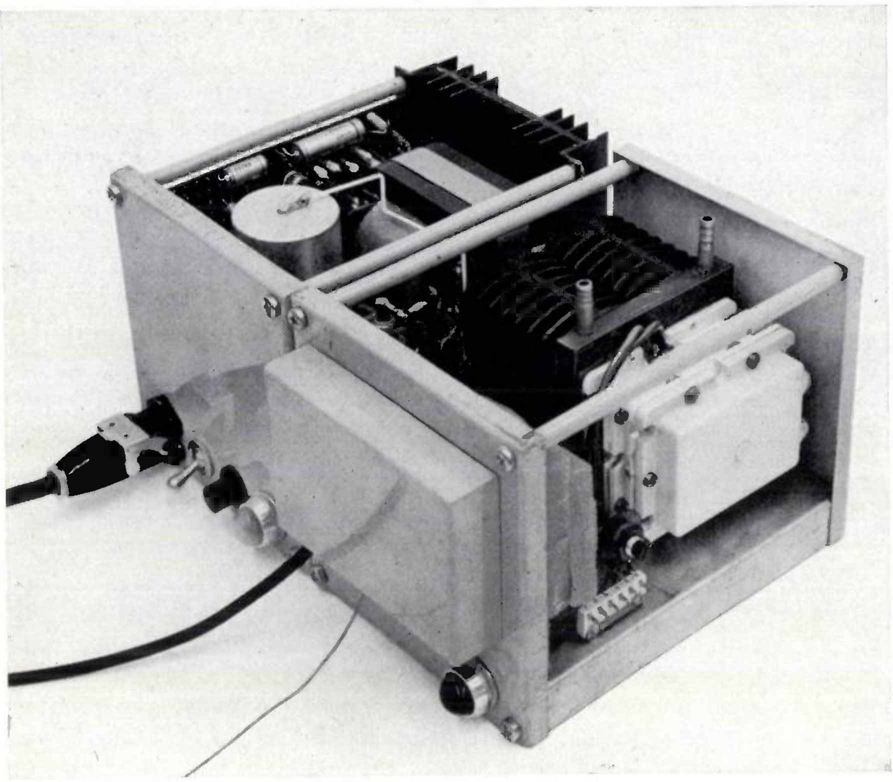


Fig. 22. Example of a freezing-point thermostat after Polnitzky [4]. The cover has been removed.

Fig. 23 shows a cooling and heating device for microscopic specimens, in which the application of Peltier cooling is particularly suitable. With this instrument the cooling or heating rate can be very accurately controlled, which is very important for example in the making of a crystal film, where the cooling rate is critical. An opening is provided to admit light, permitting all microscopic methods of examination with transmitted and incident light and polarized light. Convection and misting-up of the glasses can be prevented by fitting a small cover. The working temperature of this device, which is comparable in capacity with the PT 11/20 type, is from about —40 to +100 °C at the specimen table. The temperature can best be measured with a thermocouple soldered to a metal plate and placed as close as possible to the specimen on the table. The calibration is carried out with substances of known melting point. Reference may be made to the relevant literature [7] for the many possible applications in microchemistry, pharmacology, mineralogy and biochemistry, e.g. in the determination of melting points and in the thermo-microanalysis of organic substances.

The use of Peltier elements for cooling the mirror of a dew-point hygrometer may be described as an almost classical application. As a further development of a device described some years ago in this journal [8], E. Andrich [9] of this laboratory has designed a transistor-operated dew point hygrometer which is independent of the mains voltage. This is a simple device in which the misting of the mirror is detected by a cadmium sulphide photoconductive cell which responds when the reflected light becomes diffuse. The dew-point is continuously measured, in so far as the mirror mists up and clears again 100 times a minute. The dew-point is determined to within 0.1 °C; the measuring range is from about —30 to +50 °C. Possible applications of this device include the monitoring of the humidity of gases or vacua, for example in chemical engineering processes.

For cooling problems where the required cooling power has to be evenly distributed at pre-determined temperatures over a large volume, Peltier cooling again offers the simplest solution. The possibility of arranging the elements in any way required is made use of, for example, in electrophoresis. In this technique small particles suspended in a non-conducting liquid are passed through an electric field between two electrodes; particles of different charge or size receive different deflections and are thus separated. The current of charged particles is often accompanied by the evolution of considerable Joule heat (sometimes a few hundred watts), and convection due to this would upset the separation process. This heat therefore has to be compensated by

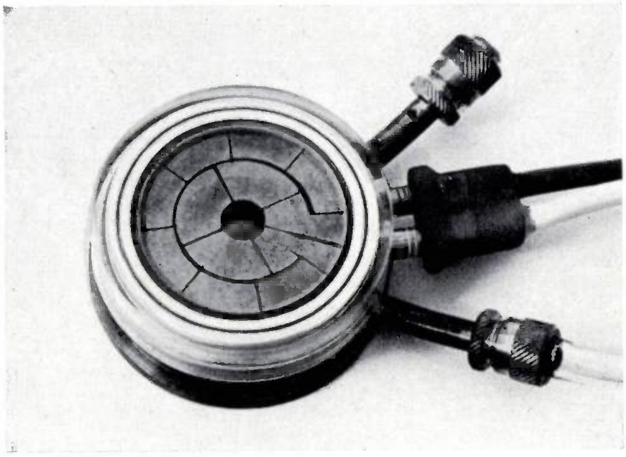


Fig. 23. Cooling and heating device for a microscope. It contains twelve elements and is water-cooled.

cooling. W. Grassmann and K. Hannig[10] have devised a very effective method for the preparative separation of proteins, which is now known as "continuous" or "deflection" electrophoresis. An apparatus that works on this principle is shown in *fig. 24*. This shows the back of the apparatus (the Peltier-cooled side). In a flat separation chamber (not visible) which extends over the front, a fine stream of the mixture to be separated is injected into a dispersion agent that flows continuously through the separation chamber. Two electrodes on the sides of the apparatus set up an electric field perpendicular to the direction of flow, and this field deflects the electrically charged particles at a certain angle from the original direction. The various separated substances are then recovered through different outlets. Measurements of the temperature distribution in the separation chamber (the temperature was constant to within 0.5 °C) proved the superiority of Peltier cooling to other previously tried methods, which showed excessive local deviations. With the more uniform temperatures, and with the more accurate adjustment of the temperature to the particular separation problem in hand, a sharper separation was achieved.

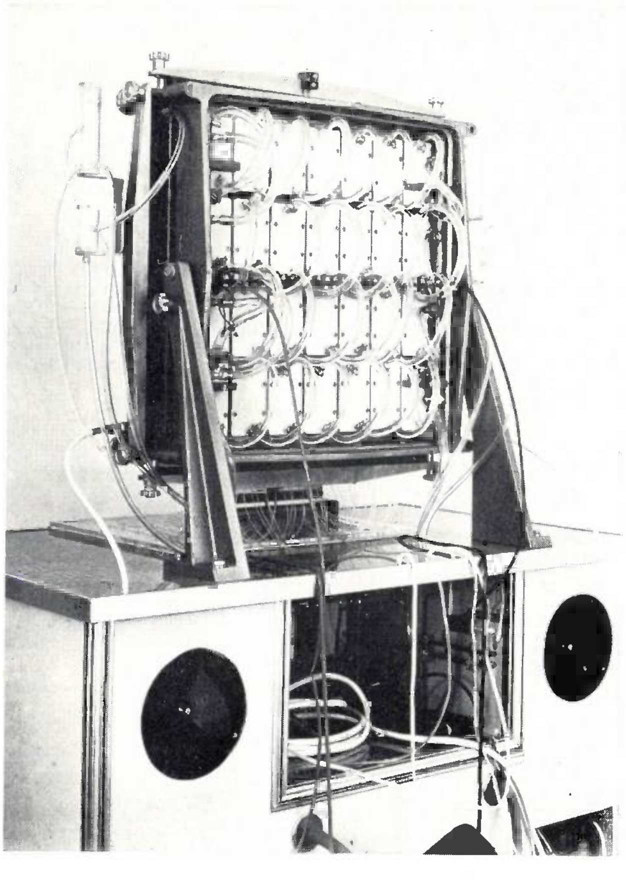


Fig. 24. Electrophoresis apparatus, after Hannig [10], for protein separation. Temperature stabilization is accomplished by means of 28 type PT 20/20 Peltier batteries.

An example of an industrial use of Peltier cooling is illustrated in *fig. 25*. In the quantity-production of television picture tubes the glass stump of the cone is snapped off automatically before fitting the neck. This is done by first heating a zone at the bottom of the rotating cone and then bringing it into contact with a knife which is cooled to —5 °C or lower. This produces a clean break, and the neck (the lower tube in the photograph) can be fitted to it straight away.

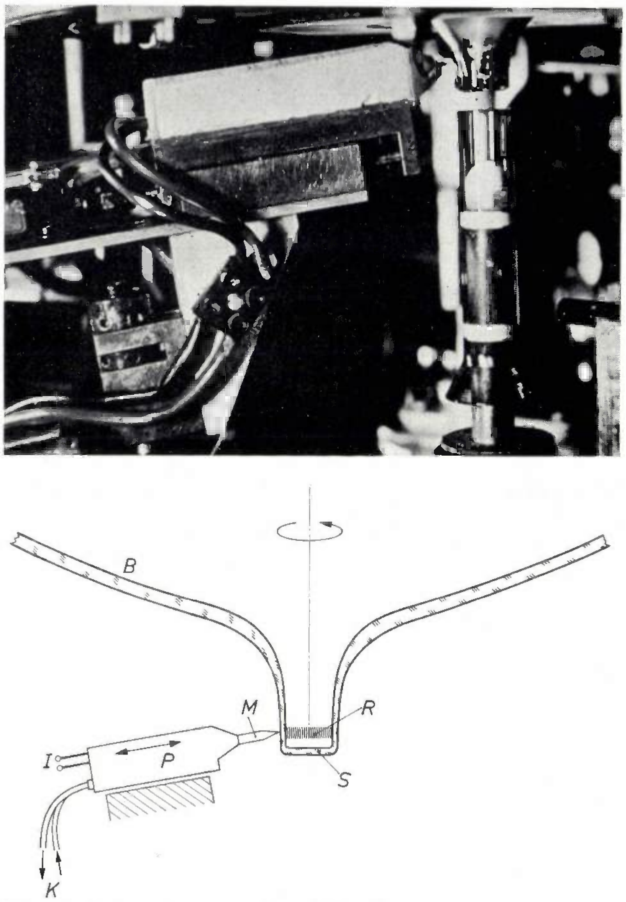


Fig. 25. Automatic separation of the glass stump of the cone in the manufacture of television picture tubes. In the sketch, *B* is the cone, and *S* the glass stump. The knife *M* is cooled to between —6 and —8 °C with a Peltier battery *P* (type PT 20/20), and is brought into contact with the annular zone *R* after this has been heated. *I* current connection. *K* cooling water.

The foregoing examples have given a brief survey of the range of applications of Peltier cooling. These applications are indicated by a number of advantageous features of the Peltier element; it is simple in design,

[7] See e.g.: A. Kofler and L. Kofler, Thermo-Mikro-Methoden zur Kennzeichnung organischer Stoffe und Stoffgemische, Verlag Chemie, Weinheim 1954.

[8] P. Gerthsen, J. A. A. Gilsing and M. van Tol, An automatic dew-point hygrometer using Peltier cooling, Philips tech. Rev. **21**, 196-200, 1959/60.

[9] Not published.

[10] W. Grassmann and K. Hannig, Präparative Elektrophorese, in: Houben-Weyl, Methoden der organischen Chemie, 4th edition, Vol. I/1, 681-751, 1958.
K. Hannig, Eine Neuentwicklung der trägerfreien Ablenkungs-elektrophorese und ihre Anwendung auf zytologische Probleme, Habilitationsschrift (Thesis), Munich, June 1964.

adaptable, reliable, easy to regulate and to replace, and requires no maintenance. It is worthy of mention that life-tests, some involving more severe requirements, have already shown that Peltier batteries can be expected to operate for years without any deterioration in performance. For all applications where, apart from the cooling power, special features such as those mentioned above are required — and particularly where the temperature has to be controlled with high precision — Peltier cooling will undoubtedly assume an important position alongside other methods of refrigeration.

Literature

A number of review articles and books selected from the numerous publications that have appeared on this subject are mentioned below. These deal with the physical problems involved in the development and preparation of the materials, and with the methods of measurement employed; applications are also described, and the design of devices is discussed.

G. Lautz, Über die technische Ausnutzung der Thermokraft von Fastmetallen und Halbleitern, Halbleiterprobleme (Vieweg, Brunswick) 4, 145-189, 1958.

P. Gerthsen, Thermoelektrische Anwendung von Halbleitern, Z. angew. Phys. **13**, 435-444, 1961.

A. F. Ioffe, Semiconductor thermoelements and thermoelectric cooling, Infosearch, London 1957.

P. H. Egli, Thermoelectricity, Wiley, New York 1960.

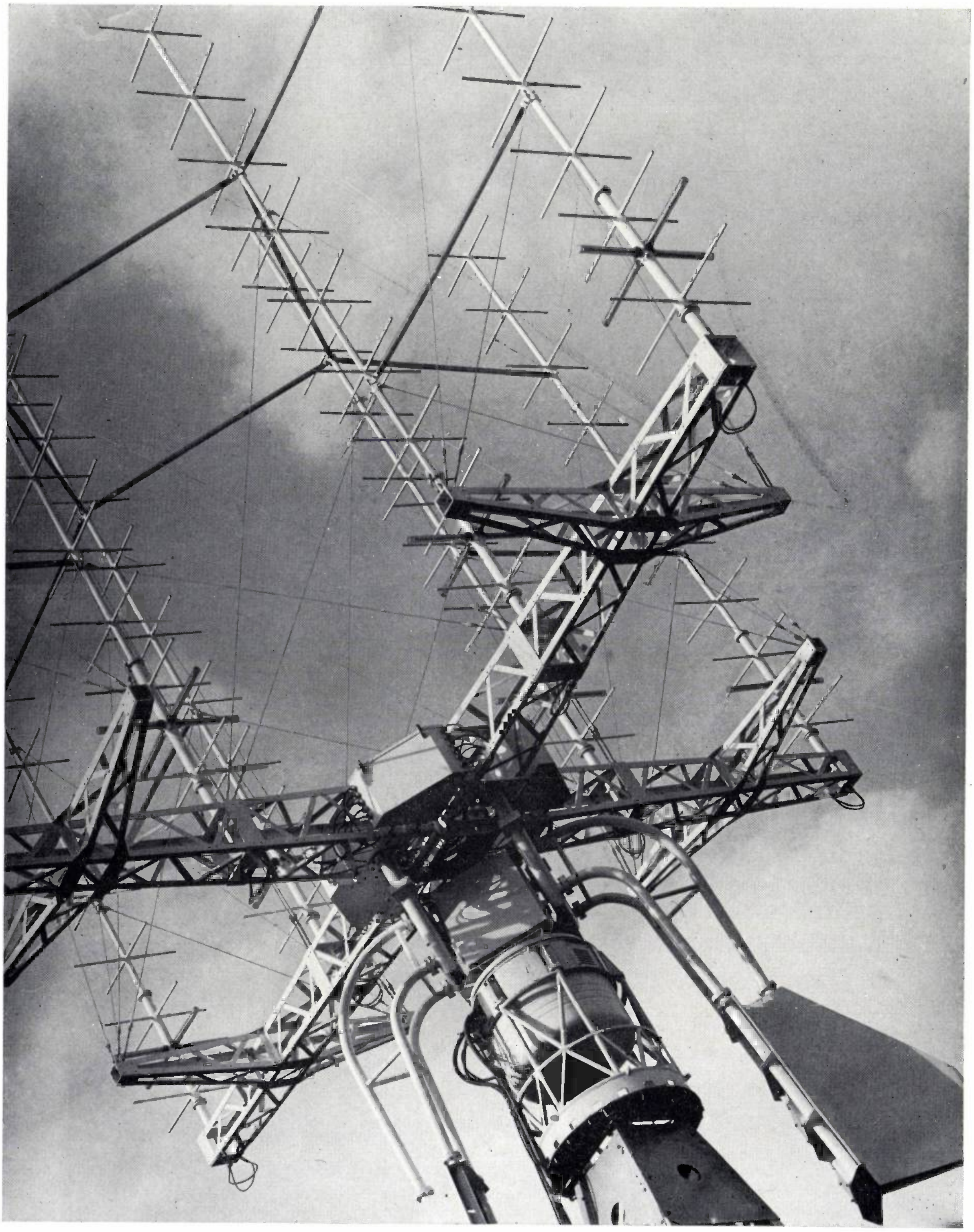
R. R. Heikes and R. W. Ure, Jr., Thermoelectricity: science and engineering, Interscience Publ., New York 1961.

Summary. When an electric current flows across the junction of two dissimilar conductors connected in series, heat is transported from one junction to the other. This "Peltier effect" is the basis of the operation of a Peltier element as a heat pump and as a refrigerator. The theory of this effect is discussed, and also that of the closely associated Seebeck effect. It is shown that the greatest effects are to be expected in semiconductors, particularly in a combination of a *P*-type with an *N*-type semiconductor, and that very minor effects are to be expected in metals.

There is an optimum current at which the cooling power of a Peltier element is a maximum. The quality (maximum cooling power, maximum temperature reduction, efficiency) is expressed by a "figure of merit", which depends almost entirely on the properties of the materials used (thermal and electrical conductivity, Peltier coefficient). The best materials for Peltier cooling are compounds derived from bismuth telluride. The preparation and properties of these materials are discussed. The adverse effect of a ripple in the supply current is analysed, and is found to be generally of little significance.

Finally, various examples of application are given to demonstrate the particular advantages of the Peltier element as a cooling device; it is small, simple in design, adaptable, reliable, easy to regulate (and hence well suited for temperature stabilization) and requires no maintenance.

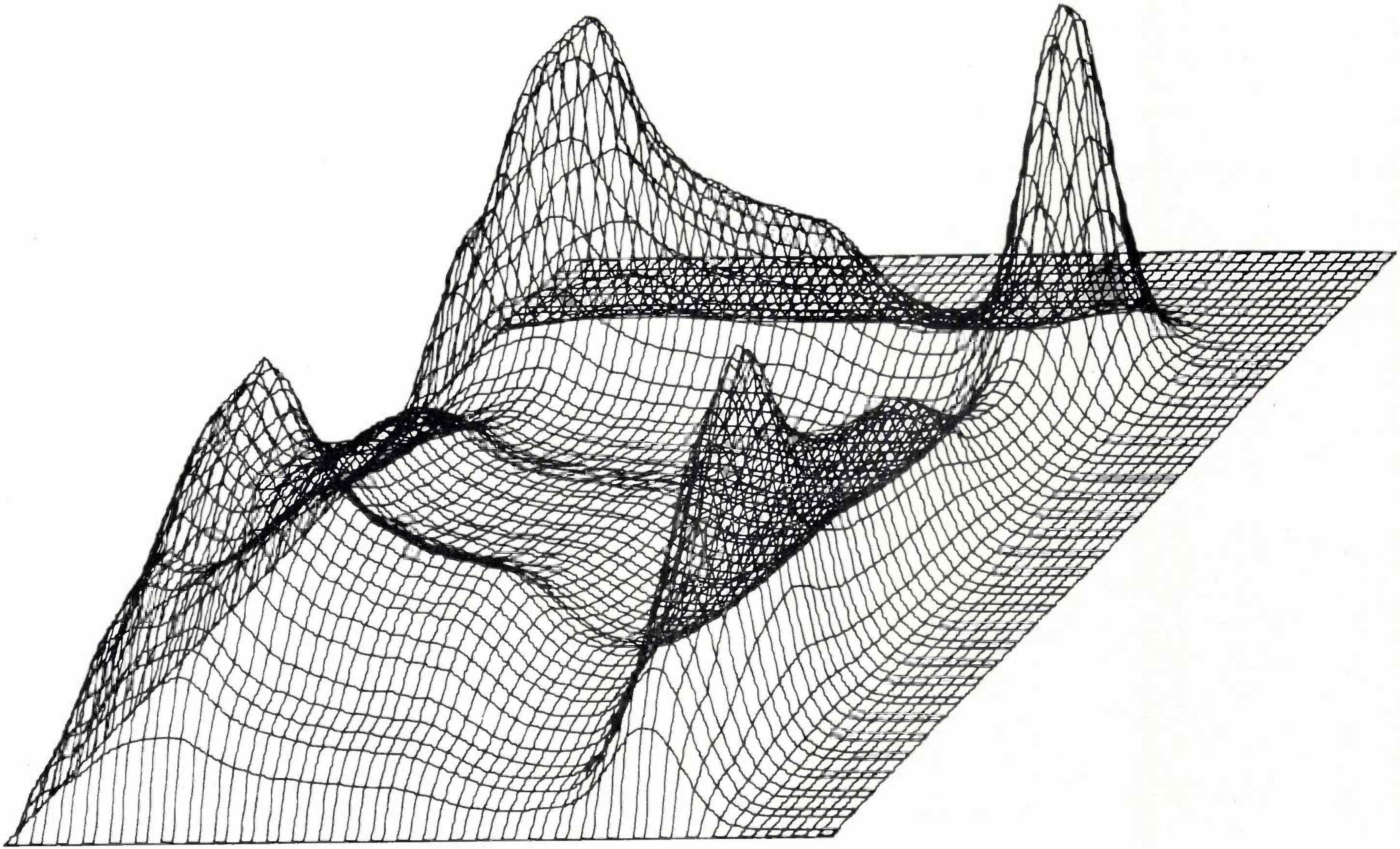
Aerial for telemetering satellite-launcher



When the European Launcher Development Organization (ELDO) fires its fourth rocket from the Woomera range, the operation of the third rocket stage will be telemetered by means of equipment developed and built by Philips' Telecommunicatie Industrie at Huizen (Netherlands). The 60 feet high receiving aerial shown in the photograph forms part of this equipment. It contains 8 vertical and horizontal dipole arrays, from which the signals are fed separately to the polarization diversity receivers; this will reduce fading effects caused by the hot gases of the rocket exhaust. Phase differences between the voltages set up in the dipole arrays give rise to difference voltages which are a measure of the aiming error of the aerial in azimuth or elevation. These voltages are used for actuating a servo-system which will lock the aerial on to the rocket.

Representations of three-dimensional energy spectra

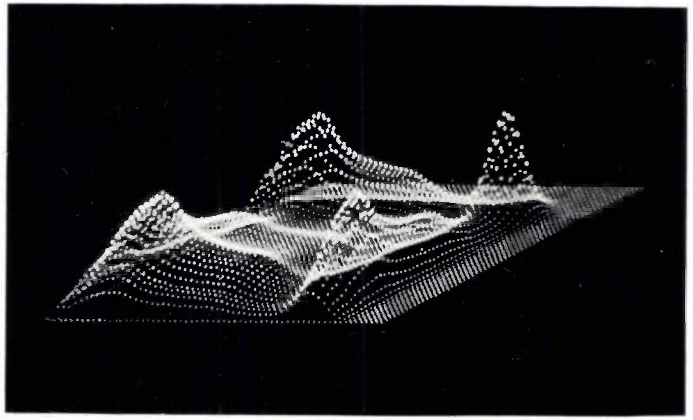
539.16.08



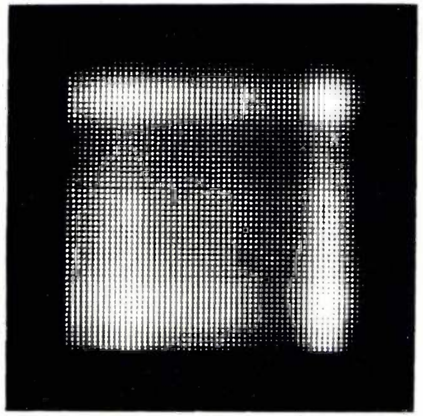
The pictures shown on these pages originate from a 4096-channel pulse-height analyser (kick sorter) built for experimental purposes at Philips Research Laboratories in Eindhoven. They represent pulse-height spectra with *two* independent variables.

Pulse-height analysers are mainly used for measuring radioactive radiation, and in particular for recording energy spectra. The radiation produces pulses in a detector, and the pulse-height is a measure of the energy of the particular quantum or particle. The number of pulses per unit time is a measure of the intensity of the radiation. The analyser sorts the pulses according to height; the number of pulses in the various height intervals (channels) are counted, and the result is retained in a store. The spectrum can now be made visible and it remains available for further processing.

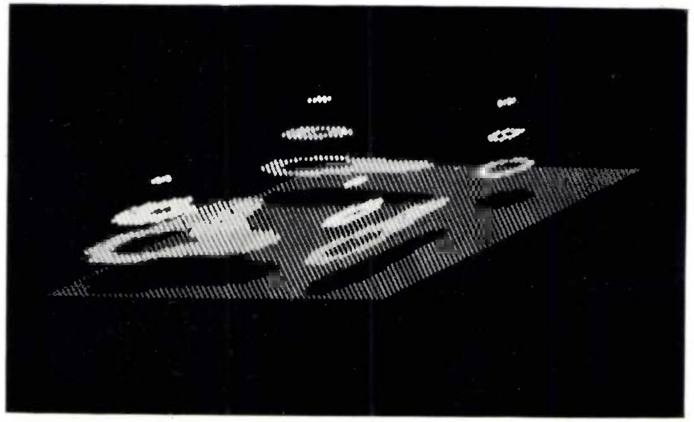
For a variety of purposes it is useful to record spectra as a function of *two* variables. The second variable may be the energy of another particle emitted by the same source and detected in a second detector; the particles are recorded only, if, for instance, they are coincident, the object being to determine energy levels and radiation transitions in the nuclide. An angle may also be used as the second variable corresponding to the radiation measured in different directions from a source.



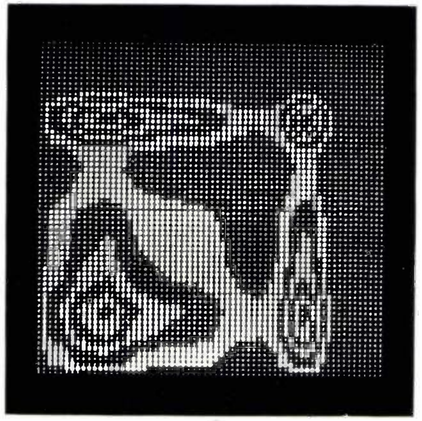
a



b



c



d

The pictorial representation of a function of two variables requires a three-dimensional figure on a plane. This can be made in different ways. The "perspective" drawing shown here was made with an *x-y* recorder. The other pictures are direct photographs of an oscilloscope trace; (*a*) is again a perspective representation of the spectrum, (*b*) is a plan view, height being indicated by intensity modulation, (*c*) is the same as (*a*), except that only a few significant (binary) digits from the numbers in the store are displayed, giving the spectrum a step-like shape. The same procedure was adopted for the plan view (*d*); the successive steps are now shown as alternately light and dark. In this last figure the location of the peaks in the spectrum can easily be measured.

F. Bregman

Ir. F. Bregman, formerly with Philips Research Laboratories, Eindhoven.

A variable inductor for r.f. heating control

K. H. Lopitzsch

621.365.5-533.66

In r.f. heating it is usually necessary to be able to adjust the power supplied to the work-piece. In this article, the writer points out the advantages of adjustment by variation of the matching between generator and work coil. He describes a matching control arrangement that requires little space and is suitable for powers up to several hundred kW.

R.f. heating generators

An r.f. heating generator is essentially a frequency converter which takes power from the 50 c/s mains supply and supplies it, at a higher frequency, to the work-piece to be heated. Distinction is made between *induction heating*, in which eddy currents are induced in a substance by placing it in the alternating magnetic field of a coil, and *dielectric heating*, in which a substance is placed in an alternating electric field, so that heating arises because of dielectric losses [1]. In induction heating, frequencies between about 100 kc/s and 10 Mc/s are used at present, while in dielectric heating the frequencies used are in the 5 to 100 Mc/s range. The actual choice of the frequency to be used depends on the dimensions and properties of the work-pieces and the heat distribution desired.

The power required from the r.f. generator depends on the size and thermal properties of the workpieces to be processed as well as on the temperature desired and the speed at which this must be reached. A power of less than 1 kW is sufficient for some uses; there are however industrial generators which will supply several hundred kW.

A generator for use in r.f. heating comprises an oscillator, which generally makes use of one or more valves specially developed for industrial application [2], and a high-voltage transformer and rectifier which derives the supply voltage from the mains. In induction heating, to which we shall confine ourselves in this article, the connection between the coil in which the work-piece is placed — the *work coil* — and the oscillator circuit is usually made by means of an r.f. transformer. If the eddy currents induced in the work-piece and, therefore, the heat generated in it, are to be confined to a relatively small zone, the work coil must also be small, so that it has a low inductance. An impedance transformation is then required for matching the work coil to the generator, in order to obtain optimum power transfer. The r.f. transformer used for this consists of a combination of the oscillator coil and a secondary coil with a small number of turns (generally only one), which is connected to the work coil.

Fig. 1 gives the block diagram of an r.f. heating installation, showing the units we have mentioned.

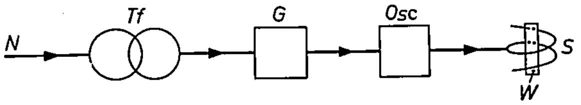


Fig. 1. Block diagram of an r.f. induction heating installation. *N* mains connection, *Tf* high-voltage transformer, *G* rectifier, *Osc* oscillator, *S* work coil, *W* work-piece.

Regulating the power

In r.f. heating it is nearly always necessary to be able to vary the power applied to the work-piece. This necessity does not only arise when successive work-pieces are of different size or shape, but can also be required during the processing of a single work-piece. More power, for instance, is required to melt a given quantity of a substance than to keep it in the molten state. The properties of a work-piece may sometimes alter markedly during heat-treatment (for example, when the Curie point of a ferromagnetic material is exceeded), or a work-piece may have to be heat-treated according to a certain programme. In cases such as these, it must be possible to adjust the power from the generator — sometimes fairly rapidly.

There are two different ways of varying the power supplied by the oscillator to the load (the work coil). If the load is *matched* to the generator, the power can be adjusted by *variation of the generator power*, for example in accordance with a particular control programme. This adjustment may be made at the mains

Ir. K. H. Lopitzsch is with Philips Industrial Equipment Division, Eindhoven.

side of the high voltage transformer at the rectifier, or at the oscillator. With a *matched* load, the power can be raised to the maximum that the generator is rated to supply. When the work-piece is changed, or its material properties change, the match is however no longer maintained, and an adjustment of the match has to be made. This requires an electromechanical device between the oscillator and the work coil. If its range is large enough, the device itself can be used to vary the power delivered to the load. This latter method permits the maximum power to be reached, even for considerable variation in loads, and gives greater freedom in the design of the work coils. Most r.f. heating generators are therefore provided with means for matching adjustment. Power variation by electronic means is however often applied in installations with automatic control.

Fig. 2 shows some of the methods used for adjusting the matching. These include variation of capacitance of the oscillator circuit (1), the connection between the oscillator valve and the circuit (2), the inductance of the oscillator coil (3), the mutual inductance of the coils of the r.f. transformer (4), the inductance of the secondary coil (5), the inductance of a coil in series with the work coil (6) and the inductance of a coil in parallel with the work coil (7). Combinations of these methods are also used.

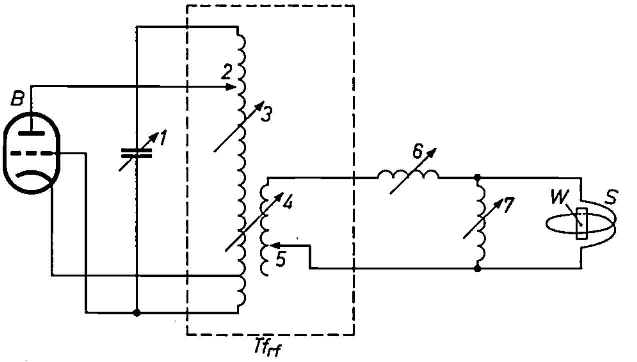


Fig. 2. Theoretical diagram of a generator for r.f. induction heating showing a few of the methods of adjusting the power. *B* is the oscillator tube, $Tf_{\rm rf}$ the r.f. transformer, *S* the work coil and *W* the work-piece.

Where the generator has to supply considerable power, the variable elements which these methods require must, of necessity, be large. This is because they have to operate with large high-frequency voltages and currents so that the distances between components and the cross-sections of the conductors must be large. These considerations apply particularly for a variable coupling between the two coils of the r.f. transformer (4), an arrangement often used because of the wide control range it gives [3]. The r.f. transformer and the matching control are often mounted in the same cabinet as the generator. Sometimes, however, it is necessary or desirable to mount these units as close as possible to the work coil. This led us to develop a solution of type (6): a variable inductor in series with the work coil. This scheme allows the dimensions to be kept fairly small, while still maintaining a sufficient range of adjustment.

The equipment concerned is mainly used in installations for welding the longitudinal seams of metal pipes. A conversion to r.f. induction welding offered considerable advantages, mainly because there would be no need for contacts, which had been required in the welding method used previously.

Because of the high power required (up to 100 kW), the generator is so large that it has to be placed some distance away from the rolling machine that shapes the metal strip into pipes. (This is usually the situation when an existing installation is converted for r.f. heating.) The lead from the r.f. transformer to the work coil, however, has to be kept extremely short, since otherwise there would have been an unacceptable power loss and inductive voltage-drop at the very high current required (up to 1600 A). To avoid making this lead too long, we divided the r.f. section into two parts: the r.f. transformer is located in a separate cabinet close to the rolling machine. The variable inductor connected in series with the work coil is also located in this cabinet.

A *series* coil is preferable to one in *parallel* with the work coil as a series coil gives less power loss when the power has to be greatly reduced. Low power with a parallel coil can only be achieved by setting it to a very low inductance, which means that the r.f. transformer must supply a very high current.

The design of the variable inductor

The frequency of the current used for pipe welding (about 0.5 Mc/s) is so high that the variable inductor does not need to have a large number of turns. A single loop is quite sufficient, and its inductance can be made adjustable by designing the loop as two parallel conductors with a movable bridge. To allow as great a relative variation in the inductance, with a small minimum value and a restricted length of the device, the conductors must be fairly close together.

[1] On this subject, see e.g. Philips tech. Rev. **11**, 165-175 and 232-240, 1949/50.

[2] See E.G. Dorgelo, Transmitting valves for use in industry, Philips tech. Rev. 20, 299-304, 1958/59.

[3] For this purpose, the coils of the r.f. transformer are usually made adjustable in relation to each other. This requires mechanical arrangements that take up a great deal of space. Spherical coils that can be rotated in relation to each other take up less room, but this arrangement is much more difficult from the design viewpoint.

The following equation applies to the inductance of two parallel cylindrical conductors of diameter d, spacing between centres D and of length l (see *fig. 3*):

$$L = k_1 l\,(k_2 \log 2D/d - D/l)\,,$$

where k_1 and k_2 are constants. For constant values of d and l, the curve of L as a function of D is roughly as shown in *fig. 4*. It can be seen that as D increases the rate of increase of L steadily decreases.

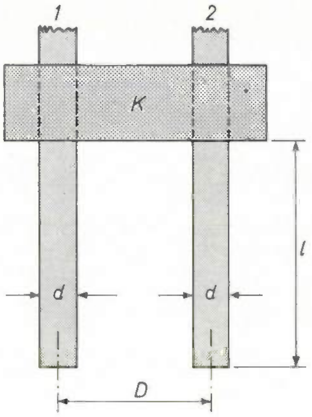


Fig. 3. The variable inductor consists of two parallel cylindrical conductors *1* and *2* and a movable bridge *K*. Currents of up to 1600 A at a frequency of e.g. 0.5 Mc/s flow through the device.

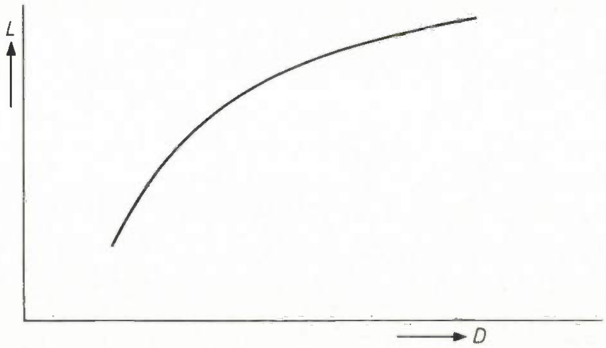


Fig. 4. An approximate curve of the inductance L of two parallel conductors as a function of their separation D at a constant diameter d and length l.

If the two conductors are arranged in a vertical position, and combined with a horizontal r.f. transformer, the whole arrangement takes up very little floor area. The connections can then be so short that, at the lowest position of the bridge, the total inductance in series with the work coil is sufficiently low. *Fig. 5* shows the arrangement adopted for 50 kW and 100 kW generators with frequencies of about 0.5 Mc/s. The 100 kW generator can be used for welding pipes with wall-thicknesses of up to 4 mm and diameters up to 80 mm, at rates of up to 20 m/min. The power has to be adjusted when the diameter or material of the pipes is changed, and also when the machine is started or if the mains voltage fluctuates. The use of variable matching allows these adjustments to be made without having to change the work coil.

The two parallel conductors and the bridge of the variable inductance in which, as stated, currents of up to 1600 A flow, have to be water-cooled. Fig. 5 shows some of the water connections.

In the design of the bridge, we have to ensure not only good contact with the conductors but also that the heat developed at the contacts can be easily dissipated by the cooling water. *Fig. 6* shows a few details of the bridge. Flat areas on the four contact blocks *4* bear against the water-cooled conductors *1* and *2*. The contact blocks are not rigidly secured to the two transverse conductors *3*, but can rotate slightly. This arrangement ensures a linear contact without placing too high a requirement on the straightness and parallel alignment of conductors *1* and *2*. The contact pressure is provided by the spring *5*: this may be made of steel, but if this is the case it must be properly screened electrically to prevent heating by induced

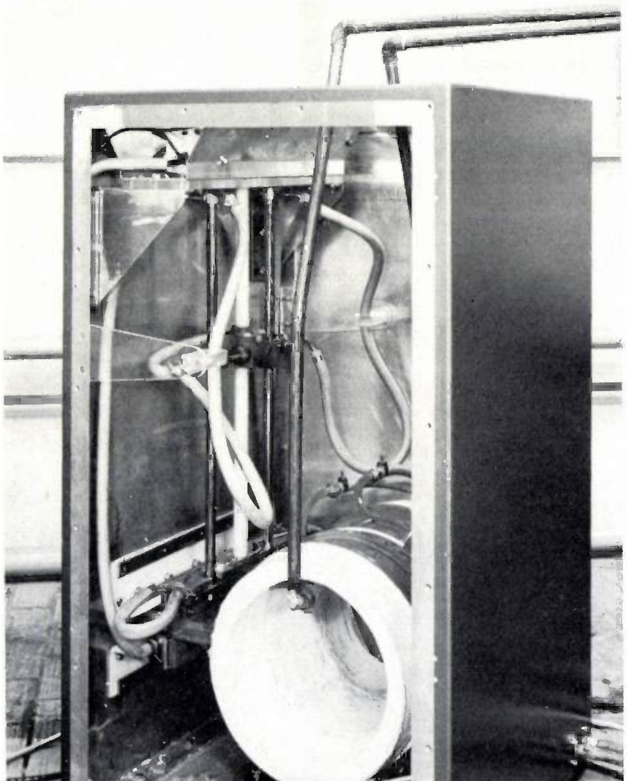


Fig. 5. Combination of a horizontally-arranged r.f. transformer and a vertically-arranged variable inductor. One of the walls has been removed from the cabinet. The various hoses and pipes carry the cooling water for the equipment. The bridge is moved up and down by a servomotor.

eddy currents; with this in mind, rubber springs are used. The contact surfaces of the blocks are made of a silver-copper alloy, which is a good conductor and is harder than copper. Thus, as the bridge is moved up and down, any irregularities in the surface of the copper conductors *1* and *2* are removed.

The power losses in the variable inductor are a small percentage of the power supplied by the generator. A notable feature is that this power loss is virtually independent of the position of the bridge; for although an increase in the inductance corresponds to an increase in the length of the operative section of the conductors, there is a decrease in the current.

Fig. 7 shows a pipe-welding installation capable of supplying a power of 100 kW. *Fig. 8* is a detail photograph showing the work coil together with the two rollers that press the edges together, thus forming the welded seam. A ferrite core is inserted in the work coil so as to concentrate the induced currents at the location of the weld. The holder for this core may be seen in the unwelded part of the pipe.

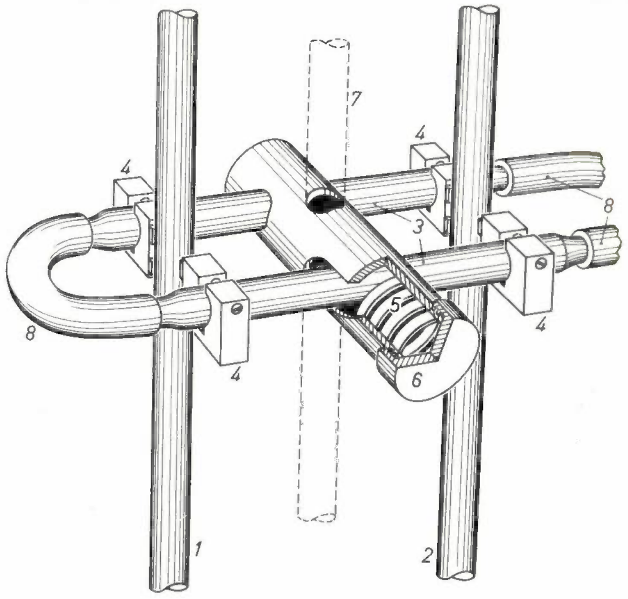


Fig. 6. The design of the bridge of the variable inductor. *1* and *2* are the conductors forming the inductor, *3* the bridge conductors, *4* the contact blocks, *5* a spring, *6* a cap for tensioning the spring, *7* a rod for moving the bridge and *8* the water hoses.

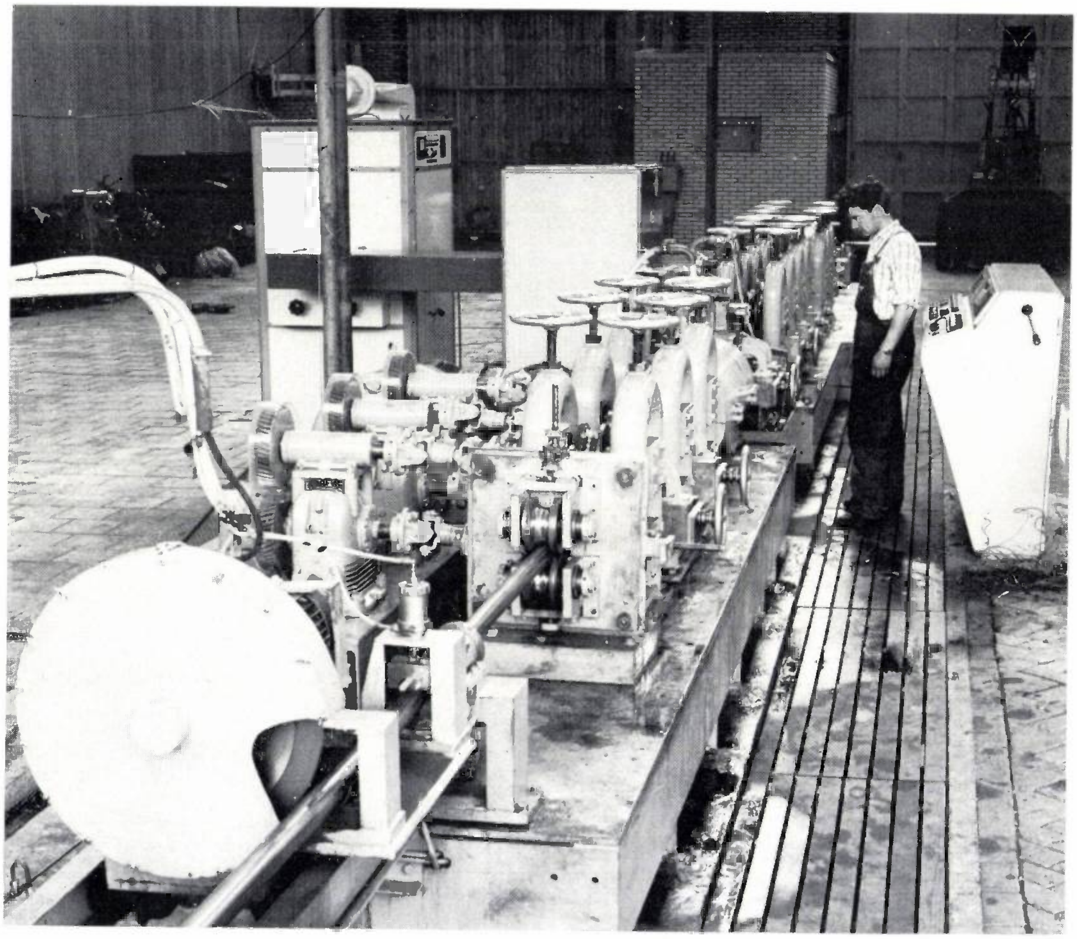


Fig. 7. An installation for the manufacture of metal pipes, in which the longitudinal seam is welded by r.f. induction heating. On the left is the generator cabinet and next to it, close to the rolling machine, is the cabinet containing the r.f. transformer and the variable inductor. The brick-built structure in the background contains the supply equipment, and the contro panel is at the extreme right. This installation can supply a power of 100 kW. In the foreground the welded pipe can be seen leaving the machine. It is cut to length by the circular saw attached to the machine.

As well as in pipe-welding, variable inductors as described here have also been used in the zone-melting of semiconductors and in a plasma-heating generator. In this latter case, the material in the work coil is an ionized gas (plasma). This may be considered as a conductor in which the r.f. alternating field generates eddy currents that heat the gas. It is thus possible to obtain a very high temperature plasmoid or "torch".

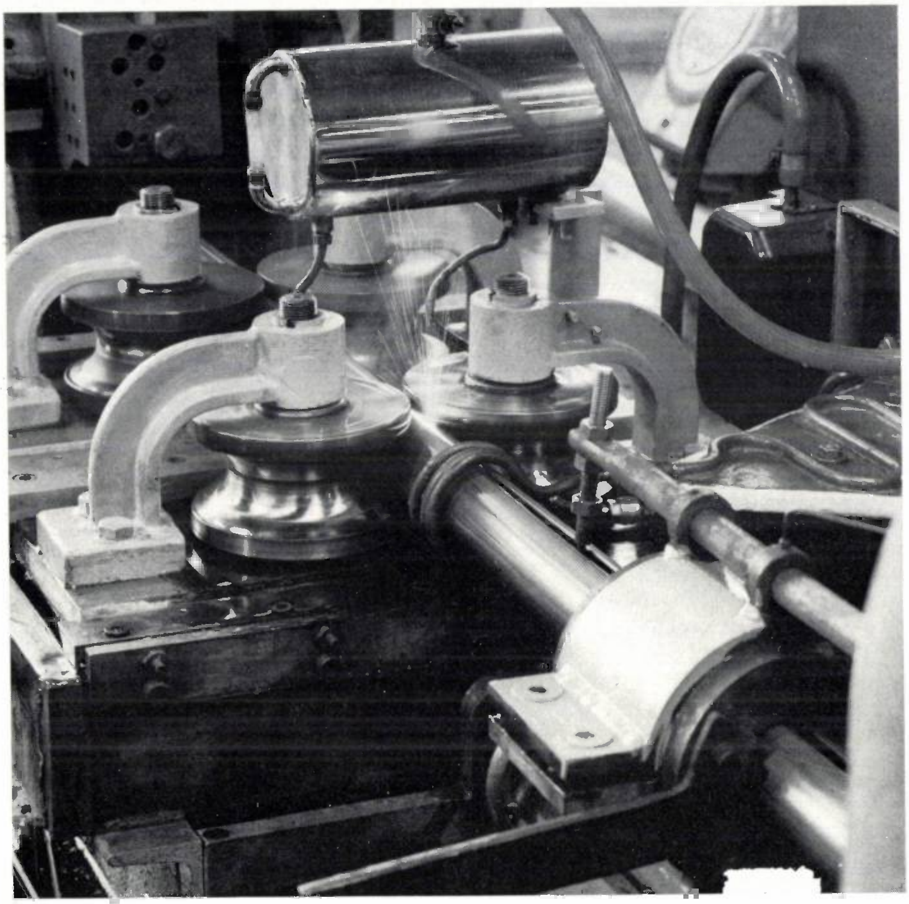


Fig. 8. A detail of the equipment. The work coil can be seen through which the pipe to be welded is fed. The heated parts are pressed together behind the coil by two rollers, thus effecting the weld. Heating takes place as follows. A current is induced in the outer wall of the pipe (not yet closed) at the work coil position. The main paths which the current may take consist of two branches in parallel, i.e. along the inner side of the pipe and along the still open edges and the weld. To ensure that most of the current flows along the weld, a ferrite core is located in the axis of the work coil to increase the inductance of the path inside the pipe. The holder for this core may be seen in the still open part of the pipe.

Summary. In an r.f. heating installation, the power supplied to the work coil may be varied in a number of ways. Adjustment by means of the matching, rather than by changing the generator power, has the advantage that both match and power may be controlled. A disadvantage is, however, that large control units are required for high power work. A relatively small control unit may be made by the use of a variable inductor connected in series with the work coil and consisting of two parallel conductors with a movable bridge. Designs on this basis have been made which can be used to control powers of several hundred kW. One of the applications is in welding the longitudinal seams of metal pipes.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands *E*
Mullard Research Laboratories, Redhill (Surrey), England *M*
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France *L*
Philips Zentrallaboratorium GmbH, Laboratory at Aachen, Weisshausstrasse, 51 Aachen, Germany *A*
Philips Zentrallaboratorium GmbH, Laboratory at Hamburg, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany *H*
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium. *B*

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

G. Arlt: Piezoelectricity in highly conducting semiconductors.
J. appl. Phys. **36**, 2317, 1965 (No. 7). *A*

A. C. Aten, J. H. Haanstra and **H. de Vries:** Fluorescence and photoconduction in tellurium-doped cadmium sulphide.
Philips Res. Repts. **20**, 395-403, 1965 (No. 4). *E*

H. G. Beljers: Resonance linewidth and spinwave absorption near the Curie temperature of a Ni-Zn ferrite and of yttrium iron garnet.
Physics Letters **18**, 248-249, 1965 (No. 3). *E*

G. Blasse: The ternary system $Li_3(Nb,Sb,Ta)O_4$.
J. inorg. nucl. Chem. **27**, 2117-2119, 1965 (No. 9). *E*

P. F. Bongers: Anisotropy of the electrical conductivity of VO_2 single crystals.
Solid State Comm. **3**, 275-277, 1965 (No. 9). *E*

J. C. Brice: The effect of inclusions on the domain coercive force of thin ferromagnetic films.
Brit. J. appl. Phys. **16**, 1523-1525, 1965 (No. 10). *M*

J. C. Brice and **U. Pick:** The resistivity of evaporated permalloy films.
Vacuum **15**, 409-412, 1965 (No. 8). *M*

A. Bril, W. L. Wanmaker and **J. Broos:** Photoluminescent properties of some europium-activated gadolinium and yttrium compounds.
J. chem. Phys. **43**, 311, 1965 (No. 1). *E*

J. J. Brissot and **R. Martres:** Surface tension of gadolinium and floating-zone single crystals.
J. appl. Phys. **36**, 3360, 1965 (No. 10). *L*

K. Bulthuis: The effect of local pressure on silicon *p-n* junctions.
Philips Res. Repts. **20**, 415-431, 1965 (No. 4). *E*

H. B. G. Casimir: Resistance of type II superconductors in the mixed state.
Physics Letters **17**, 177-178, 1965 (No. 3). *E*

P. Cornelius, W. de Groot and **R. Vermeulen:** Quantity equations, rationalization and change of number of fundamental quantities, I, II, III.
Appl. sci. Res. **B 12**, 1-17, 235-247, 248-265, 1965 (Nos. 1 and 4). *E*

P. J. Flanders and **S. Shtrikman:** Magnetic field induced antiferromagnetic to weakferromagnetic transitions in hematite.
Solid State Comm. **3**, 285-288, 1965 (No. 9). *M*

L. Fraiture: Scattering of electromagnetic waves from rough surfaces.
Ann. Soc. Scient. Bruxelles **I 79**, 144-162, 1965 (No. 2). *B*

P. Gerthsen, R. Groth and **K. H. Härdtl:** Halbleitereigenschaften des $BaTiO_3$ im Polaronenbild.
Phys. Stat. sol. **11**, 303-311, 1965 (No. 1). *A*

P. Gerthsen, R. Groth, K. H. Härdtl, D. Heese and **H. G. Reik:** The small polaron problem and optical effects in barium titanate.
Solid State Comm. **3**, 165-168, 1965 (No. 8). *A*

P. Gerthsen and **E. Kauer:** The luminescence diode acting as a heat pump.
Physics Letters **17**, 255-256, 1965 (No. 3). *A*

P. J. van Gerwen: On the generation and application of pseudo-ternary codes in pulse transmission.
Philips Res. Repts. **20**, 469-484, 1965 (No. 4). *E*

J. A. Geurst and **H. J. C. A. Nunnink:** Numerical data on the high-frequency characteristics of thin-film transistors.
Solid-State Electronics **8**, 769-771, 1965 (No. 9). *E*

P. A. Gould: The resistivity and structure of chromium thin films.
Brit. J. appl. Phys. **16**, 1481-1491, 1965 (No. 10). *M*

H. C. de Graaff: Relation between channel conductance and characteristics of thin-film transistors.
Solid-State Electronics **8**, 835-837, 1965 (No. 10). *E*

C. Haas: Phase transitions in crystals with the spinel structure.
J. Phys. Chem. Solids **26**, 1225-1232, 1965 (No. 8). *E*

F. N. Hooge, H. Kalter and **A. M. H. Hoppenbrouwers:** Influence of pulse duration on the spectra of GaAs lasers above threshold.
Physics Letters **17**, 254-255, 1965 (No. 3). *E*

B. B. van Iperen and **W. Kuypers:** Experimental CW klystron multiplier for submillimetre waves.
Philips Res. Repts. **20**, 462-468, 1965 (No. 4). *E*

B. B. van Iperen and **H. J. C. A. Nunnink:** Harmonics in velocity-modulated cylindrical electron beams.
Philips Res. Repts. **20**, 432-461, 1965 (No. 4). *E*

H. Klotz: Chemischer Transport von Lanthanhexaborid.
Naturwiss. **52**, 451-452, 1965 (No. 15). *A*

W. Kwestroo and **A. Huizing:** The preparation of ultra pure lead oxide.
J. inorg. nucl. Chem. **27**, 1951-1954, 1965 (No. 9). *E*

M. A. R. LeBlanc, W. F. Druyvesteyn and **C. T. M. Chang:** Behavior of the magnetization of hard superconductors with transport current in transverse fields, I.
Appl. Phys. Letters **6**, 189-191, 1965 (No. 9). *E*

J. Lemmrich: Frequenzanaloge Motorsteuerung mit kontaktlosen Bauelementen.
Elektr. Ausrüstung, 1965, No. 2, pp. 58-61. *H*

J. Liebertz: Hydrothermal-Untersuchungen an einigen Verbindungen.
Chemie-Ing.-Technik **37**, 830-832, 1965 (No. 8). *A*

E. J. Millett, L. S. Wood and **G. Bew:** Oxygen in germanium by vacuum fusion and infra-red absorption.
Brit. J. appl. Phys. **16**, 1593-1594, 1965 (No. 10). *M*

E. Neckenbürger: Ein Magnetometer mit oszillierender Probe.
Z. angew. Phys. **18**, 440-441, 1965 (No. 5/6). *H*

B. Okkerse: Consecutive X-ray reflexions in the same perfect crystal, Part II: Determination of the refractive index of X-rays; Part III: Measurement of the intrinsic curvature of lattice planes.
Philips Res. Repts. **20**, 377-388 and 389-394, 1965 (No. 4). *E*

D. J. van Ooijen and **G. J. van Gurp:** Measurement of noise in the resistive state of type II superconductors.
Physics Letters **17**, 230-231, 1965 (No. 3). *E*

G. W. van Oosterhout and **J. Visser:** The determination of active oxygen and iron(II) in oxide compounds.
Anal. Chim. Acta **33**, 330-332, 1965 (No. 3). *E*

O. Reifenschweiler: Fortschritte in der Entwicklung einfacher Neutronengeneratoren.
Z. Instrumentenk. **73**, 197-204, 1965 (No. 8). *E*

P. Reijnen: Het evenwicht tussen ferrieten met spinelstructuur en zuurstof.
Klei en Keramiek **15**, 127-132, 1965 (No. 5). *E*

F. C. de Ronde: A precise and sensitive X-band reflecto"meter" providing automatic full-band display of reflection coefficient.
IEEE Trans. on microwave theory and techniques **MTT-13**, 435-440, 1965 (No. 4). *E*

C. J. M. Rooymans, C. Langereis and **J. A. Schulkes:** KFe_5O_8, a new phase in the K_2O-Fe_2O_3 system.
Solid State Comm. **3**, 85-87, 1965 (No. 4). *E*

J. J. Scheer and **J. van Laar:** GaAs-Cs: a new type of photoemitter.
Solid State Comm. **3**, 189-193, 1965 (No. 8). *E*

H. Severin: Ferrites — chemical composition, crystal structure and method of preparation.
Electronic Components **6**, 312-319, 1965 (No. 4). *H*

F. A. Staas, A. K. Niessen and **W. F. Druyvesteyn:** Hall effect in type II superconductors.
Physics Letters **17**, 231-233, 1965 (No. 3). *E*

J. H. Stuy: Fate of transforming DNA in the Haemophilus influenzae transformation system.
J. mol. Biol. **13**, 554-570, 1965 (No. 2). *E*

D. R. Tilley: The temperature dependence of the nucleation fields in dirty anisotropic superconductors.
Proc. Phys. Soc. **86**, 678-679, 1965 (No. 3). *M*

C. G. Venis and **W. J. M. Carpay:** Vulcanized latex as an aid for clearing away radioactive contaminations.
Health Phys. **11**, 1091-1092, 1965. *E*

J. Verweel: On the determination of the microwave permeability and permittivity in cylindrical cavities.
Philips Res. Repts. **20**, 404-414, 1965 (No. 4). *E*

N. Warmoltz: The influence of a magnetic field, parallel to the cathode, on the second cathode sheath of a low-pressure arc discharge.
Appl. sci. Res. **B 11**, 418-422, 1964/65 (No. 6). *E*

J. D. Wasscher: Evidence of weak ferromagnetism in MnTe from galvanomagnetic measurements.
Solid State Comm. **3**, 169-171, 1965 (No. 8). *E*

G. Winkler: Über die Herstellung von Ferriten mit planarer Anisotropie.
Elektrotechn. Z. **B 17**, 769-771, 1965 (No. 23). *H*

W. J. Witteman: Increasing continuous laser-action on CO_2 rotational vibrational transitions through selective depopulation of the lower laser level by means of water vapour.
Physics Letters **18**, 125-127, 1965 (No. 2). *E*

Volume 27, 1966, No. 5 pages 113-140 Published 12th September 1966

PHILIPS TECHNICAL REVIEW

VOLUME 27, 1966, No. 6

The stellar interferometer at Narrabri, Australia

I. The principles of the intensity interferometer; general description

II. Measurement of the correlation

535.853.4:522.617.3

In 1956 Hanbury Brown and Twiss demonstrated that the fluctuations in the outputs of two photomultipliers, illuminated by coherent light beams, are correlated. This noted experiment was an important step in applying the intensity interferometer — which had been developed in radio astronomy a few years earlier — to visible stars.

Since that time Hanbury Brown, Twiss and their co-workers have developed and built the optical intensity interferometer at Narrabri, with which it is possible to measure the angular sizes of a number of stars for the first time. This is done with the aid of a refined electronic apparatus, the "correlator", which measures the correlation between the fluctuations in the outputs of two photomultipliers. This electronic correlator has been designed and built by Mullard Research Laboratories.

The present subject is presented in two articles. The first gives an introduction to intensity interferometry and a general description of the apparatus in Narrabri, the second describes the correlator in some detail.

We are pleased to welcome Prof. Hanbury Brown, who wrote the first article, as a guest author to our Review.

I. The principles of the intensity interferometer; general description

R. Hanbury Brown

Introduction

The purpose of the stellar interferometer at Narrabri is to measure the angular diameter of the bright stars. The "angular diameter" of a star is the angle subtended by the diameter of the star at the Earth.

At first sight it seems that the simplest way to measure the angular diameter of a star would be to view it with a telescope and to measure the image. In fact that is precisely how the angular diameters of the Sun, the Moon and the planets are measured. However, in practice this method cannot be used for the stars because their distances are so great that their outlines cannot be seen in a telescope. There are two principal reasons why this is so. Firstly, the ability of a telescope to distinguish between a point source of light and a source of finite angular size is limited fundamentally by the size of the principal lens or mirror of the telescope and by the wavelength of light. Thus if we look at a point source of light with a telescope the image obtained will appear, not as a point, but as a bright disc surrounded by bright rings. If the principal mirror has a diameter of d metres the angular diameter of the disc, i.e. the *apparent* angular size of the point source, is approximately λ/d radians, where λ is the wavelength of the light. Thus, in theory we cannot with a telescope of diameter d measure the angular diameter Θ of a star unless Θ is greater than λ/d radians.

Prof. R. Hanbury Brown, F.R.S., formerly of the Jodrell Bank Observatory, is now Professor of Physics (Astronomy) at the University of Sydney, Australia, and director of the Narrabri Observatory.

The largest angular diameter of any star is that of Betelgeuse, 0.047 seconds of arc, and just to distinguish this star from a point source we should theoretically need a telescope lens or mirror roughly 3 metres in diameter. The *hot* bright stars have angular sizes considerably less than that of Betelgeuse; typically they subtend angles of 0.001 seconds of arc and we should require telescope mirrors with a diameter of about 150 metres to resolve them.

However, in making this estimate we have neglected a second limitation which is set by scintillation in the Earth's atmosphere. Irregularities in the atmosphere introduce random changes in the phase and amplitude of the light reaching the ground and, roughly speaking, these changes are uncorrelated at two points separated by more than about 10 cm. In practice this means that different parts of a large telescope mirror view the star through different irregularities and in consequence the image of a star is distorted, dances about, and is increased in size. The limitations set by the atmosphere depend on a variety of factors such as the height of the Observatory and the weather; roughly speaking they set limits to the minimum angular size which can be seen clearly in a telescope of between 1 and 0.1 seconds of arc. Thus, no matter how large our telescope, we are prevented by the Earth's atmosphere from measuring the angular size of stars.

There are, however, two types of instrument with which the angular diameters of stars have been measured. In both the light is received, not on one huge mirror with a diameter of λ/Θ, but on two comparatively small mirrors which can be separated by a distance λ/Θ. The first of these types is the well-known *Michelson interferometer*. Here the light received at the two small mirrors is brought together at the centre to give interference patterns. The second type is the *intensity interferometer* which has been developed in the past ten years. In this instrument the light received at two separated mirrors is focused on to two photoelectric detectors and, in effect, is converted into electric currents. The two currents are then brought together and "interfere" electrically; in other words the currents contain high-frequency fluctuations which are compared or correlated.

The stellar interferometer at Narrabri, the subject of the present two articles, is an intensity interferometer. In the first article we shall review the general principles on which it works and compare it with Michelson's interferometer. In the second article the *electronic correlator* which compares the currents from the two photoelectric detectors will be described. The correlator, which was designed and built at the Mullard Research Laboratories, presented considerable design problems.

The purpose of the measurements

The majority of the material in the Universe is condensed into stars and an understanding of their structure and evolution is one of the major aims of astronomy. Theoretical studies show that the mass and chemical composition of a star determine to a large extent its other properties and it is the aim of these studies to relate these basic parameters to the observed properties of the star. Two important properties of a star which can be observed are its *surface temperature* and its *size*, both of which exhibit a considerable variety. The fact that stars differ in temperature can be seen clearly from their spectra which, for historical reasons, are classified in the curious order *O*, *B*, *A*, *F*, *G*, *K*, *M* in descending order of temperature. The associated scale of temperature ranges from about 30 000 degrees for stars of type *O*, to 3000 degrees for type *M*. The physical size of stars also varies greatly. If we confine our attention to the most common types of star, known to astronomers as main sequence stars, their physical diameters vary from roughly 20 times that of the Sun (type *G*) for stars of type *O*, to about one half the size of the Sun for stars of type *M*. However, much greater variations in size are found if we look at rarer types of star. For example, "supergiant stars" of type *M* are about 500 times as large as the Sun, while at the other extreme the size of the "white dwarfs" is believed to be about 100 times smaller than the Sun.

Fig. 1 shows how the observed properties of stars can be displayed in a significant form. This diagram, called a Hertzsprung-Russell diagram, shows the relationship between surface temperature, spectral type, physical radius and total radiated light flux for different types of star. The spectral type and the associated surface temperature are shown along the horizontal axis. The vertical scale is a measure of the total light radiated by the star. The total light is measured in terms of the "absolute magnitude" of the star which is the apparent magnitude of the star if placed at a standard distance of 10 parsecs (32.6 light years) from the earth. Of course, the absolute magnitude cannot be measured directly but can be calculated from observations of the distance and the apparent magnitude [1]. The straight lines across the figure represent the physical radii of the stars relative to the Sun.

Now the data in fig. 1 are based on measurements of light flux, spectra, and distance (from the parallax). The temperature corresponding to each spectral type can be estimated from the general distribution of light flux with wavelength, or in other words from the colour; alternatively it can be found from a detailed analysis of the spectral lines and hence the degree of ionization of various elements in the stellar atmosphere. Generally

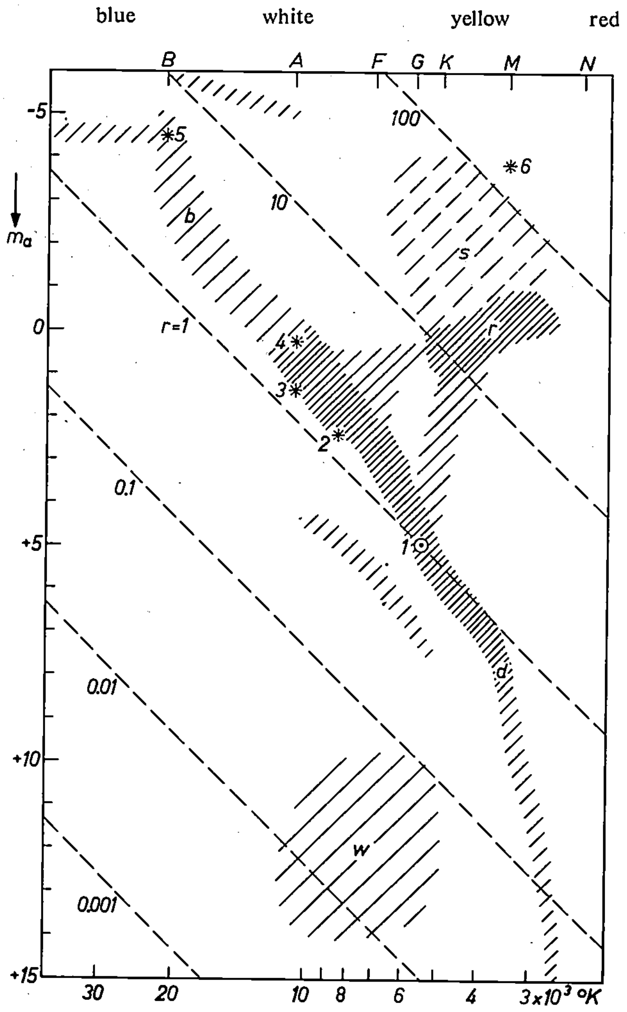


Fig. 1. The Hertzsprung-Russell-diagram, a classification of the stars according to absolute magnitude m_a (proportional to the log of the total radiated light flux) and surface temperature T. Lines of constant radius r are drawn according to Stefan's law (total radiated light flux $\propto T^4r^2$), valid for black body radiators. The values of r are relative to the radius of the sun. In the horizontal direction the stars are plotted according to their spectral type *O*, *B*, *A*, . . ., which is closely related to T. This relation is rather uncertain for hot stars.

Most stars lie on the "main sequence" (*b-d*), which includes the blue giants (*b*) and the red dwarfs (*d*). Large groups outside the main sequence are: the red giants (*r*), the super giants (*s*), and the white dwarfs (*w*). Some well-known stars are indicated: *1* Sun, *2* Altair, *3* Sirius, *4* Vega, *5* β Crucis, *6* Betelgeuse.

speaking both these methods are approximate but work reasonably well for stars cooler than type *A*. For hotter stars of type *O* and *B* there are two principal difficulties. Firstly, these hot stars are extremely blue and so the majority of their radiation is in the ultra-violet and is strongly absorbed by the earth's atmosphere; hence it is impossible, without observation from above the atmosphere, to find the true distribution of flux with wavelength. Secondly, their temperatures are so high that they show only a few weak spectral lines. As a result the temperatures of these very hot stars are uncertain.

The physical size of a star can be estimated from the flux of light received (apparent magnitude), the measured distance and the estimated temperature. If the star radiates like a black body the radiant flux M from unit area of its surface (emittance) will be given by:

$$M = \sigma T^4, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where σ is Stefan's constant and T the temperature of the star. If R is the distance and r the radius of the star, then the flux density E received at the Earth will be:

$$E = \frac{4\pi r^2}{4\pi R^2} M = \frac{r^2}{R^2} \sigma T^4. \quad . \quad . \quad . \quad (2)$$

Thus if we know T, R and E, we can find the radius r of the star.

This method requires that we know the distance of the star and this can only be measured for stars which are relatively close to us. As it happens there are very few *hot* stars for which the distance is known reliably and so their radii cannot be estimated reliably.

On the other hand, if the distance and the surface temperature are not known, we can find the surface temperature of a star by measuring its angular size Θ and the light flux density E; by substituting $\frac{1}{2}\Theta = r/R$ in eq. (2) we have:

$$E = \tfrac{1}{4}\Theta^2\sigma T^4. \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

The primary objective of the stellar interferometer at Narrabri is to measure the angular diameters of hot stars. The only existing measurements of angular diameter were made with Michelson interferometers at Mount Wilson. About 40 years ago an instrument was used with a maximum separation between the mirrors of about 6 metres. Its resolving power was only sufficient to measure six stars; all these stars are cool and relatively close. Ten years later a 15 metre instrument was used [2], but few new results seem to have been published, probably because of difficulties which will be explained later. These early observations have proved to be of considerable value and it is our aim at Narrabri to extend the work to the hot stars. The most valuable application of our data will depend on their accuracy and on what we find. At present we visualize their most important application to be the improvement of the scale of temperature for hot stars of types *O*, *B* and *A*. For reasons mentioned later the interferometer is limited to hot stars of spectral type *O*, *B*, *A* and *F* and

[1] The apparent magnitude m of a star is a measure, on a logarithmic scale, of the amount of light received from it. Thus, $m = -2.5$ log(light flux received) and the zero of the scale has been arbitrarily chosen as the mean magnitude of a group of Northern stars. On this scale the brightest star, Sirius, has a magnitude of about -1.5, while the faintest star which one can see with the naked eye has a magnitude of about $+6$. In the past stars were classified on a brightness scale and this scale has been adjusted so that a ratio of 100 to 1 in light flux corresponds to a difference in magnitude of 5.

[2] F. G. Pease, Ergebnisse der exakten Naturw. 10, 84, 1931.

cannot be applied to cooler stars; furthermore, it is limited to stars brighter than apparent magnitude +2.5. It is expected that the instrument, in its present form, will be able to measure the angular diameters of about 50 hot stars.

In deriving equation (3) we have assumed that we can measure the total flux, integrated over all wavelengths, and that the star radiates like a black-body to which we can apply Stefan's law. For hot stars neither of these assumptions can be made. As we have already noted a substantial part of their light flux is absorbed by the Earth's atmosphere; furthermore, the spectral distribution of the radiation of hot stars does not follow closely that of a black body. In these cases the measurements of light flux and angular size are combined, using Planck's law, to find the equivalent black-body temperature of the star at the wavelength λ at which the light flux has been measured; this quantity is called the "*brightness temperature*" $T_b(\lambda)$ of the star. For a black body the brightness temperature and the effective temperature are obviously equal at all wavelengths. For the hot stars more sophisticated theoretical models are needed to relate them. The measured values of brightness temperature are then used to calibrate the theoretical models of hot stars to find their effective temperatures.

It is hoped in this way to find the effective temperatures of a number of hot stars of type *O*, *B* and *A* with an accuracy of better than 5% and thus to contribute significantly to our knowledge of the stellar temperature scale. In addition to these temperatures the observational programme for determining $\Theta = 2r/R$ should also yield the radii of the few hot stars for which reliable distances are known.

There are also some more specialized results which we hope to obtain. Firstly, we expect to discover that some of the well-known bright stars are unsuspected double or multiple stars. Many of the bright stars are already known to be multiple, either from direct observation of their components (visual binaries etc.) or from their spectra (spectroscopic binaries), but it is probable that the interferometer will reveal multiple stars which cannot be found by any other existing technique. Secondly, we propose to make measurements of the angular size of some stars which show bright emission lines and to check current theories of the origin of these lines. Finally, we shall try, if the equipment works sufficiently well, to measure the distribution of brightness over the disc of the bright star Sirius. This measurement, if it can be done, would give the first direct observation of the so-called "limb darkening" law for Sirius (the darkening towards the edge of the star disc), which is at present based almost entirely on theory.

Michelson's interferometer

Before describing the new instrument, let us consider the operation of the Michelson interferometer [3] and the reason why it cannot be extended to measure more stars.

Fig. 2 illustrates the principle of the interferometer. The light from the star received on the two small mirrors is directed into the focal plane of a large mirror where the two beams are brought together. The separation between the two small mirrors can be altered by the observer. When the light at the two small mirrors is mutually coherent the two beams of light interfere in the focal plane to form a pattern of alternate bright and dark bands, the "interference fringes". The "visibility" of these fringes, i.e. the depth of the modulation of the light intensity along the pattern relative to the mean light intensity, is a measure of the mutual coherence of the light at the two mirrors. When the star is not a point source the fringes belonging to different points of the source will overlap. This overlap, decreasing the visibility, depends both on the angular diameter Θ of the star and on the mirror separation d. When Θd is much smaller than the wavelength λ of the light the overlap is small and the visibility is large; but for $\Theta d \approx \lambda$, or greater, the visibility will decrease rapidly. The measurement of the angular diameter Θ of a star thus consists in observing the visibility as a function of the mirror separation d; from the value of d where the visibility is decreasing strongly: $d \approx \lambda/\Theta$, the angular diameter is obtained. An exact analysis [4] shows that the fringe visibility as a function of mirror separation is the Fourier transform of the angular intensity distribution across the stellar disc when reduced to an equivalent line source [5].

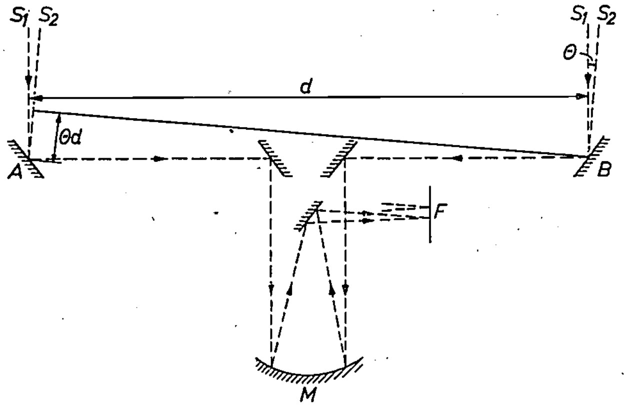


Fig. 2. The Michelson interferometer. Light from the star (or from two distant point sources S_1, S_2 separated by an angle Θ) is received at two small mirrors A and B. The two beams interfere in the focal plane F of the large mirror M. The modulation depth ("visibility") of the fringes in F is measured as a function of the mirror separation d.

The relation between visibility, mirror separation and angular diameter can readily be indicated more quantitatively for the simple case of two distant point sources S_1 and S_2 at an angular distance Θ, both emitting monochromatic light of the same wavelength λ (fig. 2). The light from S_1 only will produce fringes in the focal plane because the difference between the optical path lengths in the beams reaching a point in F by A and B respectively, depends upon the point in F considered. The centres of the bright bands of this pattern occur in points to which the optical paths differ by $n\lambda$; the dark band centres have path differences of $(n + \frac{1}{2})\lambda$. The light from S_2 will produce a second independent fringe pattern with the same band spacing. The difference of the two light paths from S_2 to a given point, however, differs by an amount Θd from the path difference for S_1. Thus the second fringe pattern is displaced by $\Theta d/\lambda$ of a fringe from the first and for $\Theta d/\lambda = 0, 1, 2, \ldots$ the visibility is a maximum, for $\Theta d/\lambda = \frac{1}{2}, \frac{3}{2}, \ldots$ it is a minimum (equal to zero when S_1 and S_2 are equally bright). Thus for two monochromatic point sources of the same wavelength the visibility is an oscillating function of Θd. When these considerations are extended to a monochromatic source of finite size results are obtained as those in *fig. 3* giving the visibility as a function of Θd for the case of an evenly radiating line source.

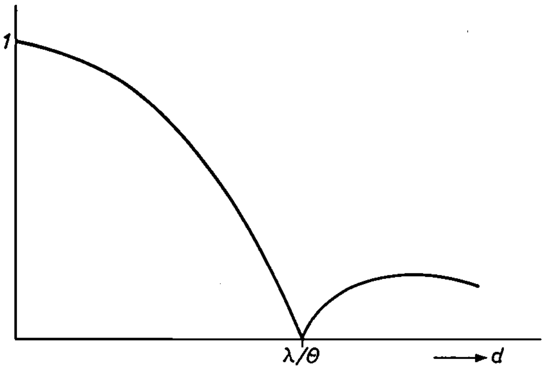


Fig. 3. The visibility of the fringe pattern in the Michelson interferometer is shown as a function of mirror separation d for an evenly radiating line source of angular size Θ emitting monochromatic light of wavelength λ. From the separation at which the visibility is zero the angular size can be derived.

The reason why the Michelson method *cannot* be used in practice with very large instruments to measure very small stars is connected with the fact that one cannot use an infinitely small frequency bandwidth. This introduces a finite critical length ΔL such that the number of wavelengths in ΔL differs by approximately one for the extreme frequencies of the band. If the path lengths from a point source to the focal plane along two beams differ by ΔL or more, these beams will no longer give rise to an interference pattern, because the fringes belonging to the individual frequencies overlap so strongly that no visibility remains. If the frequency bandwidth is $\Delta\nu$, then $\Delta L = n = (n - 1)(\lambda + \Delta\lambda)$ from which we obtain $\Delta L = \lambda^2/\Delta\lambda$ or, with $\Delta\nu/\nu = \Delta\lambda/\lambda$, $\Delta L = c/\Delta\nu$ (c = velocity of light).

For a typical bandwidth $\Delta\lambda$ of 5 nm this critical length ΔL equals 0.06 mm and the two paths from the star to the focal plane should be equal to better than, say, 0.01 mm. This raises two difficulties in extending Michelson's interferometer to larger instruments. Firstly, attaining this accuracy from the small mirrors to the focal plane sets extremely close limits to any flexure or differential expansion of the two arms of the instrument; also to equalize the paths from the star to the mirrors to the same order of accuracy requires that the instrument should be "guided" very exactly at the star. As a result it is doubtful whether a Michelson interferometer could be built with a baseline of 100 metres, at least without inordinate difficulty and expense. Secondly, as with the telescope, a major difficulty is introduced by atmospheric scintillations. Movements of the air and the resultant pressure fluctuations vary, differentially, the optical lengths of the two beams from the star to the two small mirrors and change the paths of the beams, the latter effect giving an angular, or positional, scintillation to the star. Both effects move and distort the fringe pattern in the focal plane. This movement makes the fringes difficult to observe and introduces a subjective element into the results. It is to be expected that, by the use of modern photoelectric techniques, the effects of scintillation could be reduced and the observations made more objective; nevertheless, this has not yet been demonstrated.

The intensity interferometer

We now consider the intensity interferometer itself. A simplified outline of the intensity interferometer is shown in *fig. 4*. Light from the star is received on two separated mirrors each of which concentrates the light

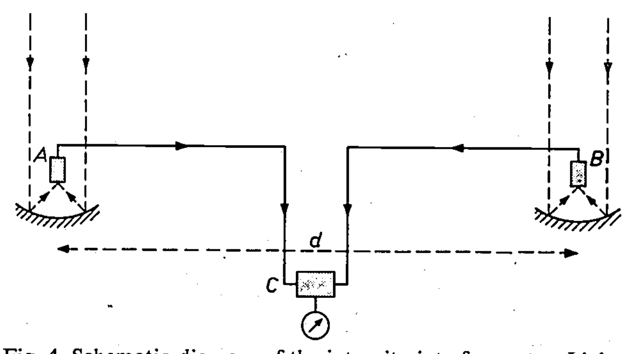


Fig. 4. Schematic diagram of the intensity interferometer. Light from the star is focused on two photoelectric detectors A and B. The correlation between the fluctuations in the output currents of A and B is measured in the correlator C, as a function of the detector separation d.

[3] C. Candler, Modern interferometers, Hilger and Watts, London 1951, p. 235.

[4] L. Mandel, Fluctuations of light beams, Progress in Optics, 181-248, 1963.

[5] "The equivalent line source" is a line, parallel to the line connecting the mirrors, for which each element has an intensity equal to the total intensity of the corresponding perpendicular strip of the disc.

on to a photoelectric detector at its focus. The output currents of these detectors contain high frequency fluctuations which are amplified and compared in an electronic correlator. The angular diameter of the star is found by measuring the correlation between these electrical fluctuations as a function of the separation between the two mirrors.

The intensity interferometer differs radically from Michelson's interferometer; the light beams received at the two mirrors are *not* brought together to produce direct optical interference but, in effect, the interference is carried out electrically. This novel technique has two valuable advantages. Firstly, it is not necessary to construct and to guide the instrument with the extreme mechanical precision which is required by a Michelson interferometer. Secondly, its operation is almost completely unaffected by atmospheric scintillation. It is therefore possible to construct an extremely large instrument which is capable of measuring the very small angles subtended by the hot stars.

A qualitative explanation

The principle of the intensity interferometer can be understood in a qualitative way as follows:

The light from a star arriving at the interferometer can be considered as the superposition of a large number of plane harmonic waves each characterized by a certain direction corresponding to one point in the light source, and a certain frequency and phase. The superposition of these waves results in light with a fluctuating intensity; that is to say, the electric vector of the light at the photocathodes has, as in the case of thermal noise, a fluctuating amplitude. Since the output currents from the detectors are proportional to the intensity of the light, the output currents will contain the same fluctuations.

In the case where the source of light is a *point source* the relative phases of all the harmonic components of the light will be the same at the two detectors, provided, of course, that the line joining the two detectors is normal to the direction of the light source. Under these conditions, it is obvious that the fluctuations in intensity at the two detectors, and therefore also in their output currents, must be completely correlated.

Let us now consider that the source of light is extended and that we can represent it by an array of point sources S_1, S_2, S_3, ... S_n, each point being separated from the next by the small angle $\Delta\Theta$; let Θ be the total angular diameter of the source (see *fig. 5*). If the line joining the two detectors is normal to the direction of S_1, then all the waves from S_1 will arrive at A and B with the same relative phases. The light from S_2, S_3, S_4 etc. however, will arrive at B *later* than at A by times $d\Delta\Theta/c$, $2d\Delta\Theta/c$ etc., so that the component waves from each of these sources have at B which are different from those at relative phases A. Thus, at any given instant they will combine to give different intensities, at the two detectors, and this will also hold for the combination of all contributions from S_1, S_2, S_3, ... S_n. The average intensities at A and B will, of course, be the same but the fluctuations of intensity will no longer be perfectly correlated.

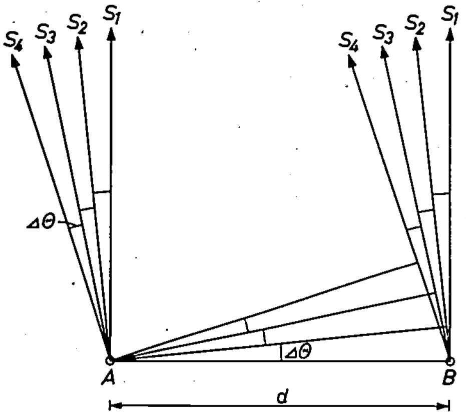


Fig. 5. The figure relates to the discussion of the fluctuations in the intensity of the light at two points A and B separated by a distance d, arriving from an array of point sources S_1, S_2, ..., separated by angles $\Delta\Theta$.

Provided that the maximum path difference Θd is small compared with the wavelength of the light, it is clear that there will be little difference between the light received at A and B. Thus, when A and B are close together ($d \ll \lambda/\Theta$) the correlation between the intensity fluctuations at the two detectors will be high; but as the baseline d increases and becomes comparable with λ/Θ the correlation will decrease and eventually will approach zero. In fact by measuring the correlation $\bar{C}(d)$ — which will be more closely defined below — as a function of d it is possible to find Θ, the angular diameter of the source of light.

The explanation given above is put in terms of light waves; however it applies to other ranges of wavelength as well, for example intensity interferometry has been used in radio astronomy.

The quantitative relationship between correlation, detector separation and angular diameter

A detailed calculation [6], which is too lengthy to be reproduced here, shows that the correlation $\bar{C}(d)$ expected when the two detectors are separated by a distance d is given by:

$$\bar{C}(d)/\bar{C}(0) = \Gamma^2(d). \quad . \quad . \quad . \quad . \quad (4)$$

where $\bar{C}(0)$ is the correlation which would be observed with the detectors close together, and $\Gamma(d)$ is the modulus of the normalized Fourier transform of the angular distribution of intensity across the light source when reduced to an equivalent line parallel to the baseline of the detectors. Thus we arrive at the strikingly simple result that the correlation observed with an intensity interferometer is proportional to the *square* of the fringe visibility in a Michelson interferometer with the same mirror separation. Thus the angular diameter of a star is determined by observing how the correlation decreases with increasing baseline and then finding the best fit with the theoretical curves of $\Gamma^2(d)$ for sources of different angular size and a given intensity distribution across the disc.

Fig. 6 shows the correlation as a function of d for two line sources corresponding to homogeneously radiating discs.

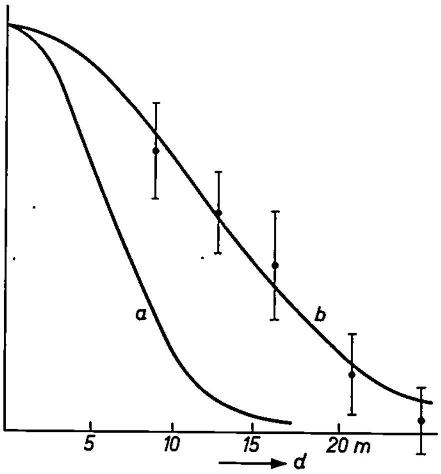


Fig. 6. The correlation as a function of detector separation d. The two curves are theoretical curves (eq. 4) for two sources corresponding to homogeneously radiating discs; *a*) disc angular size 0.008 sec; *b*) disc angular size 0.004 sec. The vertical line segments indicate typical experimental results and their probable error.

Correlation and noise

The correlation is the time-average of the cross-product of the detector current fluctuations. It is measured by applying the fluctuations, after these have been amplified and limited in frequency band, to a linear multiplier. The mean d.c. output of the multiplier is the correlation.

The correlation measurement is complicated by an unavoidable extra source of noise: the "*shot noise*" in the detector currents. Let us first consider the fluctuations in the output current of one detector. It has been shown [7] that, in the simple case where the incident light is linearly polarized and completely coherent over the surface of the detector, the mean square of these fluctuations measured at the input of the multiplying unit can be written:

$$\overline{N^2}=2e^2\left[\int_\nu A\alpha(\nu)n(\nu)\mathrm{d}\nu+\int_\nu A^2\alpha^2(\nu)n^2(\nu)d\nu\right]\int_f|F(f)|^2\mathrm{d}f, \quad . \; . \; . \; (5)$$

where e is the electronic charge, A is the area of the detector, $n(\nu)$ is the number of quanta incident on unit area in unit time and unit bandwidth, $\alpha(\nu)$ is the quantum efficiency of the photocathode, and $F(f)$ is the frequency response of the complete electrical system up to the multiplier input and including the photoelectric detectors; the respective integrations are carried out over the optical (ν) and electrical (f) bandwidths. The first of the two terms in brackets represents the shot-noise in the cathode current of the photomultiplier. The second term represents the fluctuations in the current corresponding to the fluctuations of intensity in the incident light.

Let us now consider the outputs of two identical detectors. It is clear that the shot-noise in the two detectors will be uncorrelated, but it can be shown that the fluctuations due to the intensity fluctuations of the light, represented by the second term in (5), are correlated. In the simple case when the light is completely coherent over the two detectors together — when the separation $d = 0$ —, and again linearly polarized the correlation is given by [7]:

$$\bar{C}(0) = 2e^2\int_\nu A^2\alpha^2(\nu)n^2(\nu)\mathrm{d}\nu\int_f|F(f)|^2\,\mathrm{d}f. \quad . \; . \; (6)$$

For zero separation the fluctuations in the light intensity, and thus the fluctuations in the output currents due to these, are identical for the two detectors. Therefore it is not surprising that the expression (6) is identical to the second part of (5): both expressions represent the time average of the product of two identical current fluctuations.

The signal to noise ratio

In equation (5) the second term represents the "signal", i.e. the fluctuations due to the light, which are correlated for two detectors close together; the first term represents the noise. If in the optical bandwidth the dependence of α and n upon ν can be neglected, then (5) shows that the ratio of signal power to noise power equals $A\alpha(\nu)n(\nu)$, i.e. the current signal to noise ratio is $\sqrt{A\alpha(\nu)n(\nu)}$. $A\alpha n$ is the number of photoelectrons per unit time per unit frequency bandwidth of the light. Now even for large mirrors and bright stars this number is very small (typically $\sim 10^{-4}$), and so in the output of the detectors the fluctuations due to the light are submerged in a much greater shot noise.

The output current of the multiplier is proportional to the product of the input currents. Therefore (cf. eq. 5 and 6), for the signal directly after the multiplication process, i.e. the correlation, the current signal to noise ratio now equals $A\alpha n$. However, in practice, after the

[6] R. Hanbury Brown and R. Q. Twiss, Proc. Roy. Soc. A **243**, 291, 1957.

[7] R. Hanbury Brown and R. Q. Twiss, Proc. Roy. Soc. A **242**, 300, 1957.

multiplication process the noise is considerably reduced by the time-averaging process and the current signal to noise ratio is improved by a factor $\sqrt{T\Delta f}$ where T is the total time of observation and Δf is the electrical bandwidth of all the equipment between the photocathodes and the multiplying unit.

Also the influence of the light on the signal to noise ratio is contained in the factor $A\alpha n$, which is the quantum efficiency of the photocathodes times the number of quanta received per unit time *per unit bandwidth of the light*. Thus the signal to noise ratio does not depend simply on the total light received but is also a function of the way in which the light is distributed over the optical bandwidth. In the simple case where this distribution is uniform, the signal to noise ratio is *independent* of the optical bandwidth provided only that it is much greater than the electrical bandwidth. A simplified derivation of these statements is given in the following article.

In practice, when we calculate the signal to noise ratio for an actual instrument, we find that a reasonable performance can only be achieved for even the brightest stars by using extremely large mirrors, high quantum efficiency photocathodes, wide electrical bandwidths and very long exposures. Furthermore we find that observations are restricted not only to *bright* stars but also to *hot* stars.

This last limitation is not immediately obvious and deserves a few words of explanation. In eq. (6) the correlation is given for the case where the size of the detectors, their separation and the angular diameter of the source are so small that the incident light can be considered as perfectly coherent over the surface of each detector and at the two detectors. In practice because the angular size of the source and the spacing between the detectors are not zero, the light will not be coherent over the two detectors. Furthermore the light may not be coherent over each reflector due to its size, i.e. each reflector may be large enough to partially resolve the source. Thus the correlation given by eq. (6) will be reduced by a factor which will depend on the size and spacing of the detectors and on the size and shape of the source of light. When this factor is included we find that the signal to noise ratio is a function not only of the parameters listed above and the brightness of the source, but also of its *brightness temperature*; in fact for a given time of observation there is a maximum possible signal to noise ratio, which increases as the brightness temperature increases, no matter how large the detectors are made. Thus a practical intensity interferometer is limited not only to stars *brighter* than a certain magnitude but also *hotter* than a certain minimum temperature; the exact values of these limits depending on the parameters of the instrument.

The way in which the brightness temperature enters into the signal to noise ratio, when the detectors are not small, is roughly as follows. The factor by which (6) must be reduced to obtain the correlation, if the light is not completely coherent over one detector, is determined by $\lambda/\Theta A^{1/2}$, analogous to $\lambda/\Theta d$ when the influence of the mirror separation d is considered ($A^{1/2}$ is roughly the mirror diameter). The reduction factor decreases as $\lambda/\Theta A^{1/2}$ decreases. For $\lambda/\Theta A^{1/2} \gg 1$ it equals unity (no reduction), for $\lambda/\Theta A^{1/2} \ll 1$ it can be shown [6][7] to approximate to $\lambda^2/\Theta^2 A$. In the latter case (large mirrors) the signal to noise ratio is proportional to $A\alpha n\lambda^2/\Theta^2 A = \alpha\lambda^2 n/\Theta^2$. Now $4n/\Theta^2 = nR^2/r^2$ (R distance, r radius of the star) equals the number of photons emitted per unit time per unit bandwidth *per unit area of the surface of the star* which is directly given by the brightness temperature. Thus in this case the brightness temperature and not the apparent magnitude determines the signal to noise ratio. In intermediate cases both quantities appear in the signal to noise ratio.

The tolerances in construction and alignment and the effect of atmospheric irregularities

The most significant advantages of the intensity interferometer are that it does not require precise mechanical construction or guiding and that it is substantially unaffected by atmospheric scintillation [8].

As we have seen it is essential to construct and guide Michelson's interferometer so that any difference between the light-paths through the two arms of the instrument does not exceed $c/\Delta\nu$, or, very roughly, one wavelength of the light. On the other hand in an intensity interferometer any path differences in the two arms must be small compared with the wavelength, not of the light, but of the electrical fluctuations transmitted from the photoelectric detectors to the correlator. More precisely the cross-correlation coefficient between the two sets of fluctuations produced at the two photocathodes is, as a function of relative time-delay, given by the Fourier transform of their power spectrum (Wiener-Khintchine theorem). Thus, assuming the two spectra to be identical, rectangular and extending to a maximum frequency f_{max}, the correlation will decrease considerably as soon as the difference in time delay between the two signals reaching the correlator exceeds a small fraction, say one tenth, of $1/f_{\mathrm{max}}$ seconds. In principle, as we have already noted, it is desirable to use as large an electrical bandwidth as possible since the signal to noise ratio increases as the square root of the bandwidth; but in practice an upper limit of about 100 Mc/s is set by the response of the photoelectric detectors. Hence any differential time delays must not exceed about 10^{-9} s. This limitation applies to any relative time delays in the light reaching the two detectors as well as in the paths from the detector cathodes to the correlator. For the light this means, roughly speaking, that any path differences in the beams reaching the detectors must not exceed about 30 cm in air. In practice these tolerances are relatively

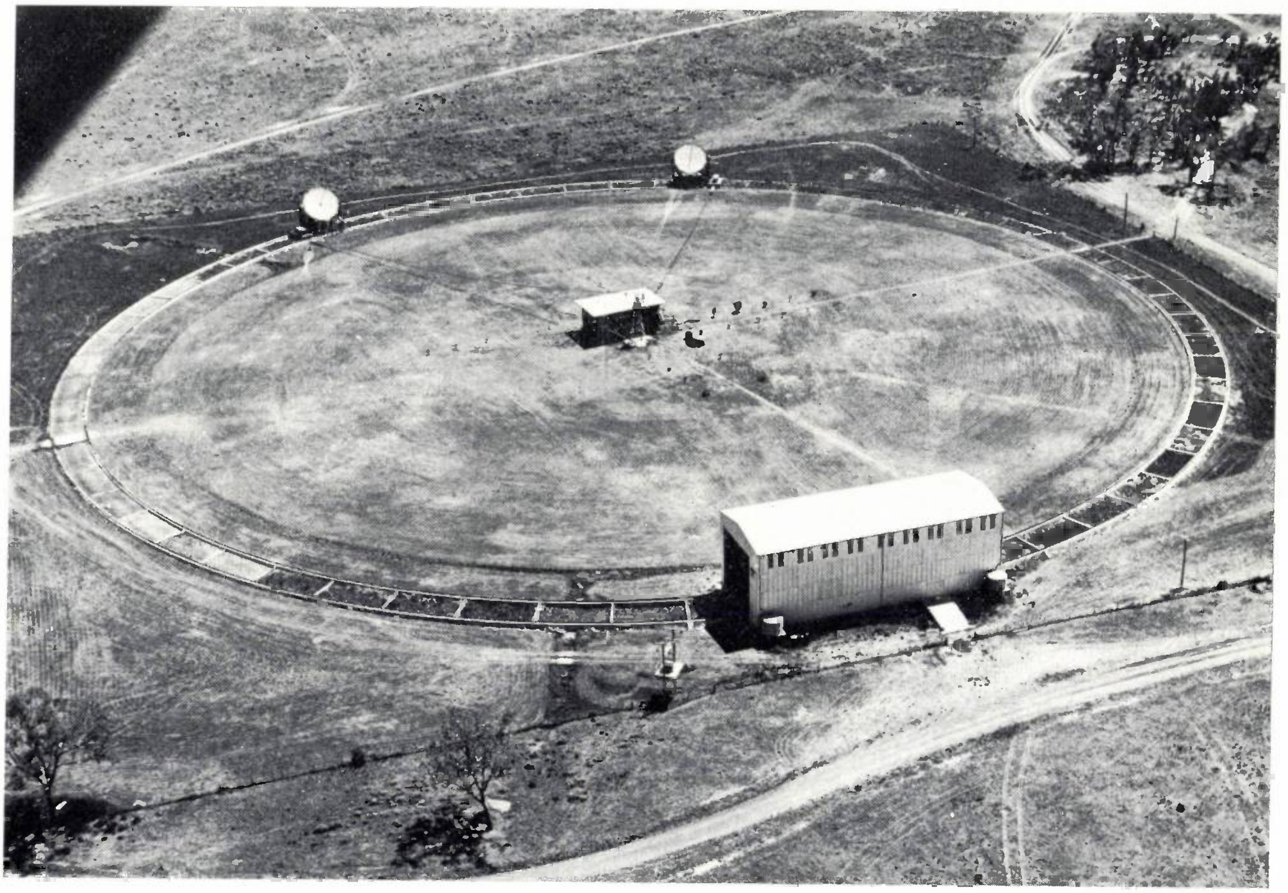


Fig. 7. The interferometer at Narrabri. The railway track has a diameter of 188 m. In the centre the control building and the cable tower, 9 m high. In the foreground is the parking garage for the reflectors.

easy to meet, even in a very large instrument with a baseline of a few hundred metres and it is therefore possible to construct an interferometer with a very high resolving power.

The second advantage of an intensity interferometer is its freedom from the adverse effects of atmospheric scintillation which are so serious in Michelson's interferometer. Briefly, the explanation of this remarkable property is contained in the previous paragraph. Thus, to affect the observed correlation significantly, the atmosphere must introduce differences in the relative time of arrival of the light at the two detectors of at least 10^{-9} s (equivalent to 30 cm in air). A detailed consideration of the types of atmospheric irregularity which are likely to exist and to be responsible for scintillation shows that the relative time delays to be expected are small compared with 10^{-9} s. Hence from a theoretical point of view it appears unlikely that atmospheric irregularities can modify the observed correlation, and this conclusion is confirmed by the limited amount of practical experience which we already have.

The interferometer at Narrabri

After a series of laboratory tests of the principle, a first attempt to make a stellar intensity interferometer was made in 1955. A pilot instrument was built, using Army searchlight mirrors with a diameter of 156 cm, and the angular size of Sirius was measured [9]. Following the experiment a larger instrument was built which is now installed and working at Narrabri Observatory. The Observatory is in flat pastoral country about 500 kilometres north of Sydney and has an elevation of about 200 metres above sea level. The site has reasonably clear weather, more than half of the nights are free from cloud, and the wind speed at night is usually less than 15 km/h. The layout of the installation is shown in *figs. 7* and *8*. The two large reflectors are mounted on trucks which run on a circular railway track with a diameter of 188 metres. The reflectors are controlled by a computer which calculates continuously the azimuth and elevation of the star. To follow the star in azimuth the reflectors move around the circle, and in

[8] R. Hanbury Brown and R. Q. Twiss, Proc. Roy. Soc. **A 248**, 199, 1958.

[9] R. Hanbury Brown and R. Q. Twiss, Nature **178**, 1046, 1956. R. Hanbury Brown and R. Q. Twiss, Proc. Roy. Soc. **A 248**, 222, 1958.

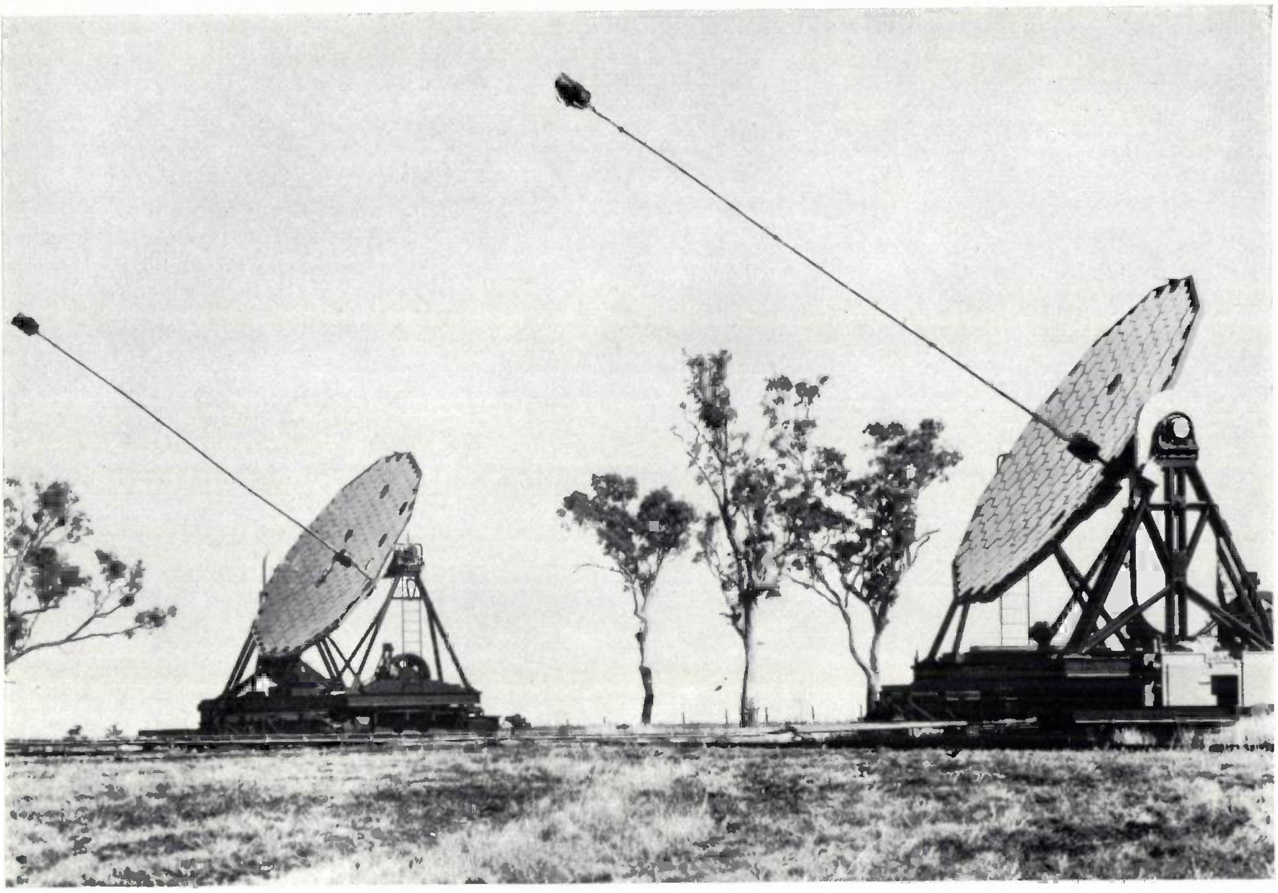


Fig. 8. The reflectors. Each reflector consists of 252 hexagonal mirrors and has a diameter of about 6.7 m. The pole carrying the photoelectric detector is 11 m long.

elevation they tilt about a horizontal axis. The separation between the reflectors can be varied from about 9 to 188 metres and they are arranged to move so that, at a constant separation, the line joining them is always at right angles to the direction of the star. This last point is an important feature of the design as it ensures that the difference in the times of arrival of the light at the two detectors is less than the limit explained above. In addition each reflector is mounted on a turntable on the truck and moves on this turntable so that it is always looking precisely in the direction of the star.

The reflectors are very large and crude by ordinary astronomical standards. Each has a light alloy framework about 6.7 metres in diameter on which 252 hexagonal mirrors are mounted. The framework is paraboloidal in shape to ensure that all of the light reaches the focus at approximately the same time, and each mirror is mounted on a three point suspension so that it can be set to reflect its beam on to the photoelectric detector. The mirrors are made of glass front-coated with aluminium and protected with silica. Each mirror has an electric heating pad cemented to it to prevent the formation of dew on the reflecting surface. Practical experience shows that 12 watts per mirror is sufficient during the winter at Narrabri. The mirrors have a spherical surface with a focal length appropriate to their position on the framework and averaging 11 metres.

At the focus the converging light is rendered parallel by a negative lens, passed through a narrow-band (8 nm) filter and focused through an iris diaphragm with a maximum aperture of 7.5 cm on to the photocathode of a 14 stage photomultiplier. The r.f. output from the photomultiplier is carried by a low loss coaxial cable to the correlator.

Although the computer is sufficiently accurate to point the reflectors at the star and keep the spot of light, which is 2.5 cm in diameter, on the photocathode, there is also a device for correcting their direction to keep the spot in the centre of the photocathode. This correction is necessary because the railway track is neither perfectly flat nor circular. The correction signal is provided by a second photomultiplier mounted alongside the main photomultiplier on each reflector. One of the 252 mirrors is set so that it reflects the light on to the centre of this guiding photomultiplier whilst the remainder illuminate the centre of the main photomultiplier. A D-shaped shutter rotates in front of the guiding photomultiplier (see *fig.* 9) and by interrupting the light produces an a.c. output from the photomultiplier. Two lamps and photo-detectors derive, from the shutter, reference waves, in quadrature, which are used to rectify the photomultiplier output synchronously and to reveal the magnitude and direction of the azimuth and elevation errors. With this correcting system the r.m.s. guiding error with the reflectors in motion is about one minute of arc corresponding to about 6 mm deviation of the light spot at the focus.

In the centre of the track is a control building which houses the computer, the correlator, motor-generators and air-conditioning plant. The trucks are connected to the equipment in the control building by cables suspended from a steel catenary wire which is itself attached to a bearing at the top of a central mast. The radial pull of this wire, roughly two tons, is taken by a tender which is towed behind each mirror truck. In the southern part of the track there is a very large garage built over the railway track in such a way that the two mirrors can be protected completely from the weather when not in use. An interesting, but expensive, feature of the design is that the trucks can be parked inside the garage without detaching any of the cables; this is achieved by slots which run from the ends almost to the centre of the garage wall on the inner side of the track.

The computer is an electro-mechanical analogue computer which solves the equation connecting azimuth, elevation, hour angle and declination. There is not space in the present article to describe it in detail and it must suffice to say that it has an accuracy of about two or three minutes of arc, and has proved to be exceptionally reliable. The correlator is described in the following article.

The whole instrument was shipped from the U.K. to Australia towards the end of 1961 and made its first measurement of a star in July and August 1963. These first tests showed that, although all parts of the equipment were functioning properly, the sensitivity was approximately one stellar magnitude less than the designed value. The instrument was designed to give an r.m.s. correlation to noise ratio of 3/1 in one hour's observation of a star with a photographic magnitude of +2.5. In its present form it will only give this performance on a star of magnitude +1.5. The reasons for this discrepancy have been investigated and it is almost certain that, by the use of improved photomultipliers, the design performance will be achieved. In the meantime the instrument is fully occupied in measuring the angular size of the brightest stars. Roughly speaking it takes 15 to 20 hours of exposure with a very clear sky to obtain a complete set of results giving the size of one star. Among the stars that have been measured are Vega [10] (type *A*0), β Crucis (type *B*1) and Altair (type *A*7). The results on β Crucis are of considerable interest by themselves; they represent the first direct measurement of the angular size of a *B* star and they involved the use of baselines up to 100 metres in length. The probable error in the measurement of the angular size of these stars is a little less than ± 5 per cent.

Finally it can be said that the first results of this novel and interesting instrument are encouraging, and we look forward to carrying out a programme of observations which we hope will make a worthwhile contribution to our knowledge of the sizes and temperatures of the hot stars.

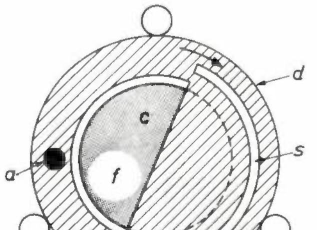


Fig. 9. Device for providing correction signals to correct the orientation of the reflectors. Light (*f*) from the star is focused by one of the 252 mirrors on to the photocathode *c* of a guiding photoelectric detector. This light is modulated by a rim driven shutter *d* if the light spot is not centred. Corrections for azimuth and elevation are distinguished by the phase of the modulation. The correction signals are obtained by synchronous rectification. The reference waves for this are obtained by interrupting two light beams (one for azimuth *a* and one for elevation *e*) passing through a slot *s* in the shutter *d*.

Acknowledgements

The interferometer was sponsored jointly by the School of Physics of Sydney University in Australia and Manchester University in the United Kingdom. It was financed jointly by the Department of Scientific and Industrial Research in the United Kingdom, the Nuclear Research Foundation within the University of Sydney, and the Office of Scientific Research of the U.S. Air Force. The design of the equipment was carried out by Messrs. Dunford and Elliott (Sheffield) and by the Mullard Research Laboratories (Redhill, Surrey).

[10] R. Hanbury Brown, C. Hazard, J. Davis and L. R. Allen, Nature **201**, 1111, 1964.

Summary. The stellar interferometer at Narrabri, Australia, has been set up to determine the angular diameters of some 50 bright stars. A knowledge of the angular diameter can be instrumental in determining the diameter and surface temperature of a star, and thus can greatly contribute to a better general understanding of the structure of stars. Up to now determinations of angular diameters have only been successful for a few stars, using Michelson's interferometer.

The intensity interferometer, which has been developed in the last 10 years, is rather similar to Michelson's interferometer. In the latter the light from the star is received at two separate plane

mirrors and the two beams are made to interfere: the visibility of the fringes decreases as the spacing of the mirrors is increased and the angular diameter can be determined from the rate of decrease of the fringe visibility. In the intensity interferometer the light from a star is received at two separate parabolic reflectors. The light is photoelectrically detected at the focus of each reflector, and the correlation between the fluctuations of the detector currents is measured. From the variations of correlation with mirror spacing it is possible to derive the angular diameter of the star. The advantage of the intensity interferometer over the Michelson interferometer is that the mechanical tolerances are very much greater; this means that a larger apparatus can be built and smaller stars can therefore be measured. Moreover, the sensitivity to atmospheric scintillations is much smaller.

The chief factor which limits the signal-to-noise ratio is shot noise in the photoelectric electrons. In each detector the fluctuations in the light intensity are very much smaller than the shot noise, which however is eliminated by the correlator. To obtain results the correlation has to be integrated over an appreciable period (minutes or hours), detectors of high sensitivity and fast response have to be used with large reflectors, and the measurements have to be limited to hot bright stars.

The reflectors of the Narrabri interferometer have a diameter of 6.7 metres, and their maximum spacing is 188 metres. Vega, β Crucis and Altair are among the stars whose angular diameters have been measured.

II. Measurement of the correlation

A. Browne

General considerations

In the previous article it has been shown that the angular diameter of a star may be obtained from the correlation between the fluctuations in the intensities of the light received at two detectors, as a function of the separation of these detectors. This article contains a description of the correlator (*fig. 10*), i.e. the electronic apparatus which measures this correlation.

A schematic outline of the apparatus is shown in *fig. 11*. The fluctuations in the output currents of the photodetectors, after being amplified and limited in frequency band, are multiplied together. By integrating the output current of the multiplying unit one obtains the average value representing the correlation. The input currents of the multiplying unit, X and Y, may be written as $X = x + z$, $Y = y + z$, where x, y and z are three independent noise currents. The components x and y are due to shot noise, and z is due to the common component in the light intensity fluctuations. The averaged product current is, because x, y and z are independent:

$$\overline{XY} = \overline{xy} + \overline{xz} + \overline{yz} + \overline{z^2} = \overline{z^2} = C\,.$$

This is the correlation, which must be measured for different separations of the detectors.

Signal to noise ratio

For an *infinite* integration time the averaging process separates the d.c. of the signal, $\overline{z^2}$, at the output of the multiplying unit, from the noise signals xy etc. because $\overline{xy}$ etc. $= 0$. For any *finite* integration time the noise is not eliminated completely and will cause an uncertainty in the measured correlation. The following simplified consideration shows how the signal to noise ratio depends on the integration time T and also on the bandwidth Δf and the mean light intensity. The component z is small with respect to x and y and the noise currents xz and yz are negligible when compared to xy. Referring to eq. (5) in the previous article, let us assume that $A\alpha n$ is independent of ν inside the optical frequency band from ν to $\nu+\Delta\nu$ and zero outside. In the same way, let F be independent of f inside the r.f. band from f to $f+\Delta f$ and zero outside. If we put $2e^2F^2\Delta\nu$ = k and $A\alpha n = n_1$, then the two terms in (5) reduce to:

$$\overline{x^2} = \overline{y^2} = kn_1\Delta f,\quad \overline{z_0{}^2} = kn_1{}^2\Delta f. \quad . \; . \quad (7)$$

The latter term represents the mean square fluctuations due to the light in one detector (second term of eq. 5) as well as the correlation for zero separation (eq. 6). The integrating (averaging) process can be considered as the selection of signals from the output of the multiplying unit with frequencies in a bandwidth $\delta f \approx 1/T$ at $f = 0$. The noise xy is distributed mainly in the frequency bands 0 to Δf and $2f$ to $2f + 2\Delta f$. The latter band is of no importance. If we now assume that the noise power, which, apart from a resistance factor, equals $\overline{(xy)^2} = \overline{x^2\,y^2} = \overline{x^2}\;\overline{y^2} = k^2n^2{}_1\Delta f^2$, is distributed evenly in the frequency band 0 to Δf, then the noise power in the band 0 to δf equals $k^2n_1{}^2\Delta f\,\delta f = k^2n_1{}^2\Delta f/T$. The correlation signal power is $(\overline{z_0{}^2})^2 = k^2n_1{}^4\Delta f^2$. Thus, for zero separation, we have:

$$\frac{\text{r.m.s. signal}}{\text{r.m.s. noise}} = \sqrt{\frac{\text{signal power}}{\text{noise power}}} = n_1\sqrt{T\Delta f}. \quad . \; . \quad (8)$$

Therefore, in a given integration time T, the signal to noise ratio increases as the electrical bandwidth Δf and the quantity n_1 (proportional to the light intensity per unit optical bandwidth, the effective cathode area and

A. Browne, B.Sc., is with Mullard Research Laboratories, Redhill, Surrey, England.

mirrors and the two beams are made to interfere: the visibility of the fringes decreases as the spacing of the mirrors is increased and the angular diameter can be determined from the rate of decrease of the fringe visibility. In the intensity interferometer the light from a star is received at two separate parabolic reflectors. The light is photoelectrically detected at the focus of each reflector, and the correlation between the fluctuations of the detector currents is measured. From the variations of correlation with mirror spacing it is possible to derive the angular diameter of the star. The advantage of the intensity interferometer over the Michelson interferometer is that the mechanical tolerances are very much greater; this means that a larger apparatus can be built and smaller stars can therefore be measured. Moreover, the sensitivity to atmospheric scintillations is much smaller.

The chief factor which limits the signal-to-noise ratio is shot noise in the photoelectric electrons. In each detector the fluctuations in the light intensity are very much smaller than the shot noise, which however is eliminated by the correlator. To obtain results the correlation has to be integrated over an appreciable period (minutes or hours), detectors of high sensitivity and fast response have to be used with large reflectors, and the measurements have to be limited to hot bright stars.

The reflectors of the Narrabri interferometer have a diameter of 6.7 metres, and their maximum spacing is 188 metres. Vega, β Crucis and Altair are among the stars whose angular diameters have been measured.

II. Measurement of the correlation

A. Browne

General considerations

In the previous article it has been shown that the angular diameter of a star may be obtained from the correlation between the fluctuations in the intensities of the light received at two detectors, as a function of the separation of these detectors. This article contains a description of the correlator (*fig. 10*), i.e. the electronic apparatus which measures this correlation.

A schematic outline of the apparatus is shown in *fig. 11*. The fluctuations in the output currents of the photodetectors, after being amplified and limited in frequency band, are multiplied together. By integrating the output current of the multiplying unit one obtains the average value representing the correlation. The input currents of the multiplying unit, X and Y, may be written as $X = x + z$, $Y = y + z$, where x, y and z are three independent noise currents. The components x and y are due to shot noise, and z is due to the common component in the light intensity fluctuations. The averaged product current is, because x, y and z are independent:

$$\overline{XY} = \overline{xy} + \overline{xz} + \overline{yz} + \overline{z^2} = \overline{z^2} = C\,.$$

This is the correlation, which must be measured for different separations of the detectors.

Signal to noise ratio

For an *infinite* integration time the averaging process separates the d.c. of the signal, $\overline{z^2}$, at the output of the multiplying unit, from the noise signals xy etc. because $\overline{xy}$ etc. $= 0$. For any *finite* integration time the noise is not eliminated completely and will cause an uncertainty in the measured correlation. The following simplified consideration shows how the signal to noise ratio depends on the integration time T and also on the bandwidth Δf and the mean light intensity. The component z is small with respect to x and y and the noise currents xz and yz are negligible when compared to xy. Referring to eq. (5) in the previous article, let us assume that Aan is independent of ν inside the optical frequency band from ν to $\nu+\Delta\nu$ and zero outside. In the same way, let F be independent of f inside the r.f. band from f to $f+\Delta f$ and zero outside. If we put $2e^2F^2\Delta\nu$ = k and $Aan = n_1$, then the two terms in (5) reduce to:

$$\overline{x^2} = \overline{y^2} = kn_1\Delta f, \quad \overline{z_0{}^2} = kn_1{}^2\Delta f. \quad . \; . \quad (7)$$

The latter term represents the mean square fluctuations due to the light in one detector (second term of eq. 5) as well as the correlation for zero separation (eq. 6). The integrating (averaging) process can be considered as the selection of signals from the output of the multiplying unit with frequencies in a bandwidth $\delta f \approx 1/T$ at $f = 0$. The noise xy is distributed mainly in the frequency bands 0 to Δf and $2f$ to $2f + 2\Delta f$. The latter band is of no importance. If we now assume that the noise power, which, apart from a resistance factor, equals $\overline{(xy)^2} = \overline{x^2\,y^2} = \overline{x^2}\;\overline{y^2} = k^2n^2{}_1\Delta f^2$, is distributed evenly in the frequency band 0 to Δf, then the noise power in the band 0 to δf equals $k^2n_1{}^2\Delta f\,\delta f = k^2n_1{}^2\Delta f/T$. The correlation signal power is $(\overline{z_0{}^2})^2 = k^2n_1{}^4\Delta f^2$. Thus, for zero separation, we have:

$$\frac{\text{r.m.s. signal}}{\text{r.m.s. noise}} = \sqrt{\frac{\text{signal power}}{\text{noise power}}} = n_1\sqrt{T\Delta f}\,. \quad . \quad (8)$$

Therefore, in a given integration time T, the signal to noise ratio increases as the electrical bandwidth Δf and the quantity n_1 (proportional to the light intensity per unit optical bandwidth, the effective cathode area and

A. Browne, B.Sc., is with Mullard Research Laboratories, Redhill, Surrey, England.

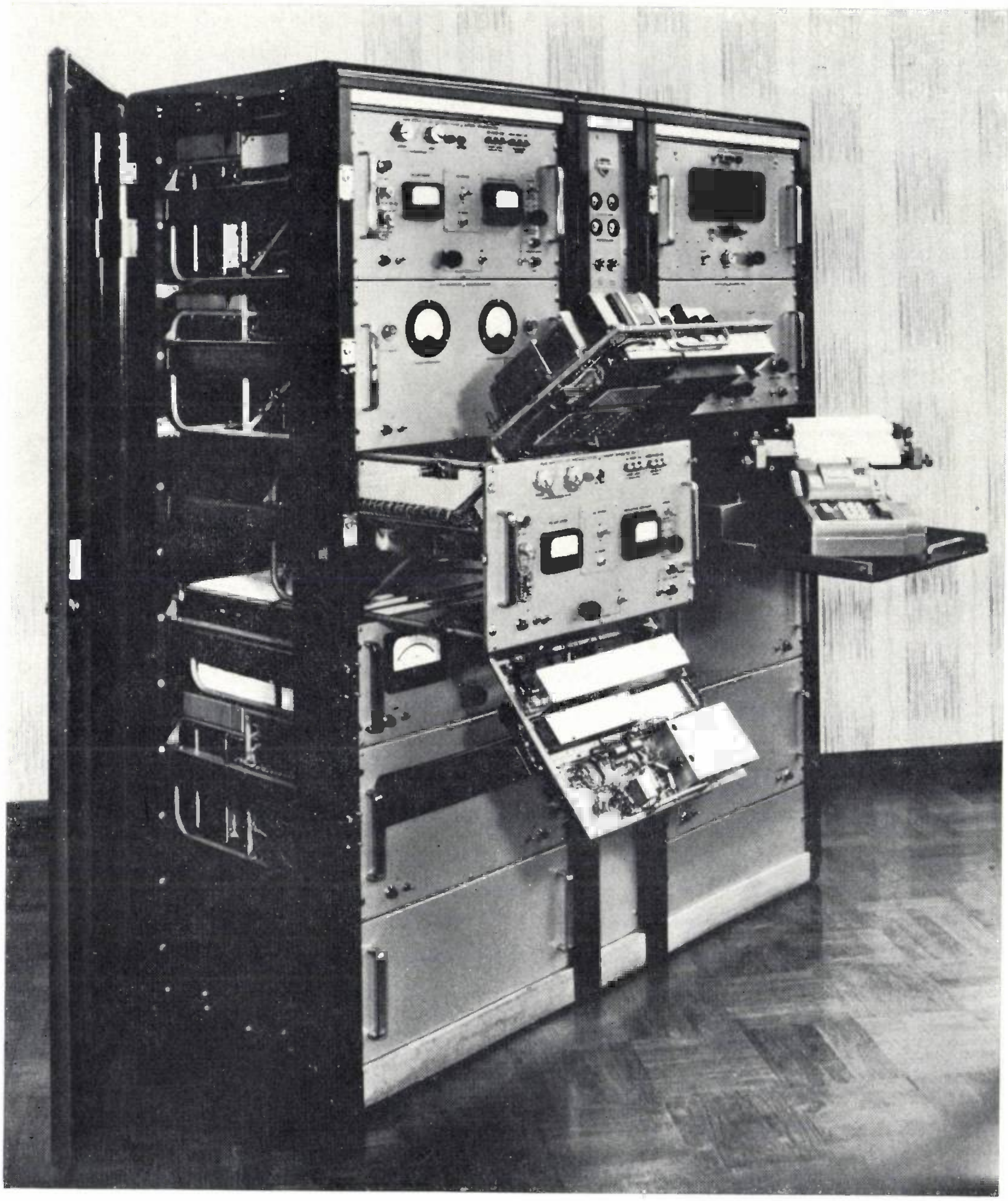


Fig. 10. The correlator. In the three upper left hand units the r.f. signals from the photomultipliers are received, processed and multiplied together. The product signal is processed further in the upper right hand units and the correlation integrals are recorded by the printer shown. The lower units contain the timing and programming units, the E.H.T. supplies for the photomultipliers and the correlator stabilized a.c. supplies.

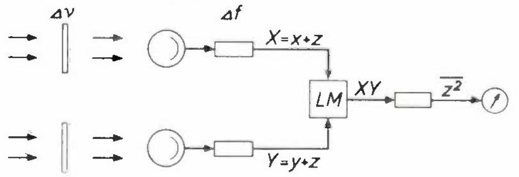


Fig. 11. Schematic outline of the interferometer. The optical bandwidth is limited to $\Delta\nu$ by an optical filter. The r.f. bandwidth of the electronic equipment between photocathodes and lincar multiplier input is Δf. The signals $X = x + z$ and $Y = y + z$ are multiplied to give XY and subsequently averaged to give $\overline{XY} = \overline{z^2}$, the correlation. x and y represent shot noise, z is the common component in the fluctuations.

the quantum efficiency) increase. Note that the optical bandwidth $\Delta\nu$ drops out of the expression.

In practice, before multiplication the signal z is around 25 to 55 dB below X and Y, thus after multiplication the signal $\overline{z^2}$ is 50 to 110 dB below XY. The correlation $\overline{z^2}$, after the multiplication, is present in the product noise XY as a direct current or, in some circuits, as a 5 kc/s alternating current. It is separated from the noise in stages by narrow band amplification, synchronous detection and integration. The integration is continued until the uncertainty due to the noise xy is significantly below the measured correlation $C = \overline{z^2}$; this may take minutes or hours.

The level of the noise signals received is high compared with the noise produced in the electronic circuits; therefore it is unnecessary to have a good noise factor in the correlator. The multiplier, which has a very poor noise factor, is preceeded by r.f. amplifiers which raise the level of the input signals. The resultant contributions from the first stages of the amplifiers and from the multiplier produce only a small degradation of the signal to noise ratio.

The integrated output is recorded every hundred seconds. *Fig. 12* shows a typical running integral of the output with the limits of the uncertainty due to the noise xy marked.

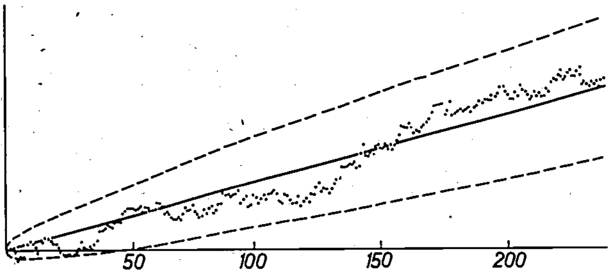


Fig. 12. Running total of the correlator output, given by the dots. The straight line is the deduced integrated correlation; the dashed curves indicate the limits of the uncertainty. The figures on the horizontal axis are numbers of hundred second periods of the observation time.

Normalization

If, during a run of measurements, the mean light intensity varies, e.g. due to changes in the atmosphere, the measured correlation, being proportional (cf. eq. 7) to the square of the light intensity (or rather: to the product of the light intensities at the two detectors) will be affected. To eliminate this effect, the correlation is normalized, i.e. the quantity $C/\overline{X^2}\,\overline{Y^2}$ is determined. This quantity is independent of the light intensities (cf. eq. 7; $X \approx x$, $Y \approx y$). The dependence upon the mirror separation is unaffected by the normalization.The quantities $\overline{X^2}$ and $\overline{Y^2}$ are amplified shot noise in the detector cathode currents. They could in principle be determined from the cathode currents and the gains and the bandwidths of the electronic sections between the photocathodes and the inputs of the linear multiplier. Provision is made for the measurement of the cathode and anode currents — and thus the gains — of the photodetectors and these currents are recorded, together with the correlation, every hundred seconds. The gains of the following sections are determined separately. In practice temperature sensitive leakage within the photodetectors makes it impossible to monitor the cathode currents over a long time. The photoelectric and leakage current can be distinguished by brief measurements with and without illumination and thus the detector gain can be determined. Assuming this gain to be constant throughout a star measurement, the correlation figures can be normalized by the anode current readings.

The E.H.T. voltages for the photomultipliers can be set independently to control the gains. The measurement of the cathode currents could in principle be used, together with the optical data of the reflector, to obtain the apparent magnitude of the star. This quantity is needed to relate the angular diameter to the temperature of the star (cf. eq. 3). At present the photocathode leakage restricts the accuracy of the measurement, and published data for the apparent magnitude of the star are being used.

Calibration

The relationship between the final integrated output and the correlation $C = \overline{z^2}$ involves many constants of the correlator. To obtain this relationship the whole correlation measuring process is calibrated in one step. To do this a noise generator is connected, instead of the photodetectors, to both inputs of the correlator, giving fully correlated noise signals of known voltage at these inputs. As the photomultiplier shot noise is excluded a correlation output far above the noise level is obtained within a short time.

Summarizing, the normalization corrects for variations in the light intensity, and the calibration corrects for gain changes within the correlator. Plots of the theoretical normalized correlation, in arbitrary units, against the reflector separation for angular diameters in the region of 0.004 and 0.008 seconds are shown in fig. 6 together with typical practical results and their probable errors. The angular diameter of the star is that corresponding to the theoretical curve which best fits the experimental results.

The correlator

5 kc/s switching

Fig. 13 shows a simplified block diagram of the correlator. The signal outputs of the photomultipliers are of the order of 1 mV r.m.s. Before applying these fluctuations to the multiplier *LM* they are amplified by 40 dB by two r.f. amplifiers A_1 and A_2. In *LM* a current proportional to the product of the instantaneous currents of the two signals is produced. The d.c. component of this product, the correlation C, may be only a millionth of the mean anode current of the multiplier. Its presence is revealed by inverting the signal from the first photomultiplier at a rate of 5 kc/s by switching the r.f. amplifier S_1. Thus for 100 μs the correlation current in the multiplier *LM* is positive and in the next 100 μs

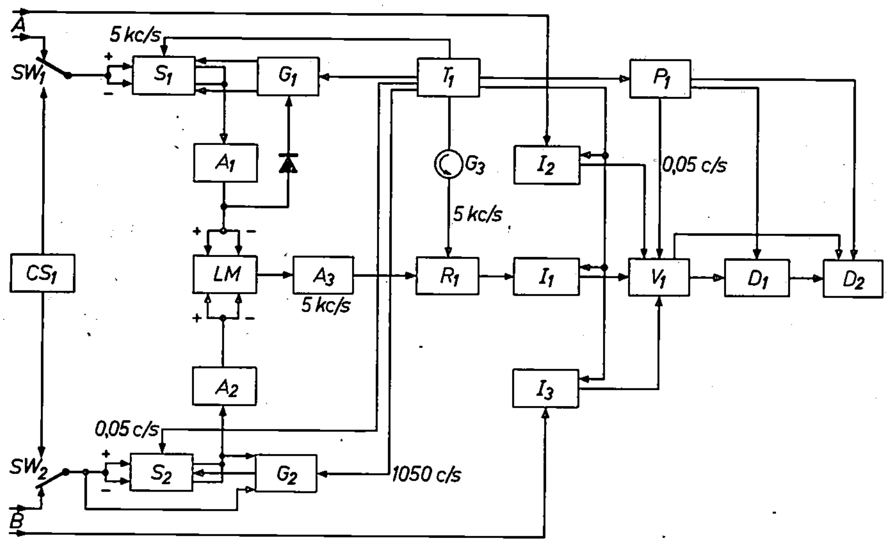


Fig. 13. The signals from the photomultipliers A and B are fed through the amplifiers S_1, A_1 and S_2, A_2 to the linear multiplier LM. S_1 and S_2 are switched at 5 kc/s and 0.05 c/s respectively; their gains are stabilized by the a.g.c. circuits G_1 and G_2. The output of LM is reduced in bandwidth by the narrow band 5 kc/s amplifier A_3 and the synchronous rectifier R_1, and integrated through 10 s in the integrator I_1. The 10 s integrals are read by the digital voltmeter V_1 (switched at 0.05 c/s) and added in the digital store D_1. One hundred second totals and a running total are printed in the printer D_2. The cathode and the anode currents of A and B are measured by the integrators I_2 and I_3; 100 s averages are read and printed by V_1 and D_2. The timing unit T_1 and the programme unit P_1 produce the required timing signals. With the goniometer G_3 the phase of the reference signal for R_1 is adjusted. The calibration source CS_1 can be connected to the correlator by the switches SW_1, SW_2.

negative, i.e. the correlation is represented by a 5 kc/s square wave. The product noise, having a wide bandwidth and containing no correlation, may be 110 dB greater than the correlation signal. The signal-to-noise ratio is increased by a narrow band 5 kc/s amplifier, A_3. The amplifier signal is reconverted to d.c. by a synchronous rectifier, R_1. In this the polarity of the signal is changed in alternate 100 μs periods once again so that the original is d.c. re-formed but now amplified. The synchronous rectification ensures that the d.c. output is derived from components of the signal at the switching frequency (or its odd harmonics) only. The transitions of the rectifier must coincide with the transitions of the signal which will have been delayed in its path from the switched r.f. amplifier S_1. The 5 kc/s switching signals for S_1 and R_1 are produced by the timing unit T_1 and a goniometer G_3 in the path to R_1. The goniometer is used to achieve the required coincidences by adjusting the phase of the switching signal to R_1.

The current from the rectifier R_1 is integrated through ten seconds by the integrator I_1 to give a total improvement of 90 dB to the signal-to-noise ratio. The integral, which is the voltage built up by the current flowing into a capacitor, is read by the digital voltmeter V_1 to give a four digit decimal number complete with polarity indication.

The inversion of the r.f. signal through S_1 is performed by splitting the signal by means of a transformer into two signals of opposite polarity which are fed through two parallel stages with a common output, and by blanking these stages alternately. The signal gains through the two parallel stages will not be equal, in general, and amplitude modulation at 5 kc/s may result. This modulation will be detected by non-linearities in the multiplier LM, resulting in a false 5 kc/s output which can be considerably greater than the 5 kc/s component from the correlation and can overload later stages. This effect is eliminated partially by balancing the multiplier valves in LM at 5 kc/s. The false signal is reduced further by detecting any 5 kc/s amplitude modulation in the output from A_1, and using the result to obtain, with an automatic gain control circuit G_1, differential d.c. biases for the two sides of S_1. The relative gains of the two sides are adjusted automatically to give the minimum modulation. The weakness in this system is that the minimum 5 kc/s output of LM and that of the detector preceding G_1 may occur for different settings of the biases in S_1. This will happen in general if a) the frequency characteristics of the two sides of S_1 are not identical, causing a slight difference in the spectra of the two output signals of S_1, and b) the frequency responses of LM and G_1 (including the detector) are not identical. Although the circuits of LM and G_1 are designed to give good similarity in this respect, these frequency effects allow a reduction of the modulation as detected by LM to only 0.05% instead of 0.005% that otherwise could be attained.

0.05 c/s switching

The remaining 5 kc/s signal due to the unbalanced gains passes through the system to give a false output from the voltmeter. Added to this are other false outputs from 5 kc/s signals picked up from the 5 kc/s switching leads by the early stages of the high gain 5 kc/s amplifier A_3, in spite of the use of doubly screened cables, and errors from the synchronous rectifier R_1, the integrator I_1, and the digital voltmeter V_1. Again an inversion technique is used to distinguish the true from the false voltmeter outputs. The r.f. signal from the second photomultiplier is passed through a switched amplifier, S_2, similar to that handling the first input but operating at 0.05 c/s. The signal is inverted in alternate ten second periods corresponding to the periods of the correlation current integrator. This changes the sign of the correlation current produced in the multiplier in alternate periods but does not invert the sign of the false signals. The synchronous rectification is performed by changing the sign of alternate outputs from the digital voltmeter V_1, when these are added in a digital store, D_1. The correlation components add together but the false outputs, being almost constant, approximate to zero for an even number of ten second periods. The summation is made through a hundred seconds period and the final total printed in the printed D_2. As a guide to the progress of the measurement the printer records a running total of these one hundred seconds totals which can be displayed in a graph as in fig. 12.

The false output due to the amplitude modulation increases with increase of the r.f. level. The other false outputs will have less, if any, dependence upon the level of the r.f. signal, but they may show slow drifts due to circuit changes. Random variations in the false outputs will extend the time required to obtain a given accuracy in the final result.

As in S_1, differences in gain between the two sides of S_2 modulate the output, in this case at 0.05 c/s. If we represent, as before, the signal from S_1 and A_1 by X, and that from S_2 and A_2 by Y, then in the multiplier XY is produced by the required multiplication and X and Y by direct amplification. Due to non-linearities in the multiplier, terms such as X^2, X^3 etc. and Y^2, Y^3 etc. are produced which form cross-products XY^2, X^2Y, X^2Y^2 etc. The even powers of X, together with the amplitude modulation of X, give rise to the false 5 kc/s signal, which is eliminated by the 0.05 kc/s switching. However, some terms, e.g. X^2Y^2, give false 5 kc/s signals modulated at 0.05 c/s. These signals do not cancel completely when alternate digital voltmeter outputs are added. They give an error in the correlation reading proportional to the product of the two modulations as seen by the multiplier. The separation by a dummy run of the true signal from this false signal is explained later.

Due to the low switching frequency it is not feasible to apply the same principle as was used in the other channel at 5 kc/s, i.e. to examine the envelope of the r.f. for modulation, and, if a technique had been devised, the time required to detect and reduce the modulation would be excessive. The method used to reduce the 0.05 c/s modulation is to compare the input and output of S_2, and, by adjusting the gain of the side that is in use, to keep the signals in a definite ratio, i.e. keep the gain of the stage constant. This applies to *both* sides of S_2, hence their gains are equalized to remove the 0.05 c/s modulation. The gain control circuit, G_2, corrects the gains within a time which is very short compared with the switching period.

The system should be able to hold the effective gains equal to 0.02% but the differences between the frequency responses of the switching stages and between the frequency responses of LM and G_2 have the same effects as those in the 5 kc/s system. The balance is disturbed further by crosstalk in G_2. For ten seconds the cross-talk is between signals of the same polarity tending to increase each other while in the next ten seconds the signals are of opposite polarity and the tendency is to reduce each other. This crosstalk shifts the bias in opposite directions for the two periods.

Photomultiplier current integrators

With the integrators I_2 and I_3 the 100 second averages of the anode and the cathode currents of the two photomultipliers are determined. These values were meant to be used in the calculation of the light flux from the star and also to give the levels of the signals, from the photomultipliers, with which the correlation outputs are normalized. As stated previously, photocathode leakage has made it impossible to take full advantage of this facility. The anode currents, in the region of 100 μA, are attenuated and fed to integrators capable of measuring hundred second current averages up to 1 μA to an accuracy of 1%. The cathode currents are in the range 10 to 3000 pA and are measured to 1%, except for the smallest currents, by integrators each with three sensitivity ranges. The outputs from the integrators are read by the digital voltmeter, V_1, and the result is printed alongside the total correlation for the same period and the running total of the correlation readings.

Sequence control

A timing unit, T_1, produces all of the frequencies and timing pulses required in the correlator. The integrators require precise timing and have their relays controlled by the timing unit. The relays operating the subsequent data handling units are controlled by the program unit, P_1, originally a mechanical sequence timer but now electronic, under the direction of the timing unit.

False signals

The effects of false signals which enter the system *after* the switching stages are reduced considerably by the 5 kc/s and 0.05 c/s switching. Signals entering both channels *before* the switching stages and having some correlation are indistinguishable from the signal obtained from the star. There are two main sources of these unwanted signals. One is crosstalk between the cables which allows the photomultiplier noise signal in one cable to enter the other. The cables from the reflectors to the tower at the centre of the circle — which have a helical membrane dielectric to give low loss — have a solid aluminium outer conductor, which eliminates the possibility of signal entering or leaving the cables through the outer. At each end it is necessary to use flexible cable and here signal can enter through the braid especially at the high frequencies. The reflectors are never close enough to allow the crosstalk to be effective at this end. On the tower where the cables have to be close together it has been necessary to feed them through separate copper pipes or to add an extra outer braid.

The second source of unwanted signals is radiation from local radio stations whose transmissions are picked up on these feeder cables. The screening mentioned above gives some improvement. Also, as the frequencies of the radio stations are below 10 Mc/s, it has been possible to reduce their effect further by raising the low frequency cut-off of the equipment from the original 5 Mc/s to 15 Mc/s.

In spite of these precautions some false correlation remains. The resulting correlation output is measured by performing a dummy run in which the light input from the star is blocked by mechanical shutters and each photomultiplier is illuminated by a lamp whose brightness is set to give the original anode current. The r.f. conditions are now exactly the same as those existing during the star measurement except that the only correlation that exists now is that of the false signals derived from crosstalk and radio signal pick-up, and of the false signal produced in the correlator itself, which has been discussed before. The result of this run is subtracted from the measurement of the star to leave a result due to the star alone.

Tests and adjustments

Calibration

When the equipment is to be calibrated, the coaxial switches SW_1 and SW_2 disconnect the photomultipliers and connect an adjustable noise diode, CS_1, acting as a wideband noise generator, to the inputs of the switched r.f. amplifiers S_1 and S_2. Long cables of equal length are used between diode and amplifiers so that a signal, e.g. noise, generated in the input of one switched amplifier and independent of the calibration source, is delayed in its path through the diode to the other amplifier in the other channel. This eliminates correlation due to cross talk of this kind and no false output is produced. Using this method of calibration, allowance must be made for the frequency response of the cables normally connecting the photomultipliers to the correlator. These cables can be included by bringing the reflectors together and connecting the noise diode to the photomultiplier ends of the cables.

Phase measurement

The correlation between two sine-wave signals of the same frequency but of different phase, varies as the cosine of the phase difference. Therefore if, for any frequencies within the band of the equipment, 15 Mc/s to 120 Mc/s, there exist *differences* in the phase shifts occurring in the two paths between the star and the multiplier, then the correlation will be reduced. In the optical section this refers to the phase shifts of the light intensity fluctuations, not to those at the optical frequencies. As has been described, the reflectors are positioned to equalize the optical path lengths. All corresponding cables in the two paths are arranged to have the same electrical length, i.e. to give the same delay to the signal, and the r.f. amplifiers are designed to give the least possible difference in the phase shifts. The delays in the photomultipliers, at the present about 70 ns, are affected by the E.H.T. voltage and when the gains have been set compensating cables are inserted to equalize the delays. If it is found that the differential phase shift involves a component in which the phase shift is proportional to frequency, this can be eliminated by inserting an appropriate length of cable in one channel.

A unit is provided to measure the phase shifts involved (*fig. 14*). One output of a 3 Gc/s oscillator, O_1, is shifted in frequency by 5 kc/s in a single sideband modulator, M_2. In this modulator a 5 kc/s signal from the timing unit T_1 amplitude-modulates the 3 Gc/s signal to produce sidebands 5 kc/s above and below the carrier. The carrier and one sideband are suppressed leaving a signal 5 kc/s away from the original frequency. This and an output taken directly from the 3 Gc/s oscillator are mixed independently with the outputs from a second oscillator O_2, variable from 3 Gc/s to 3.15 Gc/s. The outputs from the mixers, M_3 and M_4, are two signals in the band 0-150 Mc/s, determined by the setting of the second oscillator, and separated in frequency by 5 kc/s. The switching of the two r.f. amplifiers, S_1 and S_2, is stopped and these two signals are fed into the points P_1 and P_2 in the two paths to the multiplier, LM, between which the differential phase shift is to be measured. When the signals cross-multiply in the multiplier a signal at the difference frequency of 5 kc/s is produced and is converted to d.c. in the synchronous rectifier R_1. This d.c. is displayed on a meter. The goniometer is adjusted to reduce the current to zero. The goniometer reading is now a direct measure of the phase of the 5 kc/s signal from the 5 kc/s amplifier A_3, with respect to the timing unit T_1. The signals to the goniometer and the single sideband modulator have a common source, thus the setting of the goniometer is related to a) the phase relationship between the input to M_2 and the beat between the mixer outputs if they were brought together at P_1 and P_2,

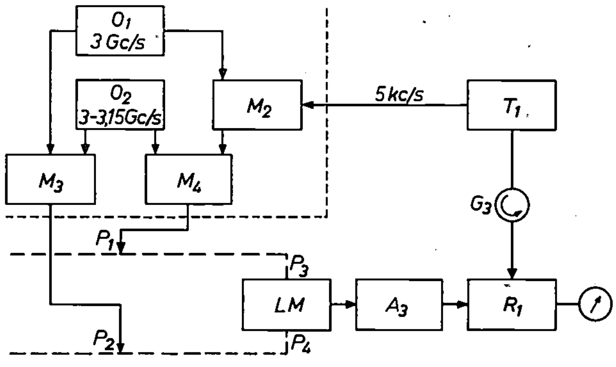


Fig. 14. Block diagram of the phase measuring unit, which is used to check for the presence of undue differential phase shifts in the correlator.

b) the differential phase shift of the r.f. signals along the two paths P_1P_3 and P_2P_4 — which appears as an equal phase shift of the 5 kc/s component of the product signal product in *LM* — and *c*) 5 kc/s phase shifts in the multiplier and the 5 kc/s amplifiers. The connections to the two inspection points are interchanged and the goniometer reset. By taking the difference between the goniometer readings, items (a) and (c) are eliminated: the difference is equal to twice the differential phase shift in (b). By starting with the connections P_1 and P_2 at the multiplier and working back to the photomultipliers it is possible to measure the differential phase shifts due to various sections of the two paths. This is repeated throughout the band of the equipment. It has been stated that the differential phase shift should not be greater than the effect of a 30 cm path in air at the highest frequency of the correlator frequency band, i.e. 40°. In fact the differential delay is less than 20° at the highest frequencies and less than 10° elsewhere.

Frequency dependence of the correlation response

Another measurement of importance is that of the dependence of the correlation response upon frequency. It is from this that the function $F(f)$ of equations (5) and (6) is derived and from these theoretical values for the normalized correlation at zero separation of the detectors and the noise in the correlator output can be calculated and compared with the practical results (cf. fig. 6). To obtain this frequency dependence the output from a signal generator is split into two nearly equal signals and fed into the photomultiplier ends of the main signal cables. The correlation output for each frequency depends upon the gain of the system and the differential phase shift existing at that frequency. With an r.f. input of constant amplitude the output is recorded for frequencies through the band; a typical response is shown in *fig. 15*.

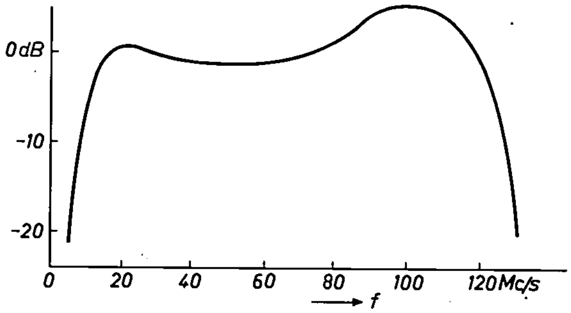


Fig. 15. Frequency response of the correlator including the cables from the photomultipliers. The response is measured in dB relative to the response at 30 Mc/s.

Summary of the measuring procedures

Let us summarize the proceedings which are followed when a star is being measured.

To start with, the equipment, including the photomultipliers, is set to the gains required for the star to be measured. Before the sky is dark, the equipment is allowed a warm-up period of about an hour, during which each photomultiplier is illuminated by the lamp which simulates the light expected from the star. In this period the reflectors are positioned on the track at the separation required, and locked on to the star.

During one night correlation measurements may be carried out at a number of reflector separations. At each reflector separation runs with the photomultipliers illuminated by the star are alternated with dummy runs in which the photomultipliers are illuminated by lamps simulating the star light. A run consists of recording the one hundred second integrals of the correlation and the running total during ten or twenty one hundred second periods. From the graphs of the running totals it is possible to see the significance of the dummy runs and to estimate the total period required to obtain a given signal to noise ratio. The correlation result of the star is corrected with the dummy run result and if necessary the net result is normalized. As an additional check on the functioning of the equipment, the noise level and the normalized correlation for zero separation may be compared with the values calculated theoretically.

The apparatus is calibrated at the end of the warm-up period and after each set of measurements performed at one reflector separation. During a calibration run the noise diode, connected to the switched amplifiers, is set so that a clear correlation result is obtained in a few one hundred second periods.

When a new star is to be measured, the first correlation measurement is made at the minimum separation of the reflectors. The measurement is repeated at other separations until the shape of the correlation separation graph is defined sufficiently [11].

Future development

As mentioned in the first article, the instrument gives a signal-to-noise ratio of 3 : 1 in one hour's observation of a star of magnitude +1.5. With a very clear sky a complete set of results giving the size of a bright star is obtained in roughly 15 to 20 hours.

To increase the scope of the instrument, using the given pair of reflectors, so that weaker stars can be measured with observation times in the same range, one must a) increase the signal-to-noise ratio and b) reduce further the false correlation outputs.

From eq. (8) it follows that to increase the signal-to-noise ratio for a given integration time one should in-

[11] After some modifications to the equipment, in particular to those the multiplying unit (cf. the last section), the general stability and freedom from drift have become sufficiently good to make possible the following simplifications to the measuring procedures. The correlator is left running during the entire observation period of about one month (a period between two full moons): in the suitable hours of the night a star is measured continuously, uninterrupted by dummy runs; during the remainder of the time a dummy run is made continuously. A three day running mean of the false correlation output, as obtained by the dummy run, is used to correct the observed correlation from the star. The mean drift over a three day period is usually less than 5% of the correlation for zero separation. The calibration has been reduced to a measurement at the beginning and the end of each night.

crease n_1 and Δf. (n_1 is the number of effective photoelectrons per unit time and per unit optical bandwidth.) For a given star and given reflectors the increase of n_1 is entirely a matter of obtaining better photomultipliers having a higher quantum efficiency (the number of electrons emitted divided by the number of photons received) and a higher collection efficiency (the number of electrons collected by the first dynode divided by the number of electrons emitted from the cathode). The major obstacles in increasing Δf (the bandwidth of the r.f. signals between photodetection and multiplication) are with the photomultiplier and the multiplying unit. Photomultipliers having greater bandwidths will be needed. In the multiplying unit in its original form (a circuit involving dual control-grid pentodes) it is not possible to increase the top frequency. Recently this has been replaced by a transistorized version giving greater scope for improvement. The change was provoked by the tendency of the pentodes to oscillate at high frequencies.

One main cause of a false correlation output is the non-linearity of the multiplying unit, combined with the remanent modulation of the r.f. signals caused by the non-perfect equalization of the gains of the two sides of each switched amplifier. It may be possible to reduce the residual modulation by obtaining a more accurate equalization of the frequency responses of the two sides of each switched amplifier so that the automatic gain controls will be more effective.

Another main cause of a false correlation output is the pick up of signals on, and the crosstalk between, the cables linking the photomultipliers and the correlator. This has been reduced by extensive double screening of the cables to about a tenth of the highest correlation reading obtained from a star. In principle this cause of trouble might be eliminated by introducing the switching stages before the cables, i.e. mounting them in the photomultiplier boxes. This has been done experimentally in the channel switched at 0.05 c/s. Two problems occur in this modification. In the first place the photomultiplier box is not temperature controlled, as is the main cabinet, and with the large changes in ambient temperature that occur, the gain of the switching stage, if it is an active device, is not stable. In the second place the major part of the A.G.C. system should also be fitted into the photomultiplier box. Both these difficulties have been overcome, in the case of the 0,05 c/s switching, by switching the signal not electronically but with the transformer that splits the signal into two signals of opposite polarities and a relay switch only in the photomultiplier box. The gains along the two paths are then fixed and any difference can be corrected with a preset correction circuit. The relay system of switching, being inherently rather slow, cannot be used in the 5 kc/s channel. Moreover it has the disadvantage of having a rather limited lifetime.

Apart from the signal-to-noise ratio and the elimination of false outputs, the general reliability of the instrument is a matter of concern. In this respect the program unit in its original form was not satisfactory. It consisted of micro-switches operated by a system of cams. Failures occurred which are believed to be due to wear in the actuating rollers after approximately a million operations. The unit has been replaced by an electronic unit.

From the above considerations it may be concluded that the most important limiting factors for the scope of the interferometer with the given pair of reflectors are the sensitivity and the bandwidth of the photomultipliers, and the bandwidth and the non-linearity of the multiplying unit.

The instrument is flexible in design, so that it is possible to incorporate new devices appearing to be desirable according to the experience gained. Recent changes include the introduction of improved photomultipliers and the transistorized multiplier. In the future these are liable to further improvement so that, with an extension of the transistorization of the equipment, greater accuracy and reliability will be obtained.

Summary. In the determination of the angular diameter of a star with the Narrabri stellar interferometer the light from the star is focused by means of two parabolic reflectors on to two photomultipliers and the correlation between the fluctuations in the output currents is measured. This measurement is performed by the correlator, which was developed and constructed at the Mullard Research Laboratories, Redhill and forms the subject of this article. The correlation is the time-average of the product of the current fluctuations. These are applied to a linear multiplier. The very small d.c. output current which represents the correlation is converted to a 5 kc/s square wave current by periodically switching the polarity at one input. Narrow-band amplification, synchronous detection, and integration (over several minutes to several hours) are applied to separate the correlation from the noise, which arises chiefly from shot noise in the photomultipliers. False 5 kc/s signals are separated from the correlation by a 0.05 kc/s inversion applied at the other input of the linear multiplier circuit. In spite of these measures, the correlator can still give a false output not due to the correlation. Moreover, crosstalk between the feed cables and pick-up of signals from near-by broadcast transmitters can cause correlated signals having nothing to do with the star. The total false result is determined by a measurement in which the star is simulated at each multiplier by means of a lamp. The result of this last measurement is used to correct the results obtained with the star. For given reflectors the factors which limit the range of measurement of the interferometer are mainly to be found in the photomultipliers and the linear multiplier.

The relation between enzyme, substrate and product

L. A. Æ. Sluyterman

577.15

The object of the investigation described in the following article is to gain a better insight into the catalytic action of an enzyme on its substrate. The investigation is still in an initial phase, in which an attempt is being made to crystallize an enzyme and a substrate in the form of a complex. The approach to this problem follows logically from a particular relation between enzyme, substrate and product, assumed on theoretical grounds, and confirmed by experiments on the enzyme, substrate and product chosen for investigation.

Introduction

The structure of an enzyme

Enzymes are the catalysts of the living cell. They make it possible for a wide variety of organic reactions to take place under the very mild conditions prevailing in the living cell.

Enzymes are proteins with a high molecular weight. This may at the very least be 10 000, but in most cases it is nearer 100 000 and sometimes considerably more. Their basic structure, however, is very simple. It is known that proteins are composed of some twenty kinds of amino acid having the general formula:

$$H_2N - \overset{\displaystyle R}{\underset{\displaystyle H}{\overset{|}{\underset{|}{C}}}} - COOH.$$

By the repetition of what are called peptide bonds between the NH_2 group of one amino acid and the COOH group of the next, as illustrated in *fig. 1*, a main chain is formed which is long and flexible and has side chains at regular intervals (the R-groups of the amino acids). A side chain can be a non-reactive paraffin chain, but most contain a functional group with specific chemical properties, e.g. of an acid or base. The presence of acid or base groups in the side chains has an important bearing on the state of charge of the protein; depending on the pH of the protein's evironment, these groups have a negative or positive charge.

The activity of an enzyme depends lastly on its *spatial* structure. As a rule this structure is fairly compact, and is brought about by the folding or coiling of the main chain. It is a rather delicate structure which is stable only under definite conditions, e.g. provided that the pH remains within certain limits. To understand this one need only think of the influence of the positive and negative groups in the molecule on the folding of the chain, bearing in mind that the distribution of these groups can be radically altered

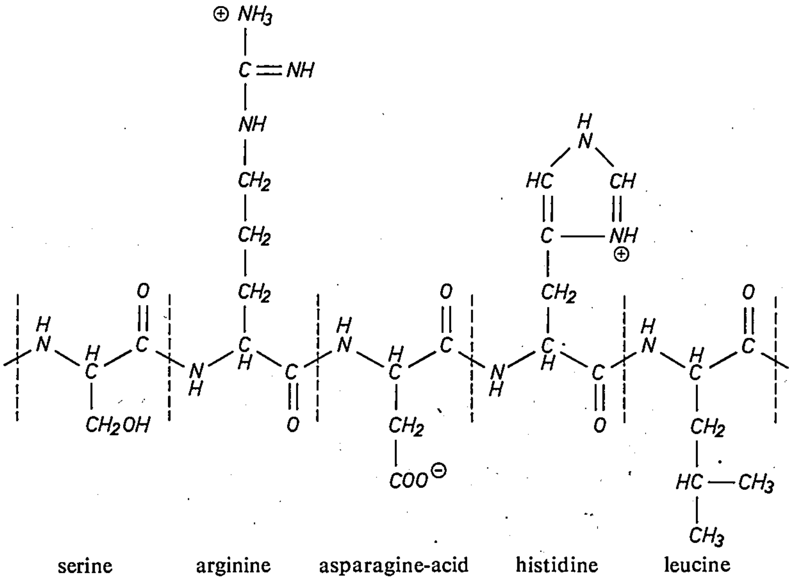


Fig. 1. Schematic representation of a fragment of a protein molecule with main and side chains. In the unfolded state, as shown here, the main chain is a zig-zag succession of regularly recurring NH-CH-CO-units. The side chains stem from the C atom indicated by bold-face type C. The fragment of a protein molecule shown here is composed of the amino acids serine, arginine, asparagine, histidine and leucine.

Dr. L. A. Æ. Sluyterman is with Philips Research Laboratories, Eindhoven.

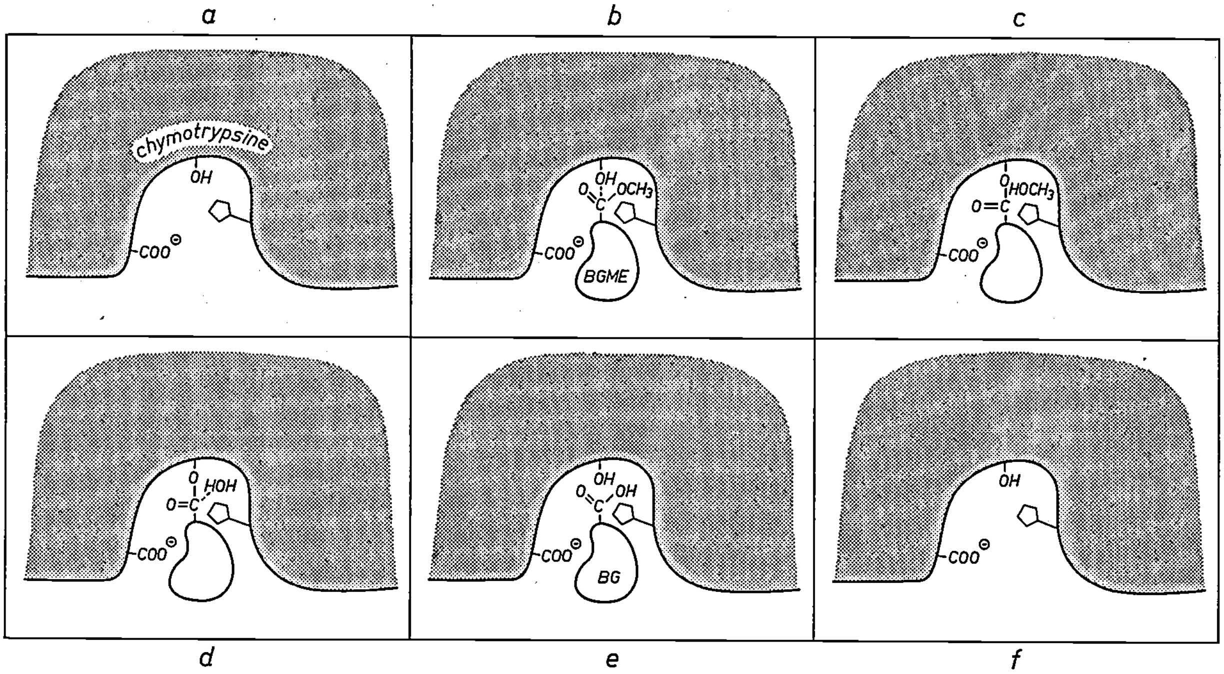


Fig. 2. Schematic and hypothetical representation of the way in which the enzyme chymotrypsin converts the substrate benzoylglycine methyl ester (BGME) into benzoylglycine (BG) and methyl alcohol. In the cavity (the active centre), a COO— group can be seen on the left, an OH group at the top and an imidazole group on the right. (The groups present in an active centre need not belong to successive amino acids, as might be thought. As the protein chain is in a folded condition, certain amino acids which would be quite a long way from each other in the unfolded condition of the chain may now be quite close to each other. This is the case here with the OH and imidazole groups.) *a*) Enzyme with "empty" active centre. *b*) The enzyme has bound a substrate molecule. *c*) Due to the action of the imidazole group the ester bond (C-O) has been broken and a new ester bond has been formed with the OH group, resulting in methyl alcohol ($HOCH_3$). *d*) The methyl alcohol molecule has made way for a water molecule (HOH). *e*) Due to the action of the imidazole group the ester bond has again been broken, resulting now in the formation of benzoylglycine. *f*) The enzyme has returned to its starting point *a*).

by changing the *p*H. Generally speaking it is therefore necessary to be very careful about changing the normal environment of the enzyme, otherwise the spatial structure may be destroyed and the enzyme "denatured".

The action of an enzyme

The action of an enzyme can be described in the simplest case by the equation

$$E + S \rightleftarrows ES \rightarrow E + P + P' + \dots$$

A molecule of the enzyme E links up with a molecule of the compound to be acted on, the substrate S, and forms with it an enzyme-substrate complex ES. This complex can either decompose into the original components, or — and in this case the catalysis is completed — it can decompose into the enzyme and the reaction products P + P', and so on.

Enzymes are characterized by their high specificity to a greater extent than ordinary catalysts. For example, the enzyme urease hydrolizes urea, but not the related compounds methylurea and thio-urea; and the proteolytic enzymes trypsin and chymotrypsin both break peptide bonds in a protein, but they do so at different sites in the molecule. Quite small changes in the spatial structure of the substrate are enough to stop the formation of the complex and thus to inhibit the catalytic action completely.

Most substrate molecules are small compared with the enzyme molecules. In the formation of a complex, therefore, only a small part of the enzyme molecule will be directly involved in the bond with the substrate molecule. This part of the enzyme molecule is called the *active centre*. The role played by the rest of the enzyme molecule is at present still uncertain; it may possibly have a stabilizing function.

The object of our investigation is to learn more about the action of enzymes; we want to obtain a more detailed and a three-dimensional picture of this process.

At the present stage we are primarily interested in what takes place in the active centre. Some insight into the processes concerned has already been gained in a few cases. As an example we give in *fig. 2* a schematic representation of a possible way in which the hydrolysis of benzoylglycine methyl ester by chymotrypsin can take place. We have already mentioned chymotrypsin as an example of a proteolytic enzyme (protease). The fact that it is also capable of hydrolyzing certain esters is not so surprising: the ester bond and the peptide bond show much resemblance and many enzymes are able to break both kinds of bond. A COO-group in the active centre of chymotrypsin has been found to be essential to the formation of the complex. Once the complex has been formed, an imidazole group in the centre exerts a catalytic action on the breakdown of the ester bond. Methyl alcohol is formed from the detached CH_3O group, while the other part separated from the substrate molecule temporarily attaches itself to an OH group present in the active centre. The last step in this process is straightforward hydrolysis, in which benzoylglycine is formed and at the same time the active centre becomes available again for a new reaction.

Although the way in which a reaction of this type takes place is broadly known, we would, for example, like to know more about the way in which the substrate binds with the enzyme, and in particular, how the reaction proceeds in three dimensions.

In order to study enzymic reactions in such detail it is desirable to have a very precise knowledge of the spatial structure of the active centre and of the substrate attached to it. The only method of acquiring this knowledge is by X-ray analysis, and so it became the first objective of our investigation to obtain a crystalline ES complex. Our attempts so far have not yet met with any success. Meanwhile Japanese investigators have been the first to report the crystallization of an ES complex [1]. The molecular weight of this complex, however, is too high for an investigation by X-ray analysis to have any chance of succeeding at the present time.

The following article is a report on the experimental and other considerations that *preceded* our attempts at crystallization [2]. Some account will be given of the kinetic methods which are often used in the study of enzymes. What is perhaps of greater interest, however, is the insight afforded into the characteristic relation that can exist between enzyme, substrate and product.

The interaction of enzyme and substrate

We decided to make the enzyme *papain* the subject of our investigations. Papain appeared suitable as it has a relatively low molecular weight (22 000), it is a fairly stable enzyme, and it can be prepared in large quantities with a high degree of purity. Like chymotrypsin, papain can hydrolyze both proteins and simple esters. The substrate we chose was the ethyl ester of benzoylarginine, which can easily be synthesized. The structure of this compound and the equation of the hydrolysis reaction are represented in *fig. 3*.

In practice the combination of this enzyme and substrate raises an important general problem, but we shall leave the discussion of this problem to the next section, in which we consider the relation between enzyme, substrate and product. We shall first discuss the determination of the "associative tendency" of enzyme and substrate and then the way in which it is affected by the pH. These data often make it possible to draw conclusions about the presence of certain functional groups in the active centre.

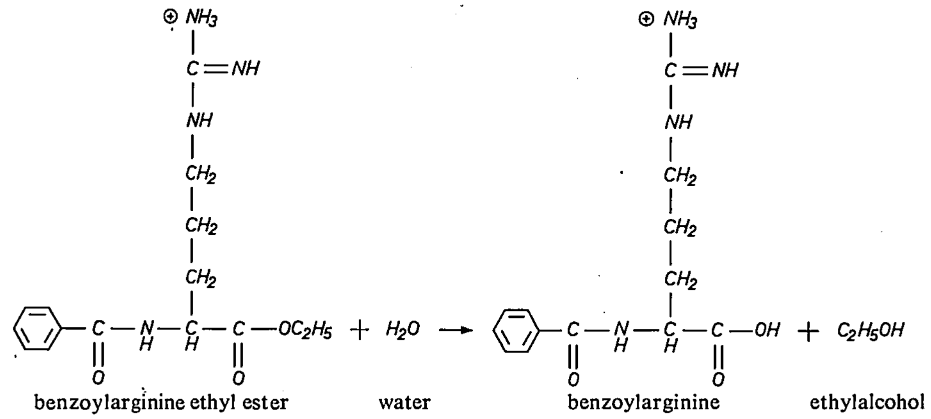


Fig. 3. Equation of the hydrolysis of benzoylarginine ethyl ester.

Determination of the association constant

In a reaction of the form

$$A + B \underset{v_{-r}}{\overset{v_r}{\rightleftarrows}} C + D,$$

the rates of the forward and reverse reactions are respectively:

$$v_r = k_r\,[A]\,[B] \text{ and } v_{-r} = k_{-r}\,[C]\,[D],$$

where [A], [B], [C] and [D] are the concentrations of the substances concerned. The proportionality constants or *reaction constants* k_r and k_{-r} are a measure of the activation energy of the forward and reverse reactions. The effect of a catalyst is, of course, to lower the activation energy.

For our case we can write the equation:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \xrightarrow{k_2} E + P + P'. \quad \ldots \quad (1)$$

The tendency of E and S to form the complex ES is expressed in the ratio k_1/k_{-1}, called the *association constant*. The substances P and P′ are here benzoylarginine and ethyl alcohol. The reaction takes place in an aqueous environment. For simplicity, water has not been included in the reaction equation, because it is present in such large quantities that any changes in the water concentration due to the reaction process have no perceptible effect on the rate of the reaction. It is not possible to derive exactly the reaction rate v of the total reaction ($E + S \rightarrow E + P + P'$). In practice, however, the conditions are usually chosen in such a way as to permit certain simplifications, making it possible to write [3]:

$$v = \frac{ek_2\,[S]}{[S] + 1/K}, \quad \ldots\ldots \quad (2)$$

where $e = [E] + [ES]$, i.e. the total amount of enzyme present, and $K = k_1/(k_{-1} + k_2)$. In most enzymic reactions of this type, k_2 is very much smaller than k_{-1} and therefore K is practically equal to k_1/k_{-1}; in this article we shall assume that this is the case. Equation (2) is usually written in the form:

$$\frac{1}{v} = \frac{1}{ek_2} + \frac{1}{ek_2\,K[S]}, \quad \ldots \quad (3)$$

so that if measured values of $1/v$ are plotted against $1/[S]$ a straight line should be obtained. The intercept of this line and the $1/v$ ordinate is equal to $1/ek_2$, and the slope of the line is equal to $1/ek_2\,K$. Both k_2 and K can be calculated in this way from the results of the measurements (see *fig. 4*).

As a straight line can be plotted from the results of the measurements with our system it can be concluded that the enzyme reaction is in fact given by eq. (1) or to put it another way, that one molecule of papain attaches itself to one molecule of benzoylarginine ester to form a complex, and that one molecule of benzoylarginine and one molecule of alcohol are then formed.

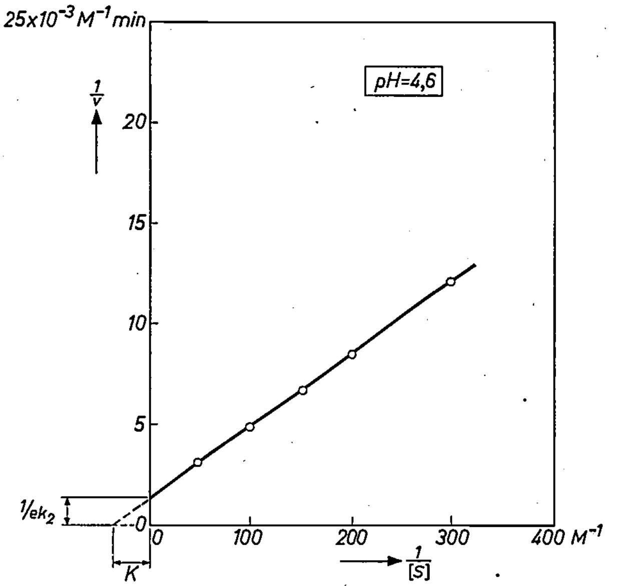


Fig. 4. Hydrolysis of benzoylarginine ethyl ester by the action of papain at pH = 4.6. The reciprocal of the hydrolysis rate $1/v$ is plotted versus the reciprocal of the concentration $1/[S]$ of benzoylarginine ethyl ester.

Influence of pH *on the associative tendency*

In *fig. 5* the association constant K of papain and benzoylarginine ester is given as a function of pH. The same measurements were made with a number of other substrates of papain and in most cases the same bell-shaped curve was obtained with a maximum at pH = 6 and half-values at pH = approx. 4 and 8.

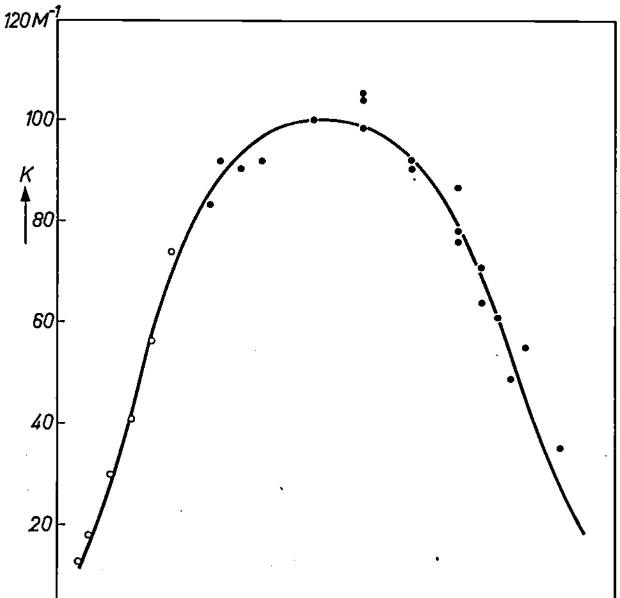


Fig. 5. Association constant K of benzoylarginine ethyl ester as a function of pH. The experimental points given by black dots on the right branch of the curve are due to E. L. Smith and M. J. Parker [4].

[1] K. Yagi and T. Ozawa, Biochim. biophys. Acta **60**, 200, 1962.

[2] See also L. A. Æ. Sluyterman, Biochim. biophys. Acta **85**, 305 and 316, 1964.

[3] M. Dixon and E. C. Webb, Enzymes, Longmans, London 1964.

[4] E. L. Smith and M. J. Parker, J. biol. Chem. **233**, 1387, 1958.

It has been concluded from the shape of the *left-hand* side of the curve that a COOH group must be present in the active centre and that this group must be in a negatively ionized state in order for the formation of the ES complex to be possible — just as for chymotrypsin. It can indeed be calculated that at pH $= 3$ only an extremely small number of such COOH groups are ionized, that at $p\text{H} = 4$ the fraction is roughly one half, and that at $p\text{H} = 6$ the ionization is virtually complete. This is therefore in very satisfactoıy agreement with the curve found for pH values lower than 6. The identity of the group governing the form of the *right-hand* side of the curve is not yet known with any certainty. It is in any case not relevant to the investigation discussed here. We shall return later, however, to the possible existence of a COO^- group in the active centre of the enzyme.

The relation between substrate and product

What is the main problem in preparing an enzyme-substrate complex in crystal form? If we let papain in an aqueous environment react with benzoylarginine ester, the complex formed will hydrolyze much too quickly to allow an ES crystal to form, because of the excess of water. The obvious answer therefore seems to be to carry out the reaction with a very small quantity of water — or even entirely without water. Upon the formation of the ES complex the reaction then *stops*, which for our purposes can only be an advantage. The difficulty, however, is that papain, like all enzymes, *must* be in an aqueous environment if it is not to become denatured. In our work there is obviously no point in investigating the enzyme in its denatured form, in which it has lost its normal spatial structure and in which it cannot exercise its normal function.

To explain the way in which we have tried to overcome thıs difficulty, it will first be useful to examine in more detail the manner in whıch a reaction can be influenced by a "true" catalyst, that is to say a substance that accelerates the reaction without itself being chemically changed in the process (as are enzymes).

The attainment of equilibrium in a reaction such as

$$\mathrm{A} + \mathrm{B} \underset{k_{-\mathrm{r}}}{\overset{k_\mathrm{r}}{\rightleftarrows}} \mathrm{C} + \mathrm{D}$$

can be treated in both kinetic and thermodynamic terms. From the kinetic standpoint the equilibrium is established as soon as the reaction rate in the forward direction $k_\mathrm{r}[\mathrm{A}]\,[\mathrm{B}]$ is equal to the reaction rate in the reverse direction $k_{-\mathrm{r}}[\mathrm{C}]\,[\mathrm{D}]$. In theımodynamic terms the equilibrium is established as soon as the free energy (actually the free enthalpy) of the whole system is at a minimum. Since, by definition, nothing of the catalyst is lost in the reaction, the presence of the catalyst can have no influence on the free energy of the system, and therefore the catalyst cannot affect the position of the equilibrium. In other woıds, the ratio between the reaction constants of the forward and reverse reactions ($k_\mathrm{r}/k_{-\mathrm{r}}$) cannot be changed by the action of the catalyst. This means that a catalyst that accelerates the forward reaction must accelerate the corresponding reverse ıeactions *by the same factor*. Thus, the terms substrate and product have merely a relative significance: a compound that is the product in the forward reaction is the substrate in the (equally accelerated) reverse reaction. For example, using papain as enzyme we can hydrolyze benzoylarginine ethyl ester in a dilute aqueous solution, and we then obtain benzoylarginine and alcohol. But we can also start from benzoylarginine in a solution of high alcohol content and in this situation, under the action of the same enzyme, benzoylarginine etbyl ester is formed.

As we have seen, the formation of a complex between the enzyme papain and the substrate benzoylarginine ester precedes the conversion into benzoylarginine and alcohol. If a product becomes a substrate by reversal of the reaction, it seems likely that this product will also form a complex with the enzyme. The idea behind our attempts at crystallization is therefore to try an opposite approach, that is to say one starting from benzoylarginine. Our aim is therefore no longer the crystallization of an ES complex but of an EP complex. An EP complex must also exist in an aqueous environment, but the presence of water can now do no harm: the danger is no longer hydrolysis but esterification. This danger can easily be avoided by allowing the complex to form in the absence of alcohol, which, unlike the omission of water, does *not* cause denaturation. It appears therefore very logical to tackle the problem from the opposite direction [5].

It should be borne in mind, however, that there ıs as yet no certainty as far as the *product* is concerned that a complex is really formed. For the substrate the existence of the complex has been demonstrated in the manner described on page 169; this method cannot, however be adopted for the product because the condition that the environment should have an excess of water (and hence little alcohol) strongly limits such experiments.

The confirmation that we wanted, that benzoylarginine forms a complex with papain, has been found from the investigations described in the next section. In all the experiments concerned the substrate is hydrolyzed in an aqueous environment and in each we study the effect of addition of certain quantities of product on the rate of the reaction.

The interaction of enzyme and product

Competitive inhibition

We begin by postulating that in a reaction accelerated by "true" catalysts the forward and reverse reactions pass through the same stages. Proof of this postulate will be found in the literature [6]. If, as we wish, both the product and the substrate react to form a complex with the enzyme, so that the reaction follows the equation:

$$E + S + H_2O \rightleftarrows ES + H_2O \rightleftarrows EP + C_2H_5OH$$
$$\rightleftarrows E + P + C_2H_5OH, \quad (4)$$

the product and substrate will both necessarily be bound in the same active centre of the enzyme. The occupation of an active centre by substrate molecules will then prevent the occupation of the same site by product molecules, and vice versa. The result is that the presence of the one will have an inhibitive action on the conversion of the other: this effect is known as "competitive inhibition".

In the experiments now to be described the benzoylarginine ester is hydrolyzed by the action of papain in the presence of a certain amount of benzoylarginine. For what we want to prove, the simplest course is to let the reaction take place without alcohol, and moreover to make the measurements only while the amount of alcohol formed duiing the reaction is negligibly small. This means in fact that we let the reverse reaction go no farther than the formation of the complex, so that we have the following reaction:

$$E + S \rightleftarrows ES \rightarrow EP \rightleftarrows E + P,$$

where P represents benzoylarginine. The reason we want alcohol to play as little part as possible in the reaction is simply that we do not wish the reaction to be more complicated than is necessary for our experiments.

The rate v_i of the reaction

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \xrightarrow{k_2} E + P, \quad \ldots\ldots \quad (5)$$

which is competitively inhibited by the reaction

$$E + P \underset{k_{-i}}{\overset{k_i}{\rightleftarrows}} EP, \quad \ldots\ldots\ldots \quad (6)$$

is found from a simple calculation [3] to be given by:

$$\frac{1}{v_i} = \frac{1}{ek_2} + \frac{1}{ek_2K_S}\left(1 + K_P\,[P]\right)\frac{1}{[S]}, \quad \ldots \quad (7)$$

where $K_P = k_i/k_{-i}$ is the association constant of the enzyme-inhibitor complex EP and $e = [E] + [ES] + [EP]$. To distinguish it more clealy, the association constant of the substrate will be represented by K_S in the following. (When [P] is zero the equation is identical with eq. 3.)

If we vary the concentration [S] of the substrate at a constant concentration [P] of the competitive inhibitor, then according to (7) we should find a straight line that intercepts the $1/v$ axis at $1/ek_2$.

As already noted in fig. 4, the intercept on the $1/v$ axis is also $1/ek_2$ in the *absence* of a competitive inhibitor. This is what one would expect, as given an infinitely high substrate concentration ($1/[S] = 0$) then the competition of the inhibitor molecules, which are present in a constant and thus limited concentration, can no longer be effective. This can be understood in another way: given an infinitely high substrate concentration *all* enzyme molecules contribute to the conversion of the substrate, whether an inhibitor is present or not, i.e. in both cases $v = ek_2$.

In *fig. 6* the circles and crosses represent the results of two of our series of measurements to determine the rate of hydrolysis of benzoylarginine ester. One series

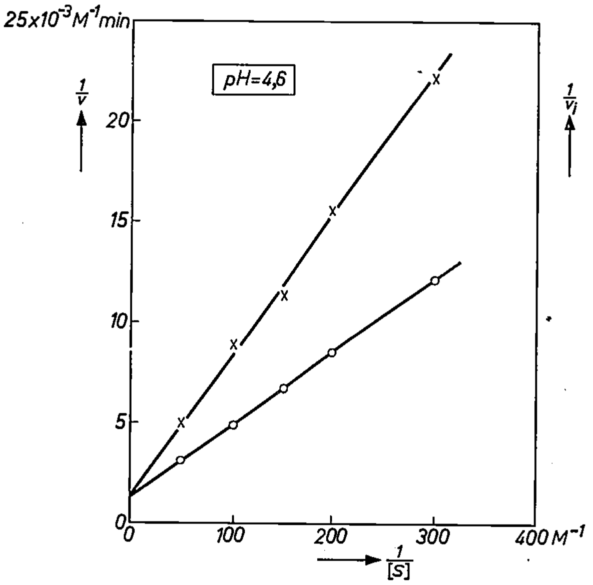


Fig. 6. Reciprocals $1/v$ and $1/v_i$ of the rates of hydrolysis of benzoylarginine ethyl ester acted upon by papain, in the absence of benzoylarginine (circles), and in the presence of 3.35×10^{-2} M benzoylarginine (crosses) at a *p*H of 4.6. The reciprocal of the concentration 1/[S] of benzoylarginine ethyl ester is plotted on the abscissa. The experimental points shown by circles are identical with those in fig. 4. From the fact that both lines intersect on the $1/v$ ordinate it may be concluded that benzoylarginine ethyl ester and benzoylarginine show competitive inhibition, which means that they are converted in the same active centre of papain.

[5] The same approach as ours has already proved successful with an enzyme other than papain; see the note at the end of this article.

[6] See e.g. A. A. Frost and R. G. Pearson, Kinetics and mechanism, p. 202-203, Chapman and Hall, London 1953.

of measurements was made in the absence of benzoylarginine and the other was made in the presence of a constant concentration of benzoylarginine. The two straight lines intersect on the $1/v$ axis, which leads us to conclude that benzoylarginine ester and benzoylarginine bind with the same active centie of papain to form a complex. This confirms oui prediction that the enzyme reaction concerned is in accordance with eq. (4).

The number of molecules involved in the complex formation

Let us now do just the opposite, and vary the concentration of the inhibitor, benzoylarginine, and we keep the concentration of the substrate, benzoylarginine ester, constant. Using eq. (3) we write eq. (7) in the form:

$$\log\left(\frac{v}{v_\mathrm{i}}-1\right) = \log\frac{[\mathrm{P}]K_\mathrm{P}v}{ek_2K_\mathrm{S}[\mathrm{S}]} = \log[\mathrm{P}] + C\,, \qquad (8)$$

wheie C is a constant. If we plot $\log\{(v/v_\mathrm{i})-1\}$ against log [P], then according to eq. (8) we should find a stiaight line of slope equal to 1. This again enables us to test our assumptions.

Equation (8) is based on the reaction equations $\mathrm{E}+\mathrm{S} \rightleftarrows \mathrm{ES} \rightarrow \mathrm{E}+\mathrm{P}$ and $\mathrm{E}+\mathrm{P} \rightleftarrows \mathrm{EP}$, and this implies the assumption that the components in both complex formations E and S and E and P occur in the ratio of one to one. The first assumption has already been confirmed by the results presented in fig. 4. The second is now confirmed by the fact that a plot of $\log\{(v/v_\mathrm{i})-1\}$ does indeed give a straight line of unit slope, as shown in *fig. 7*. (It is easily verified that a reaction equation $\mathrm{E} + n\mathrm{P} \rightleftarrows \mathrm{EP}_n$ would have given a slope n, as the concentration [P] then occurs in eq. (8) in the nth power.)

In the above, then, we have established that papain forms a complex both with the product and with the substrate through the action of one molecule per active centre. This is our second indication that the enzyme reaction is as given in eq. (4).

Influence of pH on the associative tendency

Fig. 8 shows the association constant K_P of the complex of papain and benzoylarginine as a function of the pH. The way in which this association constant is calculated at a given pH can be seen from equations (3) and (7) and fig. 6. From eq. (7) the slope of the upper line is equal to

$$\frac{1}{ek_2K_\mathrm{S}}(1+K_\mathrm{P}[\mathrm{P}])\,,$$

and from eq. (3), that of the lower line is equal to

$$\frac{1}{ek_2K_\mathrm{S}}\,.$$

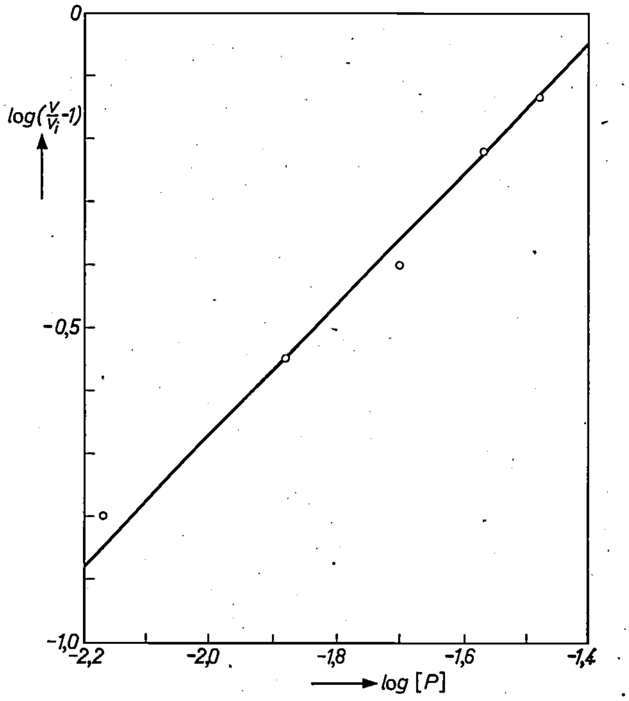


Fig. 7. By varying the inhibitor concentration [P] while keeping the substrate concentration constant and studying the effect on the rate of hydrolysis, conclusions can be drawn as to the number of molecules involved in the formation of an inhibitor enzyme complex. As can be seen, $\log\{v/v_\mathrm{i}-1\}$ is a linear function of log[P] with a slope very nearly equal to unity. This indicates a one to one ratio between the numbers of inhibitor and enzyme molecules in the EP complex.

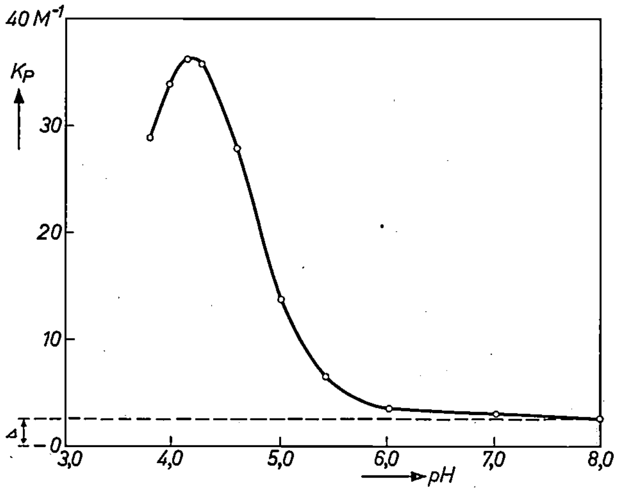


Fig. 8. Association constant K_P of benzoylarginine methyl ester as a function of pH.

In these determinations the concentration [P] of benzoylarginine is known, and therefore K_p can be directly calculated for any pH from the ratio of the two slopes. The curve thus found, which is shown in the figure, has a maximum at $p\mathrm{H} = 4$, and is quite different from the bell-shaped curve with its maximum at $p\mathrm{H} = 6$ which was found for the complex of papain and benzoylarginine ester (fig. 5). A reasonable explanation for this difference is given in the following.

We have already remarked that the shape of the latter curve in the pH interval from 3 to 6 may be attributed to the ionization of a COOH group at the active centre to a COO^- group. The ester group of the substrate remains neutral in this pH interval. The situation with respect to the action if the enzyme on the product is quite different. In addition to the COOH group of the active centre of the enzyme the COOH group of the product can also be ionized.

At pH = 8 the ionization of both COOH groups is complete. At this pH value the COO^- groups of the product, having like charges, repel each other, with the result that the tendency to form a complex is very slight. At pH = 6 the product is still largely in its ionized form, and there is no appreciable increase in the fraction of the product molecules having a neutral carboxyl group until there is a further decrease in pH. Only then may we expect the association constant to increase. This state continues until the COO^- group of the active centre is also neutralized. We already know that this has the effect of inactivating the active centre. Owing to the presence of the COOH group in the product molecule the maximum in the association curve shifts towards a lower pH. This explains, in qualitative terms, the difference between the curves in fig. 5 and fig. 8. A more quantitative comparison can however also be made.

Plainly, for quantitative comparison of the association curves of the substrate and the product, the product with the *uncharged* carboxyl group (which we shall call P_I) must be set against the substrate with its likewise uncharged ester group. We shall do this in the pH region from 3.7 to 5.5, where the inhibition is mainly attributable to P_I. At pH = 8 the inhibition is very much less and is almost entirely due to product molecules with an *ionized* carboxyl group (which we shall call P_{II}).

To calculate the association constant K_{PI} of P_I we should write eq. (7) in a modified form. In the first place K_P, which relates to both inhibitors together, has to be corrected for the inhibition of P_{II}. A rough but sufficiently accurate correction is to subtract the amount Δ indicated by the dashed line in fig. 8: instead of K_P we now write $K_{PI} + \Delta$. Secondly, we must substitute for the concentration [P] the concentration $[P_I]$ of the non-ionized form. We calculate $[P_I]$ for each pH value with the aid of the known value of the dissociation constant of the relevant COOH group:

$$K_{\rm diss} = \frac{[COO^-]\,[H^+]}{[COOH]} = \frac{[P_{II}]\,[H^+]}{[P_I]}.$$

The association constant K_{PI} calculated in this way is plotted in *fig. 9*, together with the association constant K_S of the substrate S, as a function of pH. They are shown by crosses and circles respectively. The agreement may be described as very satisfactory.

Investigation using other substrates

Finally, we shall mention two results, found with other substrates. Instead of using benzoylarginine ester as a substrate, other investigators have used benzoylglycylglycine [7], a compound which, like benzoylarginine, contains a COOH group. In agreement with oui assumptions the association constant was found to decrease as pH increased from 4 to 9 (*fig.* 10), as we found for benzoylarginine. If the investigations had been extended to pH values lower than 4, a corresponding decrease would presumably have been established.

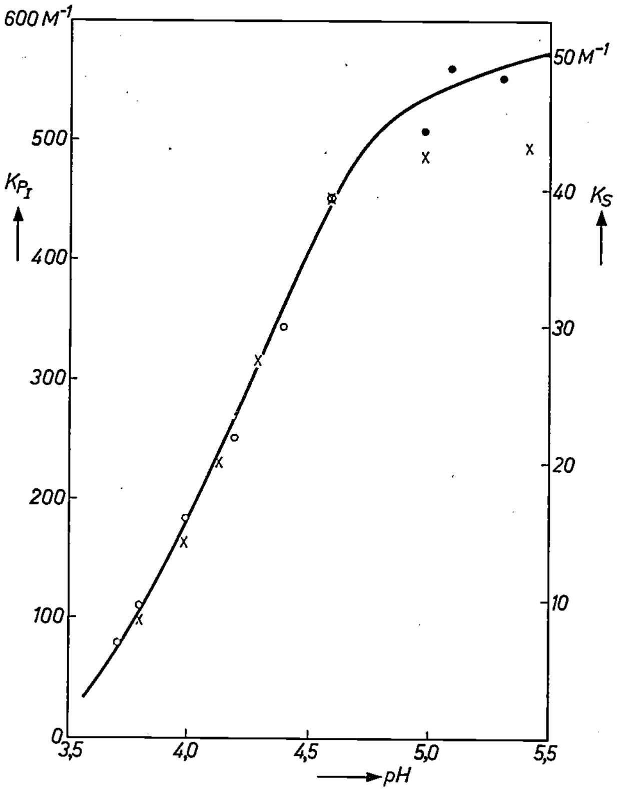


Fig. 9. Association constant K_S of benzoylarginine ethyl ester (circles) and association constant K_{PI} of benzoylarginine in non-ionized form (crosses). The black dots represent measurements on benzoylarginine ester reported by E. L. Smith and M. J. Parker [4].

We would also like to mention an investigation of our own, which used benzoylglycine ester as substrate. This substrate is converted by papain much more slowly than benzoylarginine ester. The question of interest here was whether the product also might be a poorer inhibitor than benzoylarginine (which, on the basis of our assumptions, was possible but not neces-

[7] E. L. Smith, V. J. Chavré and M. J. Parker, J. biol. Chem. **230**, 283, 1958.

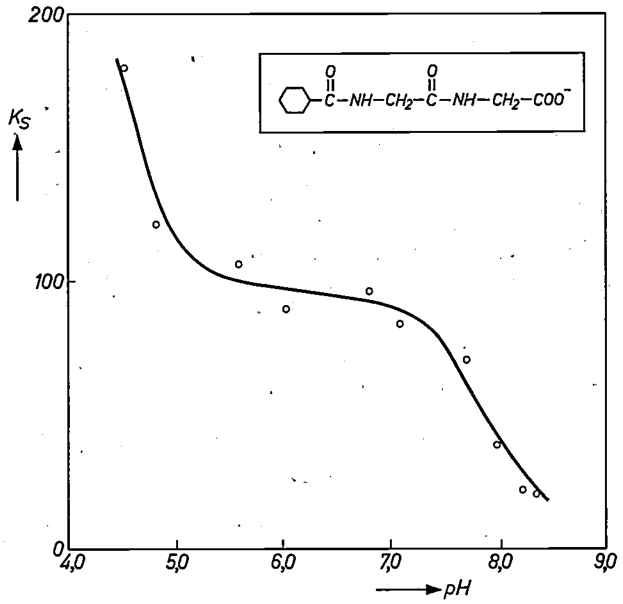


Fig. 10. Association constant K_S of benzoylglycylglycine as a function of *p*H in arbitrary units. (Due to E. L. Smith, V. J. Chavré and M. J. Parker [7].)

sarily so). This proved to be the case, the inhibitor action being six times smaller. Thus, the specificity of papain with respect to the substrate is found here to be reflected, as it were, in its specificity with respect to the product.

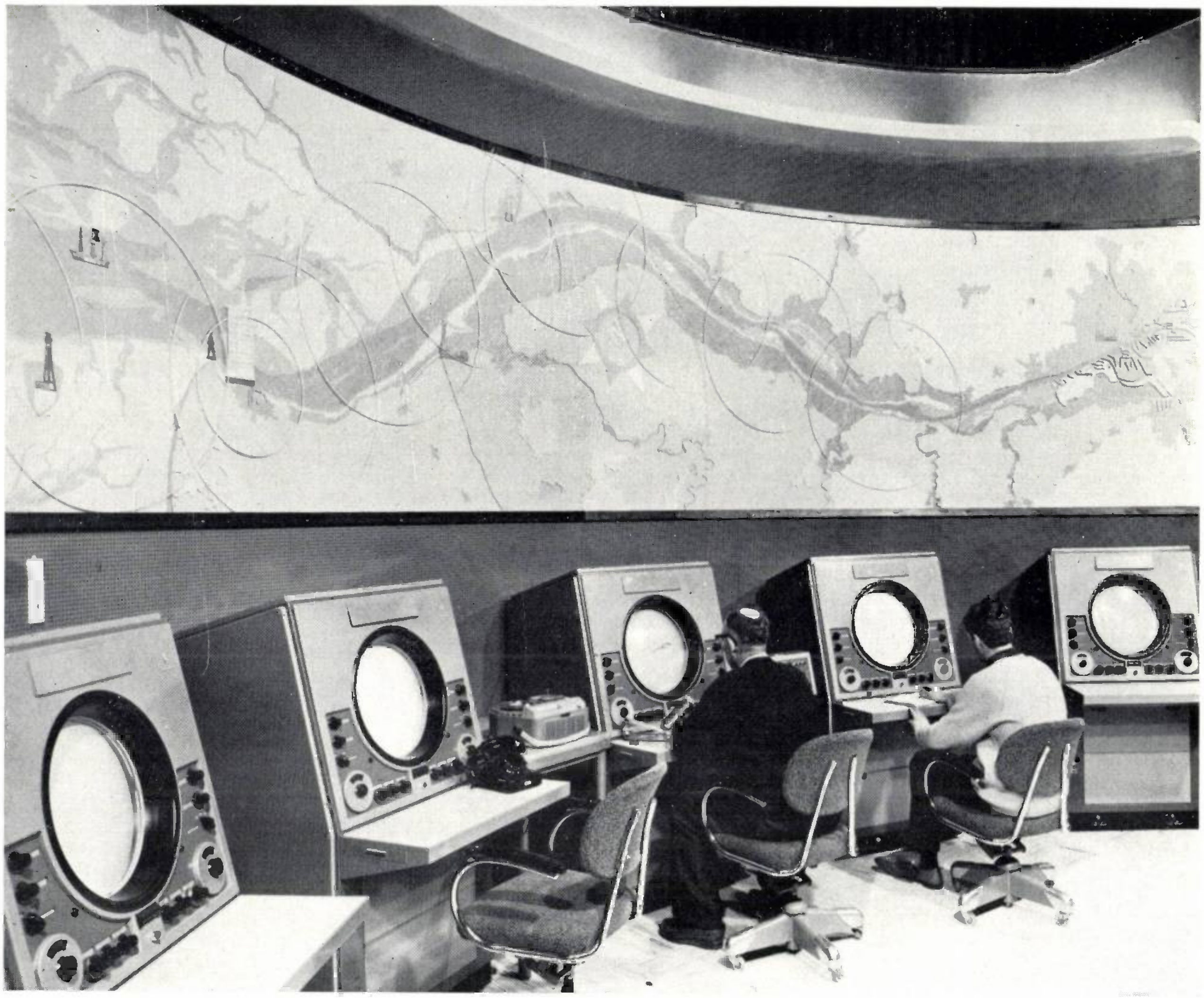
We may summarize the ıesults of our investigation by saying that we have established a certain degree of agreement between the chosen substrate and product. It would certainly be interesting to go into the question of whether this is a geneıal rule, but for the time being it remains a pure speculation. For our immediate purposes it is sufficient to have found a number of indications that we have in fact chosen an appropriate "product" as substrate for the formation of a crystalline enzyme-substrate complex.

Note upon going to press: When this article was being prepared for publication, a report appeared on the successful crystallization of EP complexes as discussed in the foregoing, using the enzyme lysozyme (L. N. Johnson and D. C. Philips, Nature **206**, 761, 1965). From the X-ray analysis results obtained so far it was possible to localize the active site and also to determine with a very high degree of certainty which groups of the enzyme lysozyme are involved in the formation of the complex. Efforts are being made to refine the procedure in order to obtain a more exact picture of the binding mechanism. From the experimentally determined fact that the relevant products are competitive inhibitors of the enzymic action, it may be concluded that the active centre investigated is also the site where the substrate is broken off by the enzyme.

Summary. Like all true catalysts, enzymes accelerate not only a forward reaction, but also a corresponding reverse reaction. The product of the forward reaction will therefore be the substrate of the reverse reaction, and it is to be expected that the interaction of enzyme and substrate will agree in certain respects with the interaction of enzyme and product. This theory was put to the test with the enzyme papain, use being made of a simple synthetic substrate (benzoylarginine ester) and the corresponding product (benzoylarginine). Five indications in support of the hypothesis were deduced from kinetic experiments.

This investigation was necessary as a preparation for attempts now being made to crystallize an enzyme-substrate complex, with a view to making possible an X-ray analysis of such a complex. The difficulty encountered when using benzoylarginine ester as substrate is that the complex hydrolyzes too quickly. The experimental indications make it seem likely that the product benzoylarginine can be made use of, in a suitable way, to obtain the required complex.

Harbour surveillance radar for the Elbe



The harbour area of Hamburg can only be reached from the sea through a narrow and winding waterway, the Elbe. As at Rotterdam, the Philips 3-cm radar is in use for navigation surveillance. The radar apparatus for this project is the same as that at Rotterdam but is more fully transistorized. The special feature of the Elbe project is that not all radar stations are manned (as at Rotterdam). The Cuxhaven radar station is accommodated in a radar tower, which can be seen indicated on the left of the chart above the radar screens. The stations up- and down-stream from Cuxhaven are unmanned, and send all data for display to Cuxhaven, by means of a microwave link. This is the first time, that radar data for civil use have been transmitted in this way.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands *E*
Mullard Research Laboratories, Redhill (Surrey), England *M*
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France *L*
Philips Zentrallaboratorium GmbH, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany *A*
Philips Zentrallaboratorium GmbH, Hamburg laboratory, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany *H*
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium. *B*

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

N. W. H. Addink: Agreement between carcinoma experiments in vivo and low-energy X-ray irradiation of the metalloprotein carbonic anhydrase.
Nature **207**, 1271-1272, 1965 (No. 5003). *E*

W. Albers and **A. C. Aten:** The preparation of single-phase single crystals of zinc telluride.
Philips Res. Repts. **20**, 556-561, 1965 (No. 5). *E*

W. Albers and **C. J. M. Rooymans:** High pressure polymorphism of spinel compounds.
Solid State Comm. **3**, 417-419, 1965 (No. 12). *E*

D. Alma: Het gebruik van doorzichtkaarten bij het vastleggen en analyseren van gegevens.
Sigma **11**, 105-109, 1965 (No. 5).

A. Baelde: Theory and experiments on the noise of transistors.
Thesis Delft, Oct. 1964. *E*

E. M. N. Baldwin, J. C. Brice and **E. J. Millett:** A syringe crystal puller for materials having a volatile component.
J. sci. Instr. **42**, 883-884, 1965 (No. 12). *M*

J. P. Baldwin: The influence of substrate material on heating effects in thin films at helium temperatures.
Bull. Inst. Int. Froid, Annexe 1965-2, pp. 293-301. *M*

K. H. Beckmann: Investigation of the chemical properties of stain films on silicon by means of infrared spectroscopy.
Surface Sci. **3**, 314-332, 1965 (No. 4). *H*

V. Belevitch and **J. Neirynck:** Response of a tuned circuit to a frequency step.
Proc. IEEE **53**, 2146, 1965 (No. 12). *B*

F. Berz: On a quarter wave light condenser.
Brit. J. appl. Phys. **16**, 1733-1738, 1965 (No. 11). *M*

G. Blasse: Antiferromagnetism of CoRh$_2$S$_4$.
Physics Letters **19**, 110, 1965 (No. 2). *E*

G. Blasse: Magnetic properties of some oxides with spinel structure.
Philips Res. Repts. **20**, 528-555, 1965 (No. 5). *E*

G. Blasse: Ferromagnetic interactions in non-metallic perovskites.
J. Phys. Chem. Solids **26**, 1969-1971, 1965 (No. 12). *E*

G. Blasse: New compositions with K$_2$NiF$_4$ structure.
J. inorg. nucl. Chem. **27**, 2683-2684, 1965 (No. 12). *E*

M. E. Bond: Accelerated life tests.
Industr. Qual. Control **22**, 171-177, 1965 (No. 4). *M*

W. Bongenaar and **N. C. de Troye:** Worst-case considerations in designing logical circuits.
IEEE Trans. on electronic computers **EC-14**, 590-599, 1965 (No. 4). *E*

A. J. Bosman, H. J. van Daal and **G. F. Knuvers:** Hall effect between 300 °K and 1100 °K in NiO.
Physics Letters **19**, 372-373, 1965 (No. 5). *E*

H. Bouma: Receptive systems.
Thesis Eindhoven, Jan. 1965. *E*

G.-A. Boutry, J. J. Brissot, R. Legoux, J. Périlhou and **G. Pietri:** Un tube convertisseur d'image pour l'infrarouge moyen "le Serval".
Philips Res. Repts. **20**, 684-706, 1965 (No. 6). *L*

C. J. Bouwkamp: Numerical solution of a nonlinear eigenvalue problem.
Proc. Kon. Ned. Akad. Wetensch. **A 68**, 539-547, 1965 (No. 4). *E*

A. Bril and **W. L. Wanmaker:** Energy transfer in CaNaBO$_3$ activated with Tb and Gd.
J. chem. Phys. **43**, 2559-2560, 1965 (No. 7). *E*

J. van den Broek: Contact barriers in red lead monoxide.
Philips Res. Repts. **20**, 674-683, 1965 (No. 6). *E*

K. H. J. Buschow: Phase relations and intermetallic compounds in the systems neodymium-aluminium and gadolinium-aluminium.
J. less-common Met. **9**, 452-456, 1965 (No. 6). *E*

B. H. Clarke: Rare-earth ion relaxation time and *G* tensor in rare-earth-doped yttrium iron garnet, II. Neodymium.
Phys. Rev. **139**, A 1944-A 1948, 1965 (No. 6A). *M*

B. H. Clarke, K. Tweedale and **R. W. Teale:** Rare-earth ion relaxation time and *G* tensor in rare-earth-doped yttrium iron garnet, I. Ytterbium.
Phys. Rev. **139**, A 1933-A 1943, 1965 (No. 6A). *M*

A. Cohen: Versprekingen als verklappers van het proces van spreken en verstaan.
Forum der Letteren **6**, 175-186, 1965 (No. 4). *E*

E. H. P. Cordfunke and **A. A. van der Giessen:** Texture and reactivity of uranium oxides.
Reactivity of solids, 5th Int. Symp., Munich 1964, pp. 456-466, Elsevier, Amsterdam 1965. *E*

H. P. C. Daniëls: Ultrasonic welding.
Ultrasonics **3**, 190-196, 1965 (Oct./Dec.). *E*

C. Z. van Doorn: Abnormal green "edge" emission in CdS due to an oxygen impurity.
Solid State Comm. **3**, 355-356, 1965 (No. 11). *E*

W. F. Druyvesteyn and **F. A. Staas:** A new kind of transverse voltage in a type II superconductor.
Physics Letters **19**, 262-263, 1965 (No. 4). *E*

C. W. Elenga and **O. Reifenschweiler:** The generation of neutron pulses and modulated neutron fluxes with sealed-off neutron tubes.
Pulsed neutron research, Proc. Symp. Karlsruhe 1965, Vol. II, pp. 609-622, Int. Atomic Energy Agency, Vienna 1965. *E*

G. Engelsma: Photo-induced hydroxylation of cinnamic acid in gherkin hypocotyls.
Nature **208**, 1117-1119, 1965 (No. 5015). *E*

J. A. Geurst: Theoretical analysis of the influence of track width on the harmonic response of magnetic reproducing heads.
Philips Res. Repts. **20**, 633-657, 1965 (No. 5). *E*

H. Groendijk, M. T. Vlaardingerbroek and **K. R. U. Weimer:** Waves in cylindrical beam-plasma systems immersed in a longitudinal magnetic field.
Philips Res. Repts. **20**, 485-504, 1965 (No. 5). *E*

E. F. de Haan and **A. G. van Doorn:** A "Plumbicon" color broadcast camera.
J. SMPTE **74**, 922-926, 1965 (No. 10). *E*

C. Haas: Phase transitions in ferroelectric and antiferroelectric crystals.
Phys. Rev. **140**, A 863-A 868, 1965 (No. 3A). *E*

C. M. Hargreaves: Corrections to the retarded dispersion force between metal bodies.
Proc. Kon. Ned. Akad. Wetensch. **B 68**, 231-236, 1965 (No. 4). *E*

E. E. Havinga and **A. J. Bosman:** Temperature dependence of dielectric constants of crystals with NaCl and CsCl structure.
Phys. Rev. **140**, A 292-A 303, 1965 (No. 1A). *E*

M. Hertogs and **J. S. C. Wessels:** Ferredoxin-stimulated photoreduction of 2,4-dinitrophenol with solubilized chlorophyll *a*.
Biochim. biophys. Acta **109**, 610-613, 1965 (No. 2). *E*

W. Hondius Boldingh: Quality and choice of Potter Bucky grids, VII. The influence of a grid on the integral absorbed dose.
Acta radiol., Diagnosis **3**, 475-480, 1965 (No. 5).

E. P. Honig: Logarithmic distribution functions for colloidal particles.
J. phys. Chem. **69**, 4418-4419, 1965 (No. 12). *E*

J. Hornstra: The interaction of grain boundaries and dislocations with vacancies in the sintering process.
Symp. sur la métallurgie des poudres, Paris 1964, pp. 97-103, Editions Métaux, 1965. *E*

K. Hoselitz: Magnetic properties.
Physical metallurgy, editor R. W. Cahn, pp. 1015-1052, North-Holland Publ. Co., Amsterdam 1965. *M*

E. Kooi: Effects of low-temperature heat treatments on the surface properties of oxidized silicon.
Philips Res. Repts. **20**, 578-594, 1965 (No. 5). *E*

E. Kooi: Effects of ionizing irradiations on the properties of oxide-covered silicon surfaces.
Philips Res. Repts. **20**, 595-619, 1965 (No. 5). *E*

W. Kwestroo and **C. Langereis:** Basic lead acetates.
J. inorg. nucl. Chem. **27**, 2533-2536, 1965 (No. 12). *E*

H. de Lang and **G. Bouwhuis:** Quasi-stationary polarization of a single mode gas laser in a magnetic field.
Physics Letters **19**, 481-482, 1965 (No. 6). *E*

H. de Lang, G. Bouwhuis and **E. T. Ferguson:** Saturation-induced anisotropy in a gaseous medium in zero magnetic field.
Physics Letters **19**, 482-484, 1965 (No. 6). *E*

F. K. Lotgering: On the antiferromagnetism of $ZnCr_2Se_4$.
Solid State Comm. **3**, 347-349, 1965 (No. 11). *E*

F. K. Lotgering: The influence of Fe^{3+} ions at tetrahedral sites on the magnetic properties of $ZnFe_2O_4$.
J. Phys. Chem. Solids **27**, 139-145, 1966 (No. 1). *E*

S. J. Lowe: Calculators for use in Hall measurements.
J. sci. Instr. **42**, 908, 1965 (No. 12). *M*

K. Minnaert: Measurement of the equilibrium constant of the reaction between cytochrome *c* and cytochrome *a*.
Biochim. biophys. Acta **110**, 42-56, 1965 (No. 1). *E*

B. J. Mulder and **J. de Jonge:** Photoconductivity of crystals of anthracene doped with tetracene and acridine.
Rec. Trav. chim. Pays-Bas **84**, 1503-1510, 1965 (No. 11). *E*

P. C. Newman: Forward characteristics of heterojunctions.
Electronics Letters **1**, 265, 1965 (No. 9). *M*

R. F. Pearson: Magnetocrystalline anisotropy of ytterbium iron garnet $Yb_3Fe_5O_{12}$.
Proc. Phys. Soc. **86**, 1055-1066, 1965 (No. 5). *M*

D. Polder and **W. van Haeringen:** The effect of saturation on the ellipticity of modes in gas lasers.
Physics Letters **19**, 380-381, 1965 (No. 5). *E*

H. Rau: Thermodynamische Messungen an SnS.
Berichte Bunsenges. phys. Chemie **69**, 731-736, 1965 (No. 8). *A*

P. Reijnen: Investigations into solid state reactions and equilibria in the system MgO-FeO-Fe_2O_3.
Reactivity of solids, 5th Int. Symp., Munich 1964, pp. 562-571, Elsevier, Amsterdam 1965. *E*

C. J. M. Rooymans: Reactivity in high-pressure transformations.
Reactivity of solids, 5th Int. Symp., Munich 1964, pp. 100-109, Elsevier, Amsterdam 1965. *E*

C. J. M. Rooymans: High pressure phase transition of europium telluride.
Solid State Comm. **3**, 421-424, 1965 (No. 12). *E*

E. Schwartz: Die Verstärkung relativer Immittanzänderungen.
Arch. elektr. Übertr. **19**, 559-565, 1965 (No. 10). *A*

G. Simon: Methoden zur Bestimmung der Gestalt der Fermi-Oberfläche in Metallen.
Z. angew. Phys. **20**, 161-172, 1965 (No. 2). *A*

A. L. Stuijts and **G. J. Oudemans:** Ceramic forming methods.
Proc. Brit. Ceramic Soc. **3**, 81-99, Oct. 1965. *E*

W. Tolksdorf: Über die Bildung von $Mg_2Ba_2Fe_{12}O_{22}$ und $Ni_2Ba_2Fe_{12}O_{22}$.
Reactivity of solids, 5th Int. Symp., Munich 1964, pp. 606-613, Elsevier, Amsterdam 1965. *H*

T. J. Turner, R. De Batist and **Y. Haven:** Relaxation modes for photon-induced reorientation of M-centers in alkali halides.
Phys. Stat. sol. **11**, 267-276, 1965 (No. 1). *E*

A. G. van Vijfeijken and **A. K. Niessen:** Longitudinal and transverse voltages in superconductors.
Philips Res. Repts. **20**, 505-527, 1965 (No. 5). *E*

H. J. Vink: Die Rolle der Chemie bei der Untersuchung von Festkörpern.
Festkörperprobleme IV, 205-244, Vieweg, Brunswick 1965. *E*

K. J. de Vos: The influence of heat treatment on the magnetic properties and the microstructure of Fe-Ni-Al alloys.
Philips Res. Repts. **20**, 667-673, 1965 (No. 6).

J. H. N. van Vucht: Note on the structures of $GdFe_3$, $GdNi_3$, $GdCo_3$ and the corresponding yttrium compounds.
J. less-common Met. **10**, 146-147, 1966 (No. 2). *E*

J. H. N. van Vucht and **K. H. J. Buschow:** The structures of the rare-earth trialuminides.
J. less-common Met. **10**, 98-107, 1966 (No. 2). *E*

J. C. Walling and **F. W. Smith:** Travelling-wave maser amplifier.
The Goonhilly Project, editor F. J. D. Taylor, pp. 104-113, Inst. Electr. Engrs., London 1964. *M*

K. Walther: Directional ultrasonic noise and kink effect in bismuth.
Phys. Rev. Letters **15**, 706-708, 1965 (No. 17). *H*

W. L. Wanmaker, A. Bril and **J. W. ter Vrugt:** Fluorescent properties of terbium-activated alkaline earth alkali borates.
J. Electrochem. Soc. **112**, 1147-1150, 1965 (No. 11). *E*

W. L. Wanmaker and **D. Radielović:** The dependence of the rate of the dissociation of strontium carbonate in some mixtures on particle size and composition.
Reactivity of solids, 5th Int. Symp., Munich 1964, pp. 529-539, Elsevier, Amsterdam 1965.

J. S. C. Wessels: Mechanism of the reduction of organic nitro compounds by chloroplasts.
Biochim. biophys. Acta **109**, 357-371, 1965 (No. 2). *E*

J. S. C. Wessels: Reconstitution by plastocyanin of the $NADP^+$-photoreducing activity in digitonin fragments of spinach chloroplasts.
Biochim. biophys. Acta **109**, 614-616, 1965 (No. 2). *E*

M. V. Whelan: Influence of charge interactions on capacitance versus voltage curves in MOS structures.
Philips Res. Repts. **20**, 562-577, 1965 (No. 5). *E*

M. V. Whelan: Graphical relations between surface parameters of silicon, to be used in connection with MOS-capacitance measurements.
Philips Res. Repts. **20**, 620-632, 1965 (No. 5). *E*

J. S. van Wieringen and **J. G. Rensen:** Paramagnetic resonance of Mn^{2+} in NaCl between 300 °C and 803 °C.
Philips Res. Repts. **20**, 659-666, 1965 (No. 6). *E*

P. C. van der Willigen: Over staal en het booglassen ervan.
Lastechniek **31**, 205-211, 1965 (No. 11 L). *E*

G. Winkler: Die Bildung und Umwandlung hexagonaler und trigonaler magnetischer Phasen im Dreistoffsystem BaO-MeO-Fe_2O_3.
Reactivity of solids, 5th Int. Symp., Munich 1964, pp. 572-582, Elsevier, Amsterdam 1965. *H*

Volume 27, 1966, No. 6 pages 141-172 Published 22th September 1966.

PHILIPS TECHNICAL REVIEW

VOLUME 27, 1966, No. 7

Gas-filled incandescent lamps containing bromine and chlorine

G. R. T'jampens and M. H. A. van de Weijer

621.326.79

It seems very likely that the first cyclical process introduced for incandescent lamps — the tungsten-iodine cycle — will be followed by several others. After the iodine lamp, there are now various types of bromine lamp on the market, and the possibilities of chlorine and fluorine lamps are at present under investigation.

The iodine lamp described in this journal a few years ago [1] was hardly out of the laboratory stage before it was put into large-scale production. The lamp very soon found many fields of application. Its initial promise has been confirmed: after the advent of Edison's carbon filament lamp in 1879 and the gas-filled tungsten lamp in 1913, the introduction of the regenerative-cycle incandescent lamp forms a new milestone in the development of the incandescent lamp.

Tungsten reacts chemically with iodine in accordance with the equations:

$$\left.\begin{array}{ll} & W + nI \rightleftarrows WI_n \\ \text{or} & W + \tfrac{1}{2}nI_2 \rightleftarrows WI_n \end{array}\right\} (n = 2 \text{ or } 4).$$

Tungsten iodide is formed at relatively low temperature and at relatively high temperature it decomposes. The result of this in an incandescent lamp is that the tungsten evaporated from the filament forms an iodide in the cooler parts of the lamp. At the temperature which holds in the operating lamp this iodide is gaseous. It diffuses towards the hot filament, where it again dissociates into tungsten and the original halogen.

This process keeps the bulb wall free of tungsten deposits, and has enabled a drastic reduction in the surface area of the bulb without the risk of blackening. The bulb must, however, be made of a glass with a high melting point, such as quartz glass, because of the higher bulb temperature. The much smaller lamp thus obtained — whose volume is only a few per cent of that of the classical version — is mechanically much stronger, so that the bulb can safely be filled with gas under high pressure. This in turn makes it possible to increase either the life of the lamp or its luminous efficiency. The lamp that results is of considerably reduced size, and gives a better performance throughout its useful life, with no loss of light output due to bulb blackening.

For a proper understanding of the technological problems arising with this new type of lamp, it must be realized that the lamp has acquired a completely new "dimension" through the introduction of a continuously operating *chemical* process in the lamp. Hitherto *physical* processes were the only ones involved in the operation of the incandescent lamp, even though a great deal of chemistry entered into their preparation. In the new type of lamp the familiar physical processes of evaporation and diffusion continue to operate, but they are supplemented by the chemical processes of the continuous formation and decomposition of tungsten compounds. This complicates considerably the events taking place inside the lamp. It is therefore not so surprising that it was soon found that the iodine lamp is not easy to manufacture reproducibly.

What exactly happens? The tungsten-iodine cycle is easily affected by traces of impurities in the lamp. The impurities may be present in the lamp right from the start, or may be released from the tungsten filament and the glass wall when the lamp is operating. The impurities can cause blackening of the bulb or seriously shorten the life of the lamp. The first effect may probably be put down to a retardation in the cycle, caused by the impurity, and the second to an acceleration of the cycle.

In the iodine lamp hydrogen is one example of the "cycle decelerators" and oxygen is among the "cycle

G. R. T'jampens, Lic. Sc., and Ir. M. H. A. van de Weijer are with Philips Lighting Division, at Turnhout and Eindhoven respectively.

[1] J. W. van Tijen, Iodine incandescent lamps (I), Philips tech. Rev. **23**, 237-242, 1961/62. For further information see E. G. Zubler and F. A. Mosby, Illum. Engng. **54**, 734, 1959; W. Schilling, Elektrotech. Z. **B 13**, 485, 1961; J. A. Moore and C. M. Jolly, G.E.C. Journal **29**, 99, 1962; J. W. Strange and J. Stewart, Trans. Illum. Engng. Soc. (London) **28**, 91, 1963.

accelerators"[2]. Since in practice most trouble is experienced from the cycle-decelerating impurities, a small quantity of oxygen is often deliberately added to the gas filling to activate the cycle. Clearly, having to strike a balance between activating and decelerating influences hardly leads to a completely controlled manufacturing process.

For the classical incandescent lamp there are various "gettering" substances which are capable of chemically combining with impurities inside the bulb. The most widely used are phosphorus and zirconium. For the iodine lamp, however, no practically effective getter has yet been found, although carbon has been mentioned as a possibility in this connection[3]. It is hardly surprising that the getter problem in halogen lamps is such a difficult one, since the getter is required to absorb the impurities but not the chemically highly active halogens. The lack of a really effective getter for iodine lamps is the more keenly felt because the impurities can have much more serious results than in the classical lamp, while moreover the filament and bulb-wall temperatures are so much higher that more impurities can be released during the operation of the lamp.

The impracticability of using a getter imposes heavy demands on the manufacturing processes. These are required to guarantee an extreme degree of purity in the lamps, which moreover has to be maintained during the whole useful life of the lamp.

Quite apart from this there are certain technological problems. Difficulties arise because of the chemical aggressiveness of iodine in its action on traditional pump materials like metals, rubber and grease, and the low vapour pressure of iodine at room temperature raises special problems in introducing the iodine vapour into the lamp.

All these difficulties were sufficient reason for trying out other perhaps theoretically less suitable halogens[1] for their usefulness as a transport gas[4]. It was soon found that halogen *compounds* could be used. In this article we shall mainly confine our attention to chlorine and bromine and their compounds[5].

Investigations of the cycle with chlorine and with bromine

Before describing the results of the experiments we shall look into the differences that may be expected between the various tungsten-halogen cycles. We shall base our considerations entirely on thermodynamic data. There are indications that the kinetic factors do not essentially alter the picture.

The reactions between solid tungsten W_s and molecular halogen X_2 can be described as follows:

$$\frac{1}{n}W_s + \frac{1}{2}X_2 \rightleftarrows \frac{1}{n}WX_n \qquad (n = 1, 2, \ldots, 6).$$

These reactions have the equilibrium constant:

$$K_p = \frac{(p_{WX_n})^{1/n}}{(p_{X_2})^{1/2}}.$$

The reactions between solid tungsten W_s and atomic halogen X can be written as:

$$\frac{1}{n}W_s + X \rightleftarrows \frac{1}{n}WX_n,$$

and the equilibrium constant for these reactions is:

$$K_p = \frac{(p_{WX_n})^{1/n}}{p_X}.$$

Fig. 1 gives a rough indication of how the equilibrium constant K_p for the halogens chlorine, bromine and iodine varies as a function of temperature.

For a proper understanding, further details should be given of the requirements that have to be satisfied to maintain the cyclical process. At the bulb-wall tem-

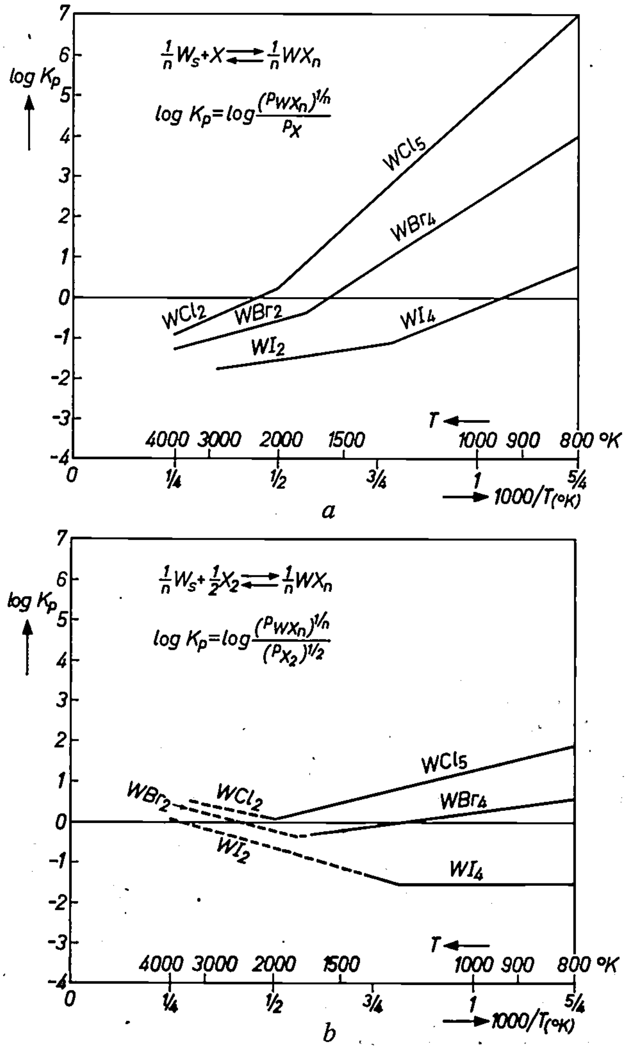


Fig. 1. A rough indication of the equilibrium constant K_p as a function of temperature T, for the reactions between solid tungsten and iodine, bromine or chlorine, in *a*) the atomic and *b*) the molecular state. The dashed part of the curves in (*b*) is only of theoretical significance, as in this temperature range the halogen can only exist in an almost entirely atomic state.

perature (about 800 to 1000 °K) the equilibrium has to be towards the tungsten halide, and at the filament temperature (approx. 3000 °K) it has to be towards the dissociation products. The atomic halogen plays the principal role. From the figure we see that all three of the halogens considered meet these requirements.

There is a separate problem with the ends of the filament and the filament supports (if these are of tungsten), since these have a lower temperature than the central part of the filament. It is quite feasible that at the central part of the filament the required dissociation of the halide and deposition of tungsten does take place, while this process is in fact reversed at the cooler parts, so that the filament is attacked here, and an early failure results.

If we now attempt to make a comparison between the various halogens (still on the basis of the data in fig. 1), we arrive at the following conclusions.

At the same temperature, the stability of the tungsten-halide increases (greater K_p) in the sequence iodine-bromine-chlorine. This means that in the cooler regions of the lamp the conversion of tungsten to halide takes place more easily, so that the danger of blackening in this sequence is reduced. In the hot parts of the lamp, however, the dissociation of the halide in this sequence is more difficult. In the immediate vicinity of the central part of the filament, however, the dissociation is still amply sufficient for all three halogens.

The problem with the chemical attack of the ends of the filament appears to be greatest with chlorine. For this halogen the equilibrium shifts fastest towards the halide when the temperature is lowered, and therefore the danger of the undesired reversal of the process is greatest.

Experimental results

For most of the experiments to be described we used 225 V and 1000 W photoflood lamps, which have been marketed as iodine lamps. These lamps have a luminous efficiency of about 32 lm/W, a colour temperature of about 3400 °K and a life in the region of 30 hours. The gas filling consists of 1 atm argon with 8 % nitrogen and about 5 torr iodine vapour. In the experiments, the halogens and halogen compounds were introduced into the lamp in place of the iodine. Chlorine was admitted in the form of carbon tetrachloride (CCl_4) and bromine as bromine gas (Br_2). CCl_4 is thermally not very stable, and dissociates during the operation of the lamp into C and 2 Cl_2 [6]. In our experiments the partial pressure of the halogen was varied between 0.1 and 10 torr.

For both chlorine and bromine at relatively low partial pressures, the cycle works quite well at the beginning of the operation: the bulb remains clear. After a certain operating time, however, fairly serious blackening suddenly appears. The life is normal, i.e. about 30 hours.

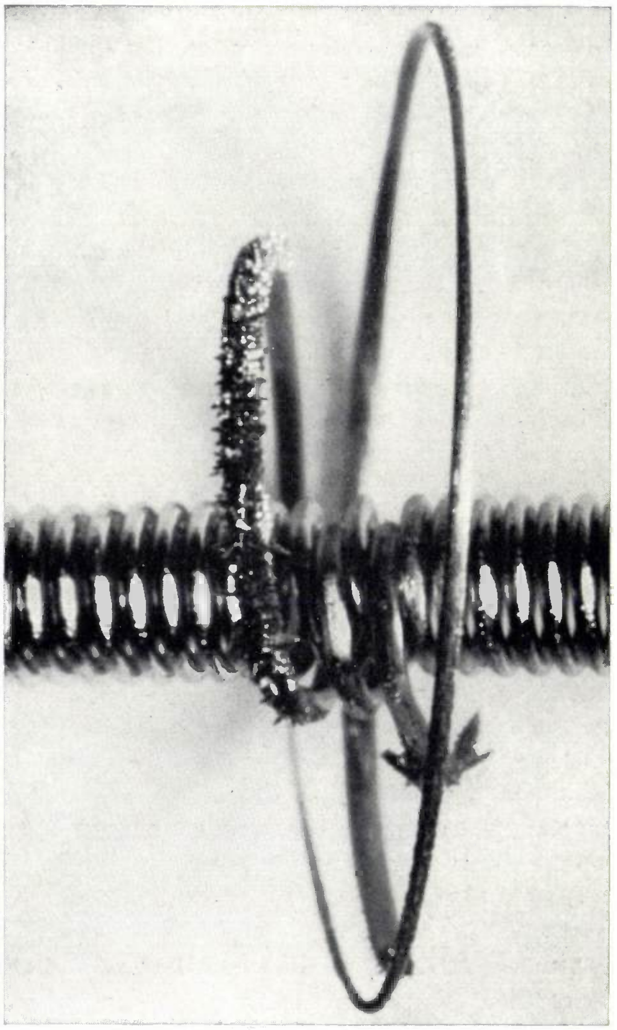


Fig. 2. Typical dendritic growth on the filament supports, due to surplus bromine. Magnification 20×.

With relatively large quantities of chlorine and bromine there is no blackening at all during the whole period of operation: the cycle is obviously working well. The life, however, is much shorter — sometimes even as short as 10 minutes with chlorine. The filament shows strong local dendritic growth (*fig. 2*) and the supports have been attacked.

[2] It is possible that oxygen plays an essential role in the reaction between tungsten and iodine. Recent investigations into this question have been published by J. Tillack, D. Eckerlin and J.H. Dettingmeyer, Z. angew. Chemie **78**, 451, 1966 (No. 8).

[3] C. Collins and E. G. Zubler, U.S. Patent 3 132 278, 1964.

[4] In the course of this investigation we had considerable assistance from R. Meijer and V. Notelteirs of the Turnhout plant of Philips Lighting Division.

[5] For a treatment of the possibilities with fluorine, see J. Schröder, Examples from fluorine chemistry and possible industrial applications, Philips tech. Rev. **26**, 111-116, 1965 (No. 4/5/6).

[6] Although certain side effects are to be expected from the carbon [3], the behaviour of CCl_4 in the lamp showed little difference from that of Cl_2. Experiments with Cl_2 have been carried out by J. Tillack (Aachen laboratory, Philips Zentrallaboratorium GmbH).

With a properly chosen partial pressure for the halogen, lamps with chlorine and with bromine could be made which showed no blackening and which also reached the required life. In the type of lamp studied these optimum pressures ranged from 0.3 to 0.6 torr for CCl_4, and from roughly 3 to 6 torr for Br_2 (*figs. 3* and *4*).

The behaviour described might be explained as follows. When large quantities of chlorine and bromine are used, the relatively cool filament ends and supports are very quickly attacked. If however a low partial pressure is chosen, the cooler tungsten parts are attacked much more slowly, while the regenerative cycle still functions sufficiently well to prevent blackening of the bulb. This is not surprising when one remembers that, in the experimental lamp, for example, the quantity of tungsten evaporating from the filament per hour is a mere few tenths of a milligramme. Pressures as low as 0.1 torr of chlorine and bromine would be sufficient to keep this quantity of tungsten in circulation. It is difficult, however, to maintain such a low pressure in practice during the whole operating life of the lamp, because the halogen is easily absorbed by impurities released during operation [7]. When this happens the cycle breaks down fairly abruptly, and there is serious blackening of the bulb.

This led to the idea of using a halogen pressure as low as this after all, while trying in one way or another to maintain this pressure during the whole operating life of the lamp.

Before going further into this, it is worth noting that more could probably be achieved with iodine lamps if a better approach were made to the optimum choice of iodine content. Although this has not been thoroughly investigated, it was found for example that the tungsten-iodine cycle was far less likely to break down at iodine pressures of about one atmosphere than at the usual pressures of 5 to 10 torr. This line of investigation has not been pursued because at such concentrations far too much light would be absorbed in the violet iodine vapour.

Investigations on the cycle with hydrogen halides

The basis of our further investigations had thus become the maintenance of a very low partial chlorine or bromine pressure during the whole operating life of the lamp. We therefore looked for compounds that partly dissociate at high temperature and, in so doing, supply the required quantity of free halogen. Compounds that exist in such a dissociation equilibrium with free halogen behave as halogen "dispensers", i.e. they replenish the free halogen absorbed by impurities. They act as a chemical buffer. As the halogen pressure has to be low, these buffer compounds have to be thermally very stable.

This also means that free halogen will only be formed to any appreciable extent at the hottest parts of the filament, while at the cooler places, such as the filament ends and supports, the halogen will occur mainly in a bound form, and the concentration of free halogen can

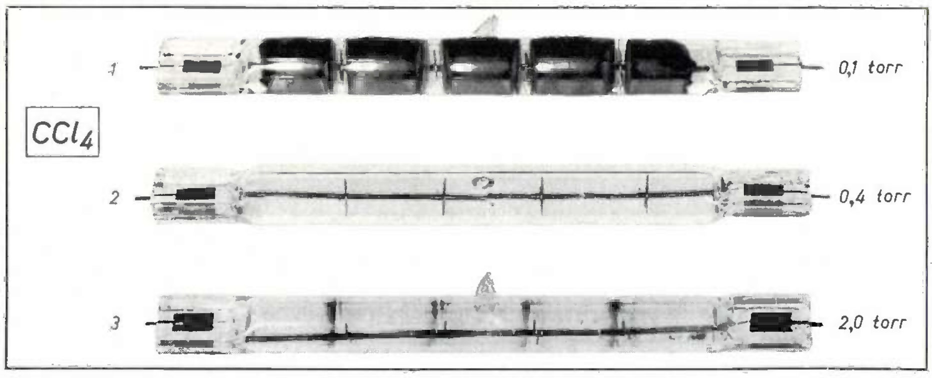


Fig. 3. Experimental lamps (225 V, 1000 W, 12 cm length) containing various amounts of carbon tetrachloride.
Lamp 1: after operating for 5 hours the lamp is still intact but shows severe blackening.
Lamp 2: failure after 26 hours.
Lamp 3: failure after 3 hours.

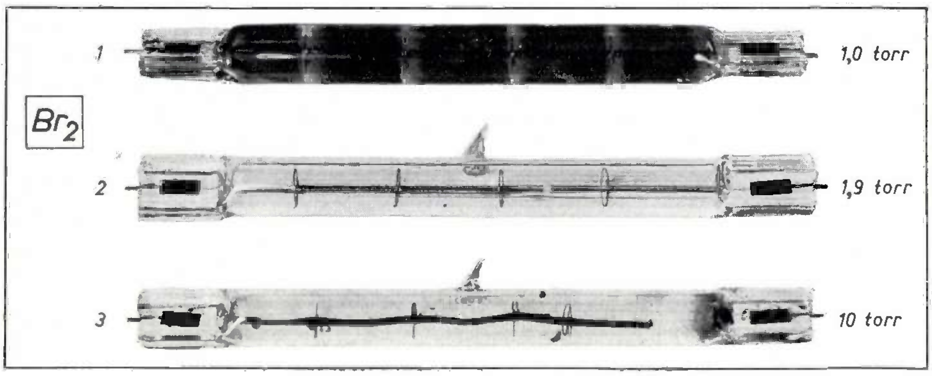


Fig. 4. Experimental lamps containing various amounts of bromine.
Lamp 1: after operating for 15 hours the lamp is still intact but severely blackened.
Lamp 2: failure after 31 hours.
Lamp 3: failure after 15 hours.

thus be kept very low at such places. This distribution of free halogen is favourable for long lamp life, which, as we have seen, is principally threatened by chemical attack at these cooler areas. "Halogen buffers" therefore kill two birds with one stone.

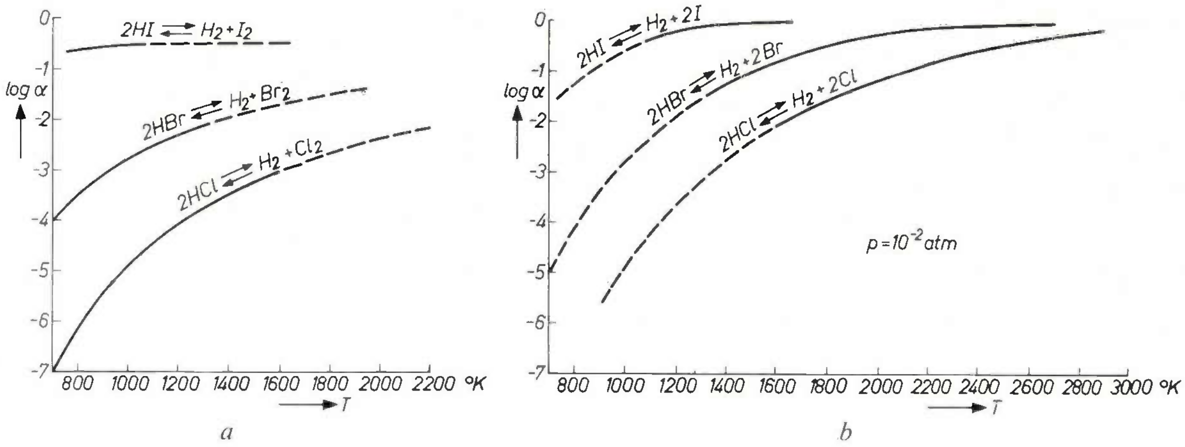


Fig. 5. *a*) The degree of dissociation *α* of hydrogen halides upon decomposition in accordance with the reaction $2\ HX \rightleftarrows H_2 + X_2$, as a function of temperature. The degree of dissociation here is independent of the pressure (equal numbers of gas molecules to the left and right of the arrows). The dashed parts of the curves are only of theoretical significance, since at the high temperatures concerned the molecular halogen X_2 dissociates almost completely into atomic halogen. *b*) The same for the reaction $2\ HX \rightleftarrows H_2 + 2\ X$. The degree of dissociation is now pressure-dependent. The curves relate to a pressure of 10^{-2} atm, which is approximately the hydrogen halide pressure used in the lamp. The dashed part of the curves indicates that in the relevant temperature range the atomic halogen will preferably recombine to form molecular halogen. Owing to the small concentrations and the consequently small collision probability, the atomic state is however certainly possible in this temperature range.

Hydrogen chloride and hydrogen bromide are halogen compounds that meet these conditions fairly well. *Fig. 5* shows the logarithm of the degree of dissociation *α* of these compounds, and of hydrogen iodide, as a function of temperature. At temperatures of about 3000 °K the first two substances have almost completely dissociated ($\log \alpha = 0$) but only a slight fraction dissociates at temperatures of about 1000 °K.

The behaviour of hydrogen halides as a regenerative transport gas was experimentally investigated in the photoflood lamps mentioned above. Using HBr it proved possible to make lamps that worked well in the pressure range from about 4 to about 50 torr; the life of the lamps was more or less normal and there was no blackening. At pressures lower than 4 torr blackening occurred, and at pressures higher than 50 torr the life was shorter (*fig. 6*). The optimum for the experimental lamps was found to lie at about 10 torr, though this value was not critical.

Lamps that worked well were also made using HCl. The establishment of the correct HCl content and the manufacture of the lamps were much more difficult, however, than with HBr. Useful results were achieved

[7] It should be remembered that a halogen with a pressure of 0.1 torr in this lamp volume of 4 cm^3 represents a total quantity by weight of only about 5×10^{-6} g, compared with a filament weight of about 0.5 g. Possible impurities in the filament of the order of 10×10^{-6} are therefore already of the same order of magnitude as the total available quantity of halogen.

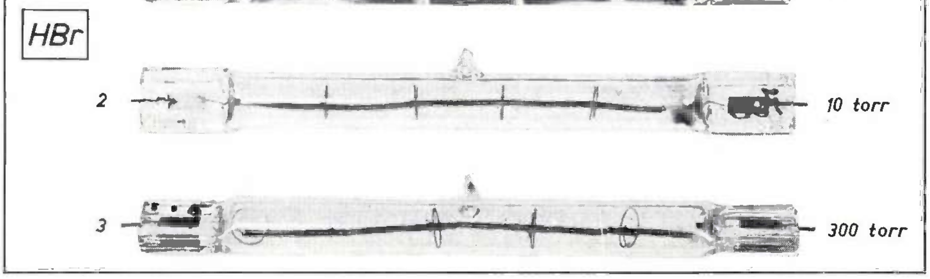


Fig. 6. Experimental lamps containing various amounts of HBr.
Lamp 1: after 17 hours the lamp is intact but blackened.
Lamp 2: failure after 28 hours.
Lamp 3: failure after 17 hours.

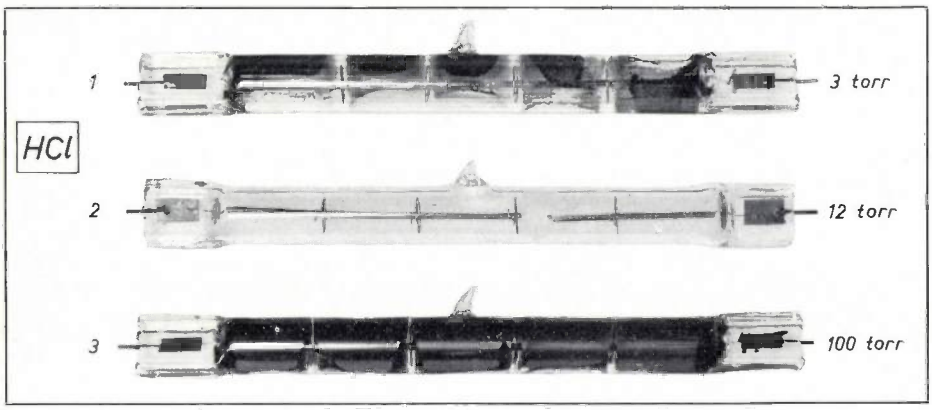


Fig. 7. Experimental lamps containing various amounts of HCl. Lamp 1: after 15 hours the lamp is intact but blackened. Lamp 2: failure after 33 hours. Lamp 3: after 15 hours the lamp is intact but blackened.

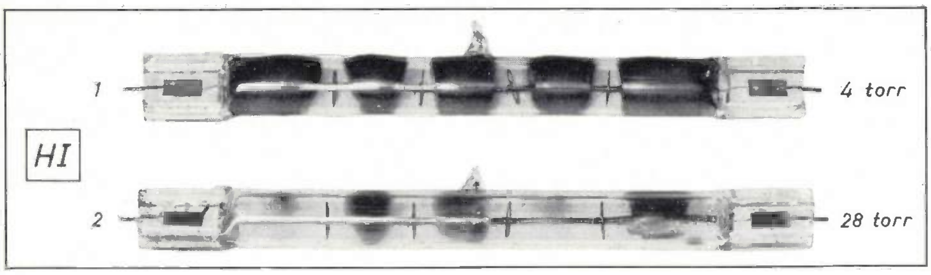


Fig. 8. Experimental lamps containing various amounts of HI. The photograph shows the situation after 15 hours of operation. Both lamps are still intact. As can be seen, the lamps are badly blackened. The blackening decreases with increasing addition of HI.

with partial HCl pressures between about 7 and 15 torr. Both at lower and at higher pressures the bulbs showed some blackening (*fig. 7*). The life was more or less normal over the whole pressure range investigated.

With HI it was not possible to make lamps that worked well in the pressure range from 0 to 100 torr. Intolerable blackening invariably occurred during the operating life of the lamps, although the blackening was slightly less serious at higher HI pressures (*fig. 8*). The life was not clearly influenced by the HI content and was more or less normal.

It is difficult, owing to the complicated nature of the events, to give a complete explanation of the differences in behaviour described above. The following features play an important part:

a) The affinity between tungsten and the relevant halogen at different temperatures, as explained earlier.
b) The competitive affinity between hydrogen and the relevant halogen at different temperatures; this is possibly the main cause of the retarding effect of hydrogen on the cycle, mentioned above.
c) The distribution in the concentrations of free halogen atoms, hydrogen and tungsten in the gas inside the lamp. This distribution depends among other things on the composition and pressure of the gas filling, and also on the local temperature conditions.

Long-life effects

When the experiments with HCl and HBr were extended to types of lamp which were required to have lives of several thousands of hours, complications were encountered. While the lamp is operating, hydrogen escapes through the quartz wall, resulting in a continuous increase in the concentration of free halogen. The behaviour of the lamp *over a longer period* then corresponds in many ways to that of the lamp with the *pure* halogen filling *over a shorter•period*, the result being serious attack of the ends of the filament and the filament supports.

These effects might be dealt with in various ways. For example:

a) By making the bulb of material that lets very little hydrogen through. Hard glass is the only material that can be considered at the present time. A difficulty, however, is that it has a lower softening point than quartz glass.
b) By making the bulb of quartz glass, but surrounding the bulb with a second one made of a material relatively impermeable to hydrogen, such as ordinary glass. The space between the two bulbs then has to be filled with hydrogen under a certain pressure to compensate losses from the quartz bulb. This appears to be an attractive solution for lamps, such as certain reflector types, which already have an extra outer bulb.
c) By adding a surplus of hydrogen to the gas inside the quartz lamp, thus creating a reserve to cover the hydrogen losses. The addition of extra hydrogen, however, affects the dissociation equilibrium of the hydrogen halide, and therefore has its limits. The hydrogen deficit in the lamp is not basically overcome by this method, but simply delayed.

The extent to which the above methods can be used with success depends to a great extent on the required life and the nature of the type of lamp [8].

Investigations on the cycle with halogen-hydrocarbons

In the same way as chlorine can be introduced into the lamp in the form of carbon tetrachloride, which dissociates in the lamp, it is possible to introduce a hydrogen halide in the form of a halogen-hydrocarbon compound, which dissociates in the lamp into hydrogen halide and carbon.

For manufacturing purposes some halogen hydrocarbons have considerable technological advantages over hydrogen halides, as they are chemically not very active. This makes it possible to work with conventional pump materials. Moreover, by the appropriate choice of compound a simple control over the hydrogen-halogen ratio can be obtained. Substances such as $CHCl_3$, CH_2Cl_2, CH_3Br and CH_2Br_2 have been used with good results. CH_2Br_2 and CH_3Br have proved to be particularly useful. Lamps containing these bromine-hydrocarbons are already in production. Owing to its surplus of hydrogen, CH_3Br contains in itself a reserve against the loss of hydrogen through the bulb wall during the operation of the lamp. This makes it suitable for long-life types of lamp. For types with a short specified life, CH_2Br_2 gives the best results.

The carbon released upon the dissociation of the halogen-hydrocarbons possibly acts as a getter [3]. The carbon is released in a very finely distributed state and spreads to all parts of the lamp. An indication that tends to verify this gettering action is the established fact that lamps with a CH_2Br_2 filling show a smaller spread in lifetime than lamps with an HBr filling.

Practical results

After various types of lamp containing the above-mentioned bromine-hydrocarbons had been put into production, comparison with the corresponding iodine lamps from production led to the following conclusions.

The tendency towards bulb blackening has very nearly disappeared. The spread in life values is smaller. About 5% more light output is obtained, since there is no longer any absorption of light in the violet iodine vapour. The lamp is less sensitive to departure from the horizontal operating position. The production equipment is much simpler.

The investigations described show that there is a fairly wide choice of materials for the transport gas in an incandescent lamp with a regenerative cycle. Not only is there a choice between the various halogens, but these can also be combined either with oxygen, as in the iodine lamp, or with hydrogen, as in the bromine and chlorine lamp.

It is therefore likely that there will be some diversity in the transport gases employed, the gas being matched to the nature of the particular type of lamp. The choice of transport gas will be determined to a considerable extent by the filament temperature and the inert gas filling, as these very largely determine the rate of evaporation from the filament and hence the demands on the regenerative cycle.

[8] The use of hydrogen halides in incandescent lamps is also mentioned in U.S. Patent 3 091 718 of 28th May 1963. Here, however, HI is applied in the "preferred embodiment of the invention". It is also specified in the patent that the bulb wall should be permeable to hydrogen for the very purpose of allowing hydrogen to escape from the lamp.

Summary. Experience with iodine lamps has shown that the tungsten-iodine cycle is easily affected by impurities inside the lamp. Moreover, iodine presents various technological problems in the manufacture of the lamp. An investigation was therefore started on the possibilities of using other halogens as transport gas so as to avoid blackening of the bulb. The investigation showed that useful results can also be achieved with chlorine and bromine if these halogens are introduced in very small concentrations and the concentration is maintained throughout the operating life of the lamp. This insight led to a study of various halogen compounds which, at high temperature, are in dissociation equilibrium with free halogen and are therefore able (by a buffer action) to maintain a constant partial halogen pressure in the lamp while it is operating. By choosing stable compounds this partial halogen pressure in the lamp can be kept fairly low. For technological reasons preference is given to gaseous compounds that are chemically not veıy active. Good prospects are opened up by hydrogen halides and particularly by various halogen-hydrocarbons, including CH_2Br_2 and CH_3Br. The latter are now being used in many lamp types produced by Philips. In future there is likely to be a choice between various transport gases for a given lamp. The choice will be determined by the special characteristics required for the lamp.

Integration of electronic circuits

539.234:621.3.049.7

In this number the reader will find three articles on integrated circuits. The following brief introduction attempts to bring the subject into the right perspective.

As electronic equipment became more and more complex, a growing need was felt for more rational methods of building up the circuits from their components. At first, components were separately soldered together by means of connecting wires. A great step forward came with the introduction of "printed wiring", where a photo-etching process is used to form the conductors on an insulating board, usually of resin-bonded laminated paper [1]. The connecting leads of the component are inserted, either by hand or automatically, into holes in the printed wiring board. The connections between the components and the printed network of conductors are made by dipping the underside of the board in a bath of molten solder.

An obvious next step was to use the photo-etching process not only for making the conductors, but for making the resistors as well. This entails numerous problems, particularly in regard to the choice of resistive material and processing methods. The technique employed is based on the evaporation of a suitable metal alloy in a high vacuum so as to deposit thin films of the alloy on to an insulating substrate, usually a special type of glass. In this procedure the connections in the circuits no longer need to be soldered but are directly formed during the evaporation process. This is the basic idea underlying "integrated" circuits; components and connections are fabricated *as a whole*.

The following article by Munk and Rademakers [2] deals with the thin-film technique referred to and describes the results obtained.

Although resistive networks may in themselves be interesting, there is of course a greater demand for circuits containing other elements as well, both active and passive. A development that has resulted in an active circuit element which can be evaporated on to a glass substrate, and which can be used like a transistor for amplification, is described in the article by De Graaff and Koelmans [3]. A welcome feature is that the techniques are compatible, that is to say the processing and heat treatments needed for making the active elements do not impair the quality of the deposited resistors and conductors. Larger-scale industrial application of these thin-film transistors can only commence, however, when certain troublesome effects have been overcome.

Various methods are also being tried for making capacitors by the evaporation process. The best results have been obtained using silicon monoxide as dielectric. A disadvantage of these capacitors is that they cover a relatively large area in the circuits. The same objection applies even more with evaporated inductors, and, as these moreover have low Q values, one may even say that inductors are hardly suitable for integration: as a rule, the circuit should be designed so as to avoid their use, or they should be replaced by sub-circuits of inductive impedance [4].

The foregoing implies that for the present, diodes and transistors, as well as capacitors, have to be made separately and attached later to the evaporated circuit. In order to preserve the advantage of the greatly reduced dimensions in thin-film circuits — miniaturization has long been a traditional pursuit of electronic engineering and is another leading objective in integration — the capacitors are fabricated with the aid of thin titanate layers as small wafers or "chips", which are soldered into the circuit. More or less the same applies to diodes and transistors: these are encapsulated in glass or plastic, and attached by suitable connector contacts.

Although this solution is satisfactory for many applications and the thin film technique offers the advantage of being able to make resistors which have close tolerances, good temperature coefficient and low stray capacitance, the need to solder in additional components does rather detract from the idea of integration.

The solid circuits, dealt with in the article by Schmitz [5], are based on an entirely different principle. Instead of a glass substrate, a wafer of monocrystalline silicon is used. From investigations started in the Bell Laboratories and also pursued by others, what is known as the *planar technique* has been developed. By an ingeniously contrived succession of oxidation, photo-etching and diffusion it is possible to form the transistors, diodes and resistors in a thin layer of the crystal. The connections are made by evaporating the conductors on to the SiO_2 layer produced by oxidation, into which holes have been etched at the places where contact is to be made with the underlying components. In the article referred to [5] this procedure is described in some detail.

The quality of the resistors produced in this way is much poorer than that of the evaporated resistors, but is sufficient for many purposes. If the resistors have to meet higher requirements, they can if necessary be made by the evaporation technique, that is to say by evaporating resistive layers on to the oxide layer — a possibility also discussed in the article by Munk and Rademakers [2].

The planar technique can also be used to form capacitors in the crystal, although the capacitance values that can be realized are relatively low.

Circuits of exceptionally small dimensions can be obtained with the planar technique. Dozens of components, which together may form a specified electrical function, can be accommodated in an area of 1 mm^2. This makes it possible to produce hundreds of these circuits together on a single silicon wafer of perhaps 2.5 cm diameter. Given good control of the techniques, ensuring a high yield, these circuits may be expected to become cheaper to make than those built up from separate components. Moreover, it already appears from experiments that circuits made in this way are very reliable. One advantage is that the simultaneous manufacture of hundreds of such circuits offers much less possibility of human error than the building up of circuits piece by piece. One might compare this to some extent with the printing of a book. Once the printing errors have been removed from the type, one can be certain that all books produced with this type will be free from errors. There is no such certainty in manuscripts copied by hand, in which each letter is separately joined to the next. The article by Schmitz [5] does indeed bring out the correspondence with the technique of colour printing: this is also based on the use of photographically produced masks, which have to fit exactly upon each other in the proper sequence.

The possibility of using this method for extremely reliable production of large numbers of components, with their interconnections, has come just in time. Without it, it would be impossible to produce large electronic equipment such as computers, electronic telephone exchanges, process control systems, etc., all of which contain enormous numbers of components. Moreover, the extremely short connections of the integrated circuits permit a reduction in the signal transit time in large machines. In spite of the fact that stray capacitances adversely affect the speed of the individual circuits, a new approach to the field of extremely high frequencies has become possible. Finally for space flight applications the importance of minimal dimensions and extreme dependability is self-evident.

Integrated circuits thus create new possibilities, but at the same time new problems. One may be surprised to learn, for example, that, as opposed to the situation with separate components, transistors are now cheaper than resistors. This calls for a new attitude on the part of the electronics engineer, who is now much freer in the use of active components. On the other hand, it remains difficult to obtain high capacitances by this technique, and virtually impracticable to make inductors. Wherever possible their use has to be avoided. The best answer to these problems will only be found by close co-operation between electronic engineers and technologists.

The gap between the circuit designer, who up till now has been able to work almost independently of the component designer, and the technologist who made the components, will be considerably narrowed in the years ahead by the use of these new techniques.

Up to the present, with a few exceptions, integrated circuits have been little more than translations of conventional circuits. It is not difficult to trace in the silicon chip the individual components contained in the circuit diagram. It is to be expected, however, when we have learned to make better use of the physical properties of the solid state, that it will prove possible to realize the required electrical functions with fewer "components".

P. W. Haaijman

Dr. P. W. Haaijman is a deputy director of Philips Research Laboratories, Eindhoven.

[1] See R. van Beek and W. W. Boelens, Printed wiring in radio sets, Philips tech. Rev. **20**, 113-121, 1958/59.

[2] E. C. Munk and A. Rademakers, Integrated circuits with evaporated thin films, Philips tech. Rev. **27**, 182-191, 1966.

[3] H. C. de Graaff and H. Koelmans, The thin-film transistor, Philips tech. Rev. **27**, 200-206, 1966.

[4] Circuits of this kind, which might also be desirable in normal electronic circuits when the inductors required would be too large, have been described by G. Klein and J. J. Zaalberg van Zelst, Some simple active filters for low frequencies, Philips tech. Rev. **25**, 330-340, 1963/64.

[5] A. Schmitz, Solid circuits, Philips tech. Rev. **27**, 192-199, 1966.

Integrated circuits with evaporated thin films

E. C. Munk and A. Rademakers

539.234:621.3.049.7

In this article we shall discuss a form of integrated circuit in which the conductors, the resistors and the capacitors (and in some cases the inductors as well) are deposited as thin films in a pattern of strips and rectangles on an insulating substrate. The advantages usually considered in the integration of circuits have been made sufficiently clear in the previous article [1], and need not therefore be reiterated here.

The thin film technique we shall deal with in this article may be regarded as an extension of the techniques in use for making conventional resistors and capacitors, in which thin layers of carbon or metal are applied to an insulating body or dielectric foil. There are various methods of applying the layers, but we shall be concerned only with those where the material is *evaporated* on to the substrate in a high vacuum.

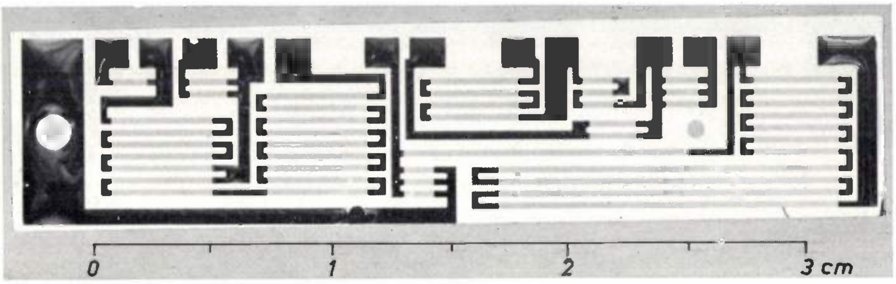


Fig. 1. Example of a pattern of conductors and resistors in an integrated circuit. The conductor layers (dark patches) are tinned.

The substrates are small flat plates, or wafers. This is important both in the manufacturing process, as it facilitates work with masks and photographic processes, and in the assembly of the circuits in larger systems, since the plates can be stacked and interconnections made at the edges.

It should be emphasized that the circuits treated in this article are not completely integrated but of a hybrid nature. In a completely integrated circuit the active elements as well (diodes and transistors) would have to be made by evaporation techniques. This is possible in principle, as explained elsewhere in this number [2], but such elements are still at the research stage. In practice, the active elements are therefore added as separate components to the thin film circuit. Moreover, it is quite often necessary to fit separate capacitors (and possibly inductors), since the capacitance or inductance value and the Q that can be achieved with the thin film components may be too limited.

An example of a thin film circuit is shown in *fig. 1.*

General considerations

The integrated circuit — unlike a circuit built up from individual components — is not flexible, and subsequent correction is acceptable only in exceptional cases. This means that it is necessary to control the production process so thoroughly as to make later correction unnecessary; in other words, it should be cheaper to scrap a product that fails to meet the specifications of the design than to correct it.

This lack of flexibility becomes particularly evident if one compares a resistor in an integrated circuit with a resistor as a separate component. In both the resistive material is applied in the form of a thin film to an insulating substrate (which is cylindrical in the separate resistor). The value of the resistance is given by:

$$R = \frac{\varrho}{t}\,\frac{l}{b},$$

where ϱ is the resistivity, and l, b and t are the length, width and thickness of the resistive film. The factor ϱ/t, the sheet resistivity (also referred to as the resistance per square, $R_\square$), is a property of the film itself, and the length-to-width ratio l/b is called the aspect ratio. The relative error in R is the sum of the relative errors in the sheet resistivity and the aspect ratio. A conventional resistor is made by first applying the resistive material to the body. The accuracy of the sheet resistivity is not critical, since the resistor is given its value with the required precision by grinding a spiral groove into the surface layer. This corrects the aspect ratio after the evaporation process. Precisely the opposite applies to a

Drs. E. C. Munk is with Philips Research Laboratories, Eindhoven, and Dr. A. Rademakers is with the Nijmegen Semiconductor Plant, Philips Elcoma Division.

resistor in a thin film circuit: the dimensions of the resistor are fixed before the film is deposited on the substrate. If there are to be no subsequent corrections, the aspect ratio and the sheet resistivity are therefore established independently of one another, which means that this must be done with tolerances which *together* do not exceed the rated tolerance of the resistor. The requirements for the process of applying the resistive layer are therefore much stricter than in the manufacture of separate resistors. These requirements become even harder to meet if the films for the resistors in a large number of circuits are to be evaporated simultaneously. The resistive layer must then not only be extremely homogeneous over the entire surface of each substrate, but must also be identical from one substrate to another. These requirements can be met by using the evaporation technique.

The problem can be solved in an entirely different way if subsequent individual corrections to the sheet resistivity are allowed. The methods developed by IBM and the Bell Laboratories are examples of this approach [3][4]. In both methods relatively thick films are used, and the thickness is reduced after deposition in the required pattern. In the IBM method the resistive material, in the form of a paste, is screen-printed in the required pattern upon ceramic wafers. After heat-treatment, the resistance is corrected by sand-blasting. In the second method the resistive material (tantalum with tantalum nitride TaN) is applied to wafer substrates by sputtering in a gas discharge. After deposition, part of the film is converted by electrolytic oxidation into non-conducting Ta_2O_5. The latter process can now be so well controlled that a very high accuracy can be achieved. During the process, however, each resistor has to be individually checked. Moreover, optimum properties can only be obtained provided the sheet resistivity is given a relatively low value (e.g. 40 Ω). This means that resistors with a high nominal value can only be produced if the aspect ratio has a high value, and this makes it difficult to meet the requirement for small dimensions.

By way of introduction to the subject, let us make an estimate of the thicknesses of the films used in these techniques. Suppose that we wish to form a combination of resistors with a tolerance of 5%, the highest nominal value among them being 10 000 Ω. We choose for this a strip 1 cm long and 100 μm wide. The aspect ratio is then 100 and so the sheet resistivity is 100 Ω. If we now unsuspectingly put in the bulk resistivity value for say aluminium (2.5×10^{-8} Ωm), we arrive at a film thickness of 0.25 nm! The tolerance of 5% now has to be shared between the aspect ratio and the sheet resistivity. This implies that we have to reproduce not only the width of 100 μm but also the thickness of 0.25 nm with a tolerance certainly below 5%. This is scarcely practicable, and it follows that materials of higher resistivity must be used. These are to be found in alloys having resistivities of the order of 100×10^{-8} Ωm. Even with such alloys, the film thicknesses are still extremely small, i.e. of the order of 10 nm.

Various other considerations besides those already mentioned play a role in the fabrication of evaporated resistive films. In particular we should mention the efforts to achieve films of low temperature coefficient and high stability: this will be discussed later.

If capacitors are to be built up from evaporated thin films (a dielectric layer between two metal layers) similar considerations indicate the requirements to be met in the control of the evaporation process, in particular for the deposition of the dielectric layer. For the conductors, the process is of course less critical.

For inductances the thin-film technique is by its nature not particularly suitable. Without going into detail, two points may be mentioned in passing: the three-dimensional character is of more significance with the inductor than with other elements, and it is difficult to concentrate magnetic energy in a small volume. For simple requirements, inductors may be evaporated in the form of spirals. In practice this method cannot be used for inductances greater than a few tens of μH at the most, and to achieve a fairly reasonable Q the films have to be thickened, e.g. with a layer of gold.

Thin-film effects

Before going further into the production of thin-film circuits, we shall first discuss some physical peculiarities of thin metal films.

The resistance of thin metal films is usually greater than the value calculated on the basis of the film thickness and the resistivity of the metal; in other words the "effective resistivity" of the film is greater than that of the bulk metal.

An effect of this nature is to be expected in principle from the theory of electrical conduction in metals. Here if the thickness of the films becomes smaller than the mean free path of the conduction electrons, the scattering of the electrons at the surface will make a perceptible extra contribution to the resistance. It has in fact proved possible to determine fairly accurately the mean free path from resistance measurements on thin films, at least for very pure metals in virtually ideal thin films.

Ideal films, that is to say films that are continuous, uniform and homogeneous, are however very rarely encountered. Usually resistance anomaly in thin films

[1] P. W. Haaijman, Integration of electronic circuits, Philips tech. Rev. 27, 180-181, 1966.

[2] H. C. de Graaff and H. Koelmans, The thin-film transistor, Philips tech. Rev. 27, 200-206, 1966.

[3] E. M. Davis, W. E. Harding, R. S. Schwartz and J. J. Corning, Solid logic technology: versatile, high-performance microelectronics, IBM J. Res. Devel. 8, 102-114, 1964.

[4] D. A. McLean, N. Schwartz and E. D. Tidd, Tantalum-film technology, Proc. IEEE 52, 1450-1462, 1964.

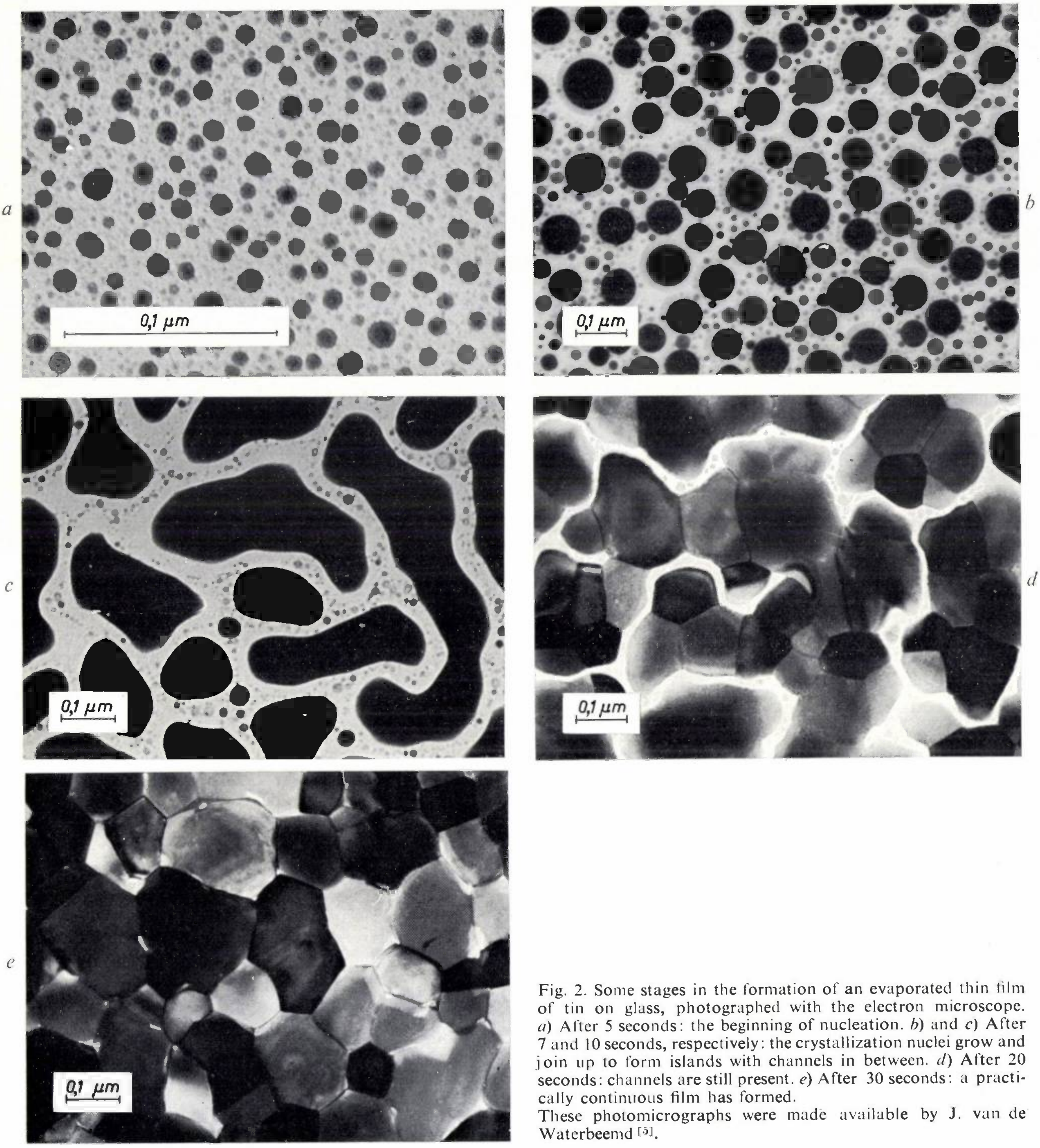


Fig. 2. Some stages in the formation of an evaporated thin film of tin on glass, photographed with the electron microscope. *a*) After 5 seconds: the beginning of nucleation. *b*) and *c*) After 7 and 10 seconds, respectively: the crystallization nuclei grow and join up to form islands with channels in between. *d*) After 20 seconds: channels are still present. *e*) After 30 seconds: a practically continuous film has formed.
These photomicrographs were made available by J. van de Waterbeemd [5].

arises from the very fact that the film is *not* ideal.

It has been found from electron-micrographs that a thin film forms by a process of nucelation and growth of crystallization nuclei during the evaporation (see *fig. 2*). In the initial stages the film consists of a large number of separate islands. As the deposition continues, this number increases and the islands already formed grow gradually larger. If two of them come into contact they may join together to form a larger one. In the stage where the islands cover practically the whole substrate an extensive labyrinth of channels remains however for some time. These channels also gradually disappear, so that finally a continuous film is formed. This process can to some extent be checked by continuously measuring the conductivity of the film during the evaporation process. At first the conductivity remains close to zero for some time, and then it increases linearly with time. In some cases, e.g. tin on glass, the channel network is extremely persistent, and a continuous layer does not form until the film has reached an appreciable thickness, of the order of a few tenths of a micron [5].

A structure may also be formed in other ways. If, for example, an unbroken, homogeneous thin film is heated in air, recrystallization takes place in the film. It is possible that oxidation at the grain boundaries causes some electrical separation between the various domains.

The presence of such structures obviously has considerable influence on the resistance. Quite often electrical conduction occurs even before the island system has completely closed. How this conduction comes about is not yet quite clear. It is probable that thermionic emission or tunnel processes play some part in the transport of electrons from one island to another. Where the grain boundaries are oxidized the conduction must take place through the oxide films.

These phenomena have a considerable effect on the temperature coefficient of resistance. In general, thermal excitation of the electrons is necessary to (or at least assists) the kinds of conduction we have referred to, so that more electrons take part in the conduction as the temperature increases. This makes a negative contribution to the temperature coefficient. The result is that the temperature coefficient, which is positive in the bulk metal (higher resistance at higher temperatures), is smaller for thinner films and in extremely thin films may even be negative.

For practical purposes the thin-film effects are therefore advantageous in two respects: the effective resistivity is greater than in the bulk metal and the temperature coefficient is smaller.

Choice of material and evaporation technique

Substrate

The first requirement of a substrate is that it must be sufficiently *smooth.* In order to deposit a film having a thickness of the order of 10 nm with adequate precision, the roughness of the substrate surface should at the most be of the order of 10 nm. A further requirement is that the material should be sufficiently *heat-resistant,* for during the vacuum evaporation process it has to be heated to 250-300 °C to obtain a stable and firmly bonded film. Other requirements that may have to be met depend on the heat generation in the circuit. Because of the miniaturization the dissipation per unit volume may be substantially greater than in conventional circuits. If this results in a high operating temperature, the *electrical quality* must be correspondingly high, so that for example ion transport damage in the substrate does not occur. On the other hand, high operating temperatures can be avoided by measures to ensure adequate heat removal. This sets requirements on the *thermal conductivity* of the substrate.

Glass is a useful material for this purpose. It is sufficiently smooth and heat-resistant. A type of glass with adequate electrical quality must be chosen; cheap soda-lime glass, for example, has limitations in this respect above 100 °C. Normal organic materials are ruled out because they are not heat-resistant. Ceramics are not as a rule smooth enough; even after thorough lapping they cannot be made much smoother than 0.5 μm. Ceramic substrates coated with a thin layer of glass can, however, be used. If thermal conductivity is important, one might consider using ceramics such as porcelain and steatite, which are a few times better than glass in this respect. Sintered Al_2O_3 or sintered BeO, whose thermal conductivities are about ten and a hundred times better than glass respectively, could also be employed.

Resistors

We shall now consider the resistive films in somewhat more detail, confining ourselves to a type made by Philips. The starting material is a nickel-chromium alloy. In wire form and with the composition 80% Ni, 20% Cr, this has a resistivity of roughly 100×10^{-8} Ωm. Due to oxidation and the thin-film effects mentioned above, the effective resistivity in the vacuum-deposited form is considerably higher, say 400×10^{-8} Ωm. Moreover, if the composition is suitably chosen the temperature coefficient of NiCr is small.

NiCr has two further important properties that strongly indicate the choice of this material. Firstly, it adheres well to glass, — clearly an important feature. Secondly, the material begins to evaporate quite rapidly at temperatures below the melting point. This means that a solid source can be used for evaporation. A linear source can therefore be used, and this is of considerable advantage in the uniform coating of large areas. This is much more difficult if the source is a molten metal.

The vacuum-evaporation equipment, which is completely contained in a bell-jar, is illustrated in *fig. 3.* The substrates are fixed to the inside of a cylinder which rotates at a uniform speed around its axis during the evaporation. Inside the cylinder, parallel with its axis, is a linear NiCr source in the form of a wire or strip, which is heated by an electric current to a temperature just below the melting point. The extent of the source allows a reasonable rate of deposition. For instance, with an 80 cm nickel-chromium wire of 2 mm diameter as the source, a film of sheet resistivity $R_{\square} = 300\,\Omega$ can be deposited on an area of 0.3 m² in 10 minutes. The source is partly screened, so that only a few at a time

[5] J. van de Waterbeemd, Philips Res. Repts. **21**, 27-48, 1966, (No. 1).

of the substrates on the cylinder wall receive the sublimed material. In addition to the NiCr source there are a number of radiant heaters inside the cylinder, which heat the glass substrates to 300 °C prior to evaporation, in order to ensure well-bonded and stable films, and also a nickel source for depositing the conductor films.

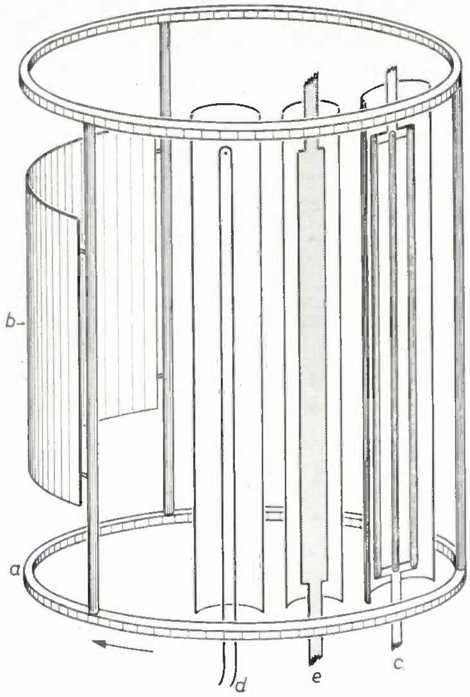


Fig. 3. A sketch of the evaporation system. The cylinder *a*, with detachable shields *b* to which the substrates are fixed, rotates during the evaporation. The substrates are pre-heated by radiant heaters *c*. The resistive films are evaporated from the nickel-chromium source *d*, the conductors from the Ni source *e*. The cylinder is mounted inside a bell-jar.

The speed of rotation of the cylinder and the evaporation rate are chosen so that the required sheet resistivity is reached after a large number of revolutions, say 100. This enables a high degree of uniformity to be obtained for the sheet resistivity of substrates all round the circumference of the cylinder. Uniformity in the vertical direction is achieved by shielding the centre of the linear source rather more than the top and bottom. With this method it proved possible to cover the 0.3 m^2 glass areas mentioned above with a uniformity within 5%. During the evaporation the resistance of one of the substrates is continuously monitored.

A point of importance in the reproducibility of the sheet resistivity $R_\square$ is that in the evaporation of nickel-chrome the constituents are not equally volatile, so that the composition at the surface of the wire changes in the course of time. As a result, the composition of successively evaporated films is not identical. This difficulty could be circumvented by using a sufficiently thick wire and pre-heating it for a sufficiently long time; if this is done a steady state is ultimately reached [6]. In practice, however, it proved more attractive to use a wire once only, for a relatively short period and after an appropriate pre-heating time. This makes it possible to control both the temperature coefficient and the stability of the film, by choice of the pre-heating time — and hence the composition.

The properties of the film are also determined by the quantity of oxygen it contains; this can be controlled by means of the pressure in the bell-jar, which as a rule is about 10^{-5} torr.

Thin films are in general unstable, in the sense that their resistance increases in the course of time, in particular at high temperatures; the resistance increases quickly at first, then more slowly. The films are therefore artificially aged by heating in air. The change of resistance which then occurs — at the most a few per cent — is at the same time a measure of the stability. This change is taken into account in the evaporation process. Generally speaking, the thicker the film, the better its stability.

The film finally obtained is the result of a compromise between efforts to achieve high sheet resistivity, low temperature coefficient and high stability; this compromise is arrived at by an appropriate choice of pre-heating time, evaporation time and pressure. A typical practical example is a film with a sheet resistivity of 300 Ω, a temperature coefficient smaller than 10^{-4}/°C, and a resistance change of less than 1% after a thousand hours of heavy-duty operation.

Conductors

The problems involved in depositing the conductor films are not so great as with the resistive films: the main consideration is that the conductivity should be sufficiently high. There is no question of close tolerances. The obvious choice is a material that has a low resistivity, i.e. a pure metallic conductor. Gold is often used for this purpose. Some care does have to be exercised to obtain a sufficiently high conductivity. If, for example, a conductor with an aspect ratio of 10 (i.e. 10 times as long as it is wide) is required to have a resistance less than 0.1 Ω, then if gold is used the film must be at least 2.5 μm thick. In evaporation technique this is a considerable thickness. At Philips the conductor connections are made by depositing a nickel film and then tinning the complete nickel conductor pattern. The tinning is used to overcome the disadvantage of the relatively poor conductivity of nickel, which in other respects, for various technological reasons, is a very suitable material. The conductor pattern can be selectively tinned because the tin does not adhere to the nickel-chrome resistor pattern.

Dielectrics

Dielectric films as well can be deposited as the dielectrics for the capacitors in the circuit. The object is to obtain the highest possible capacitance per unit area while ensuring that the breakdown voltage remains at a safe margin above the working voltage. Low dielectric losses and not too high a temperature coefficient are also desirable.

The film that has been most investigated and used is a silicon oxide film produced by evaporating silicon monoxide. The composition of the film again depends on what exactly happens in the vacuum during evaporation. It is certain that there is more oxygen present in the deposited film than indicated by the formula SiO. A safe field-strength for films of this kind is 10 V/μm. At an operating voltage of a few volts the minimum thickness must then be 0.5-1 μm, which, for $\varepsilon_r = 4$ to 5, gives a capacitance of 40 to 80 pF per mm^2.

Dielectric films are much more sensitive to imperfections than resistive films. If there is a hole in the dielectric film, the result will be a short-circuit when the next electrode is deposited, whereas in the resistive film the current would simply flow around the hole. It is therefore particularly important to avoid dust when making dielectric films. Experience has shown that the chance of short-circuiting increases sharply with the size of the surface. In practice the capacitances used are limited to a few thousand pF.

Various ways of increasing the capacitance per unit area have been tried, so far without conspicuous success. One is to use much thinner dielectric layers such as those found in the electrolytic capacitor. This is done for example by evaporating tantalum, anodically oxidizing it and then, after it has been dried, evaporating an upper electrode on to it [4]. By this method 1000 pF/mm^2 can be achieved for voltages of the same order as mentioned above (5 to 10 V). As yet, however, reasonably reliable capacitors have been made by this method on only a modest scale. Much less progress has been made in efforts to deposit materials that have a relatively high dielectric constant, such as TiO_2 and $BaTiO_3$.

It has proved to be a practical proposition to use thin ceramic wafers of the latter materials (e.g. 0.1 mm thick) which have been provided with evaporated electrodes. If these are soldered flat in the circuit they take up little space, and the existing choice of materials and thicknesses makes it possible to achieve capacitances up to 200 pF/mm^2.

Generating the pattern

In addition to the problem of producing films with specific properties there is also the problem of how to produce the required pattern of strips and rectangles, as illustrated for example in fig. 1.

An obvious method, and one which is very widely used, is that of evaporation through masks. The films are then formed directly in the pattern required. The whole process takes place in a number of stages, in which each successive material is evaporated through an appropriate mask. *Fig. 4* shows the steps required for making various elements. For making a circuit with resistors, capacitors and crossed connections, as many as six steps may be needed. If the production is to be reasonably efficient, the operations corresponding to these steps, including changing of the sources and masks, should be performed while maintaining the vacuum. Moreover, the masks must be aligned with the utmost precision. All this calls for machines of such complexity that this method has not yet in practice been used for economic production runs.

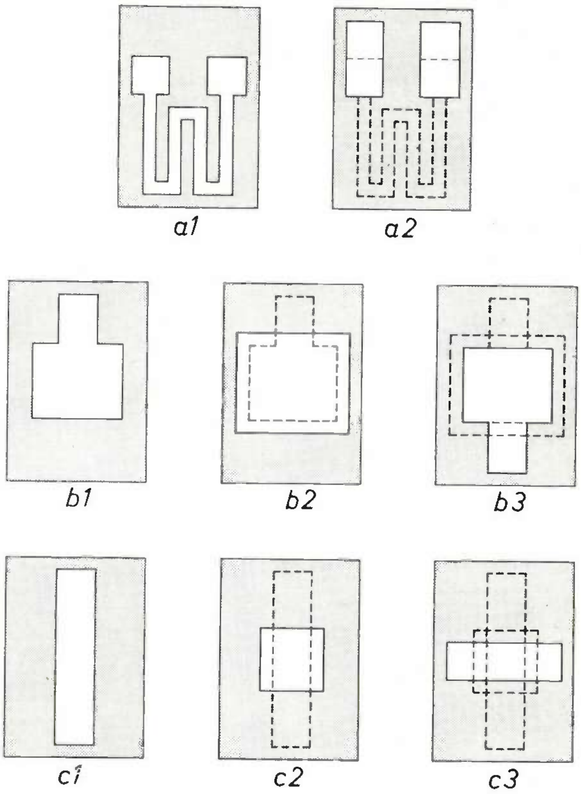


Fig. 4. The formation of a pattern by means of masks during evaporation. The masks that have to be used successively to form a given element are shown for three different elements.
a) Resistor with conductor contacts: *1* mask for the resistive film, *2* for the conductive layer.
b) Capacitor: *1* mask for the first electrode, *2* for the dielectric, *3* for the second electrode.
c) Crossed connection: *1* mask for the first conductor, *2* for the insulator, *3* for the second conductor.
b) and *c*) can be carried out simultaneously.

[6] P. Huijer, W. T. Langendam and J. A. Lely, Vacuum deposition of resistors, Philips tech. Rev. **24**, 144-149, 1962/63.

An entirely different solution of the problem consists in evaporating a film that completely covers the substrate and then, having coated the film with a resistant lacquer in the pattern required, etching away the superfluous parts.

A particularly suitable etching method is the photo-etching process (see *fig. 5*), which has long been used for making printing-blocks and printed wiring, and which is also employed in the fabrication of solid circuits, as described elsewhere in this number [7]. The lacquer used becomes soluble after illumination, in certain solvents (or insoluble; both types are available). The lacquer is applied to the whole surface of the wafer, illuminated through a photographic mask and "developed", i.e. the exposed parts (or the unexposed parts, as appropriate) are dissolved. Finally, the parts of the deposited material not protected by lacquer are etched away.

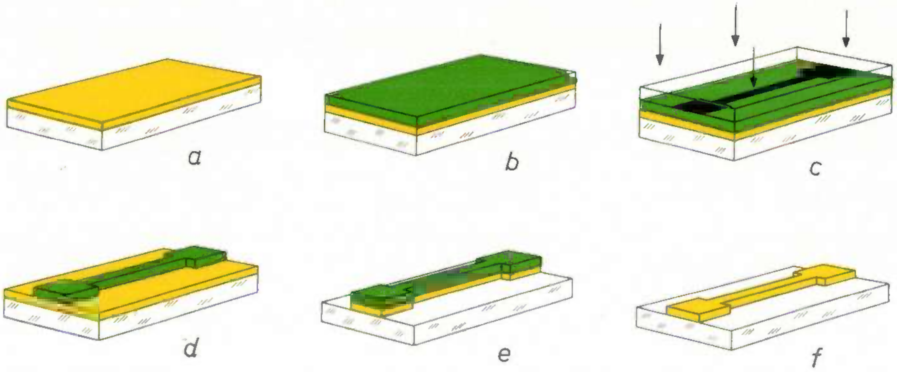


Fig. 5. The photo-etching method. *a*) Substrate with metal layer (yellow); *b*) the metal layer covered with a lacquer layer (green); *c*) the lacquer layer is exposed through a photographic mask; *d*) the lacquer is photographically developed; *e*) the uncovered metal is etched away; *f*) the residual lacquer is removed.

If the film is not readily susceptible to chemical attack, use can be made of an underlayer of an easily etchable material, e.g. copper. The copper film is first evaporated on to the substrate, the pattern is etched out, and then the desired material is evaporated. Next, the copper is removed by etching, the material deposited on it being removed with it at the same time. An advantage of this variant is that the details of the pattern can be checked before evaporating the resistive film.

In practice the techniques described can be combined in various ways for producing a circuit. One example is the following method of making a combination of resistors and conductors [8]. In a single vacuum cycle a double layer is deposited on the substrate, first the resistive layer and then the conductive layer on top. Next, the double layer is etched away so as to leave the required pattern, and finally the conductive material is selectively etched away at the places where the resistors are to be located. This method obviously sets special requirements on the materials; the resistive film must be etchable, but it must be resistant to the etch that removes the conductive layer.

If the element is a "meander" of strips (see fig. 1) of width b separated by insulating strips of the same width, the resistance that can be produced *per unit area* is equal to $\frac{1}{2}R_\square/b^2$, where $R_\square$ is the sheet resistivity. This value thus increases as the strips get narrower. In a given fabrication process, however, the relative accuracy in the width, and hence in the resistance, then decreases. The strip width is therefore generally chosen to give a compromise between a high resistance per unit area and a high resistance accuracy. With the methods described above it is possible to reproduce strip widths with tolerances as small as 0.01 mm. Strips 0.3 mm wide can then easily be made with a tolerance of ± 5%. With $R_\square = 300\ \Omega$ the expression quoted shows that a value of 1600 Ω/mm^2 can be obtained.

The complete circuit and the encapsulation

To complete the circuit the active elements, diodes and transistors, have to be added to the network of resistors and capacitors. The active elements should preferably be adapted to the thin-film technique; they should be small and easy to mount in the flat circuit. The semiconductor devices are generally small enough, but the method of attaching them calls for some attention.

In practice conventional transistors are often used. Their connections are wires or tags, which can be connected appropriately, for example by pressing the connectors into the tinned conductors of the circuit with a heated pin. In a refinement of this procedure the connectors can be soldered either ultrasonically or by means of localized current pulses. Better prospects for fast production in long runs are offered by the use of transistors mounted on a ceramic wafer, which carries the connectors in the form of solid, pre-tinned metal contacts (see fig. 8). If the pattern of the contacts matches that of the circuit, these transistors can be directly mounted and soldered on the circuits. This can be done automatically, and for a large number of circuits at a time, by using alignment jigs, and heating in a furnace. This method is often used with planar silicon transistors. In a promising variant of these transistors the ceramic wafer is completely dispensed with, and the silicon crystal is soldered directly into the circuit. The crystals, which are about 0.6 × 0.6 mm, are pre-coated with an extremely thin layer of glass. The contacts are minute tinned beads mounted in small holes in the glass film [9].

As a protection against atmospheric effects, in particular against moisture, the whole circuit is encapsulated. The requirements to be met by the encapsulation depend on the conditions in which the circuit is to be used. The semiconductor devices are always the most sensitive elements in the circuit. They can be fitted hermetically sealed into the circuit, so that less exacting requirements have to be made on the encapsulation of the circuit as a whole. There is, however, a tendency to use semiconductor devices that are covered merely with plastic or lacquer and to enclose the entire circuit in a hermetically sealed can. An example can be seen in *fig. 6*. The can is fitted with pins, so that it can be connected to a printed wiring panel.

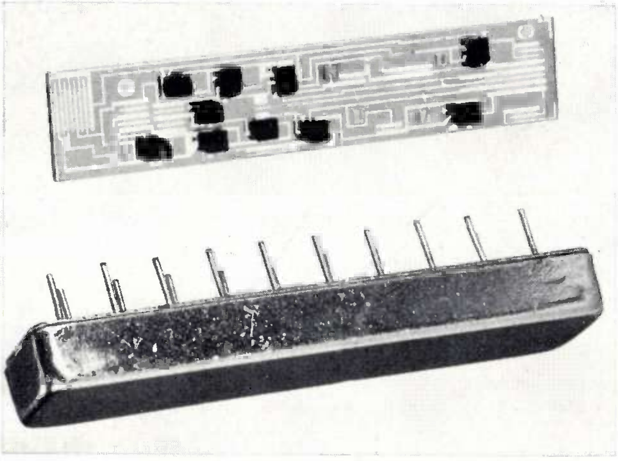


Fig. 6. Example of a hermetically sealed thin-film circuit.

Some applications

We shall now briefly describe two examples of circuits in which the foregoing technique has been successfully employed. Both are hybrid types of circuit in which the conductors and resistors have been deposited as thin films and the other elements — capacitors, transistors, and, in the first example, inductors — are added later.

The first example is a stage of a broad-band amplifier for the frequency range from 40 to 230 Mc/s (*fig. 7*). This amplifier is used for the simultaneous distribution of a large number of signals from television and FM broadcast aerials. To make this possible the amplifier is provided with negative feedback which reduces the distortion to a level at which signals do not interfere with one another. Apart from the transit times in the transistors, capacitive and inductive elements also contribute to the parasitic phase shift in the feedback loop. To prevent oscillation tendencies the parasitic phase shift should be kept a great deal smaller than 180° up to that frequency at which the loop gain has dropped to unity. This frequency is much higher than 230 Mc/s, the highest frequency to be transmitted. The phase shift is kept small by making the feedback loop very short. This means that the circuit must be *small*.

A second point is that the pattern should be sharply defined dimensionally and be *reproducibly* fabricated. This is important because unwanted couplings — in spite of the screening effect of a "ground-plate" (see fig. 7) — cannot entirely be avoided; these unwanted couplings can be taken into account in the design, but since the tolerances are small they must be sharply reproduced.

In these respects the thin-film technique is superior to other manufacturing methods. Although printed-wiring assembly is excellently reproducible, it is not so small (the capacitance of two neighbouring soldered joints for example, is nearly 1 pF). Miniature circuits can indeed be made by the direct interconnection of separate components, but this method is not adequate where reproducibility is concerned.

Finally, the resistors and capacitors must be purely *linear* in view of the linearity requirements which the amplifier has to meet. In this respect the thin-film circuit is preferable to the solid type, in which the resistors tend to be non-linear owing to their semiconductor nature.

The amplifier contains a few small isolating inductances, which need only have a low Q (about 20). The possibility of producing these in the form of evaporated spirals has been considered but rejected because the spirals take up too much space (20 mm diameter for 1.2 μH). Moreover, even though only a low Q is needed, special measures would have to be taken to improve the conduction sufficiently.

The second example (*fig. 8*), a digital (double NAND) circuit, will be dealt with briefly. Circuits of this kind are used in large numbers in electronic computers. Here again, the principal virtue of the thin-film technique is that it makes *miniature* circuits possible, so that compact computers can be built. Moreover

[7] A. Schmitz, Solid circuits, Philips tech. Rev. **27**, 192-199, 1966.

[8] See C. W. Skaggs, Photo-etching thin-film circuits, Electronics **37**, No. 18, 94-98, 1964.

[9] E. M. Davis, W. E. Harding and R. S. Schwartz, An approach to low cost, high performance microelectronics, 1963 WESCON tech. Papers, Part 2, publ. No. 13.1.

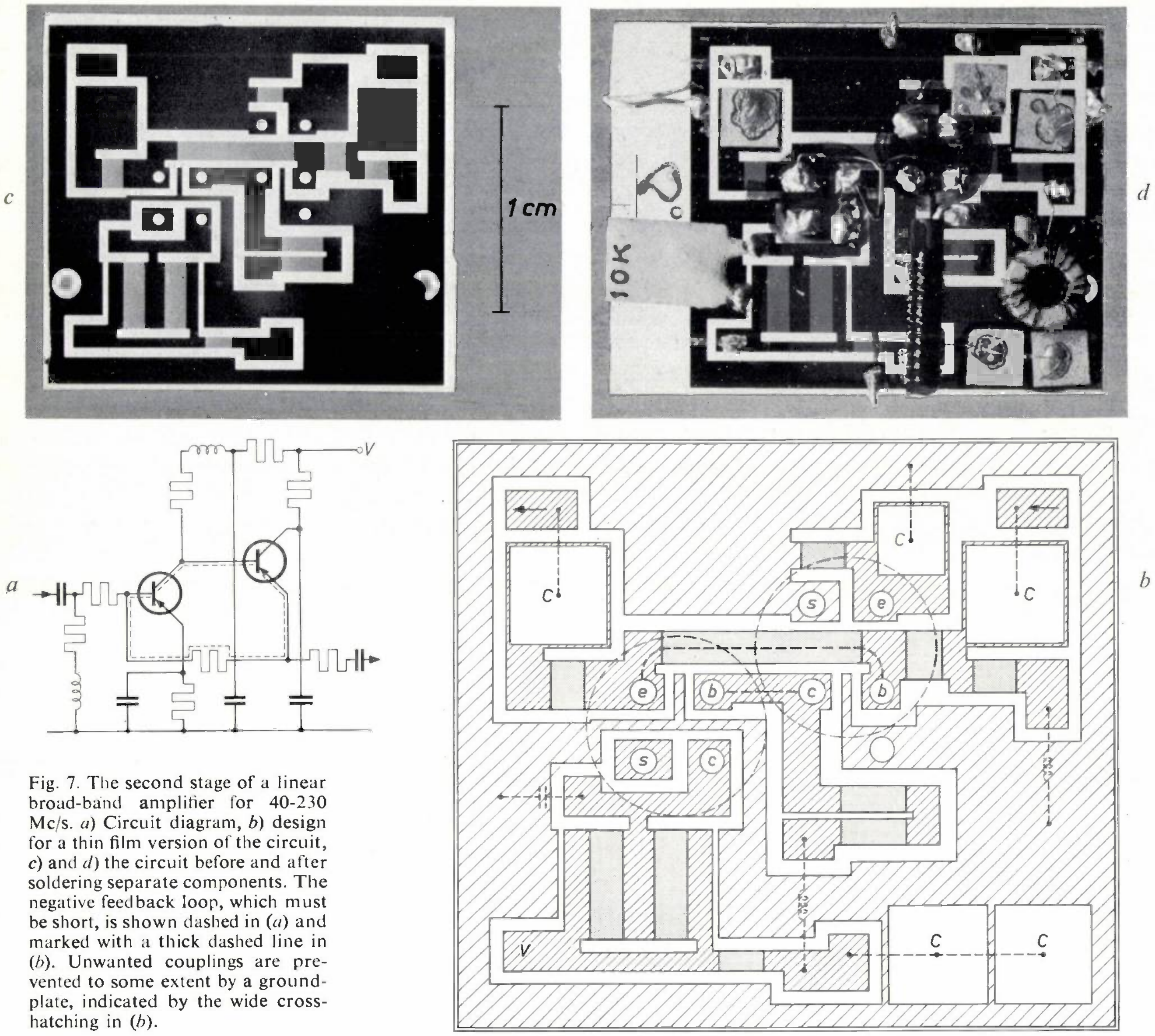


Fig. 7. The second stage of a linear broad-band amplifier for 40-230 Mc/s. *a*) Circuit diagram, *b*) design for a thin film version of the circuit, *c*) and *d*) the circuit before and after soldering separate components. The negative feedback loop, which must be short, is shown dashed in (*a*) and marked with a thick dashed line in (*b*). Unwanted couplings are prevented to some extent by a groundplate, indicated by the wide cross-hatching in (*b*).

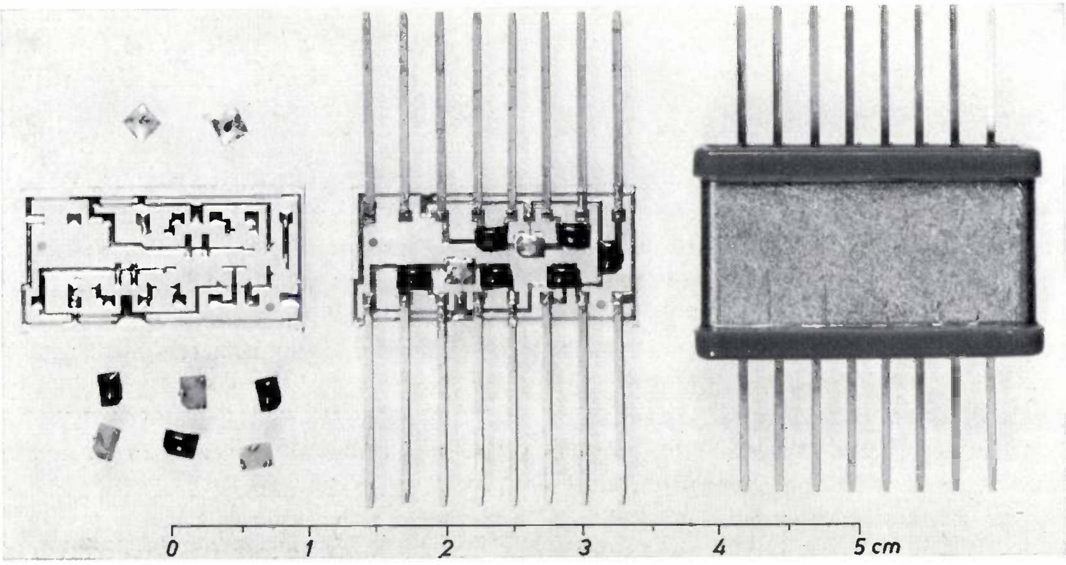


Fig. 8. A double NAND circuit made by thin-film technique.

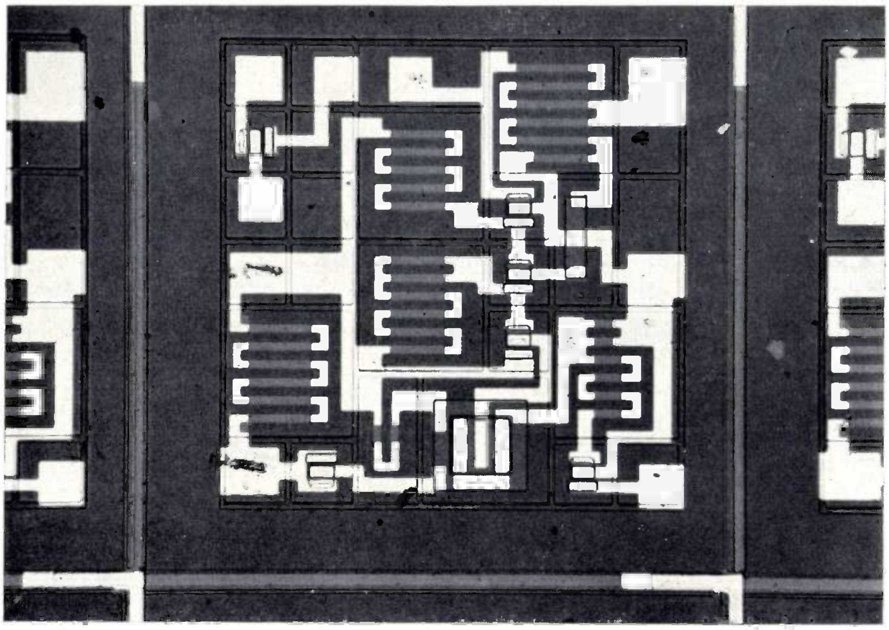


Fig. 9. A solid circuit with evaporated films. A number of these circuits are made at the same time on a single crystal wafer. Each circuit covers an area of dimensions 1 × 1 mm.

— partly because the circuits are so small — large numbers of them can be produced simultaneously in a few stages, but to take advantage of this it must also be readily possible to solder large numbers of transistors and capacitors into the circuits simultaneously. This is why the special transistors with pre-tinned contact surfaces are used here (see previous section); they are mounted by means of jigs, many at a time, at the appropriate places, and soldered simultaneously.

Thin films on solid circuits

To conclude we shall mention briefly a development in which the solid circuit discussed in the following article [7] is combined with a thin-film technique. An insulating oxide film is made on a silicon crystal, in which various functions have been incorporated by the diffusion method; the circuit is then completed with resistive and conductor films evaporated on top of the insulating film. This may also be considered a hybrid type of circuit, but compared with the hybrid type of circuit referred to in the introduction the situation has been more or less reversed. While the thin film circuit described above had separate semiconductor elements added to it, the solid circuit here is supplemented with thin evaporated films.

This system makes it possible to use the thin-film technique for circumventing problems arising in the production of solid circuits. A few such problems may be mentioned here. Exacting requirements for the insulation between two components in a circuit are difficult to meet in a single crystal. The difficulty is esolved if one of those components can be applied, in the manner indicated, as a thin film. With this method it is also possible to introduce high resistances (of the order of 1 MΩ), which cannot be done in a purely monolithic circuit. Another problem arises if, owing to the semiconductor nature of all the components in the purely monolithic circuit, the sensitivity to temperature variations is too great. If the resistors are the critical elements, the obvious answer is to deposit the resistors in the form of thin films, which have a low temperature coefficient. Furthermore, since the sheet resistivity for thin films is readily reproducible, specified tolerances may be more readily maintained in production runs of these circuits.

The main problem in this combination of a solid circuit with thin films arises from the tolerances in the strip-widths: the strips have to be much narrower than in the thin-film circuits discussed above, because the crystal on which the pattern is to be deposited is always extremely small. The strip widths for such circuits are between 10 and 25 μm. In order to reproduce these with tolerances within a few per cent, the etching technique has been perfected to give ten times greater precision. *Fig. 9* shows a laboratory example of a circuit produced by this technique.

Summary. In thin-film circuits the circuit elements are deposited in the form of narrow strips and rectangles of vacuum-evaporated materials. In practice the deposited elements are generally confined to a number of passive elements; transistors and diodes, and frequently capacitors and inductors as well, are added separately. This article devotes particular attention to evaporated resistors. For reproducible quantity production, the sheet resistivity and the aspect ratio, which are independently chosen, must both be accurately reproducible. The characteristics of thin films as such, in particular the relatively high bulk resistivity and the low temperature coefficient, are attributed to their island structure. A suitable resistive material is NiCr, which can be evaporated from a wire source. The properties of the film are determined by the pre-heating time of the source, the evaporation time and the pressure in the bell-jar. Dielectric films for capacitors are often made from SiO. Conductors are produced by evaporation of a highly conductive material (e.g. gold) or by tinning a moderately conductive nickel film after evaporation. The photo-etching method is particularly suitable for forming the pattern in a layer; this method can be applied selectively to different films one on top of the other. The technique is illustrated with two applications: a broad-band amplifier for 40-230 Mc/s and a double NAND circuit. The advantages of thin-film and monolithic techniques can be combined by directly evaporating thin films on to a solid circuit coated with an insulating oxide layer.

Solid circuits

A. Schmitz

621.3.049.7:548

This article deals with integrated circuits whose elements, both active and passive, are incorporated in a *single crystal of silicon* by forming zones of *P*-type and *N*-type silicon in the crystal at certain places and in a certain sequence. Such circuits are therefore referred to as solid or monolithic circuits, as distinct from integrated circuits made by evaporating thin films on to a substrate (thin-film circuits, see the preceding article [1]).

The *P*-type and *N*-type zones are formed by a *diffusion process*: the crystal is heated in an atmosphere containing the substance to be incorporated (donors or acceptors), which then diffuses into the crystal. The diffusion is confined to previously determined places by masking the rest of the surface with a layer of silicon dioxide. By subjecting a crystal to successive operations of this kind, various circuit elements can be generated in it. These elements are interconnected to form a more or less complete electronic circuit by means of contacts and conducting strips all in one plane on the SiO_2. Hence the name "planar technique" commonly used for this method. The volume of these circuits can be kept exceptionally small. The circuits discussed in this article are made from a crystal slice or wafer about 200 μm thick, covering an area of less than 1 mm^2.

The various elements that can be made in this way are diodes, transistors, resistors and capacitors. Inductors cannot be made by this technique; if in a solid circuit application an inductive impedance is required, some other means of providing it must be found.

After some general introductory remarks on the processes employed, we shall discuss the method of manufacture and the characteristic features of the elements in a solid circuit. A few examples of solid circuits made in this way will be described.

The masking and diffusion process

The starting point is always a single crystal consisting of a wafer of *N*-type silicon. As mentioned, the thickness is of the order of 200 μm. The wafer is first coated with the masking layer of silicon dioxide by heating it to about 1150 °C in an oxidizing atmosphere. Oxygen might be thought to be the obvious oxidizing agent, but experience has shown that the oxidation process goes much faster when steam is used. An oxide film from 0.4 to 0.6 μm thick, which is sufficient to mask the crystal surface in the diffusion process, can be formed in about half an hour at 1150 °C in steam; if oxygen were used it would take about eight hours. In *fig. 1*, which illustrates the various stages in the process, the result of this initial operation is represented by *a*. Openings are introduced in the appropriate places in the oxide layer by means of a photo-etching process; the wafer is therefore coated with a photosensitive lacquer after oxidation (fig. 1*b*). This lacquer has the property that on exposure to light it ceases to be soluble in a particular solvent. An exposure of the lacquer film is made with a mask which transmits light to all places where diffusion is not wanted, and the unexposed areas of lacquer are dissolved away (fig. 1*c*). Next, the remaining lacquer film is hardened, and it is then used as a mask for etching the required openings into the SiO_2 layer (fig. 1*d*). After this, the remaining layer of lacquer, which is of no further use, is removed by chemical agents. The wafer is then cleaned.

In order to set up zones of *P*-type silicon in the *N*-type crystal, the wafer is heated to about 1170 °C in an atmosphere containing boron. At the places where there is no SiO_2 the boron then diffuses into the *N*-type silicon, forming zones of *P*-type silicon (fig. 1*e*). By again subjecting the wafer to oxidation, the zones are sealed off with a layer of SiO_2 (fig. 1*f*). The thickness of the oxide layers is of the same order of magnitude as the wavelength of light, and beautiful interference effects can be observed. These can be put to good use, for since the first and the later formed layers are not equally thick, the zones of *P*-type silicon remain clearly distinguishable (see fig. 12*b* and fig. 13*b*).

By repeating the photo-etching process an *N*-type layer can in turn be produced in the *P*-type silicon now formed. Contacts can then be applied to the various layers so as to obtain the required circuit elements. These elements will be reviewed in turn below.

The manufacture of various circuit elements

Diodes

A zone of *P*-type silicon obtained as described and the encircling *N*-type material form a diode. To make a connection to the *P*-type material an opening is etched into the SiO_2 layer covering it (see fig. 1*f*). It is necessary to proceed here very carefully; the places where the *P-N* junction reaches the surface of the crystal must not be affected by the etching process. The contact is then applied by evaporating aluminium on to the surface. The aluminium is deposited over the whole surface of the wafer, and is later removed at the

Drs. A. Schmitz is with Philips Research Laboratories, Eindhoven.

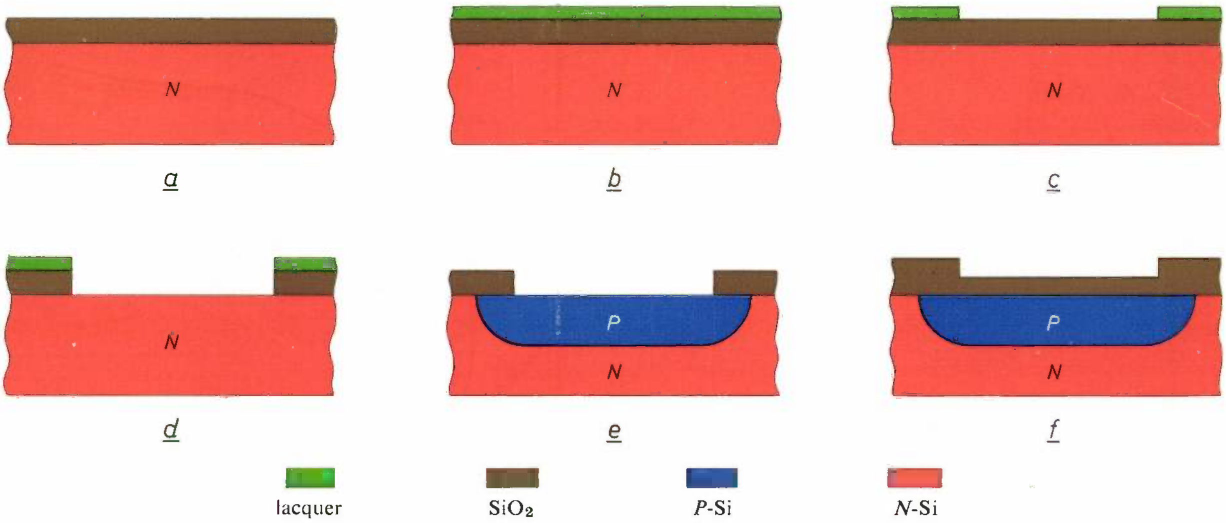


Fig. 1. Schematic representation of various steps in making an integrated circuit by the planar technique, the processes being oxidation, masking and diffusion (the diagrams are not to scale).

a) A silicon dioxide film is formed by oxidation on a crystal wafer of *N*-type silicon.

b) A coating of photosensitive lacquer is applied.

c) An opening is etched into the coating of lacquer by a photolithographic process.

d) The oxide underlayer is etched away.

e) The remaining lacquer film is removed by chemical agents, after which the wafer is heated in an atmosphere containing boron. This diffuses into the crystal to form a zone of *P*-type silicon.

f) The opening in the SiO_2 layer is sealed off by a further oxidation process.

places where it is not required, again by a photo-etching process. If only a diode is required, the connection to the *N*-type silicon is made by alloying the crystal wafer to a gold-plated underlayer. If, however, the diode is part of an integrated circuit, it usually has to be completely insulated from its surroundings, as will be discussed later. In this case a contact has to be applied to the upper side of the *N*-type silicon as well. Since aluminium does not form an ohmic contact [2] with *N*-type silicon that contains the quantity of donors necessary for diode operation, a zone first has to be formed which has a higher donor content (N^+ material). With this the aluminium does form an ohmic contact. The required connection is then made at the same time as that to the *P*-type material, by the evaporation of aluminium. The result is represented in *fig. 2*.

Resistors

A zone of *P*-type material which is reverse-biased with respect to the *N*-type material around it can be considered to be insulated from its surroundings. In this material Ohm's law applies, and the resistivity is such that useful values of resistance can be obtained with practical dimensions. The openings in the oxide layer for making the connections are again introduced by the photo-etching process described above. *Fig. 3* shows such a resistance in cross-section and seen from above. For the *P-N* junction to act as an insulator the whole resistor must thus be given a negative potential with respect to the surrounding *N*-type material.

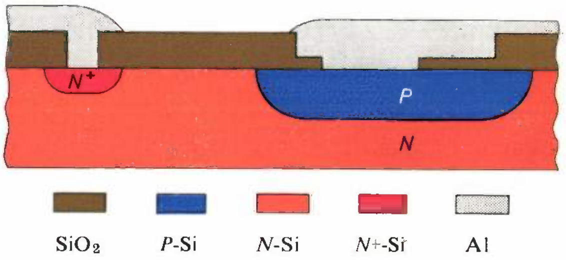


Fig. 2. Manufacture of a diode. A hole is etched in the SiO_2 layer formed on the *P*-type silicon (see fig. 1*f*), and aluminium is then deposited by evaporation to make a connection. In an integrated circuit the other diode connection is formed on the upper side of the original crystal, after first depositing a layer of N^+ silicon.

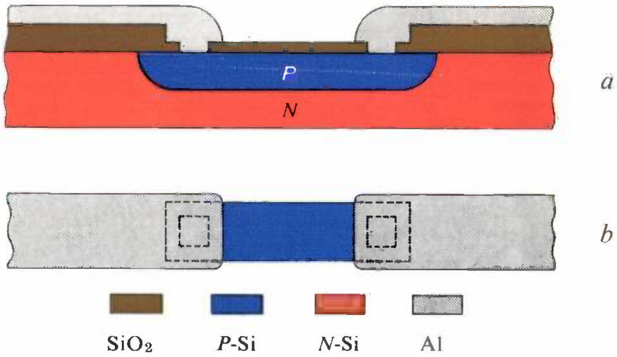


Fig. 3. Cross-section (*a*) and plan view (*b*) of a resistor formed by a layer of *P*-type silicon. The whole resistor must have a negative potential with respect to the surrounding *N*-type material.

[1] E. C. Munk and A. Rademakers, Integrated circuits with evaporated thin films, Philips tech. Rev. **27**, 182-191, 1966.

[2] By an "ohmic" contact we mean a contact that passes current linearly in both directions, i.e. has no rectifying action.

In the design of solid circuits it is necessary to bear in mind that resistors made in this way show a fairly considerable spread in value. Although the resistors in a single circuit can be produced with an accuracy of about 5%, the differences between different circuits may be as high as 20%. The designer of a solid circuit therefore has to allow for the use of resistors with a very large tolerance. Other limitations arise because the resistivity of semiconductors has a high temperature coefficient (about 2×10^{-3}/°C) and because it is difficult to make resistors with values greater than 30 to 50 kΩ by this technique. The area taken up by such resistors is so large that the chance of an irregularity in the crystal at the location of the resistor is considerable (see also page 196).

Finally, it should be noted that a *P-N* junction has a certain capacitance, so that it is not really a pure resistance but a combination of resistance and capacitance (*fig. 4*). This sets an upper limit to the frequency range in which a resistance formed by this technique can be considered as a substantially pure resistance.

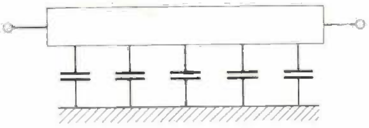


Fig. 4. Equivalent circuit of a resistor produced by diffusion in a crystal.

Capacitors

Because a *P-N* junction has a certain capacitance, it can also be used as a capacitor when biased in the reverse direction. This approach allows capacitors with values up to 200 pF to be made with an accuracy of 3 to 5%. A difficulty with this kind of capacitor is that the thickness of the barrier layer, and hence the capacitance as well, is not constant but depends on the applied voltage. Moreover, the equivalent impedance of a *P-N* junction is not a pure capacitance but a combination of capacitance with a series and a parallel resistance (*fig. 5*). The applications of these capacitors are therefore limited.

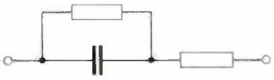


Fig. 5. Equivalent circuit of a capacitor formed by a *P-N* junction.

Transistors

A transistor is made by introducing a layer of *N*-type silicon into a layer of *P*-type silicon formed in the manner illustrated in fig. 1. This is done by first etching away, as described, some of the oxide layer covering the *P*-type silicon (*fig. 6a*), after which the crystal wafer is heated in an atmosphere containing phosphorus. The resulting *N*-type silicon, which, to function properly as an emitter, must contain more donors than there are acceptors in the base (N^+ material), is then covered with another film of SiO_2 (fig. 6*b*). We now have an *N-P-N* transistor in which the original crystal wafer serves as the collector. The connections to emitter and base are again made by etching openings into the oxide layer and depositing aluminium. The collector connection can be made in two ways. A separate transistor, like a separate diode, is made by alloying the crystal wafer to a gold-plated underlayer. In an integrated circuit, however, a connection to the collector almost invariably has to be made on the upper face. This again makes it necessary to form a zone of N^+-type material (fig. 6*c*). This can be formed at the same time as the emitter.

It is very difficult to use this method to make both *N-P-N* and *P-N-P* transistors in a single crystal wafer. It is, however, possible to obtain a *P-N-P* transistor by forming two layers of *P*-type material *next* to one another, although such a transistor has a very low current amplification factor. Such a transistor can nevertheless be combined with an *N-P-N* transistor, to produce a circuit which in many ways corresponds to a *P-N-P* transistor with a high current amplification factor [3].

Other elements

Various other circuit elements can be made by the planar technique we have described. An example is given in *fig. 7*, which shows a *field-effect transistor* [4]

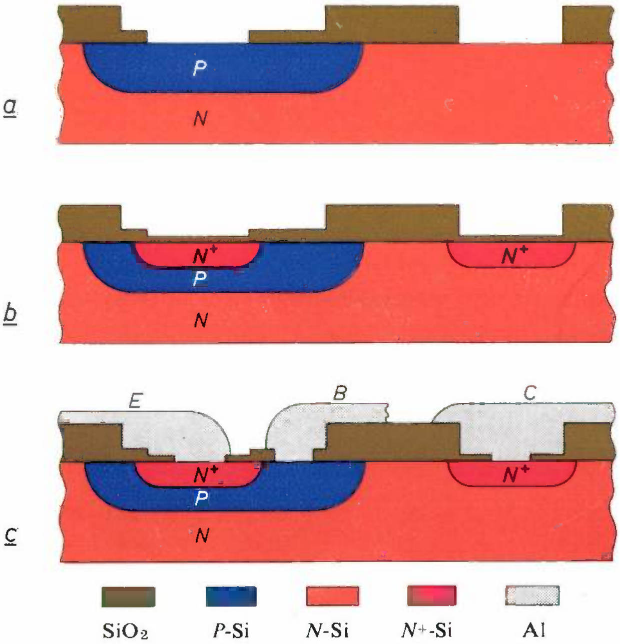


Fig. 6. Manufacture of an *N-P-N* transistor.
a) Openings are etched into the SiO_2 layer sealing off the *P*-type silicon (see fig. 1*f*) and above the *N*-type silicon.
b) Layers of N^+ silicon are formed under these openings by diffusion. The whole surface is then sealed off again by oxidation.
c) After once more etching holes into the SiO_2 layer, connections are made to the emitter (*E*), base (*B*) and collector (*C*).

in cross-section and seen from above. Here again, successive layers of P and N^+ semiconductor are formed in a wafer of N-type silicon. The last layer formed is now, however, made long enough in one direction for it to make contact with the original crystal, while *two* connections are applied to the P-type material, one on each side of the N^+ layer. The current through the P-type layer can now be controlled by the voltage between the N-type and P-type material.

Other elements that can be made by the planar technique are *N-P-N-P* rectifiers (silicon controlled rectifiers, thyristors) and elements in which a layer of silicon dioxide acts as dielectric. These include MOS capacitors (MOS = metal-oxide semiconductor) and MOS transistors (MOST devices).

Insulation of circuit elements from each other

When several circuit elements have been introduced in a crystal wafer by the method described, they all have an N-type layer in common; i.e. they are connected to one pole. In nearly all circuits, however, the elements have to be completely insulated from each other. This can be done by means of an extra P-N junction, using a procedure illustrated in *fig. 8*. Instead of starting from a wafer of N-type silicon, we now start from a single-crystal layer of N-type material which is grown epitaxially on a crystal of P-type silicon. Again, a silicon dioxide film is formed on this layer and a pattern of channels is etched following lines where an insulating separation is required (fig. 8*a*). P-type silicon is now formed along these lines by diffusion. The diffusion process is continued until the newly-formed P-type material extends to the original P-type crystal. After the diffusion the channel pattern is once again sealed off by a layer of SiO_2 (fig. 8*b*). The N-type silicon now forms a number of "islands" in the P-type material, and the islands can be insulated from each other by applying a reverse voltage between the P- and N-type silicon. (This insulation has a fairly high capacitance.) The processes described above can be used to make a circuit element in each of these islands.

Fig. 8*c* shows a cross-section of an N-P-N transistor formed in this way. The collector connection was made by the method described above.

A transistor made by this technique has several special features which have to be taken into account in application. In the first place, the collector-series resistance is fairly high. This is because the epitaxial N-type layer from which the collector is formed is

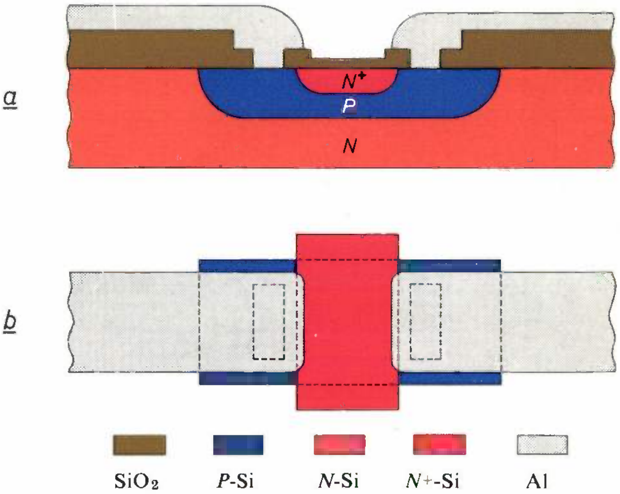


Fig. 7. Cross-section (*a*) and plan view (*b*) of a field-effect transistor. A layer of N^+ silicon is formed on the strip of P-type silicon, which has two connections. The N^+ silicon strip is long enough to connect with the N-type silicon of the underlayer.

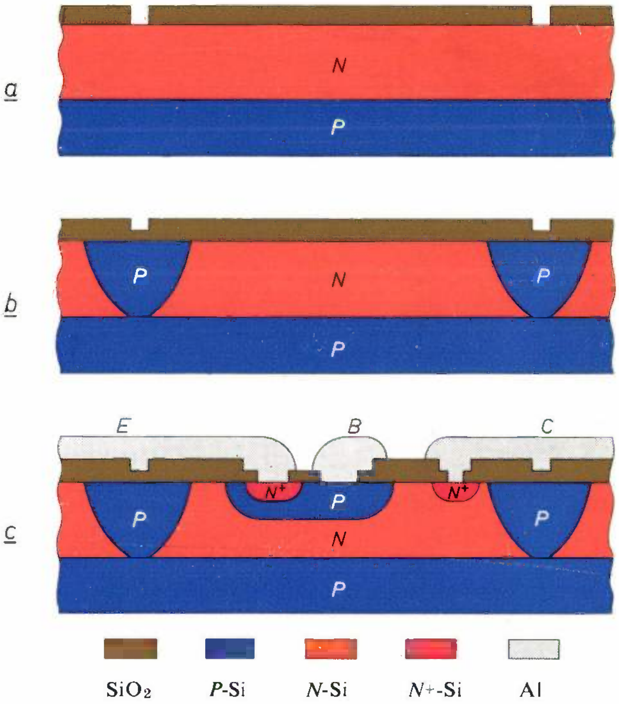


Fig. 8. Insulation of elements in a solid circuit.

a) An epitaxial N-type layer is formed on a substrate of P-type silicon. The N-type layer is covered with a layer of SiO_2 by oxidation. A pattern of lines is etched into the SiO_2 layer by the method previously described.

b) P-type silicon, penetrating to the substrate, is produced along the lines of the pattern by diffusion. This results in "islands" of N-type silicon, which, when a voltage is applied in the reverse direction, can be insulated from the surrounding P-type silicon.

c) Formation of a transistor which is insulated from the other elements in the circuit. In the island of N-type silicon, which acts as collector, a layer of P-type silicon is formed, the base. In this a layer of N^+ silicon is formed, the emitter. To make a connection to the N-type material of the collector, a layer of N^+ silicon is formed in it at the same time as the emitter. The connections to emitter, base and collector, which are produced by the evaporation of aluminium, are indicated by *E*, *B* and *C*.

[3] See H. C. Lin, T. B. Tan, G. Y. Chang, B. van der Leest and N. Formigoni, Lateral complementary transistor structure for the simultaneous fabrication of functional blocks, Proc. IEEE **52**, 1491-1495, 1964.

[4] The field-effect transistor and its operation are described in the next article in this number: H. C. de Graaff and H. Koelmans, The thin-film transistor, Philips tech. Rev. **27**, 200-206, 1966.

extremely thin: it is about 10 μm thick. The actual collector region, situated under the base (see fig. 8*c*), is even thinner, about 7 μm. Because of the high collector series resistance the knee voltage is always greater than that of transistors made by "normal" methods. It is therefore difficult to use the planar technique in this integrated form to produce transistors of high efficiency and relatively high power output.

The collector resistance can be reduced and the knee voltage lowered if a layer of N^+ material (the "buried layer", see *fig. 9*) is formed in the *P*-type silicon before the first operation, i.e. the growth of the epitaxial *N*-type layer on the *P*-type silicon. The buried layer, which has a higher conductivity than the *N*-type silicon, is thus situated under the *N*-type material of the collector and can appreciably reduce the collector resistance.

Another feature of transistors of this type is that, because of the presence of several regions of *P*- and *N*-type material, transistor characteristics other than those o the desired transistor may be found. For example, the base, the collector and the underlayer of *P*-type material constitute a *P-N-P* transistor, the base being formed by the collector of the desired transistor. Since this "base" is thicker than that of a normal transistor, the amplification factor of such a *P-N-P* transistor is fairly low (e.g. 4 to 5). Nevertheless, this may give rise to unwanted effects in some circuits, and preventive measures have to be taken.

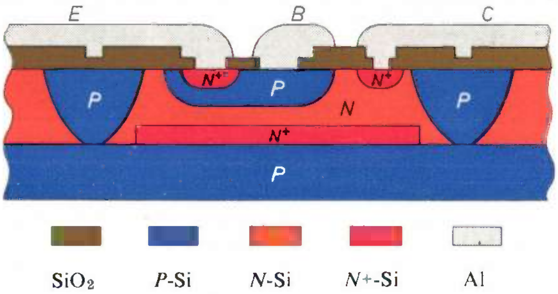


Fig. 9. The collector resistance can be reduced by means of a layer of N^+ silicon under the *N*-type silicon of the collector.

Finally, we would point out that in the process described a combination of four layers of silicon is formed (emitter, base, collector and underlayer), and that these are alternately *N*-type and *P*-type. Such a combination may show characteristics comparable to those of a thyristor. Any unwanted effects arising through this can nearly always be avoided by giving the voltages between the layers an appropriate polarity.

Another method of insulating the elements from each other consists in etching grooves into a crystal wafer and providing an insulating SiO_2 layer by oxidation. The manufacture of such circuits, known as "Epitaxial Passivated and Isolated Integrated Circuits" (EPIC), however involves considerable difficulties.

Production of large numbers of circuits

In the foregoing we have considered the method of manufacture for a single solid circuit. An economically acceptable production yield is possible with this rather laborious technique as a large number of circuits can be made simultaneously on a single wafer of silicon. Hundreds of similar circuits can be made on a wafer of 2.5 cm diameter. For a single circuit, greatly magnified drawings (200 times) are first made of the masks to be used in all successive operations. These drawings are photographically reduced in two steps and the final mask is obtained by repeating the last photographically reduced drawing many times on a single negative. To make sure that the masks used in the successive processes exactly cover the appropriate parts of the crystal, the printing has to be done with extremely high precision. The maximum permissible deviation is of the order of 1 μm.

Limitations of the planar technique

Slight imperfections in the crystal may give rise to impermissible deviations in the leakage current or reverse breakdown voltage of one or more of the diffused *P-N* junctions. The probability that such a crystal defect will be encountered in a circuit increases with the area occupied by the circuit. In a number of simultaneously manufactured solid circuits the percentage of rejects therefore increases with the area of the individual circuit. It is thus also desirable for reasons other than that of microminiaturization to keep the size of the circuit as small as possible. The need for small dimensions also sets a limit to the number of elements that can be incorporated in a circuit and also to the maximum values of the resistances and capacitances.

We have already mentioned a maximum resistance of 50 kΩ and a maximum capacitance of 200 pF.

Another cause of faufts which is also closely bound up with the small dimensions may be the presence of dust during masking operations on the crystal. These should therefore be carried out in a dust-free room.

Since, in a number of simultaneously produced solid circuits, the percentage of rejects due to faulty operation increases with the area of the circuit, the costs involved in incorporating an element are determined more by the area occupied than by the type and the required characteristics of the element. For example, a 10 kΩ resistor, which takes up nearly as much space as 5 or 6 transistors, is 5 to 6 times more expensive than a transistor. A rule that nearly always applies in the economical design of other kinds of circuit, namely that resistors are much cheaper than transistors, thus ceases to apply in solid circuits. It may be advantageous to design these circuits in such a way that relatively few large resistances are needed; there is much less objection to the use of more transistors than in a conventional circuit.

For completeness we shall mention some of the factors that limit the realization of some of the requirements that may be asked of solid circuits: the high temperature coefficient of the resistors, the high capacitance between elements insulated from each other, the variation of capacitance with voltage, the large collector series resistance of the transistors and the fact that inductors cannot be formed in a solid circuit.

Some of the resultant problems can be overcome by combining monolithic and thin-film techniques. If the resistors are not formed by diffusion in the crystal but by metal films evaporated on top of the oxide layer, then the temperature coefficient of the resistors is much smaller. This brings us to a combination of diffusion technique and thin-film technique, and the "hybrid circuits" thus produced have been mentioned in reference [1] (page 191).

Assembly

The form in which a solid circuit is supplied depends on how it is to be mounted in the equipment in which it is to be used. A method still widely used is to mount the circuit in a holder of the same dimensions as those used for transistors (type TO 5). This means, of course, that full advantage is no longer taken of the extremely small volume of a solid circuit (e.g. 0.25 mm^3). After assembly the space thus occupied may be as much as 0.3 cm^3, which is more than 1000× as large, without taking into account the space needed for the connectors.

In another method of assembly the solid circuit is fixed in a flat holder fitted with connectors at both ends (a "flat package"). The total volume of this arrangement is smaller, about 0.05 cm^3. Both forms of assembly are shown in *fig. 10*.

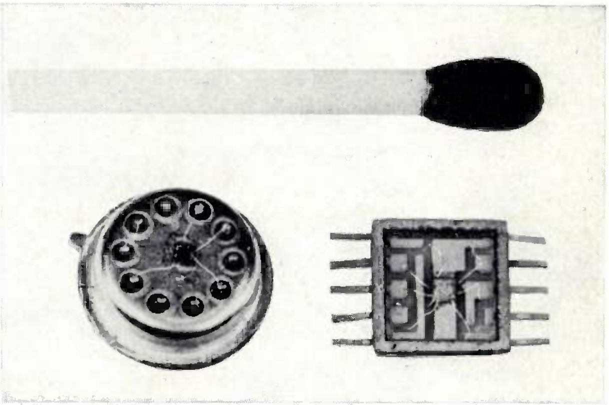


Fig. 10. Two mounting methods for solid circuits. Left, transistor mount TO 5, right "flat package".

In view of the small dimensions, high precision is also required in making the connections between a solid circuit and the connecting wires. The material generally used for these connections is gold: the procedure is illustrated in *fig. 11*.

After all connectors have been fitted, the holder is hermetically sealed in a dry nitrogen atmosphere.

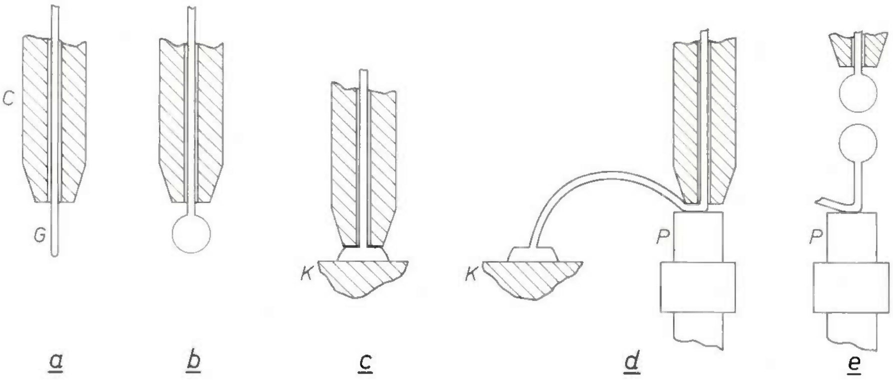


Fig. 11. Making the connections between a solid circuit and the connections to the mount.
a) There is a 20 μm gold wire *G* inside the capillary tube *C*.
b) By means of a hydrogen flame, a ball is formed at the end of the gold wire (diameter roughly 75 μm).
c) A connection is made between the gold wire and the contact by pressing it on to the contact *K* of the solid circuit, which is heated to 320 °C.
d) The capillary tube is raised, releasing the gold wire. The tube is then pressed down again and connection is made with the contacts *P* in the assembly mount, which is heated as well.
e) After the capillary tube is raised, the gold wire is melted in two, producing a ball again at both ends. The ball formed on the contact pin *P* can be removed if necessary.

Examples of solid circuits

Figures 12 and 13 show two solid circuits as examples of the technique we have been discussing.

Fig. 12 gives the circuit diagram and an enlarged photograph of an amplifier (with two transistors, one diode and four resistors) which has strong negative feedback to produce an accurately defined gain (operational amplifier). The location of the various compo-

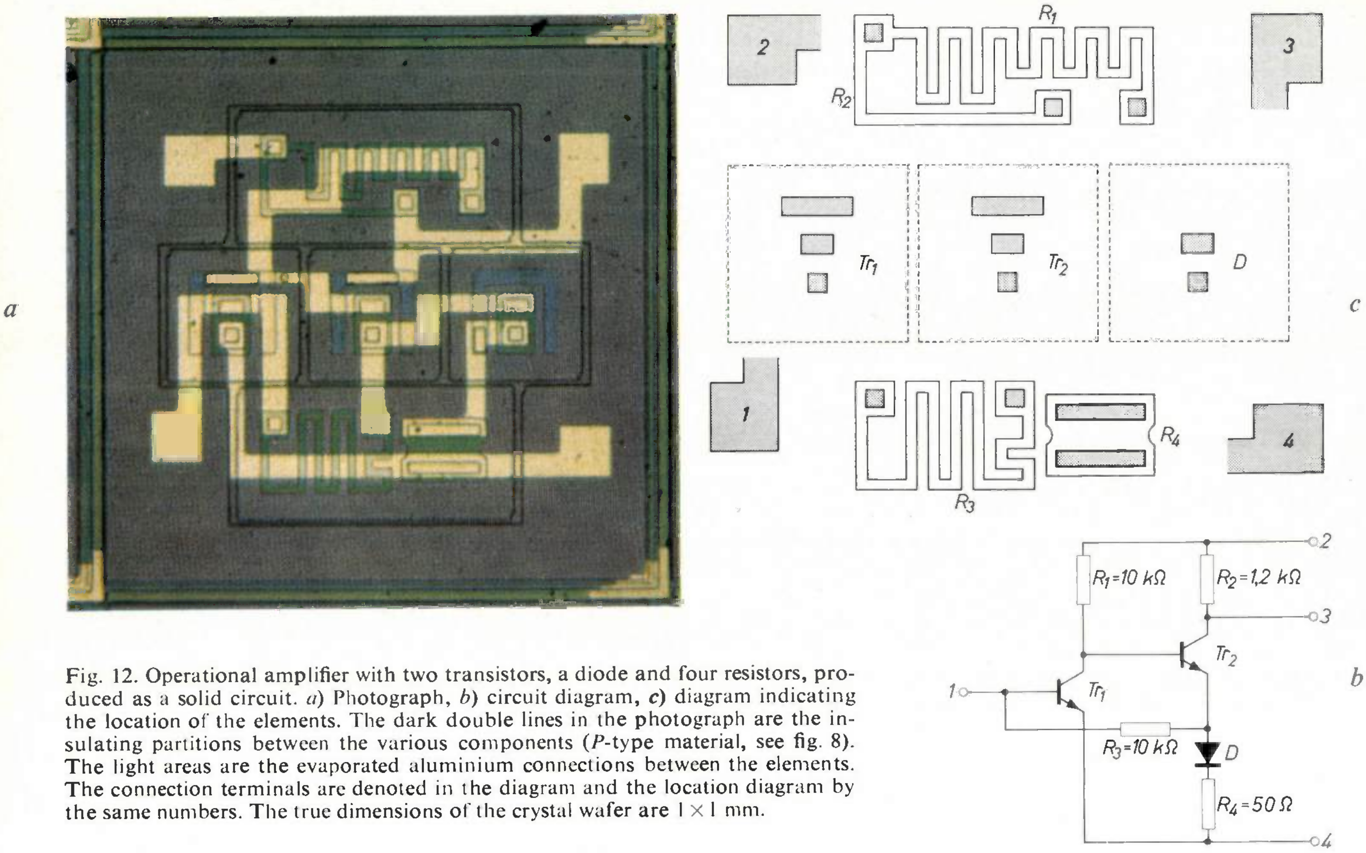


Fig. 12. Operational amplifier with two transistors, a diode and four resistors, produced as a solid circuit. *a*) Photograph, *b*) circuit diagram, *c*) diagram indicating the location of the elements. The dark double lines in the photograph are the insulating partitions between the various components (*P*-type material, see fig. 8). The light areas are the evaporated aluminium connections between the elements. The connection terminals are denoted in the diagram and the location diagram by the same numbers. The true dimensions of the crystal wafer are 1 × 1 mm.

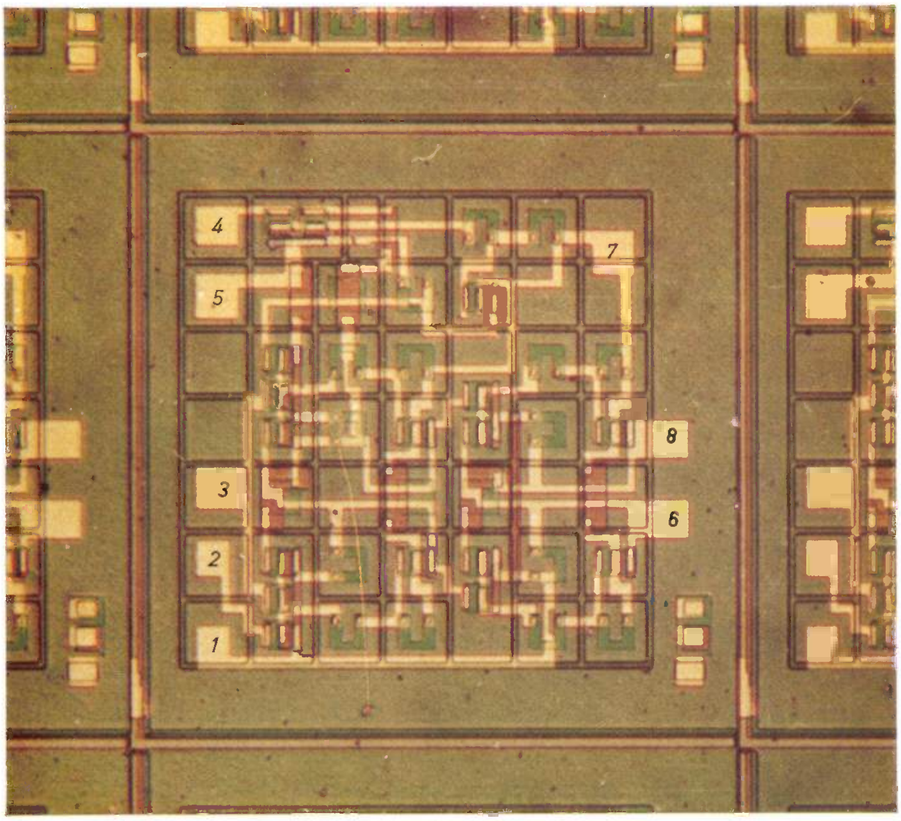


Fig. 13*a*

Fig. 13. Photograph (*a*) and circuit diagram (*b*) of a shift register for a computer. The circuit contains 17 transistors and 15 resistors. Here again the dark lines are the insulating partitions between the elements and the light areas are the connections between them. Connections are indicated by the same numbers on the diagram and the photograph. In the lower right-hand corner of the crystal wafer (1 × 1 mm) is a monitoring transistor, which is used for checking the diffusion process.

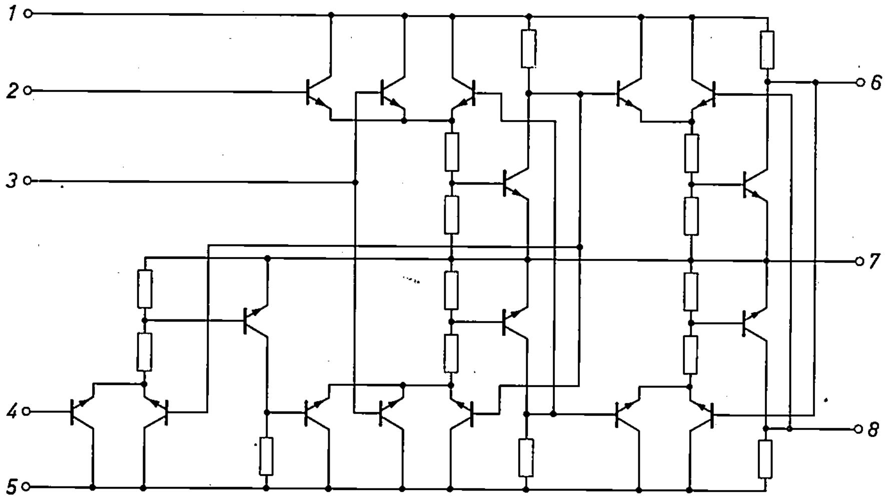


Fig. 13*b*

nents in the diagram is indicated in the diagram of fig. 12*c*.

To illustrate the fact that more elaborate circuits can be produced in solid form, *fig. 13* shows the circuit diagram and a photograph of a shift register for a computer. This circuit contains 17 transistors and 15 resistors, and, like the previous one, is contained on a crystal wafer with a surface area of 1 mm^2. The area occupied by the circuit itself is of course even smaller.

Summary. Solid circuits are integrated circuits in which both the active and the passive elements are formed in a silicon crystal. This is done by subjecting a silicon wafer to a succession of masking and diffusion processes to form alternate zones of *P*-type or *N*-type (or N^+) silicon. The article describes how diodes, resistors, capacitors and transistors are made by these processes. The circuits thus formed have extremely small dimensions, i.e. less than 1 × 1 mm. Reverse-biased *P-N* junctions can be used to provide insulation between the various elements. An economically acceptable production yield is made possible by the fact that large numbers of circuits can be made simultaneously on a single crystal wafer. Various limitations are mentioned which affect the design of solid circuits. Finally, two methods of assembly are described, and examples of typical solid circuits are given

The thin-film transistor

H. C. de Graaff and H. Koelmans

621.382.3:539.234

Nature and operation of field-effect transistors

In 1907 the discovery was made that a current of electrons flowing in a vacuum can be controlled by a relatively weak electrical signal. The application of this discovery has made radio telephony, radio broadcasting and many other things possible. In 1930 a device was invented — by J. E. Lilienfeld of New York — by means of which, in an analogous way, it was possible to control a current flowing through a *solid*. This device, however, was never put to practical use, nor was a similar circuit element which was discovered a few years later by O. Heil of Berlin.

The device invented by Lilienfeld functioned in the same way as the one which is the subject of this article: it may be regarded as the first field-effect transistor, and in particular, as the first thin-film transistor [1]. The current in a field-effect transistor is controlled by utilizing the well-known electrostatic induction effect, nowadays briefly referred to as the *field effect*. The principle of this method of control will be discussed with reference to *fig. 1*.

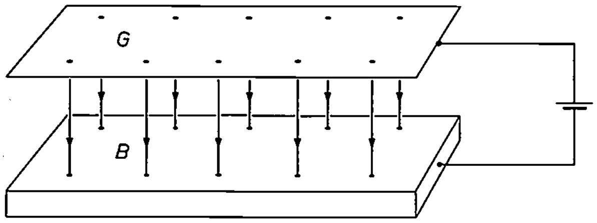


Fig. 1. Illustrating the way in which the current is modulated in a field-effect transistor. When a voltage is applied to electrode G, a charge induced in plate B affects the conductivity of B.

When a potential difference is applied between a flat electrode G — the control electrode, or gate — and a conducting plate B, which is electrically insulated from G and is situated approximately parallel to it at a certain distance away, lines of force run from G to B (or vice versa), which must terminate at induced charges in B. These induced charges can have an appreciable influence on the conductivity of B. This is the case when:

1) the induced charges are mobile, and
2) the magnitude of the induced charge is not negligibly small compared with the mobile charge naturally present in B.

Requirement (1) can be met both by metals and semiconductors. In practice, however, only semiconductors are considered: in metals the free charge naturally present is usually too high to make it possible to meet requirement (2). The technology of manufacturing semiconducting materials has now advanced to a stage where the concentration of the natural free charge can be made very small. The reduction in this charge can be brought about not only by using an appropriate material but also by reducing the thickness of B; this does not affect the induced charge, which is present in a thin layer under the surface of B.

Unlike the situation in ordinary transistors, the minority charge carriers — e.g. the holes in an N-type semiconductor — play no part whatsoever in field-effect transistors; the field effect only affects the concentration of the majority charge carriers.

Modern field-effect transistors; the thin-film transistor

There are now several kinds of field-effect transistor; they differ mainly in the way in which the gate is insulated from the semiconducting layer, called the channel, through which flows the current to be controlled. In the oldest type of modern field-effect transistors, described by Shockley (*fig. 2a*), there is really no insulation in the ordinary sense of the word; the gate, like the channel, is of semiconducting material, but with the opposite type of conduction, and its potential is such that the P-N junction between gate and channel is biased in the reverse direction. A much more recent device is the MOS transistor (fig. 2*b*). This has a metal gate (M = metal) which is insulated by an oxide layer (O = oxide), from the semiconducting channel (S = semiconductor). One of the most recent versions of the field-effect transistors now being studied is the thin-film transistor. As we have said, this may be looked upon as the modern version of the Lilienfeld device. The channel here consists of a polycrystalline semiconductive layer of say, CdS or CdSe, deposited by vacuum-evaporation on to an insulating substrate. One form, which has been described by P.K. Weimer [2], is shown schematically in fig. 2*c*; the latest version, developed at the Philips Research Laboratories in Eindhoven, is represented in fig. 2*d*. The MOS transistor and the thin-film transistor are often known by the collective name "insulated-gate field-effect transistor", and the type described by Shockley is referred to as the "junction-gate field-effect transistor".

Since thin-film transistors are made by the evaporation process, it is logical to combine them with circuits containing evaporated passive elements. Experi-

Ir. H. C. de Graaff and Dr. H. Koelmans are with Philips Research Laboratories, Eindhoven.

ments in our laboratories have shown that the production processes for both kinds of circuit element are compatible. If the limitations which still appear in thin-film transistors can be sufficiently overcome — and there are good prospects of this — it will then be possible in principle to make completely integrated circuits by evaporation on to an insulating substrate. A very attractive solution will then have been found for cases where the tolerances laid down for the passive elements (particularly resistors and capacitors) are closer than can be realized with solid circuits made by the planar technique [3] or where the insulation method used in solid circuits [4] is not adequate. In very large circuits as well, where the production yield from the planar technique is lower owing to statistically distributed surface errors, it may be a much better proposition to use circuits made entirely by the evaporation process.

In the following section we shall deal with the principal characteristics of the thin-film transistor. In the concluding section we shall consider more deeply the various types of design and methods of production, with emphasis on our own production method.

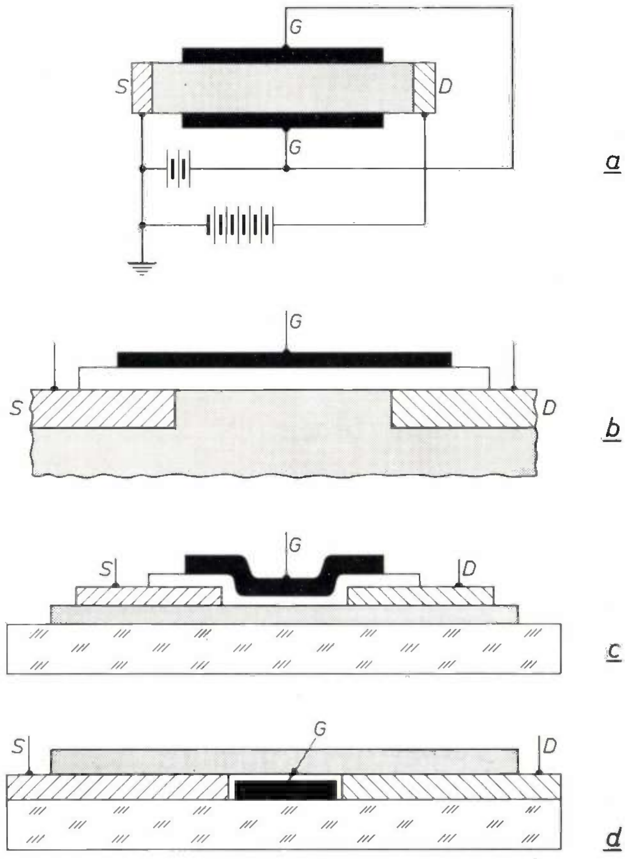


Fig. 2. Schematic cross-section of *a*) a field-effect transistor with two gate electrodes *G* (symmetrical transistor) which are not insulated from the channel but form with it a *P-N* junction, which has to be reverse-biased; *b*) a MOS transistor; *c*) one of the early forms of an insulated-gate thin-film transistor; *d*) the most recent type of thin-film transistor made in our laboratories. In all figures *S* is the source, *D* the drain and *G* the gate. The source and drain electrodes are shown cross-hatched, the channel is shaded, the gate is black and the insulating layer white. The thin-film transistor has a glass substrate.

Properties and characteristics of thin-film transistors

Fig. 3 shows a lateral cross-section of a thin-film transistor. *S* is the source and *D* the drain and between them (shaded) is the channel. Its length, i.e. the distance from *S* to *D*, is denoted by *l*. Above the channel there is an insulating layer and above that is the gate *G*. The potentials of *S*, *D* and *G* are V_s, V_d and V_g respectively. It is customary to put $V_s = 0$, i.e. to relate all potentials to that of the source. By V_g we therefore mean $V_g - V_s$, and by V_d the potential difference between the ends of the channel, in other words the channel voltage. Since the conductivity of the channel is largely determined by the charge present in it when a gate voltage V_g is applied, the capacitance *C* of the gate is also an important quantity.

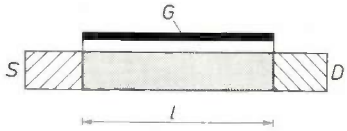


Fig. 3. Model of a thin-film transistor, used for the explanation of the characteristics. The letters and hatching have the same significance as in fig. 2. The length of the channel is *l*.

In an approximate treatment due to Shockley — his "gradual approximation" — it is assumed that the induced charge produced by V_g at a given point in the channel is not dependent on the potential difference between source and drain. The relation between the current *I* and the voltages V_g and V_d in a certain interval of V_g values is then given by the equation:

$$I = \frac{\mu C}{l^2}\left\{(V_g - V_0)V_d - \tfrac{1}{2}V_d^2\right\}, \qquad (1)$$

where V_0 is a constant and μ the mobility of the charge carriers. As can be seen, the current *I* varies linearly with the gate potential, and as the square of the channel voltage. The relation between *I* and V_d is found from

[1] The Lilienfeld device was described in three American patents, granted in 1930, 1932 and 1933 respectively. The Heil device was described in a British patent, granted in 1935. For particulars relating to the invention of the various kinds of transistors see e.g. C.T. Sah, IEEE Trans. on electron devices **ED-11**, 324, 1964, and also the literature quoted in this article. A detailed bibliography (up to September 1963) on the field-effect transistor is to be found in J. T. Wallmark, IEEE Spectrum **1**, No. 3, 182, 1964.

[2] P. K. Weimer, Proc. IRE **50**, 1462, 1962. This author gave an extensive treatment of the insulated-gate thin-film transistor in Physics of thin films **2**, 147, 1964.

[3] See the article by E. C. Munk and A. Rademakers in this number, page 182.

[4] See the article by A. Schmitz in this number, page 192.

eq. (1) to take the form of a parabola as shown in *fig. 4*. At V_d values greater than the value $V_g - V_0$ corresponding to the maxima, eq. (1) is no longer valid. It is assumed that the remainder of each characteristic is horizontal to a first approximation (*fig. 5*). For this region, the saturation region, it can be derived from (1) that:

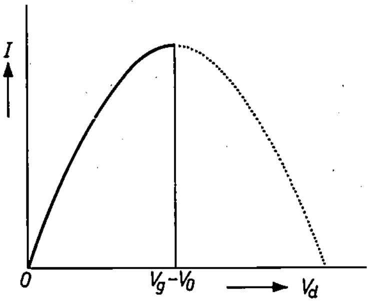


Fig. 4. In the region $0 < V_d < V_g - V_0$ the theoretical I-V_d characteristic is parabolic. The maximum of the parabola occurs at $V_d = V_g - V_0$.

$$I_{\max} = \frac{\mu C}{2l^2}(V_g - V_0)^2. \quad . \quad . \quad . \quad . \quad (2)$$

The physical significance of the constant V_0 is as follows. If V_s and V_d are both zero, V_0 is the value of V_g at which the induced charge is equal and opposite to the charge initially present. At $V_g = V_0$ the total charge density is therefore exactly zero. It should be added that we assume in this section that the channel is so thin that the induced charge is more or less equally distributed over its thickness. This may be regarded as a schematic representation of the extremely thin layer opposite the gate, at the surface of the semiconducting layer, which carries most of the induced charge and supports the greater part of the conduction.

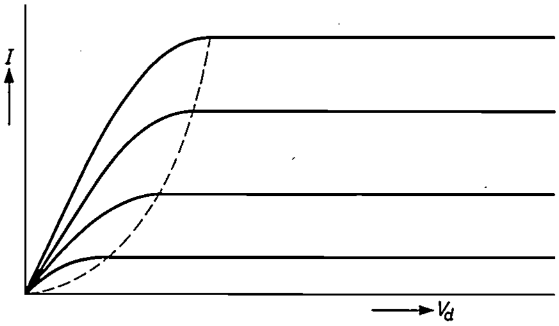


Fig. 5. Family of theoretical I-V_d characteristics. Each curve relates to a specific value of V_g. The dashed line (likewise a parabola) passes through the points $V_d = V_g - V_0$ (the maxima of the parabolae); to the right of the dashed line the characteristics are horizontal.

Now of course in any element of the channel a current dependent on V_g and V_d can flow only if the density of the total charge in that element is negative (electron concentration positive). Its potential V must therefore fulfil the condition $V_g - V > V_0$, or $V < V_g - V_0$. Equation (1) is valid only if this condition is satisfied for all the elements of the channel. Since the highest potential occurring in the channel is V_d, it therefore follows that $V_d < V_g - V_0$; the right-hand side is the upper limit of the range in which equation (1) is valid, as mentioned previously. The value $(V_g - V_0)$ of V_d is called the pinch-off voltage. (The "pinch-off" is not comparable with the cut-off caused in a thermionic valve by applying a strong negative bias to the control grid: a current still flows through the channel at $V_d \geqq V_g - V_0$; see fig. 5.)

In some transistors V_0 is positive and in some it is negative. This depends on the nature of the channel material. If there are already mobile electrons in the channel at zero gate bias ($V_g = 0$), then of course the voltage needed to displace them must be negative. If the channel contains no electrons but does have empty traps, then V_0 is positive. At relatively low positive values of V_g all that happens in this latter case is that some of the traps are filled; this does not increase the conductivity. Only after all traps have been filled, i.e. when $V_g > V_0$, does any increase in the conductivity occur. This point is dealt with in more detail in small print at the end of this section.

We shall now turn to the I - V_g characteristics. The transconductance S of the transistor — the partial derivative of I with respect to V_g at constant V_d—can immediately be calculated from the equations given above for I. In the region $0 < V_d < V_g - V_0$:

$$S = \frac{\mu C}{l^2} V_d, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3a)$$

and in the saturation region:

$$S_{\max} = \frac{\mu C}{l^2}(V_g - V_0). \quad . \quad . \quad . \quad (3b)$$

It can be seen from these equations that a high transconductance is obtained if the channel length l is small and the capacitance high — i.e. if the insulating layer is thin — and if a material is chosen in which the electron mobility is high.

Figs. 6 and *7* show characteristics of a transistor made in our laboratories. Fig. 6 gives the I - V_d characteristics for five different values of V_g. The parabolic region of the characteristics can be seen on the left (cf. fig. 4 and 5) and the "horizontal" region on the right. It will be seen that the latter region is not in reality completely horizontal. The hysteresis effect in the upper two characteristics will be discussed below.

Fig. 7 shows two I - V_g characteristics derived from the characteristics in fig. 6. At $V_d = 10$ V the maximum transconductance of this transistor is seen to be about

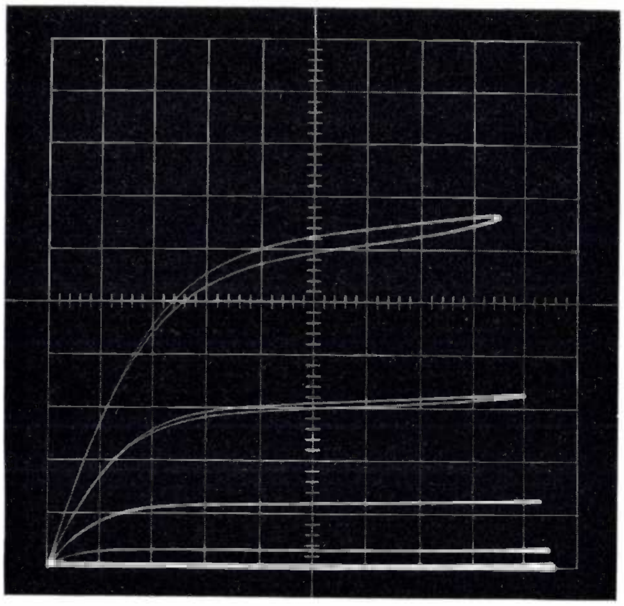


Fig. 6. The I-V_d characteristics of one of our thin-film transistors of the type illustrated in fig. 2*d*. One division on the I axis corresponds to 2 mA, and one division on the V_d axis corresponds to 1 V. The curves (from top to bottom) relate to V_g = 6 V, 5 V, 4 V, 3 V and 2 V respectively.

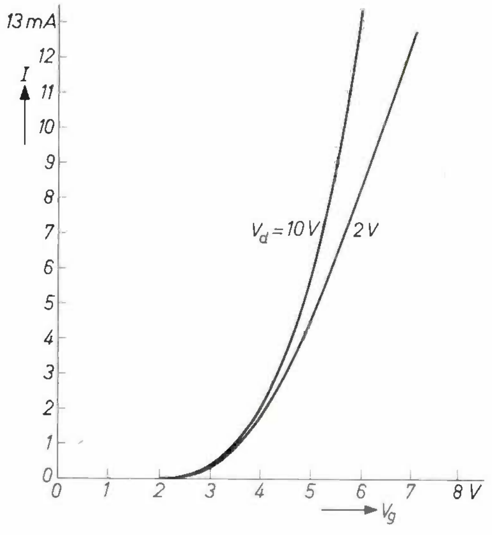


Fig. 7. Two I-V_g characteristics derived from fig. 6. The transconductance is about 5 mA/V.

5 mA/V. The capacitance C is about 20 pF. In the common source configuration the input resistance at low frequencies is about 10^4 MΩ. The surface area is approximately 0.02 mm^2. The length l of the channel is about 8 μm.

Thin-film transistors, like MOS transistors, have a fairly high noise level. The noise intensity is inversely proportional to the frequency f (this is referred to as $1/f$ noise). The noise of our thin-film transistors is sufficiently low, however, for them to be used in gramophone amplifiers. At frequencies higher than 10^6 c/s the $1/f$ noise is negligible, thermal noise then being virtually the only noise present.

In the foregoing, for the sake of simplicity, several points have not been taken into account. In the first place we have tacitly assumed that the channel-to-gate capacitance is constant. This is not entirely true: apart from the insulating layer between channel and gate, the space charge region must be taken into account. In other words, the region in which the energy bands are bent must also be considered as part of the dielectric. The thickness of this layer, however, depends on the applied voltage. In a transistor with an N-type channel and positive V_0, C can only be considered as a constant if V_g is very high; in this case the thickness of the space charge layer is small compared with that of the insulation, and the capacitance of the space charge layer is therefore very high, so that the total capacitance is almost identical with the contribution from the insulating layer. If, in our example, we let V_g decrease from a high positive value — or V_d increase from zero — then C will gradually decrease. The same effect is found in MOS transistors.

A second simplification was our suggestion that when a positive voltage V_g is applied to transistors with a positive V_0, the extra induced charge will be completely contained in traps as long as $V_g < V_0$. In reality in a certain region of V_g values the induced charge is partly trapped and partly free. Only at high values of V_g are nearly all the traps filled, and any further increase of V_g induces, almost exclusively, charge carriers that are free [5].

This can be taken into account quite simply by a suitable choice of μ in the equations above. An effective mobility of $\mu = a\mu_H$ must be chosen, where a is the free fraction of the induced charge and μ_H is the mobility found in Hall effect measurements. Since a depends on V_g, so also does the effective mobility μ; at high V_g the value of μ is virtually constant. If V_g is very high, μ may theoretically be smaller again owing to scattering of moving charge carriers at the surface of the layer.

The third and last comment concerns the fact, already mentioned, that the I - V_d characteristics for $V_d > V_g - V_0$ are not strictly horizontal but rise slightly. Various explanations for this have been put forward. The effect has, for example, been attributed to the presence of donor atoms which are assumed to give rise to a number of impurity levels of different energy [6]. The electrons in the "shallow" impurity levels would already be free when V_d is low whereas those at deep levels are not free until V_d is high; using this hypothesis complete saturation is never entirely reached, because this would call for such a high V_d that the transistor would break down long beforehand. Measurements have shown, however, that this explanation does not apply to the transistors made by us [5]. An entirely different hypothesis is that the mobility of the charge carriers depends on the strength of the electric field [7], but this explanation is also not completely satisfactory.

It is however certain that, even in the absence of physical effects such as those just mentioned, some increase of I with V_d is to be expected as the insulating layer is not infinitely thin. The effect of this has been exactly calculated [8] for a hypothetical transistor with a gate electrode on each side of the channel (symmetric field effect transistor), all four electrodes being infinitely long and the channel, source and drain infinitely thin. It is found that the behaviour of the current is entirely governed by the parameter h/l, where h is the thickness of the insulation, and l the length of the channel. The curves actually measured, are usually

[5] H. C. de Graaff, Solid-State Electronics **8**, 835, 1965 (No. 10).

[6] J. E. Johnson, Solid-State Electronics **7**, 861, 1964.

[7] S. R. Hofstein and G. Warfield, IEEE Trans. on electron devices **ED-12**, 129, 1965 (No. 3).

[8] J. A. Geurst, Theory of insulated-gate field-effect transistors near and beyond pinch-off, Solid-State Electronics **9**, 129-142, 1966 (No. 2).

considerably steeper; in most transistors the thickness of the insulating layer is evidently not the only cause of the rising characteristics. In any case, the slope can never be less than that given by the calculations of reference [8] for the appropriate h/l value.

Alternating-current behaviour

A rough idea of the behaviour of a field-effect transistor when V_g contains an *a.c. component* — the same applies incidentally to all semiconductor devices — can be obtained from the following reasoning. At relatively low frequency the behaviour of the transistor is quasi-static: the situation at any instant can be derived from equation (1) by substituting the appropriate value of V_g. This approximation is no longer valid if the frequency is so high that the period of the signal is approximately equal to the mean transit time τ of the charge carriers, or smaller. This transit time is given roughly by:

$$\tau = l^2/\mu V_d. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

At not too high frequencies it is possible to obtain a good indication of the behaviour by determining the product of gain and bandwidth, as in thermionic valves. The bandwidth of an insulated-gate field-effect transistor, such as the thin-film transistor, is approximately equal to $1/2\pi RC$, where R is the load resistance. The product mentioned is therefore $SR/2\pi RC$ or $\mu V_d/2\pi l^2$. If the bandwidth is expressed as an angular frequency, the gain-bandwidth product is $\mu V_d/l^2$. (As can be seen, this is the reciprocal of the transit time τ.) To obtain good high-frequency characteristics it is therefore necessary to make the channel as short as possible and to choose a semiconductor with the highest possible mobility.

At very high frequencies the gain-bandwidth product as calculated above gives only a very rough indication. The main reason for this is that the transconductance S is a complex quantity whose imaginary component is no longer negligible at high frequencies. Calculations have shown that the absolute value of S at frequencies several times higher than the frequency at which, according to the gain-bandwidth product, the gain should have dropped to unity, is not much lower than the quasi-static value [9]. The gain-bandwidth product of our thin-film transistors is a good 30 Mc/s, but power amplification is possible to up to 200 Mc/s.

Types of design and production methods

In the discussion in the previous section on the physical significance of the potential V_0 we mentioned that its value is connected with the nature and number of the traps present. The first attempts to make an insulated-gate field-effect transistor were only moderately successful, because of the high concentration of surface traps which immobilized most of the induced charge carriers. If the gate is replaced by a reverse-biased P-N junction this effect does not occur [10].

Partly as a result of the progress made in the manufacture of ordinary transistors, thin-film technology has advanced so far that the presence of surface traps need no longer be purely a matter of chance. Although complete control of the process still remains to be achieved, it has nevertheless proved possible in recent years to make field-effect transistors with reasonable success. Two types of design have already been shown in fig. 2. *Fig. 8* again shows a configuration of the type in fig. 2*c*, but with the electrodes on either side of the channel.

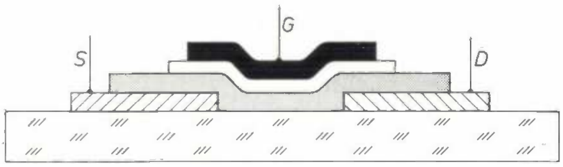


Fig. 8. Configuration of the type in fig. 2*c*, but now with the electrodes on either side of the channel.

As explained above, the channel should be made as short as possible to obtain a high gain-bandwidth product: the lengths that have to be considered are of the order of magnitude of microns. Further, to minimize feedback, the gate electrode should not overlap the other two. With configurations of the type shown in fig. 2*c* and fig. 8 it has indeed proved possible to make transistors meeting these severe requirements [11], but highly sophisticated production methods have to be used.

Not long ago good results were achieved at Philips Research Laboratories in Eindhoven with experimental production methods that do not require such very high precision and are therefore much easier to employ. In the earliest of these methods, which we shall not discuss here, we used a photo-etching technique in which the location of the next electrode was determined by the shadow of the preceding one [12]; this evidently rules out overlapping. The process now mainly used by us, and by which the transistor in fig. 2*d* was made, is illustrated and described in *fig. 9*. A characteristic feature of this process is that the etching away of the aluminium layer undercuts the edge of the lacquer layer [13]. Another important aspect is the use of a special photosensitive lacquer, which, although it can be "developed" in certain aqueous solutions — the exposed part of the lacquer dissolves and the unexposed part does not — will dissolve in organic solvents whether exposed or not. The advantages of the process in fig. 9 are evident. In the first place, overlapping of electrodes is ruled out without the need for precision alignment. Secondly, anodic oxidation produces a layer which, even though extremely thin, is a very good insulator — much better than an evaporated film, as pin-holes quickly close up in the anodic oxidation process. Thirdly, the gate electrode and the layer opposite the channel are entirely enclosed, so that these are not exposed to atmospheric effects. This means that thin-film transistors made by the process of fig. 9 can be stored for a long time without any deterioration in quality. *Fig. 10* shows two photomicrographs of thin-film transistors made in our laboratories.

Although these methods now make it possible to produce thin-film transistors, not everything hoped for has yet been achieved. For example, the nature and number of the surface traps — and hence V_0 — are not yet fully under control, so that transistors identically processed in the same batch may still show individually a fairly wide spread in electrical characteristics. This spread is not due to differences in the properties of the contacts; our process for applying contacts has proved very satisfactory.

A second shortcoming exhibited by all insulated-gate

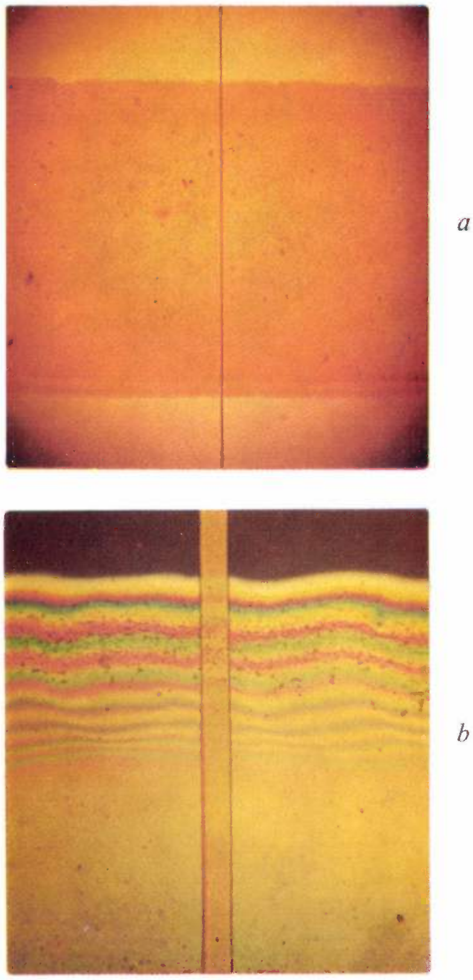


Fig. 10. *a*) Photo-micrograph of thin-film transistor of the type in fig. 2*d* with channel of CdSe. The thin vertical line is the gate electrode. The two other electrodes and the channel cannot be separately distinguished; they form together the dark horizontal strip. *b*) Detail of the edge of the channel adjoining the gate (i.e. the vertical golden strip) looked at through the substrate. The coloured wavy lines are interference fringes in the thin region at the edge of the channel (magnification 475 ×).

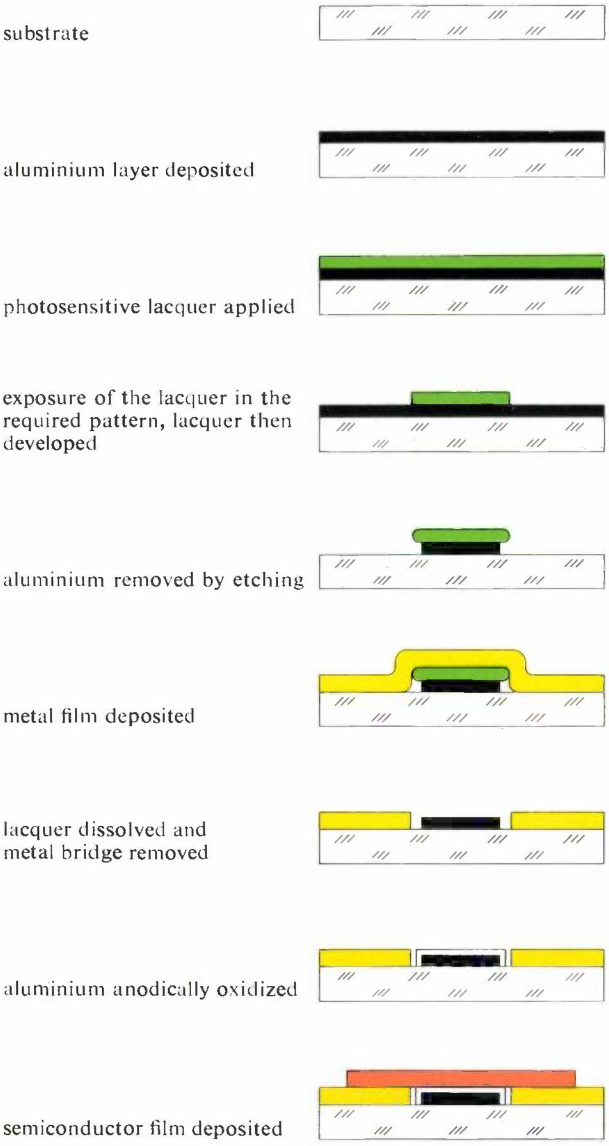

substrate
aluminium layer deposited
photosensitive lacquer applied
exposure of the lacquer in the required pattern, lacquer then developed
aluminium removed by etching
metal film deposited
lacquer dissolved and metal bridge removed
aluminium anodically oxidized
semiconductor film deposited

Fig. 9. Steps in the production of a thin-film transistor of the type in fig. 2*d*, by the photo-etching method developed at Philips Research Laboratories in Eindhoven.

[9] J. A. Geurst, Solid-State Electronics **8**, 88, 1965 (No. 1). Numerical results have been given by J. A. Geurst and H. J. C. A. Nunnink, Solid-State Electronics **8**, 769, 1965 (No. 9).

[10] W. Shockley, Proc. IRE **40**, 1365, 1952.

[11] See articles quoted in reference [2].

[12] H. A. Klasens and H. Koelmans, Solid-State Electronics **7**, 701, 1964.

[13] This method was first used by R. de Werdt of this laboratory for making transistors for ultra-high frequencies.

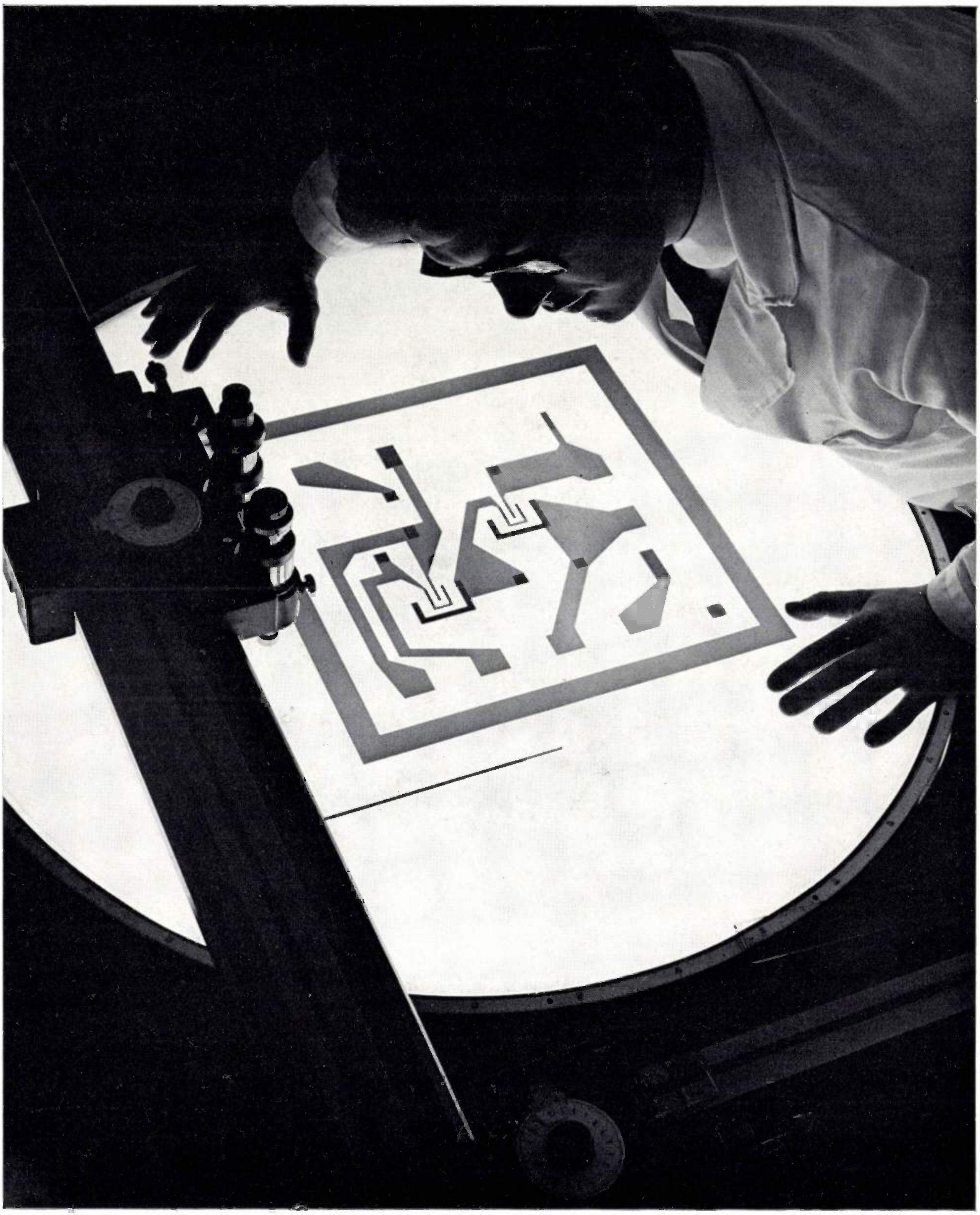
field-effect transistors is that, even when all the voltages are constant, the drain current initially shows a slight drift and only becomes approximately constant after a few hours of operation. In the MOS transistor the current increases, in the thin-film transistor it decreases. The current drift in the thin-film transistor is often attributed to the presence of traps that exchange charge-carriers only very slowly. This hypothesis has not yet been proved. There are indeed other conceivable causes, such as the diffusion of traps in the semiconducting layer under the influence of the field produced by the gate electrode; this would give rise to a gradual change in the concentration of the traps at the surface.

Finally, the characteristics of most thin-film transistors show a certain hysteresis effect (see fig. 6). This is not a result of the effect just discussed — it is also found when the characteristics are recorded at an alternating voltage of 50 c/s or more. The cause is not yet known. The curves in fig. 6 are more or less typical of the magnitude of the effect.

Owing to the limitations we have mentioned, thin-film transistors have not yet come into large-scale production and use. As soon as these difficulties have been sufficiently overcome, the thin-film transistor, particularly in integrated circuits, will undoubtedly find many and various applications.

Summary. The thin-film transistor is a type of field-effect transistor. A strip of CdS or CdSe about 10 μm wide (the channel) with a source and drain electrode on either side of it is deposited on an insulating substrate (glass). The conductivity in the channel is controlled, via the field effect, by the potential V_g applied to an insulated gate electrode. The channel and the electrodes are thin evaporated films; this makes the thin-film transistor ideally suited for use in integrated circuits with evaporated passive elements. The current varies with the square of the voltage V_d between source and drain until V_d reaches a value where the channel is "pinched-off"; above the pinch-off the current is virtually independent of V_d. The value V_0 of V_g at which, for $V_d = 0$, the total charge density is exactly zero, can be either positive or negative. A simple but very effective method of manufacture is a photo-etching process developed at Philips Research Laboratories in Eindhoven. Transistors made by this process have, at low frequencies, an input resistance of about 10^4 MΩ, a transconductance of about 5 mA/V and a surface area of about 0.02 mm^2. The gain-bandwidth product is about 30 Mc/s. Power amplification is possible up to about 200 Mc/s. Individual transistors still show a certain spread in electrical properties, some hysteresis in their characteristics and some current drift: thin-film transistors are therefore not yet quite ready for large-scale production and use. They have many prospective applications, especially in very large integrated circuits.

Manufacture of integrated circuits



In the planar technique several hundred integrated circuits are manufactured from one silicon wafer measuring a few cm. The successive operations in the manufacturing process require the use of a number of masks (the process is described in detail in this issue on page 192). In each mask the pattern required for one operation is repeated several hundred times. Each pattern is first drawn on a very large scale (200 : 1) by means of a drawing machine. The mask is then produced by repetition photography of the drawing. The drawings for the successive patterns must register very accurately; this is being checked here by stacking a number of drawings and viewing them against the light.

Determination of small dimensions by diffraction of a laser beam

M. Koedam

537.525:535.339

Lasers deliver coherent radiation which is concentrated in very narrow spectral regions and contained within an extremely small solid angle. These properties make laser light ideally suited for demonstrating diffraction and interference effects. A method based on this can be successfully applied to a problem long present in the manufacture of incandescent lamps and thermionic valves, for which a solution based on the diffraction of light was proposed many years ago [1]. This is the problem of measuring the diameter of very thin wires.

The arrangement is illustrated in *fig. 1.* A gas laser with He-Ne filling is used as the light source, and this radiates about 5 mW at a wavelength $\lambda = 632.8$ nm (its radiation in the infra-red can be neglected in this context). The diameter of the radiated beam leaving the source is about 1 mm, and its solid angle no more than about 5×10^{-6} steradians (aperture angle approximately 9 minutes of arc). The wire of diameter d is placed transversely in the laser beam, thus setting up a diffraction pattern, which is displayed on a projection screen or a plate of frosted glass. The pattern has minima corresponding to angles a_n ($n = 1, 2, 3, \ldots$), given by:

$$\sin a_n = n\lambda/d. \quad \ldots\ldots \quad (1)$$

By measuring the distance between the minima on the screen and the distance from wire to screen one can calculate $\sin a_n$, which can be used to determine the diameter of the wire from eq. (1). It is thus possible to achieve an accuracy of 0.5% with relatively simple means.

In order to obtain a sharply defined diffraction pattern that can be immediately interpreted, it is essential that the angular distribution of the intensity of the laser beam should show only one maximum. Conventional gas lasers do not meet this requirement, as different modes may occur in the laser tube [2]. For our measurements we constructed a laser in which a particular configuration of the mirrors and the use of a narrow discharge space allowed only single mode operation. (In fact, this configuration unavoidably causes some divergence in the radiated beam, but the beam aperture is still quite small enough for the purpose.)

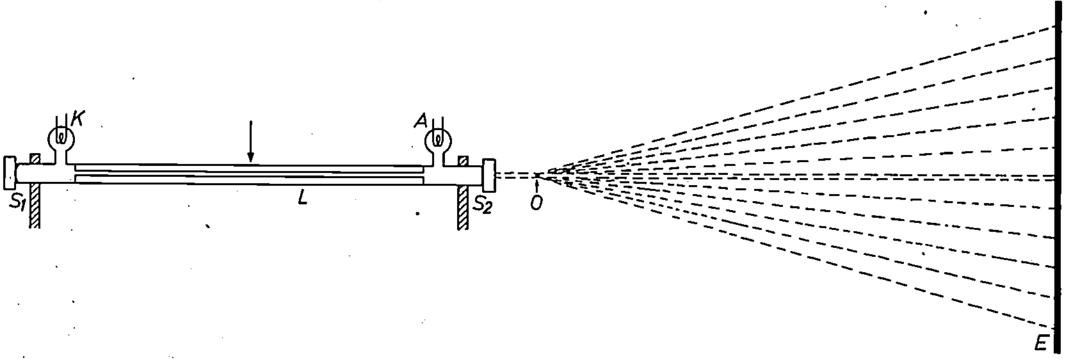


Fig. 1. Arrangement for the measurements. L gas laser, with electrodes A and K, with a potential difference of about 2000 V between them, and with a concave mirror S_1 and a flat mirror S_2. The laser delivers about 5 mW at $\lambda = 633$ nm (in a single mode). The mirrors can be adjusted very easily by bending the tube in the middle, the tube being clamped at both ends. The diffraction pattern of the object O in the laser beam is projected onto a screen E. The dotted lines in the figure indicate the maxima in the diffraction pattern of a wire placed transverse to the beam, and perpendicular to the plane of the drawing.

We used the method for wires with diameters ranging from 5 μm (the diameter of the thinnest wire we use) to 500 μm. The diffraction patterns of wires of various diameters can be seen in *fig. 2a-d.* The clearly perceptible fine structure in the central light-spot, particularly for the thicker wires (*c* and *d*) is the diffraction phenomenon caused by the finite width of the laser beam. To ensure that this structure does not appear in the successive diffraction maxima of the wire, which would interfere with the measurements, the beam diameter must be made several times greater than the diameter of the wire. This is simply done by placing the

Dr. M. Koedam is with Philips Lighting Division, Eindhoven.

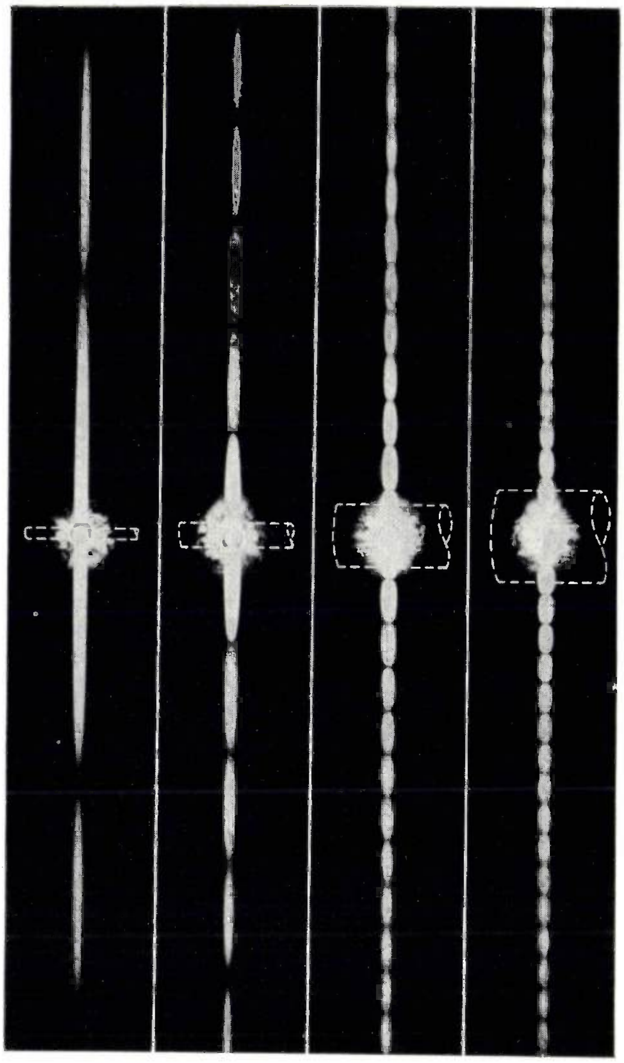


Fig. 2. Diffraction patterns for four tungsten wires of different thicknesses, the wire-to-screen distance being the same in each case. The position of the wires for these photographs is marked by dashed lines. The distances between the minima in each pattern are inversely proportional to the diameter of the wire.

wire at a sufficient distance from the laser. This makes the method less applicable to relatively thick wires: with a diameter of 500 µm a distance of several metres is required. Moreover, with such relatively thick wire the angles α_n are so small that it is also necessary to make the distance between the filament and the screen very considerable (in this case at least 3 m).

The following should make it clear that only the advent of the laser has made this method of measurement a practical possibility.

In order to obtain a sufficiently sharp diffraction pattern with a normal (non-coherent) light source, the solid angle ω of the beam of radiation and the area A of the radiating surface must fulfil the condition [3]:

$$\omega A < \frac{\pi^2}{4} \lambda^2. \qquad \qquad (2)$$

This at the same time limits the luminous intensity of the diffraction pattern, because the luminous flux in the beam is ωA times the given luminance of the light source. With one of the best light sources hitherto available for this purpose, the CS 100 W mercury-vapour lamp (luminance 1.7×10^9 cd/m², $A = 0.2$ mm²), and using the spectral line at $\lambda = 546.1$ nm, one finds from eq. (2) that ω must be $< 4 \times 10^{-6}$ steradians, i.e. about equal to the solid angle of our laser beam (see above). Because of the larger beam diameter, but chiefly owing to the laser's enormously greater luminance, the luminous flux for the diffraction pattern provided

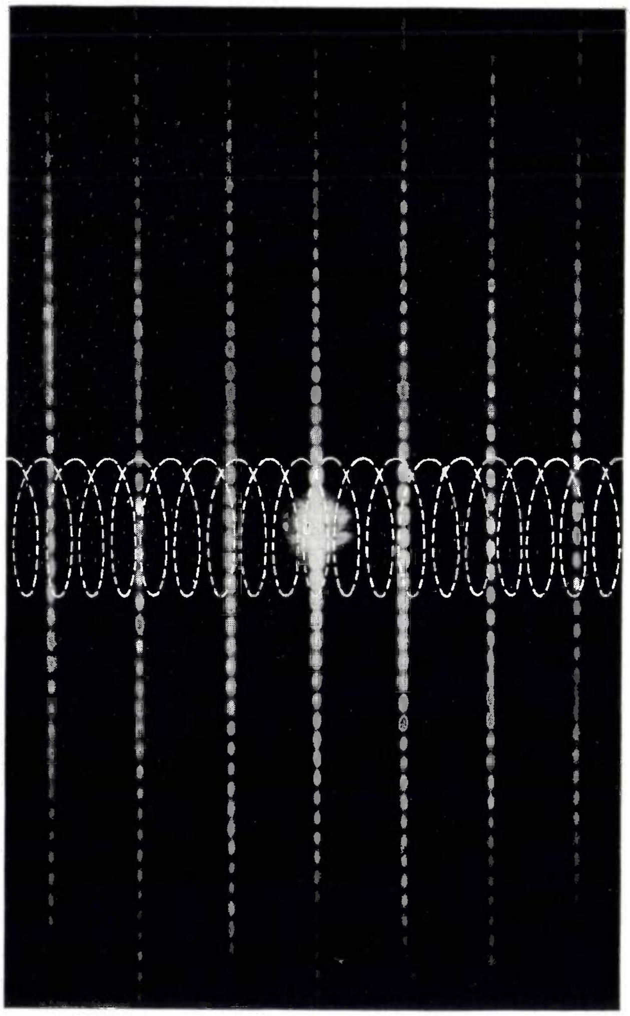


Fig. 3. Diffraction pattern of a helix of 113 µm diameter and 26 µm pitch. The helix is marked by dashed lines.

[1] I. Runge, Technisch-wissenschaftliche Abhandlungen aus dem Osram-Konzern 1, 165 and 170, 1930.

[2] The various maxima thus produced in the transverse cross-section of the beam, forming a regular pattern in it, could also be clearly seen with the small gas laser previously described in this journal; see J. Haisma, S. J. van Hoppe, H. de Lang and J. van der Wal, Philips tech. Rev. 24, 95, 1962/63, particularly fig. 3.

[3] A. C. S. van Heel, Inleiding in de optica, Nijhoff, The Hague 1958, page 83, eq. (132). We have adapted this equation to the problem dealt with here.

within this angle by the laser is about 5000 times greater (making due allowance for the sensitivity of the human eye, which at $\lambda = 546$ nm is substantially greater than at our laser wavelength of $\lambda = 633$ nm).

As the pattern obtained with the old light sources was so very dim, making it necessary to work in the dark and to use photographic aids, the method has hitherto not been attractive. The measurements can now be performed visually in daylight, so that only now can full benefit be taken of the many advantages of diffraction measurement compared with other methods of measuring the diameter of thin wires. The method is simple, non-destructive, fast, and accurate, it requires no calibration, and, if necessary, can be used for small lengths of wire (about 1 mm). It is also possible to measure continuously the diameter of a moving wire by photo-electric recording of the diffraction pattern.

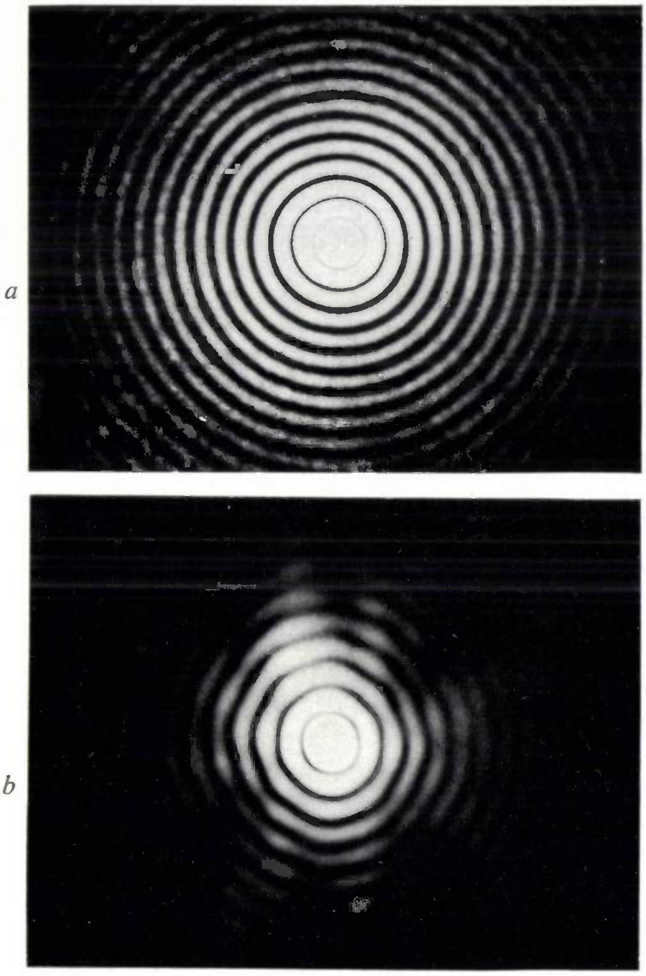


Fig. 4. Diffraction patterns for die bores.
a) Bore of diameter 136 μm, reasonably circular.
b) Bore of diameter 50 μm, somewhat octagonal.
(The magnification of photographs (*a*) and (*b*) is not identical.)

Instead of a straight wire a *helical* wire may also be placed in the laser beam. A *double* diffraction effect is then obtained (*fig. 3*): the periodic structure of the helix acts like a diffraction grating, producing an interference pattern in the form of a series of equidistant bright lines perpendicular to the axis of the helix. This may serve for measuring the pitch of the helix. Along each of these lines, however, diffraction maxima appear which prove to correspond to the diffraction which would be caused by a solid straight filament of a thickness equal to the helix diameter.

Our last figure shows diffraction patterns obtained from a *diamond die*, placed axially in the laser beam. The diffraction at the circular hole formed by the bore of the die results in a series of concentric rings (*fig. 4a*). The bore diameter can be calculated very accurately from the distances between the rings. This measurement, made possible in practice only by the laser method, enabled us to confirm a fact we had long suspected: that the wire drawn through a die has a slightly greater diameter than the bore. If, after considerable use, the bore is no longer circular, the imperfection can immediately be seen from the appearance of irregularities in the diffraction pattern (fig. 4*b*).

Summary. A wire is placed in the beam of a gas laser which radiates about 5 mW at a wavelength of 633 nm. From the diffraction pattern projected on to a screen behind the wire the diameter of the wire can be derived in a simple way with an accuracy of 0.5%. The method has only become a practical proposition through the use of the laser which, because of its extremely high luminance, gives a diffraction pattern 5000 times brighter than that obtained with the best light sources previously available. This method of measurement, which offers considerable advantages over conventional methods for very thin wire, has been used for wires from 5 μm to 500 μm in diameter. The same arrangement can be used for determining the diameter and the pitch of a wire helix as well as the bore diameter of diamond dies.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands *E*
Mullard Research Laboratories, Redhill (Surrey), England *M*
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France *L*
Philips Zentrallaboratorium GmbH, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany *A*
Philips Zentrallaboratorium GmbH, Hamburg laboratory, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany *H*
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium. *B*

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

W. Albers: Preparation of single-phase crystals of compounds with extremely small existence regions.
J. chem. Phys. **43**, 4401-4402, 1965 (No. 12). *E*

C. Albrecht and **J. Proper:** Detail rendition in X-ray images: theory and experimental results.
Medicamundi **11**, 44-48, 1965 (No. 2). *E*

K. Bethe: Unterdrückung von Mantelströmen auf geschlossenen Hochfrequenzleitungen.
Int. elektron. Rdsch. **20**, 137-142, 1966 (No. 3). *H*

G. Blasse: Ferromagnetism and ferrimagnetism of oxygen spinels containing tetravalent manganese.
J. Phys. Chem. Solids **27**, 383-389, 1966 (No. 2). *E*

G. Blasse: Antiferromagnetism of the spinel $LiCuVO_4$.
J. Phys. Chem. Solids **27**, 612-613, 1966 (No. 3). *E*

R. Bleekrode and **W. C. Nieuwpoort:** Absorption and emission measurements of C_2 and CH electronic bands in low-pressure oxyacetylene flames.
J. chem. Phys. **43**, 3680-3687, 1965 (No. 10^{I}). *E*

K. H. J. Buschow and **J. H. N. van Vucht:** Die binären Systeme Cer-Aluminium und Praseodym-Aluminium.
Z. Metallk. **57**, 162-166, 1966 (No. 2). *E*

J. Cayzac: Problèmes posés par la détection des faibles différences de température en télévision infrarouge.
Acta electronica **9**, 7-18, 1965 (No. 1). *L*

B. H. Clarke: Resonance relaxation in $Mn_xFe_yO_4$ by slow relaxing manganic and ferrous ions.
J. Phys. Chem. Solids **27**, 353-362, 1966 (No. 2). *M*

R. C. French: A wideband transistorized power amplifier.
Electronic Engng. **38**, 8-11, 1966 (No. 455). *M*

J. A. Geurst: Veldeffecttransistoren.
Ned. T. Natuurk. **31**, 436-453, 1965 (No. 12). *E*

O. Glemser, J. Schröder and **J. Knaak:** Notiz zur Darstellung von Stickstofftrifluorid durch Elektrolyse von geschmolzenem Ammoniumhydrogenfluorid.
Chem. Berichte **99**, 371-374, 1966 (No. 1). *A*

C. A. A. J. Greebe: The influence of trapping on the acousto-electric effect in CdS.
Philips Res. Repts. **21**, 1-15, 1966 (No. 1). *E*

G. S. Gruintjes and **G. J. Oudemans:** Continuous hot pressing.
Special Ceramics 1964 (Proc. Symp. Brit. Cer. Res. Ass.), pp. 289-299, Academic Press, London 1965. *E*

G. E. G. Hardeman and **G. Gerritsen:** Suppression of nuclear dynamic polarization by RF radiation.
Low temperature physics LT 9, Part B (Proc. 9th int. Conf., 1964), pp. 921-924, Plenum Press, 1965. *E*

E. E. Havinga: Ferroelectric perovskites containing manganese ions.
Philips Res. Repts. **21**, 49-62, 1966 (No. 1). *E*

B. R. Hearn: Mechanism of heat loss from ferrite storage cores in still air.
Electronics Letters **1**, 276-277, 1965 (No. 10). *M*

J. C. M. Henning, H. van den Boom and **J. Dieleman:** Electron-spin resonance of Co^{2+} in cubic ZnSe.
Philips Res. Repts. **21**, 16-26, 1966 (No. 1). *E*

T. Klein and **J. R. A. Beale:** Simultaneous diffusion of oppositely charged impurities in semiconductors.
Solid-State Electronics **9**, 59-69, 1966 (No. 1). *M*

S. R. de Kloet: Studies on protein biosynthesis in yeast, I. The effect of cycloheximide on cell free protein formation.
Proc. Kon. Ned. Akad. Wetensch. **B 68**, 266-283, 1965 (No. 5). *E*

J. T. Klomp and **Th. P. J. Botden:** A new sealing method for high-purity alumina.
The use of ceramics in valves, Symp. Brit. Cer. Res. Ass., Stoke-on-Trent 1965, pp. 129-139. *E*

J. W. L. Köhler: The Stirling refrigeration cycle.
Sci. American **212**, No. 4, 119-125 and 127, 1965 (April). *E*

H. Kunze: On the mobility of electrons in bismuth doped with tellurium.
Physics Letters **20**, 469-470, 1966 (No. 5). *A*

H. de Lang and **G. Bouwhuis:** Nonlinear preference for circular mode polarization observed on various lasing neon transitions.
Physics Letters **20**, 383-384, 1966 (No. 4). *E*

M. C. Lefranc: Le refroidissement des détecteurs de rayonnement infrarouge.
Acta electronica **9**, 47-90, 1965 (No. 1). *L*

G. Marie: Préamplificateurs paramétriques adaptés à la détection de l'infrarouge.
Acta electronica **9**, 35-46, 1965 (No. 1). *L*

M. K. McPhun: Practical stability criterion for tunnel-diode circuits.
Electronics Letters **1**, 289, 1965 (No. 10). *M*

H. A. Meijer: Quantitative analysis of ferrite in austenitic stainless steel.
Brit. Welding J. **13**, 12-17, 1966 (No. 1). *E*

E. A. Muyderman: Constructions with spiral-groove bearings.
Wear **9**, 118-141, 1966 (No. 2). *E*

J. Neirynck: Analyse de circuits non-linéaires par la méthode des récurrences; application à deux circuits de thyristors.
Rev. MBLE **8**, 191-203, 1965 (No. 3/4). *B*

K. H. Nicholas: Studies of anomalous diffusion of impurities in silicon.
Solid-State Electronics **9**, 35-47, 1966 (No. 1). *M*

G. W. van Oosterhout: Voorstadia van het breken van korrels tijdens het maalproces.
Chem. Weekblad **62**, 16-17, 1966 (No. 2). *E*

A. J. W. M. van Overbeek and **W. A. J. M. Zwijsen:** Stable variable-frequency *RC* filters with positive feedback.
Proc. IEEE **53**, 1758-1759, 1965 (No. 11). *E*

C. Picot: Préamplificateurs pour appareils de prise de vues de télévision infrarouge (domaine spectral situé au voisinage de 4 microns).
Acta electronica **9**, 19-34, 1965 (No. 1). *L*

H. Rau: High temperature equilibrium of atomic disorder in SnS.
J. Phys. Chem. Solids **27**, 761-769, 1966 (No. 4). *A*

D. A. Schreuder: Über die Beleuchtung von Verkehrstunneln.
Lichttechnik **17**, 145 A-149 A, 1965 (No. 12).

D. A. Schreuder: Verlichting voor nachtelijk wegverkeer.
T. soc. Geneesk. **44**, 230-234, 1966.

G. Schulten: Resonatoren für Millimeterwellen und ihre Verwendung zur Beobachtung von Gasresonanzen.
Frequenz **20**, 10-22, 1966 (No. 1). *H*

J. van Suchtelen, J. Volger and **D. van Houwelingen:** The principle and performance of a superconducting dynamo.
Cryogenics **5**, 256-266, 1965 (No. 5). *E*

J. Tillack, P. Eckerlin and **J. H. Dettingmeijer:** Preparation and properties of tungsten dioxide diiodide, WO_2I_2.
Angew. Chemie, Int. Edition in English **5**, 421, 1966 (No. 4). *A*

D. R. Tilley: Cylindrical Josephson junctions.
Physics Letters **20**, 117-118, 1966 (No. 2). *M*

M. G. Townsend and **O. F. Hill:** Tetravalent cobalt ion in α-Al_2O_3.
Trans. Faraday Soc. **61**, 2597-2602, 1965 (No. 12). *M*

E. Tscholl: Effect of nickel on thermally stimulated conductivity spectra in CdS-crystals.
Solid State Comm. **4**, 87-90, 1966 (No. 2). *E*

T. J. Turner, R. De Batist and **Y. Haven:** Photon-induced reorientation of M-centers in potassium chloride.
Phys. Stat. sol. **11**, 535-552, 1965 (No. 2). *E*

J. Volger: Thermo-elektriciteit en haar toepassingen.
Electrotechniek **44**, 45-49, 1966 (No. 2). *E*

J. van de Waterbeemd: Kinetics of growth and structure of thin films of tin on an amorphous carbon substrate.
Philips Res. Repts. **21**, 27-48, 1966 (No. 1). *E*

H. W. Werner: Mass and energy dependence of the sensitivity of photographic plates as ion detectors in mass spectrometers.
Philips Res. Repts. **21**, 63-70, 1966 (No. 1). *E*

J. S. van Wieringen and **J. G. Rensen:** Internal magnetic fields at iron sites in "Ticonal" permanent magnets determined by Mössbauer measurements.
Solid State Comm. **4**, 1-2, 1966 (No. 1). *E*

H. Zimmer: Harmonic generation of microwaves produced by a field-emission cathode in a superconducting cavity.
Appl. Phys. Letters **7**, 297-298, 1965 (No. 11). *H*

H. Zimmer: Improved stability of field-emission current at microwave frequencies.
Appl. Phys. Letters **7**, 298-300, 1965 (No. 11). *H*

Volume 27, 1966, No. 7 pages 173-212 Published 2nd November1966

PHILIPS TECHNICAL REVIEW

VOLUME 27, 1966, No. 8

Phase theory

III. Ternary systems

J. L. Meijering

541.12.01

This third article in the series on phase systems sets out to give an understanding of the theory and practical uses of ternary diagrams. The phase rule is subjected to a critical scrutiny and a slightly modified formulation is proposed. A method of calculation is given which enables useful predictions about ternary systems to be made using only data from binary systems.

Introduction

The concept of "component"

In a single-phase region of a unary system the pressure and temperature can be varied independently. If two phases are in equilibrium with each other, only one of these quantities is independently variable, and if three phases are in equilibrium there is no variable left. If the number of phases is P and the number of degrees of freedom F, the equilibrium state of a unary system is given by $F + P = 3$.

For a binary system the analogous relation $F+P = 4$ applies. For example, the binary system acetonitril-water shows a four-phase equilibrium (two liquid + two solid phases) at one specific temperature and pressure (—24.2 °C, 1240 atm) [1].

We thus arrive at the definitions that a system is unary, binary or ternary if, consisting of one phase, it has two, three or four independent variables respectively. In the unary system the pressure and the temperature are the only independent variables; in the binary system there is one additional concentration which can be independently varied, and in the ternary system there are two (*fig. 1*).

It may look rather as if we have gone about our definitions in a rather roundabout way, and that it might have been sufficient to define the respective systems as consisting of one, two or three components. The use of the term component brings with it however certain difficulties, which become apparent if we ask the following questions.

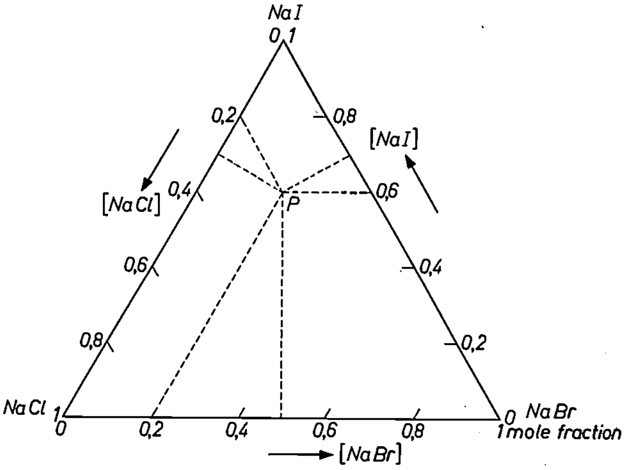


Fig. 1. Diagram of the ternary system NaCl-NaBr-NaI, where the concentrations are plotted along the sides of an equilateral triangle (in mole fractions). Concentrations can also be plotted and read off in a diagram of this kind by drawing lines perpendicular to the three sides of the triangle (see P). The mole fraction of each component is then indicated by the line perpendicular to the side opposite the vertex representing that component, taking the height of the triangle as equal to 1.

How many components does the system NaCl-NaBr-NaI possess? This is easily answered: three. How many components are there in the system NaCl-KCl-NaBr-KBr? Again, the answer is three — this time because of the occurrence of the reaction $NaCl + KBr \rightleftarrows NaBr + KCl$. What are the three components? We can choose any three of these four compounds. Which components of the system they may be is not relevant, but their number is. When the concept of component is used, this fact is still too easily overlooked. Moreover, as the foregoing shows, this number will by no means always be so easy to determine. There are various other reasons for this, which we shall deal with in some detail later.

We state here as our provisional conclusion, to be amplified later, that it is still too often assumed that

Prof. Dr. J. L. Meijering, formerly with Philips Research Laboratories, Eindhoven, is now Professor of Inorganic Chemistry and Metallurgy at the Technical University of Delft, Netherlands.

[1] G. Schneider, Z. phys. Chem. N.F. **41**, 327, 1964.

the number of components of a system can always be found from a purely theoretical analysis of that system. The way in which the well-known phase rule is formulated may also be seen to originate from this approach. We prefer to work with the number of independent variables in a system, a number that can be found by experiment. Partly in this connection we shall, in one of the following sections, propose a slightly modified formulation of the phase rule.

In the considerations which follow on the ternary system we shall repeatedly refer back to parts I and II of this series of articles, dealing with unary and binary systems [2].

Ternary diagrams

The ternary system, as stated, has four independent variables: the pressure p, the temperature T and two concentration variables, x and y. The gas phase, and with it the effect of the pressure, are often left out of consideration. In order to investigate the effect of the three other variables, what is really needed is a representation in three dimensions, e.g. with T plotted vertically above the triangle in fig. 1. The practice is to use *cross-sections* through the three-dimensional diagram; these can be either *isothermal cross-sections*, where the temperature is kept constant (fig. 1), or *"vertical" cross-sections*, where a concentration or concentration ratio is kept constant. Both cross-sections are discussed in the following pages.

Two-dimensional *survey diagrams* are also used, which indicate what happens to all possible *three-phase equilibria* in a ternary system when the temperature varies. These diagrams, an example of which is given in Appendix I, provide a good general picture of ternary systems.

As regards the isothermal cross-section it should be added that the two concentration variables can simply be plotted at right angles to each other instead of at an angle of 60°, as in fig. 1. A right-angled triangle is then obtained. The only advantage of the usual equilateral triangle is that the three components are treated alike. In many cases, however, the right-angled representation has definite advantages, particularly if one of the components is of a differing character, as for example in H_2O-$CaCl_2$-$MgCl_2$. This applies all the more if the investigation does not extend to the anhydrous salts, but no farther, say, than to $CaCl_2.6H_2O$ and $MgCl_2.6H_2O$, so that the triangle is not "complete" anyhow. Another advantage of the right-angled representation is that one can choose any arbitrary scale for the x and y axis. This is an obvious advantage in systems such as Fe-Ni-C, which in practice involve the investigation of alloys which contain up to only one percent by weight of carbon.

It should be noted that all methods of representation discussed here, no matter whether percentages by weight or mole fractions are used, obey the rule that the overall concentration of heterogeneous mixtures of two phases is always located on the *straight line* which joins the concentration points of the individual phases.

The isothermal cross-section

The way in which the binary phase diagram can be derived from the behaviour of the Gibbs' free energy was relatively easy to visualize (see page 17 in I). In the ternary case we are at once compelled to think in more three-dimensional terms. For example, we encounter double-tangent planes which can be "rolled" in various directions against particular surfaces. This does not, however, introduce anything really new, as we shall try to show by considering a simple binary example.

The G-x diagram of the binary system A-B in *fig. 2* shows the Gibbs' free energy curve of the phases α and β as a function of x, the mole fraction of the component B. The curves chosen are "convex"; the central portion has been omitted since it is not for the moment relevant. What we want to know is: at what values of x is the phase α stable? To find the answer we draw a tangent to the G-x curve for α at the poitn $x = 0$, and we "roll" this tangent along the curve until it also becomes tangent to the G-x curve for β. In the concentration region which we have thus covered, the phase α is stable. (As soon as the tangent to the α curve becomes a tangent to the β curve as well, or cuts it, heterogeneous mixtures of α and β become stable.)

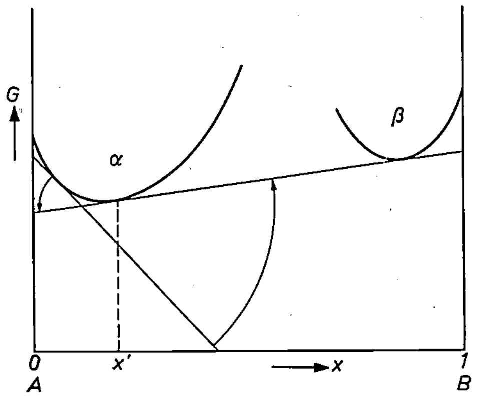


Fig. 2. Part of a G-x diagram of a binary system with phases α and β. If a tangent is rolled around the G-x curve of the α phase until it becomes tangent to the G-x curve of the β phase, the concentration range is found (from $x = 0$ to $x = x'$) at which the phase α is stable with respect to heterogeneous mixtures of α and β.

The same procedure can be adopted for a ternary system A-B-C. Instead of G-x curves we then have G-x-y surfaces, and instead of two "convex" curves for α and β we then have three convex surfaces, or "pockets", for the phases α, β and γ, which link up with the corners A, B and C respectively of the concentration triangle: see *fig. 3*. Tangential *planes* now play

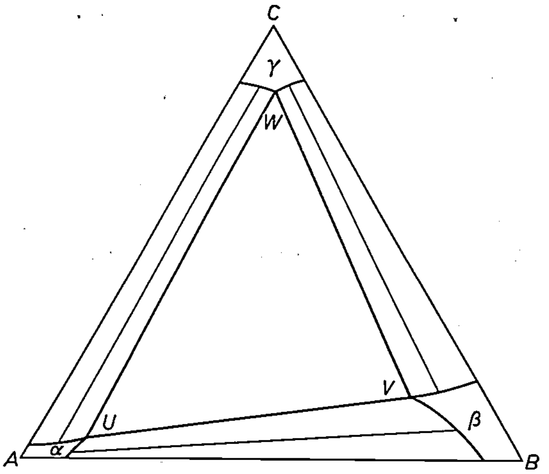


Fig. 3. Isothermal cross-section with three phases. In the corner zones of the triangle the homogeneous phases α, β and γ are stable; there are also three two-phase regions $\alpha + \beta$, $\alpha + \gamma$ and $\beta + \gamma$ (in which a few tie-lines are drawn) and a triangle within which three-phase equilibria are stable. The tie-lines need not run parallel; knowledge of a ternary diagram is therefore not complete unless the direction of the tie-lines is also known. The three phases of which the three-phase equilibria consists each have, at a given pressure and temperature, a very specific composition, which is indicated by the respective vertices U, V and W of the "three-phase triangle".

the same part as the tangential *lines* in the binary system, double-tangent planes have the function of double-tangent lines, and so on.

If we apply a tangent plane at some point or another on the α surface, and this plane continues on under the two other G-x-y surfaces, then the α phase is stable with the concentration at the chosen point. This tangent plane can be rolled in the direction of B or of C until it becomes tangent to a second surface, the β or the γ surface. We then have two associated tangent points which indicate the concentrations of two phases in equilibrium with each other. These equilibrium concentrations are plotted in the phase diagram. In the three-dimensional diagram they form boundary planes, and in isothermal and vertical cross-sections they mark the boundary lines between the homogeneous regions (α, β and γ) and the two-phase regions ($\alpha + \beta$, $\alpha + \gamma$ and $\beta + \gamma$); see, for example, the isothermal cross-section in fig. 3. We have to think of each point of a boundary line as being connected by a tie-line to a point on the other boundary line of the two-phase region. Only a few of these tie-lines are drawn. Unlike those in the T-x cross-sections of binary systems, these tie-lines do not have to run parallel with each other. Usually they tend rather to fan out.

The method of forming of a ternary diagram can be represented rather more easily by imagining a *double-tangent* plane being rolled against *two G* surfaces. If, for example, we roll such a double-tangent plane along α and β in the direction of C, the γ surface will at a given moment be met and a triple-tangent plane will be obtained. The same triple-tangent plane is obtained by letting a double-tangent plane roll on β and γ, or on α and γ. The three sides of the triangle UVW in fig. 3 are the "last" tie-lines of the three two-phase regions $\alpha + \beta$, $\beta + \gamma$ and $\gamma + \alpha$: UV, VW and WU. At all concentrations *inside* the triangle UVW the most stable state — at the temperature in question — is an equilibrium of three phases, whose composition is given by U, V and W. The G surface of α, β, γ and their stable heterogeneous combinations thus consist in this case of a plane triangle, three ruled surfaces (in which at any point a straight line can be drawn, i.e. a tie-line) and three completely curved surfaces.

Let us now include the middle part of the G-x-y surfaces in our considerations. The three "pockets" corresponding to α, β and γ *may* form part of one and the same surface (which is then not "convex" in the middle). In that case α, β and γ have the same structure, and we can speak of immiscibility in the strict sense of the word as used in the previous articles.

As we know, the tendency towards immiscibility generally decreases when the temperature is increased. In *fig. 4*, which is identical with fig. 17 in I, this is illustrated by the change in the G-x curve of a binary

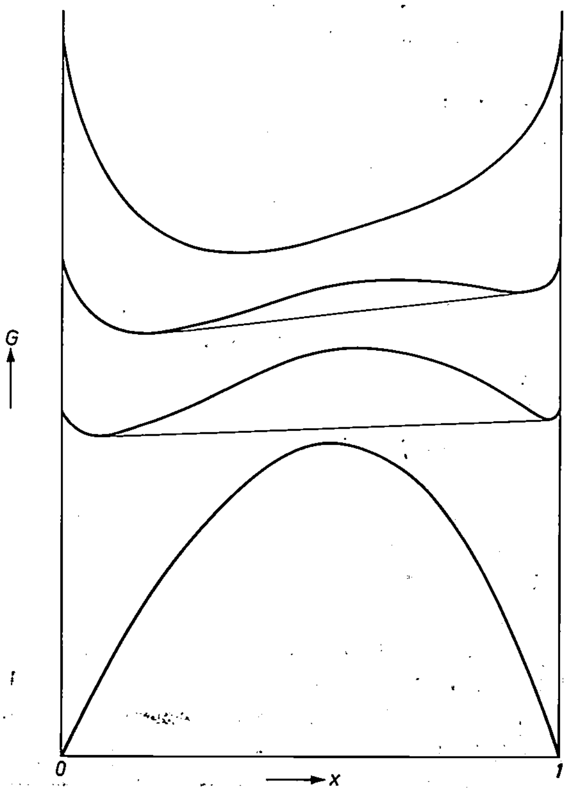


Fig. 4. Change in miscibility as a function of temperature, demonstrated by the change in the binary G-x curves. The temperatures for which each curve is calculated increase on going up the diagram. (Cf. fig. 17 in part I.) In the lowest case the miscibility gap extends from $x = 0$ to $x = 1$, in the topmost case there is no separation into two phases.

[2] J. L. Meijering, Phase theory, I. Introduction to unary and binary systems; II. Quantitative considerations on binary systems, Philips tech. Rev. **26**, 12-26 and 52-60, 1965.

system. If the curve has become completely convex (the upper curve) then immiscibility no longer occurs. Thus, a *G-x-y* surface of the kind referred to will generally turn into a completely convex surface (a pocket) with increasing temperature. The isothermal cross-section then indicates only one homogeneous phase.

We shall now consider some intermediate forms. Let us assume that the *G-x-y* surface is not perfectly convex, but has a fold between *P* and *Q*, as shown in *fig. 5*. In the binary boundary system a double-tangent line can then be drawn and from this we can roll a double-tangent plane backwards until finally the tangent points coincide at *R*. Beyond *R* the surface is completely convex and therefore no double-tangent planes can be introduced. Corresponding to this *G-x-y* surface is an isothermal cross-section as shown in *fig. 6*, with a *critical point* at *R*. Such folds need not of course be limited to one side, and so a diagram can be obtained which has two or three distinct miscibility gaps, each ending in a critical point. Folds that reach from one boundary system to the other give rise to a strip, as in *fig. 7a*. A fold, or rather an indentation in the surface that nowhere reaches the edges, gives a miscibility gap that begins and ends in a critical point (fig. 7*c*). A further possibility is a diagram with a three-phase triangle and two miscibility gaps with a critical point; see *fig. 8*.

In demonstrating these variations in the type of isothermal cross-section we have so far confined ourselves to the case where there is only one continuous *G-x-y* surface. If we let several such surfaces, relating to different phases, run through one another, the number of possible variants is virtually inexhaustible. We shall consider here just a few of the most important types.

First of all we consider the possibility that a fourth phase is present, or even a fifth, sixth, etc. An example is shown in *fig. 9*. Referring to *fig. 10*, one can have a cross-section of this type at a temperature just above a ternary *melting point minimum* (this is rather rare, but see fig. 10 in I for the binary analogue). Imagine that the *G-x-y* surface of the solid phase has the form of a shallow bowl. Through this bowl there penetrates a deeper one relating to the liquid phase. We then have a series of double-tangent planes forming a closed ring. If the temperature is lowered, then at a specific temperature the bowls will only just touch. At

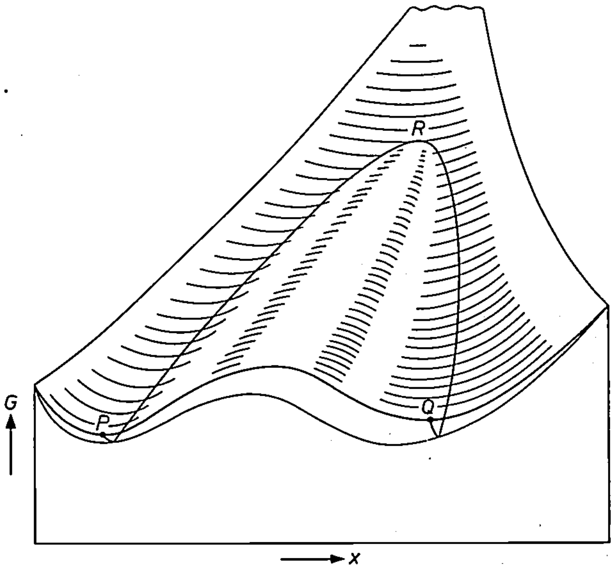


Fig. 5. Perspective sketch of part of a *G-x-y* surface with a fold at one edge. The curves *PR* and *QR* are described by the tangent points of a double-tangent plane rolled against the underside of the surface. The two tangent points meet at *R*.

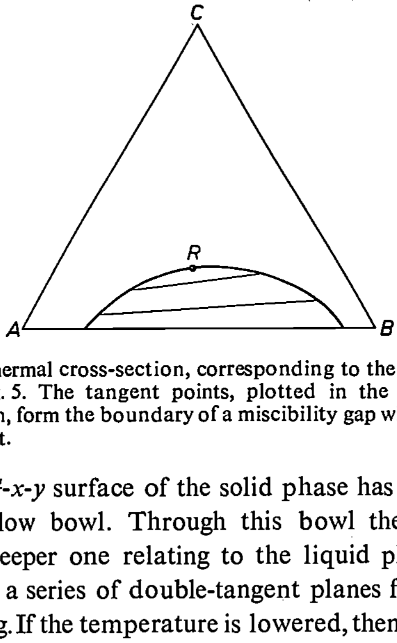


Fig. 6. Isothermal cross-section, corresponding to the *G-x-y* diagram in fig. 5. The tangent points, plotted in the isothermal cross-section, form the boundary of a miscibility gap with *R* as the critical point.

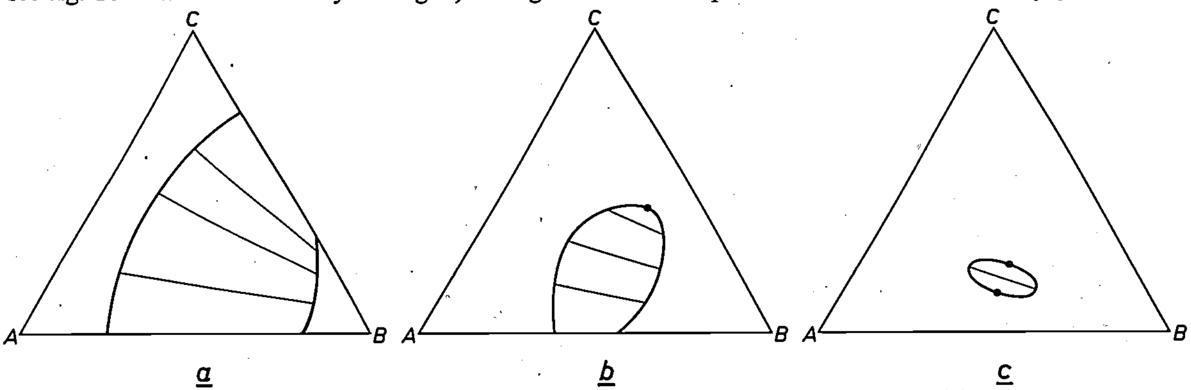


Fig. 7. Three isothermal cross-sections of the same system, obtained at successively higher temperatures. *a*) Miscibility gap in the form of a strip. *b*) Miscibility gap at a boundary. Unlike the type of miscibility gap in fig. 6, the mutual miscibility of *A* and *B* decreases when a small amount of *C* is added. *c*) Closed miscibility gap, with two critical points.

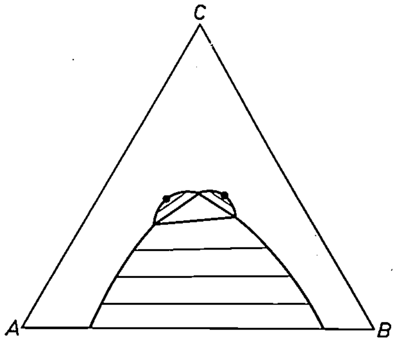


Fig. 8. Miscibility gap with three-phase triangle and two critical points.

this melting point minimum the two closed curves disappear simultaneously at one point.

Let us now take a closer look at the isothermal cross-section in fig. 3. This diagram can be taken as applying to a system of three solids which form mixed crystals when mixed on a limited scale, and equilibria of two or three phases under more extensive mixing. We are at liberty to make further assumptions about the system, for example that each of the binary boundary systems is of the eutectic type (cf. fig. 12 in I). This then implies that the isothermal cross-section has been taken at a temperature lower than the lowest of the three eutectic temperatures — otherwise the diagram would certainly contain a liquid phase as well. What now interests us is how a diagram of this kind changes character if we raise the temperature sufficiently to stabilize the liquid phase as well, in other words, what the diagram will look like above the temperature at which the G surface of the liquid phase first comes into contact with the G surface of α, β and γ and their heterogeneous combinations. This surface, as we have seen, consists of a planar triangle, three ruled surfaces and three fully curved surfaces. Usually the first contact with the G surface of the liquid phase will be in the triangle. At that temperature there is a *quadruple*-tangent plane, at α, β, γ and the liquid phase. We thus have a four-phase equilibrium: in the triangle UVW there is then a point that denotes the composition of a *ternary eutectic*, joined by three tie-lines to the points U, V and W. When the temperature is raised somewhat higher, the G surface of the liquid phase breaks through the $(\alpha + \beta + \gamma)$ triangle. There are now no longer any stable $\alpha + \beta + \gamma$ equilibria. We have instead the three new phase equilibria: $\alpha + \beta + 1$, $\beta + \gamma + 1$ and $\gamma + \alpha + 1$; see *fig. 11a.*

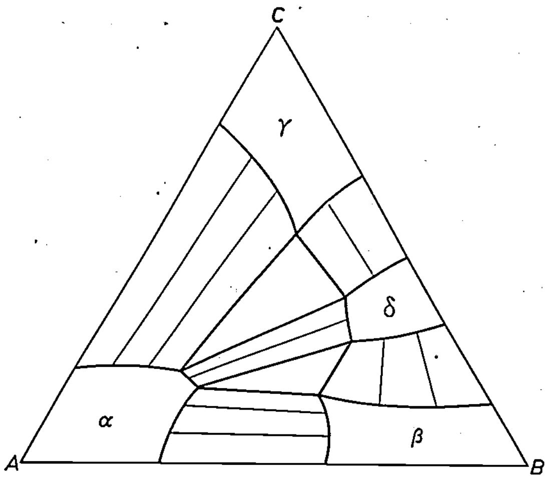


Fig. 9. Isothermal cross-section with *four* phases and two three-phase triangles.

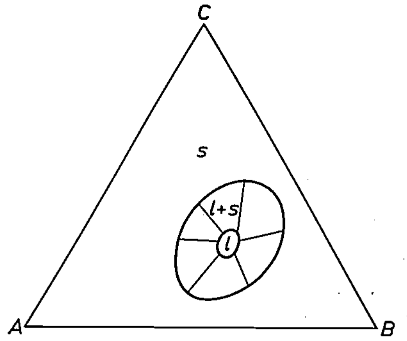


Fig. 10. Isothermal cross-section at a temperature just above a ternary melting point minimum.

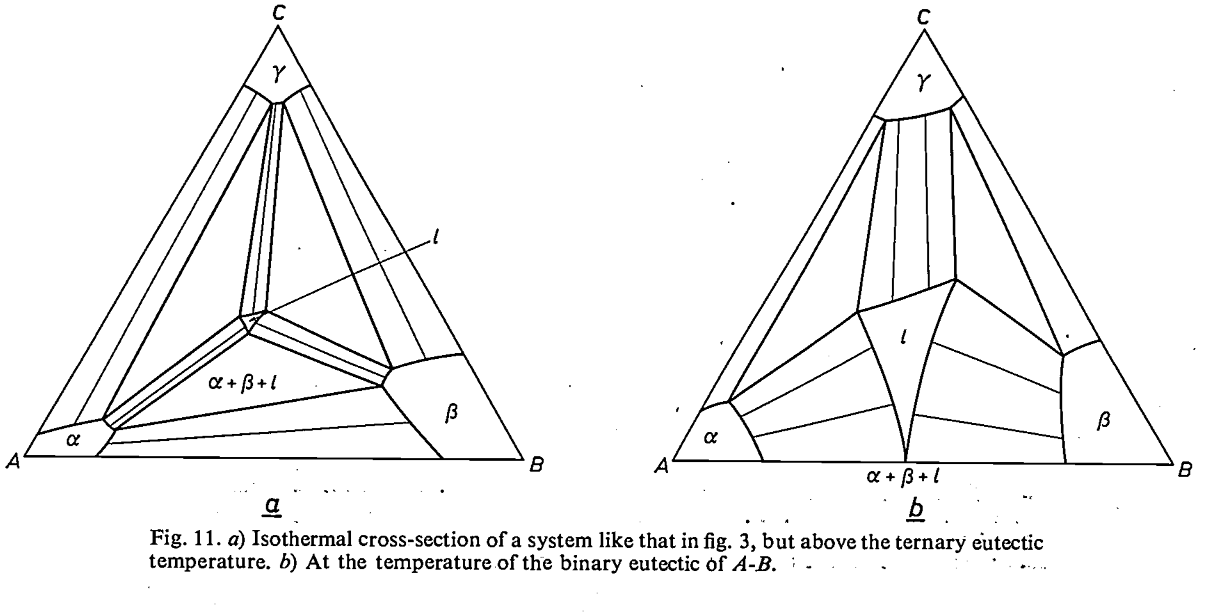


Fig. 11. *a*) Isothermal cross-section of a system like that in fig. 3, but above the ternary eutectic temperature. *b*) At the temperature of the binary eutectic of A-B.

If no complications arise, these three-phase equilibria change one after the other into binary equilibria upon further increases of temperature. If the eutectic temperature of the *A-B* system is lower than that of the two other boundary systems, the cross-section at this temperature is as shown in fig. 11*b*. The liquid zone has here just reached the side *A-B*, and the three-phase triangle $\alpha + \beta + 1$ has degenerated to the binary three-phase equilibrium $\alpha + \beta + 1$.

In the four-phase equilibrium discussed above, one of the concentration points was located in the triangle formed by the three other phases. The other possibility is that the four points lie on a tetragon. To understand how such a four-phase equilibrium comes about, let us look at fig. 9. For convenience we assume that the minima of the *G* surfaces of α and δ on one hand and of β and γ on the other have the same value at every value of *T* considered. (We may do this because the differences in *G* between *A*, *B* and *C* are insignificant, provided that *A*, *B* and *C* are not mutually convertible, and thus can be chosen freely.) In fig. 9 the minima of α and δ are lower than those of β and γ, and therefore we can roll the double-tangent plane on α and δ between two end points, so that it becomes tangent on one hand to β and on the other to γ as a third phase. We now change the temperature and assume that the minima of the *G* surfaces of β and γ then fall with respect to the minima of α and δ. At a given temperature we then obtain a quadruple-tangent plane. The two-phase region $\alpha + \delta$ has then degenerated to a diagonal of the tangent point tetragon. If we change *T* still further, the *G* minima of β and γ become lower than those of α and δ. The other diagonal of the tetragon then develops into a two-phase region $\beta + \gamma$. The two three-phase regions $\alpha + \delta + \gamma$ and $\alpha + \delta + \beta$ are now no longer stable, but their place has been taken by $\beta + \gamma + \alpha$ and $\beta + \gamma + \delta$.

We have now dealt with a few of the chief basic types of isothermal cross-section.

Possible transitions of ternary three-phase equilibria

There are five different ways in which a ternary three-phase equilibrium can begin (or come to an end) under changing temperature:

A: it is present right from the beginning, i.e. at $T = 0$;

B: at one side: changing into a binary three-phase equilibrium (e.g. eutectic);

C: in a four-phase equilibrium; here 3 + 1 or 2 + 2 three-phase equilibria join;

D: at a "critical end-point";

E: at a *T* minimum or maximum (e.g. melting point minimum).

A closer analysis of this will enable us to discuss various types of diagrams not dealt with. By following the transitions of three-phase equilibria it is possible to arrive at convenient schemes for ternary systems: see Appendix I.

We need go no deeper into A, B and C. In case D a three-phase triangle comes to an end because two of the three vertices coincide, so that the triangle degenerates into a tie-line between a critical point and a single-phase region (see *fig. 12a* and *b*). Imagine a triple-tangent plane at the fold in fig. 5, and at another surface farther back (not shown). The critical point *R* is now not stable. We change *T*, causing only a quantitative change in the fold. If the other surface moves to a relatively higher position, then the two tangent points on the fold will approach one another and finally coincide at the critical point at a particular value of *T*. Upon further changes of temperature the entire "edge" of the fold becomes stable, and both two-phase regions break away from each other.

Fig. 13 shows the companion diagram: here the critical point of the miscibility gap is stable longest.

In case E two three-phase triangles appear simultaneously from one tie-line. This case arises, for example, when during a temperature change, a new phase first becomes stable in a two-phase region, as a point on

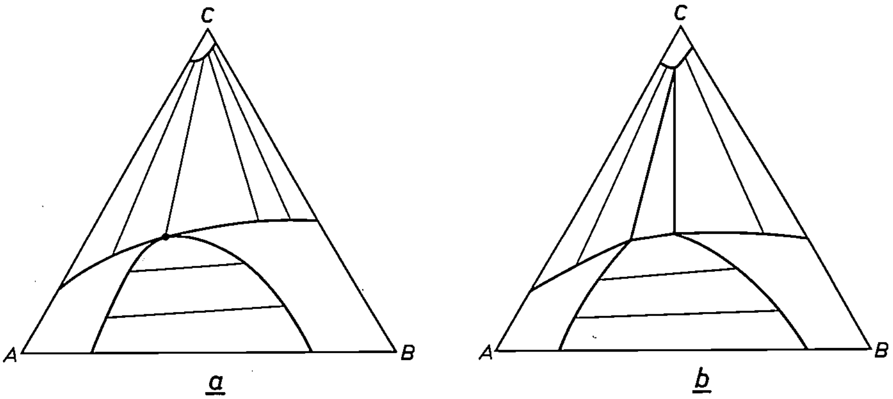


Fig. 12. Two miscibility gaps meeting at a critical point. The three-phase triangle in *b* is formed at a slightly higher temperature from a tie-line in *a*.

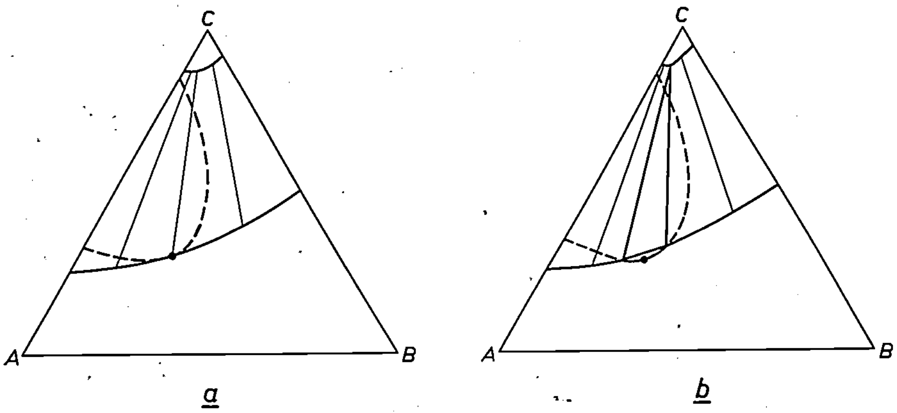


Fig. 13. Companion diagram to fig. 12 with metastable miscibility gap.

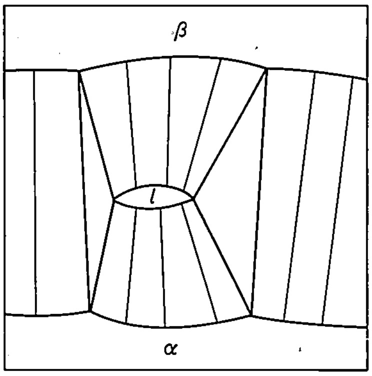


Fig. 14. Two three-phase triangles $\alpha + \beta + l$ are formed upon a temperature change from a tie-line $\alpha + \beta$, on which a liquid point has appeared at a particular temperature.

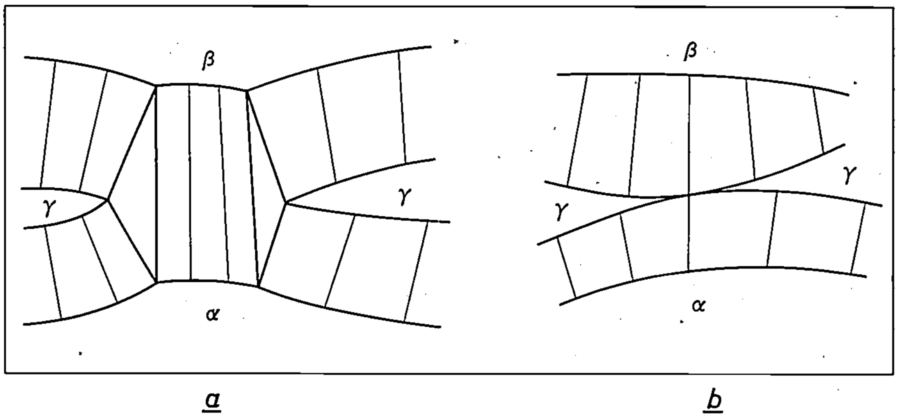


Fig. 15. Companion diagrams to fig. 14. Two three-phase triangles (*a*) arise from two tie-lines (*b*) each situated in the extension of the other.

a tie-line. Shortly afterwards we then have the situation of *fig. 14*.

The companion diagram to fig. 14 is *fig. 15*. Here the two triangles $\alpha + \beta + \gamma$ have averted their flat tops from each other. Fig. 15*b* represents a cross-section at the temperature where the two triangles disappear, in the last tie-line $\alpha + \beta$. If T is changed slightly further, the two-phase regions $\alpha + \gamma$ and $\beta + \gamma$ break away from each other.

It should further be noted that *binary two-phase equilibria* can also begin in five ways upon a change in temperature, comparable with the above-mentioned cases A to E:

A′: at $T = 0$;

B′: in a unary two-phase equilibrium (melting point, transition point);

C′: in a three-phase equilibrium, here 2 + 1 two-phase equilibria join;

D′: in a critical point;

E′: in a T minimum or maximum.

The Zr-Ta system represented in *fig. 16* shows all five possibilities.

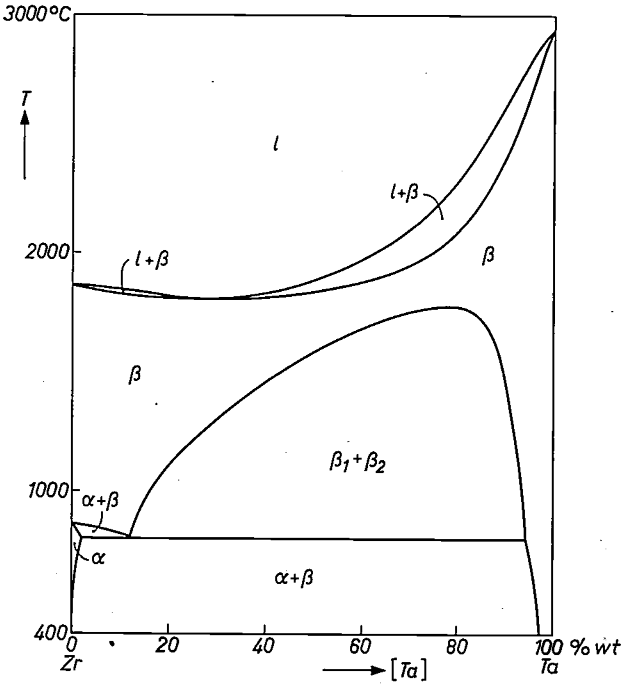


Fig. 16. Phase-diagram of the system Zr-Ta.

The vertical cross-section

Fig. 17 shows a vertical cross-section of a system of the same type as that for which the isothermal cross-sections were given in figs. 3, 11*a* and *b*. This cross-section of the triangular T-x-y prism is a centre plane perpendicular to the line AC. The heterogeneous structure of a mixture with for example 60 at.% B, 20 at.% A and 20 at.% C (see dashed line) can now be read off as a function of temperature; with decreasing temperature the mixture consists of l, $l + \beta$, $l + \beta + \gamma$, $l + \alpha + \beta + \gamma$ (at the four-phase temperature) and $\alpha + \beta + \gamma$, respectively.

We *cannot* read from this diagram, as we can from a T-x diagram of a binary system, the composition of the co-existing phases, nor therefore can we find the quantity of each phase.

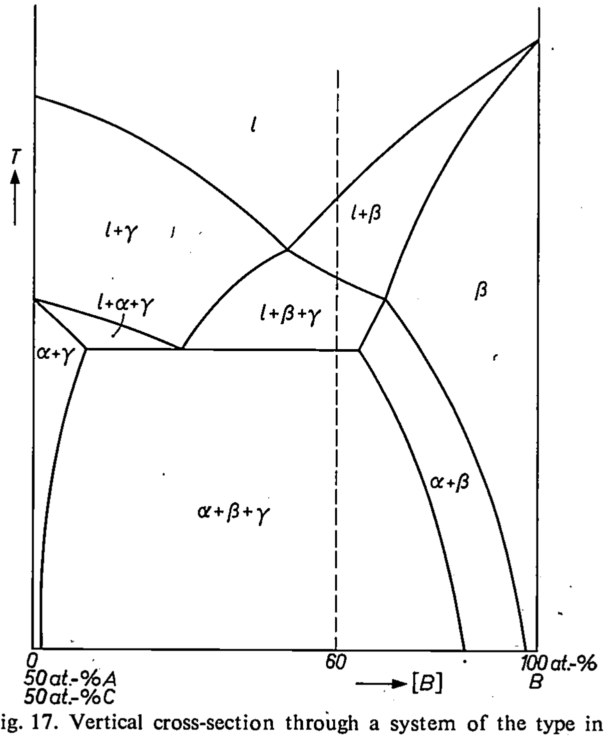


Fig. 17. Vertical cross-section through a system of the type in fig. 11*a* and *b*, following the centre plane perpendicular to *AC*.

This becomes clear if we remember that the cross-section generally cuts the tie-lines at an angle. The compositions of the co-existing phases lie, as it were, on opposite sides of the plane of the paper, and they vary with the mole fraction B in a manner unknown to us.

The ternary three-dimensional diagram

Fig. 18 shows a three-dimensional T-x-y diagram with a ternary eutectic. The four-phase equilibrium is determined by *four points* related to each other, which are the concentrations K, L, M and N of the phases α, l, β and γ respectively. A three-phase equilibrium is determined by *three* (identically coloured) *lines* related to each other, a two-phase equilibrium by *two* related *surfaces*, to be thought of as connected by tie-lines, and a single-phase region is given by a *three-dimensional body*. The surfaces, lines and points in question are boundary surfaces, edges and vertices of the single-phase bodies. It should be borne in mind, however, that "two related surfaces" may sometimes join smoothly via a critical line (see the isothermal cross-section in fig. 6). Also two of three "related lines" may join smoothly (see page 218, case D).

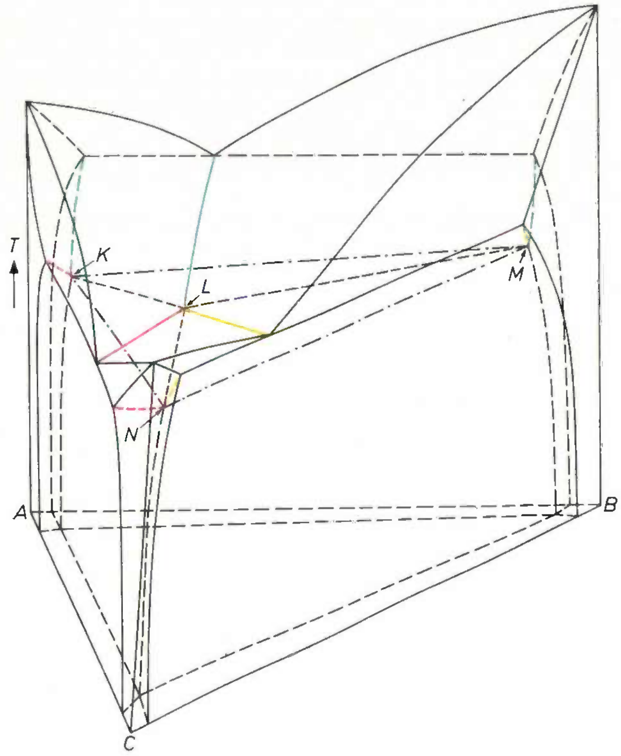


Fig. 18. T-x-y diagram with ternary eutectic. The points K, L, M and N indicate the concentrations of the four phases in equilibrium with each other at the eutectic temperature. The concentrations of the possible three-phase equilibria with liquid are found on *lines* (yellow, red and green), the concentrations of the possible two-phase equilibria form *surfaces*, and the concentrations of the homogeneous phases form *three-dimensional bodies*.

Diagrams in three dimensions (or more!) have their practical drawbacks, and the use of projections will be preferred. Although, for example, in the binary four-phase equilibrium mentioned on page 213 the four phases in the system acetronitril-water have different compositions, p and T are necessarily the same. A projection on the p-T plane thus gives a four-phase point at which four three-phase lines meet. Similarly, a ternary five-phase equilibrium, which, as discussed below, is possible if the pressure is not kept constant — becomes a point in the p-T projection, where five four-phase lines meet. This point may be compared with the well-known triple point in a unary system.

The phase rule

From the previous section it is evident that in a ternary system under constant pressure the sum of the number of degrees of freedom F and the number of phases P is 4, i.e. $F + P = 4$. Without the restriction of a constant pressure, however, $F + P = 5$, because the surfaces, lines and points mentioned in that section are displaced if we change the pressure, for which fig. 18 holds. Let the number of components of a system be C, then we have the well-known phase rule:

$$F + P = C + 2. \qquad (1)$$

This is a useful rule, but in our opinion it is wrong to regard it as the fundamental law of phase theory. There are, after all, exceptions to it, as will appear from the following considerations.

If we roll a double-tangent plane in the manner described on two G-x-y surfaces (at constant p and T) then at a given moment it can become tangent to a third surface, but it would be a very great coincidence if it became tangent to a fourth one *simultaneously*: as discussed for example in connection with fig. 11, this happens only at one specific T (or p). If, however, we have mirror-image isomers as components, a system can be rigorously symmetrical, and in that case a ternary four-phase equilibrium may well occur in a *finite* p-T region. A simple example where the phase rule does not apply is the occurrence of the two-phase region left-quartz + right-quartz in the unary system SiO_2.

The use of equation (1) for calculating the number of degrees of freedom F sometimes presents difficulties if one is not sure of the number of components C. Whether the system acetaldehyde-paraldehyde-water behaves as a ternary or a binary system [3] is exactly what must be found from the number of degrees of freedom determined by experiment. If the number is three in a two-phase region, we then conclude that $C = 3$; if, however, we find $F = 2$, then $C = 2$. After substituting for C in equation (1) we can then calculate F for various values of P. In fact we have then not used (1) but:

$$F + P = \text{constant}. \qquad (2)$$

In part I it was stated on page 26 that if transformations between isomers behave "ambiguously" — too

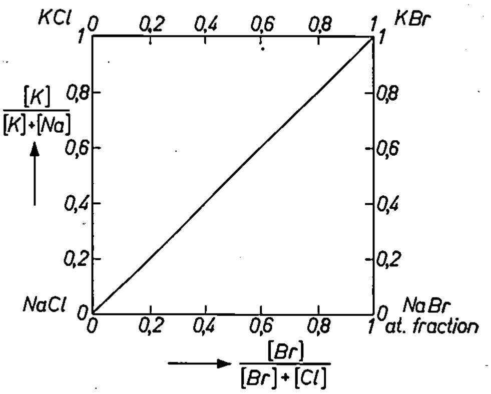


Fig. 19. Isothermal cross-section of the system NaCl-KCl-NaBr-KBr in the form of a square. The system is ternary since it can be described by two independent concentration variables. Its ternary nature may also be deduced from the fact that all compositions of the system can be obtained by mixing three components. Any three of the four compounds considered can be chosen for this purpose. Half of the system must then, however, be described with the aid of *negative* concentrations.

slow for unary behaviour and too fast for binary—, it is no longer readily possible to apply thermodynamic principles and phase theories in the normal way. The effective number of components may then be influenced by the time available, the presence of impurities acting as catalysts, and the temperature. For example, the acetaldehyde-paraldehyde-water system mentioned above might show ternary behaviour at low temperature and behaviour more of a binary nature at high temperature.

Irrespective of kinetics, C may in some cases be doubtful in connection with deviations from the stoicheiometric composition (these may be difficult to demonstrate). These difficulties, and also the fact that pressure and/or temperature are often kept constant, make it advisable to use (2) instead of (1).

To illustrate the difficulty of establishing the value of C, let us consider the systems NaCl-NaBr-NaI and NaCl-NaBr-KBr-KCl mentioned in the introduction. These ternary systems can be thought of as cross-sections of quaternary systems, Na-Cl_2-Br_2-I_2 and Na-K-Cl_2-Br_2 respectively. In these systems we have three concentration variables and — at constant p and T — we can plot these along the edges of a tetrahedron. Because of the requirement that the sum of the metal atoms must be equal to the sum of the halogen atoms, the tetrahedron in the first example is cut to form an equilateral triangle, and in the other case to form a square (*fig. 19*). The fact that the cross-section in the one case is a triangle and in the second case a square is not relevant to the question of whether or not the system concerned is a ternary system. In both cases there are two independent variables, in the first case plotted at an angle of 60°, in the second case at an angle of 90°. What decides whether the system is ternary or not is whether each tie-line of the system really lies in the cross-section. This means that the requirement that the number of metal atoms and halogen atoms should be equal applies each to phase individually and not only to the heterogeneous mixture as a whole. In our two halide systems this will no doubt in fact be the case. In many oxides and sulphides, however, stoicheiometric deviations occur that are sometimes difficult to detect, as in NiO, and sometimes very pronounced, as in wüstite (e.g. $Fe_{0,9}O$). These deviations are connected with the presence of trivalent Ni and Fe alongside the bivalent metals.

Verification of ternary diagrams; Schreinemakers' rule

The verification of a ternary diagram is no simple matter. The diagram only offers checking points for quick verification at few places. In isothermal cross-sections the points to note are the vertices of three-phase triangles, and in particular the way in which the boundary lines of the homogeneous zones meet there (see for example the vertex U in fig. 3).

The first rule is that the angle enclosed by two boundary lines must never be greater than 180°. This requirement — which is completely analogous with the characteristic kink at a binary three-phase point, mentioned on page 19 in part I — is not difficult to understand. A two-phase equilibrium $\alpha + \beta$ is metastable with respect to the formation of γ in the region where the double-tangent plane at the G-x-y surfaces of α and β cuts that of γ. The tangent plane at α has to be rolled back from such a position in order to become tangent to γ. If we follow a straight line in the homogeneous α region, beginning at the vertex, we thus meet the stable boundary α-($\alpha + \gamma$) — where contact is made with γ — earlier than the extension of the boundary line α-($\alpha + \beta$) — where γ is cut. The "forbidden" situation — first a meeting with the metastable part of the boundary line α-($\alpha + \beta$) and only then with the stable boundary α-($\alpha + \gamma$) — would arise if the angle enclosed by the two *stable* boundary lines were greater than 180°. This is the required proof that the angle in question must be smaller than 180°.

The region entered by the extension of the α-($\alpha + \beta$) boundary may be either the two-phase region ($\alpha + \gamma$) or the three-phase triangle ($\alpha + \beta + \gamma$). Now there is a *second* general rule which states that the metastable extensions of α-($\alpha + \beta$) and α-($\alpha + \gamma$) either enter the three-phase triangle both together or not at all. This is the well-known *Schreinemakers' rule*.

Schreinemakers' rule is often sinned against when working out diagrams in which a number of measured points are known, or when drawing hypothetical diagrams, even — though by implication — in otherwise sound text books. Often, for example, one finds a figure like that shown in *fig. 20a*, representing two two-phase regions meeting at a critical point. Schreinemakers has

[3] See page 26 in I.

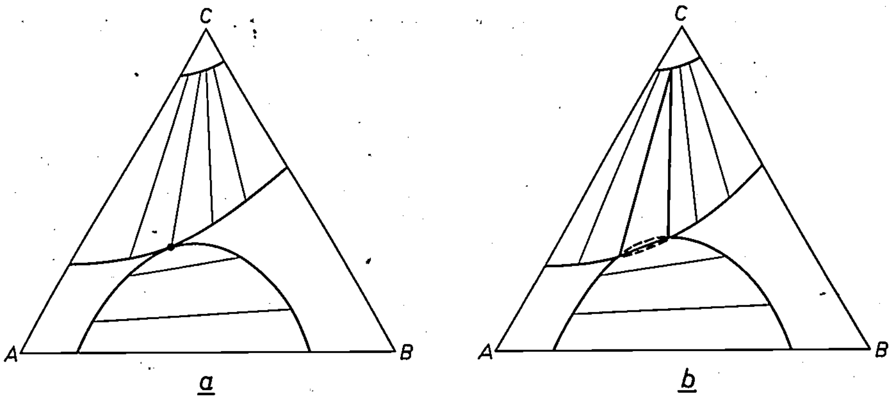


Fig. 20. *a*) Two two-phase regions meeting at a critical point. The tangential boundaries of the two-phase regions are shown here with the wrong curvature with respect to each other, as appears from *b*), the same system at a somewhat higher temperature. The latter does not obey Schreinemakers' rule. The correct diagrams can be seen in fig. 12*a* and *b*.

already pointed out that in such a case the curves meeting at the critical point must both be curved in the same direction [4]. This can best be understood as follows. By changing the temperature we can cause the curves to intersect in the isothermal cross-section. The critical point then becomes metastable and a narrow three-phase triangle is formed (fig. 20*b*). This, however, does not obey Schreinemakers' rule. The correct diagrams can be seen in fig. 12*a* and *b*.

Derivation of Schreinemakers' rule

It will be instructive to give a derivation here of Schreinemakers' rule, if only because this is seldom to be found in text-books.

In *fig. 21*, which shows part of an isothermal cross-section, the lines through P and R and through Q and S are the boundary lines of a two-phase region $\alpha + \beta$. The tie-lines PQ and RS are interrupted in the figure; PR and QS should be imagined to be infinitely small compared with PQ. Therefore PQ and RS must be considered to be parallel. In the plane of the paper we imagine there to be a double-tangent plane which is tangent at P to the "pocket" of the G-x-y plane of α, and at Q to that of β. We then imagine there to be a second double-tangent plane, at a small angle, which is tangent to the pockets at R and S. Parallel to the first double-tangent plane we can construct a plane which cuts the pockets of the α and β surfaces, forming ellipses [5] with P and Q as the centres and cutting the second double-tangent plane at the tie-line RS. The latter is thus *tangent* to both ellipses.

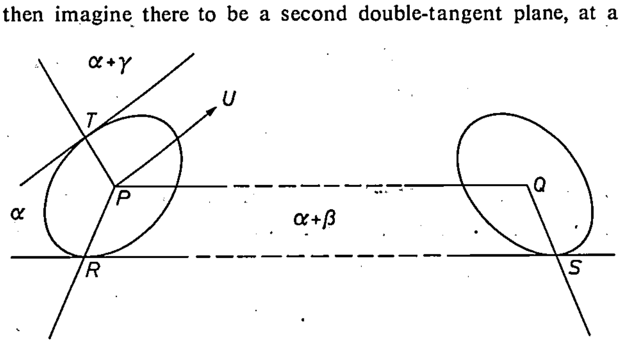


Fig. 21. Figure used for the derivation of Schreinemakers' rule.

Now let PQ be one side of a three-phase triangle PQU. From P there now runs a further line that marks the boundary α-(α+γ), the line through T. The tangent to the ellipse at T is parallel to the side PU of the three-phase triangle, as similarly the tangent to the ellipse at R is parallel to PQ. Schreinemakers' rule is thus reduced to a relation that must apply between the directions of two arbitrary lines tangent to an ellipse and the directions of the lines connecting the tangent points with the centre point of the ellipse.

The proof by analytical geometry of the proposition in this form is not difficult. Here we shall make it even easier. There is nothing to prevent us from taking the x and the y axis of the concentration variables in the direction of the axes of the ellipse and from choosing the scales along the axes in such a way that the ellipse becomes a circle. The two boundaries of the α region then terminate at right angles to the appropriate sides of the phase triangle. If the α angle of the three-phase triangle is obtuse, then both boundaries when extended enter the triangle: if, however, the angle is acute, both boundaries pass in front of the triangle. If we transform the circle back into the original ellipse, then the angles change but not the sequence of the four directions. It will also be evident that the validity of Schreinemakers' rule is independent of the use of percentages by weight or mole fractions, or of rectangular co-ordinates or co-ordinates at an angle of 60°.

The regular approximation

If we had analytical expressions for the Gibbs' free energy G of the various phases it would be possible to calculate a ternary phase diagram. Such expressions can only be approximations, and the mathematical work involved is usually prohibitive. Nevertheless, something can be achieved in this direction. An interesting problem, and one of considerable practical importance, is the following: in how far can the broad outlines of an unknown ternary system be predicted on the basis of known data of the binary boundary systems?

We shall confine ourselves first to separation into phases of similar structure, for instance to miscibility gaps in the liquid phase. We then have only one G-x-y surface at each temperature. The simplest approach is what is termed the *regular approximation*:

$$\begin{aligned} G &= H - TS \\ &= axy + bxz + cyz + RT(x \ln x + y \ln y + z \ln z), \qquad (3) \end{aligned}$$

where x, y and z ($= 1 - x - y$) are the mole fractions of the three components. For the binary boundary systems — obtained, for example, by putting $z = 0$ — we arrive at eq. (1) from part II:

$$G = ax(1-x) + RT\{x \ln x + (1-x) \ln (1-x)\} \,. \quad (4)$$

If a is positive, this approximation leads to a symmetric miscibility gap given by eq. (2) in part II:

$$T = \frac{a(1-2x)}{R \ln \dfrac{1-x}{x}} \,. \quad . \; . \; . \; . \quad (5)$$

The regular approximation (eq. 3) is thus a radically formalized expression, and can only be used semi-quantitatively. What one really does is to take parameters a, b and c for the energetic interaction of the three kinds of molecular or atomic pairs from the binary systems and to substitute them in the ternary equation. To make the treatment more general, one would have to characterize each binary phase by at least *two* parameters, which amounts to introducing an enthalpy of mixing of the form:

$$H = px^2y + qxy^2. \quad . \; . \; . \; . \quad (6)$$

For $p = 3a$ and $q = 2a$, and $y = 1 - x$, we obtain eq. (3) in part II, which was used in that part for obtaining the asymmetric miscibility gap in fig. 4. Equation (6) means in fact that energetic interactions between groups of *three* atoms each, but still atoms of only *two* kinds, are added together. In the ternary system there must therefore logically be a term in xyz; the appropriate parameter cannot of course be taken from the binary boundary systems. For our purpose — that of making predictions for the ternary system — we shall therefore not get far with a refined application of the equations for the boundary systems.

The regular approximation is found to answer this purpose reasonably well. By varying the values a/RT, b/RT and c/RT it is possible to calculate ternary diagrams of all kinds, and in many cases qualitative agreement can be found with a diagram established from experimental data. We shall deal with this only in broad outline and refer the reader to the literature for a more detailed treatment [6].

A systematic survey of the diagrams for the various combinations of a, b and c parameters — particulars of which are also given in the article quoted — can be obtained more easily by studying the isothermal spinodal curves. For the asymmetric binary system mentioned above the spinodal equation (page 54 of part II) is also much simpler than eq. (6) and (7) of part II, which *together* define the miscibility gap. As in fig. 4 of part II, the spinodal equation in ternary systems indicates the temperature range within which the miscibility gap occurs, and critical points can also be derived from it.

Miscibility gaps at the boundary

If $RT < \frac{1}{2}a$, a miscibility gap adjoins the X-Y boundary, because immiscibility is then found in the binary boundary system; see fig. 1 in part II. This miscibility gap may simply close at a critical point (cf. fig. 6), but the regular approximation is also capable of yielding the more complicated form shown in fig. 8. It is, of course, possible for immiscibility to occur in two or in all three binary boundary systems, namely if RT is smaller than $\frac{1}{2}b$ and/or $\frac{1}{2}c$. The strip-shaped miscibility gap (fig. 7*a*) and the type in fig. 3 are also calculated by the regular approximation, and many other types of diagram not mentioned. In this way, also, a discovery is made which is by no means obvious, namely that a four-phase equilibrium occurs when a, b and c are positive and approximately equal. This four-phase equilibrium occurs of course — if one thinks of the phase rule — at one very specific temperature [7]; at a higher temperature one has a diagram of the type in fig. 11*a*, with a closed homogeneous zone inside. Upon a further increase of temperature this flows into the three homogeneous regions in the corners, forming three separate miscibility gaps as shown in fig. 8, with a total of six critical points. A system of this kind has not yet been found experimentally.

Closed miscibility gaps

Apart from miscibility gaps at a binary boundary, various examples of closed miscibility gaps are known, nearly all of them in the liquid phase. Taking eq. (3) as our starting point, we find that in order for this to occur at least one of the interaction parameters must be negative, and more strongly negative than the difference between the other two. Taking the parameters by definition $a \geqq b \geqq c$, then:

$$c < (b - a). \quad . \; . \; . \; . \; . \; . \; . \quad (7)$$

If, however, all three parameters are negative, this is not sufficient and the expression must then be:

$$\sqrt{-c} > \sqrt{-b} + \sqrt{-a} \quad . \; . \; . \; . \; . \quad (8)$$

[4] H. W. Bakhuis Roozeboom, Die heterogenen Gleichgewichte, Brunswick 1913, part III, 2, page 119.

[5] The directions and relative lengths of the axes of the "indicatrix-ellipse" of α are determined by the principal radii of curvature of the G surface of α in P. The third derivatives of G with respect to x and y enter into the picture only when a larger part of the surface is cut off and the section is then no longer an ellipse.

[6] J. L. Meijering, Philips Res. Repts. 5, 333, 1950 and 6, 183, 1951.

[7] At this four-phase temperature one plane is tangent at four places to one and the same G-x-y surface, unlike the situation with a ternary eutectic.

for a system with a closed miscibility gap. At first sight it seems rather odd that immiscibility should arise when all binary combinations are readily miscible.

To make it clearer, let us take as an example a simple symmetrical combination: $a = b = -1$ kcal, $c = -9$ kcal. The enthalpy is $H = -xy - xz - 9yz$. Let us consider a section through the X-vertex and the middle of the Y- Z side. In this section $y = z = \frac{1}{2}(1-x)$, which, substituted in the above equation, gives: $H = \frac{1}{4}(-9 + 14x - 5x^2)$. This expression, although everywhere negative (except for $x = 1$), represents a parabola which lies entirely above the straight line through ($x = 1$, $H = 0$) and ($x = 0$, $H = -9/4$). At a sufficiently low temperature one may therefore expect separation into a high-X and a low-X mixture. It is precisely when Y and Z are so readily miscible ($c = -9$ kcal compared with $a = b = -1$ kcal), that the dilution of their mixture by the addition of X is energetically unfavourable, provided at least the X-Y and X-Z interactions do not overcompensate this factor.

The classical example of a system with a closed miscibility gap is water-acetone-phenol, and for metallic liquids it is Bi-Cu-Sb [8]. In the first system acetone-phenol is the boundary system with the strong negative interaction, and in the second it is Cu-Sb.

A closed miscibility gap can often be regarded as consisting o two miscibility gaps at boundaries. This can be illustrated with the system Bi-Cu-Sb. Bound up with the strong negative interaction of Cu-Sb is the existence of the compound Cu_3Sb. The section Bi-Cu_3Sb goes through the middle of the miscibility gap and is virtually binary. The ternary system Bi-Cu-Sb can therefore be divided into two ternary systems Cu-Bi-Cu_3Sb and Sb-Bi-Cu_3Sb, each of which exhibits an ordinary miscibility gap along the side Bi-Cu_3Sb.

As an example of a successful prediction with the aid of the regular approximation, we mention the prediction of the closed miscibility gap in the liquid phase of the system Bi-Zn-Ag [9].

The only closed miscibility gaps so far found experimentally in *solid phases* [10] are those in Au-Ni-Cu and Cr-Mo-W [11]; the phases in question are face-centered cubic and body-centered cubic respectively. The boundary systems of Au-Ni-Cu are sufficiently known to enable us to use the regular approximation for investigating whether the closed miscibility gap can be understood. This is indeed the case. The strong negative interaction here is that between Cu and Au [6].

Further analysis of miscibility: the Bancroft and Timmermans rule

According to an old qualitative rule put forward by Bancroft and Timmermans [12], when a small amount of a third component Z is added to two components X and Y, the mutual miscibility of X and Y becomes *poorer* if there is a marked difference between the affinities of Z with respect to X and with respect to Y, but *better* if the third component shows no great preference.

Using the regular approximation one finds [6]:

$$\left(\frac{\mathrm{d}|x_1 - x_2|}{\mathrm{d}z}\right)_{z=0} \begin{cases} > 0 \\ < 0 \end{cases} \quad \text{according as} \quad |b - c| \begin{cases} > a \\ < a \end{cases} \quad . \; . \; . \quad (9)$$

Here x_1 and x_2 are the x co-ordinates of the co-existing phases. The parameter a must be positive, otherwise there would be no miscibility gap in X-Y. If the difference between the two other interaction parameters is greater than a, then according to (9) the width of the miscibility gap ($x_1 - x_2$) must become greater upon addition of Z and smaller in the opposite case. The Bancroft-Timmermans' rule is thus confirmed for regular systems and at the same time cast into a simple quantitative form.

Examples of the two cases are to be seen in fig. 6 and fig. 7*b*. It is interesting to examine what will be the effect of a temperature increase, the result of which is generally greater miscibility. In the first case a temperature increase may be expected to result in a smaller miscibility gap which disappears into the binary critical point of X-Y. In the second case one may expect that when the temperature rises the shrinking miscibility gap will break away from the X-Y side, and at a still higher temperature will finally disappear into a ternary critical point. In that case there is a closed miscibility gap between the binary and the ternary critical temperature, in accordance with (7).

Another result yielded by the regular approximation is that the isothermal miscibility gap diminishes most rapidly upon addition of Z if $b - c = 0$, that is to say if Z interacts as strongly with X as with Y. Notice here the symmetrical arrows "0" in *fig. 22*, which indicate the initital slopes of the miscibility gap. If we let $b - c$ increase, the two initial slopes diverge. The parallel arrows "1" indicate the initial slopes for the case where $b - c = a$.

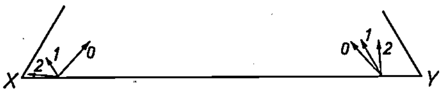


Fig. 22. Initial slopes of the boundaries of a miscibility gap calculated [6] for $(b - c)/a = 0$, 1 and 2, indicated by the arrows *0*, *1* and *2* respectively.

It is instructive to establish by a different process of reasoning that the mutual miscibility of two components, in the presence of a small amount of a third component, is greatest when the affinities with respect to the third component are equal ($b = c$), in other words that "asymmetry" with respect to the third component is not conductive to mutual miscibility.

Let us draw a cross-section through the *G-x-y* surface (3) at a constant — not too large [13] — value of z ($= z_0$) and then draw a double-tangent line to the resulting *G-x* curve. This means that we cause a homogeneous phase to separate into two phases on the subsidiary condition that z must not change. We act as if the Z atoms were immobile, and as if only the X and Y atoms can exchange places and so separate into two phases with different x and y but identical z. In this "special" miscibility gap the tie-lines are thus parallel to the *X-Y* boundary.

Substitution of $z = z_0$ in (3) gives:

$$G = axy + RT(x \ln x + y \ln y) + \\ + bz_0x + cz_0y + RTz_0 \ln z_0 \,. \quad . \quad . \quad (10)$$

Since $y = 1 - z_0 - x$, the terms that contain b and c are linear in x, and they therefore have no effect on the concentrations of the tangent points of the double tangent. The shape of the special ternary miscibility gap is therefore independent of the parameters b and c.

If we now let the Z atoms move freely, as well, then we must construct *planes* tangent to the *G-x-y* surface in order to see whether a homogeneous phase is stable. If we do this at a concentration lying on the boundary of the "special" miscibility gap, we know that the tangent plane — since it contains the double-tangent line on the cross-section at constant z — will meet the surface further on at a point. There are two possibilities: either the tangent plane is tangent to the *G-x-y* surface at that point, or it cuts it. In the first case the tangent plane is a double-tangent plane and the tie-line of the "true" miscibility gap is identical with the tie-line of the "special" miscibility gap, parallel to the *X-Y* axis. Each tangent plane considered will be a double-tangent plane if $b = c$, for then the ternary system is symmetrical. In the second (asymmetric) case the true miscibility gap is obviously greater than the special one, and therefore it is also greater than the symmetric true one.

An example is to be found in the face-centered cubic system Au-Pd-H, where the metal atoms are located at the lattice points, but the H atoms are interstitial. At 250 °C, for example, separation takes place into a low-hydrogen and a hydrogen-rich phase [14], but the Au/Pd ratio in each of the phases remains equal because at that temperature the Au and Pd atoms are virtually immobile. Here, then, we have an experimentally determined "special" miscibility gap, where one of the two concentration variables is fixed.

Some remarks on salting-out

This is scarcely the place to deal with calculations of ternary diagrams in which phases of different structure occur. It may be mentioned, however, that the regular approximation can be used here too for understanding or predicting a particular *type* of diagram, e.g. that of *fig. 23* [15]. The effect in which the solution of a salt (NH_4F) in water-alcohol mixtures (for example) gives rise to separation into two liquid phases is known as salting out. An analogous case is met with Cu-Ni-Cr in the solid state [16]. Here chromium fulfils a function similar to that of the salt in the previous example. It brings about a separation into the Cu-Ni mixed crystal phase — an effect that, but for thermodynamic considerations, would be unexpected. The strong negative interaction of Ni-Cr plays a leading role in this separation. The usefulness of the regular approximation is brought out very clearly in this case, as appears from a comparison of the diagram calculated by this method [17] in *fig. 24a* with the experimental diagram in fig. 24*b*.

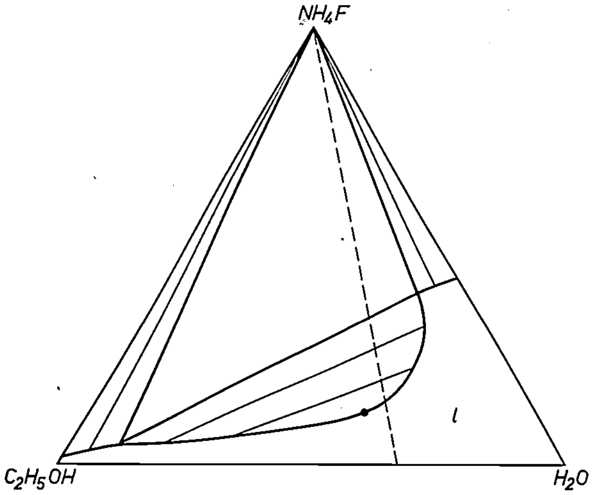


Fig. 23. Isothermal cross-section of the system H_2O-C_2H_5OH-NH_4F at 25 °C (concentrations in weight percentages). Following the dashed line upwards we see that NH_4F is initially simply soluble in a water-alcohol mixture, but at a certain concentration there is a separation into a high-water and a high-alcohol phase. This is the basis of the familiar method of separation by *salting-out*; the salt NH_4F expels the alcohol from the water-alcohol mixture.

[8] F. A. H. Schreinemakers, Z. phys. Chem. **33**, 83, 1900; N. Parravano, Gazz. chim. ital. **40**, 445, 1910.

[9] J. L. Meijering and H. K. Hardy, Acta metallurgica **4**, 249, 1956.

[10] The number is small because there are so many possible crystal structures.

[11] E. Raub and A. Engel, Z. Metallk. **38**, 11, 1947; N. V. Grum-Grzhimailo and D. I. Prokof'ev, Russ. J. inorg. Chem. **7**, 303, 1962.

[12] W. D. Bancroft, J. phys. Chem. **1**, 34, 1896/97 and **3**, 217, 1899; J. Timmermans, Z. phys. Chem. **58**, 129, 1907.

[13] This is in order to avoid the risk (if b and c are positive) of entering a three-phase triangle.

[14] A. Maeland and T. B. Flanagan, J. phys. Chem. **69**, 3575, 1965.

[15] A. K. Zhdanov and M. A. Sarkazov, Zhurnal fizicheskoi khimii **29**, 602, 1955.

[16] J. L. Meijering, G. W. Rathenau, M. G. van der Steeg and P. B. Braun, J. Inst. Metals **84**, 118, 1955/56.

[17] J. L. Meijering, Acta metallurgica **5**, 257, 1957.

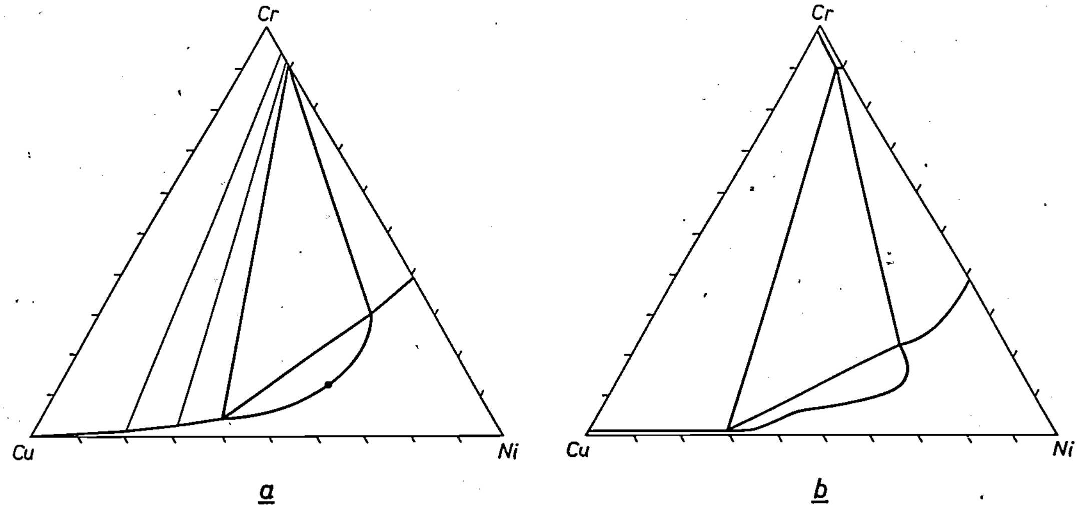


Fig. 24. Isothermal cross-section of the system Cu-Ni-Cr at 930 °C, *a*) calculated using the regular approximation, *b*) found by experiment.

Appendix I: Survey diagram of the three-phase equilibria in the system Al-In-Sb

Fig. 25 shows a survey diagram of the three-phase equilibria in Al-In-Sb [18]. We have chosen this as an example because it comprises all five ways (A to E, page 218) in which a ternary three-phase equilibrium can arise. *Fig. 26* and *fig. 27* show the binary boundary systems Al-In and In-Sb. The third, Al-Sb, is of the same type as In-Sb, but the intermetallic compound AlSb melts at a higher temperature (1065 °C) than Al (660 °C) and Sb. The section AlSb-InSb through the ternary system is also binary; it exhibits a eutectic at 523 °C from which two three-phase triangles AlSb + InSb + liquid arise when T is lowered.

The binary nature of AlSb-InSb means that there are no stable equilibria between phases that contain more and less than 50 at.% Sb respectively. It automatically implies that the ternary system can be split into two ternary systems: AlSb-InSb-Sb and Al-AlSb-InSb-In; see *fig. 28*. Similarly, the binary system In-Sb can be split into the systems In-InSb and Sb-InSb, and the compound InSb treated as a unary system.

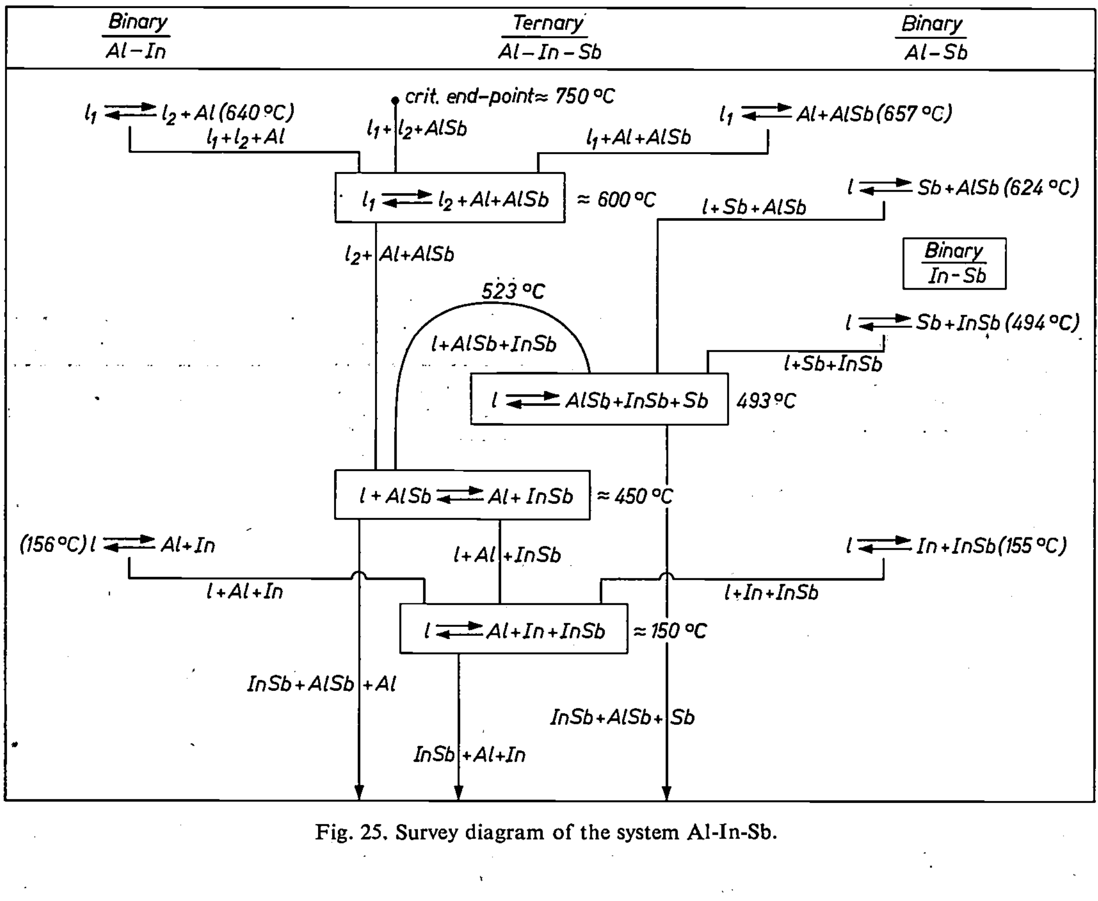


Fig. 25. Survey diagram of the system Al-In-Sb.

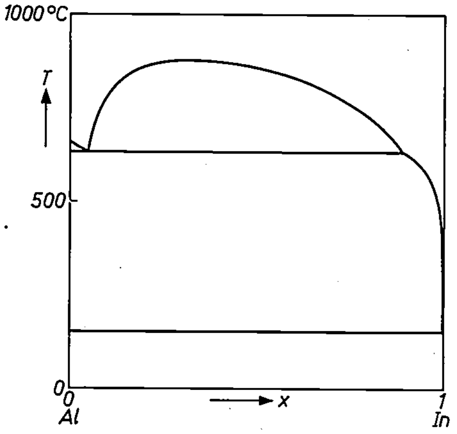


Fig. 26. The system Al-In.

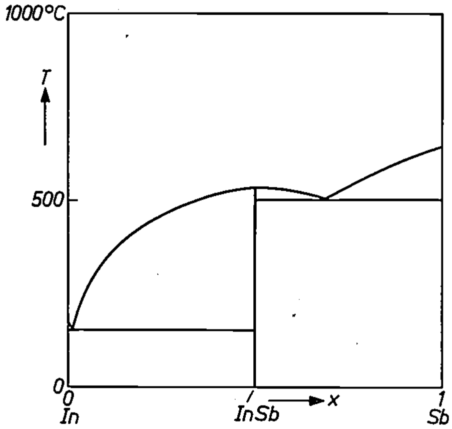


Fig. 27. The system In-Sb. An intermetallic compound InSb occurs.

The ternary system last mentioned may of course also be represented aš a square instead of a trapezium. The diagonal InSb-Al is approximately binary as well, so that the large triangle Al-In-Sb can be divided into three smaller ternary systems.

Whereas, in considering the system Al-In-Sb, we see that the formation of two three-phase triangles AlSb + InSb + liquid at 523 °C (fig. 25) corresponds to case E, it corresponds to twice case B if we divide Al-In-Sb into two ternary systems along AlSb-InSb.

Appendix II: Some remarks on sintering and precipitation hardening

To conclude this article, we.shall return briefly to a subject, touched upon in the introduction of I, which is of considerable practical importance. Hitherto, we have consistently disregarded the part played by surface free energy. It is due to this free energy that a material is less stable in a fine state of division than in a coarser one. The surface free energy supplies the driving force in the *sintering* of powders. The grain boundaries attempt to minimize their surface area as much as possible.

The surface of a phase boundary also of course has its free energy, which attempts to be as low as possible. This means that here also a fine precipitate is less stable than a coarse precipitate of the same composition. Between a homogeneous mixed crystal and a coarse heterogeneous state having the same overall concentration one can imagine there to be a series of *intermediate states*, and indeed these can often be achieved by treating (ageing) a supersaturated mixed crystal for suitable periods at appropriate temperatures. The *precipitation hardening* obtained in this way [19] is of great practical value for such uses as the making of alloys of high mechanical strength and permanent magnets.

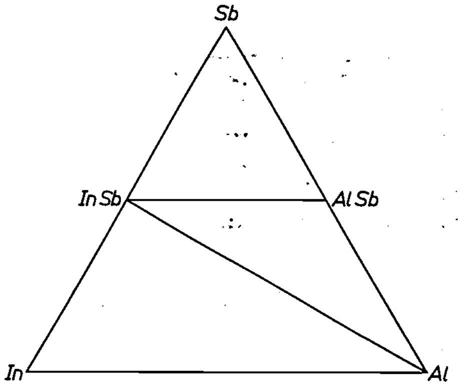


Fig. 28. A possible division of the system Al-In-Sb into three ternary systems (at low temperature).

[18] W. Köster and B. Thoma, Z. Metallk. **46**, 293, 1955.

[19] See e.g. J. L. Meijering, Philips tech. Rev. **14**, 203, 1952/53.

Summary. Ternary systems are discussed in this last article of a series of three on phase theory. The difficulties involved in the representation of three-dimensional diagrams lead to the use of cross-sections (isothermal and vertical cross-sections), with projections and two-dimensional survey diagrams. The sometimes confusing concept of component is considered in some detail and in the same connection the phase rule is subjected to a critical examination. Considerable attention is paid to a more or less quantitative approach to ternary systems, referred to as the regular approximation. With this method ternary systems can be fairly accurately predicted on the basis of data from binary boundary systems.

X-ray pictures in colour

W. J. Oosterkamp, A. P. M. van 't Hof and W. J. L. Scheren

616-073.75:621.397.132

The introduction of television techniques in X-ray diagnostics has opened up all kinds of possibilities that have proved to be of great practical value [1]. One that has recently been put to use is the *subtraction* of X-ray images [2].

The principle of the method of image subtraction has long been known: it makes it possible to eliminate certain dominant bone structures, especially bone shadows, if these are not of interest in the examination of a radiograph. This is particularly important in brain examinations where the petrous bone and the eye sockets can often be very troublesome in radiographs of the skull. In one of the methods of brain examination a contrast medium is injected into one of the arteries that supply the brain. The contrast medium is circulated with the blood stream, first filling the arterial vessels during the arterial phase, then entering the brain capillaries, which are so fine that they do not appear separately in the radiograph, and finally being carried away from the brain through the veins (venous phase). This circulation lasts from about 10 to 15 seconds, during which time a series of radiographs is taken. If a radiograph has been taken immediately prior to the injection, and measures have been taken to keep the patient's head immobile during the whole series of exposures, it is then possible to subtract this "empty" X-ray radiograph from each of the circulation radiographs, thus obtaining a very distinct picture of the vascular system in the relevant phase of the circulation.

Originally the subtraction of two radiographs was performed by photographic means, that is to say by copying the negative of one radiograph on top of a transparent positive print of the second [3]. This is a rather cumbersome procedure and great care is required; its application has therefore been limited.

The subtraction of two images can be performed far more easily with a closed circuit television system, as the video signal voltages for two images can at will be either added together (as in picture-mixing in television studios) or subtracted from each other. The following procedure is used for viewing radiographs by this method [2]. One radiograph, in this case the "empty" one in our skull series (*fig. 1a*) is placed in front of a viewing box and the television camera gives a sharp picture of it on the monitor screen; the video signal of this picture is then recorded in a magnetic store [4]. The next radiograph (fig. 1*b*) is then placed in the viewing box and observed with the television camera, but this time the signal held in the store is subtracted from the video signal from the camera before the picture is passed to the monitor. The positive and the negative image on the screen can quite easily be brought into correct register by shifting the second radiograph in the viewing box by hand. The resultant subtraction image on the monitor can also be photographed for documentation (fig. 1*c*).

We have now extended this subtraction method by the application of *colour television*, which adds as it were another dimension to the image information. In this application we record several radiographs of a series one after another on different tracks of the magnetic store. The signals of two or three of the tracks are then read out simultaneously and — after appropriate subtraction and mixing — are fed into the three colour channels of a colour monitor. In this way, in one and the same picture, we can represent the arterio-vascular system, visible in the first radiographs of the skull series, in a particular colour, say red, and the venous system from one of the later radiographs in a contrasting colour such as blue. The procedure is illustrated schematically in *fig. 2*. With the system shown in fig. 2

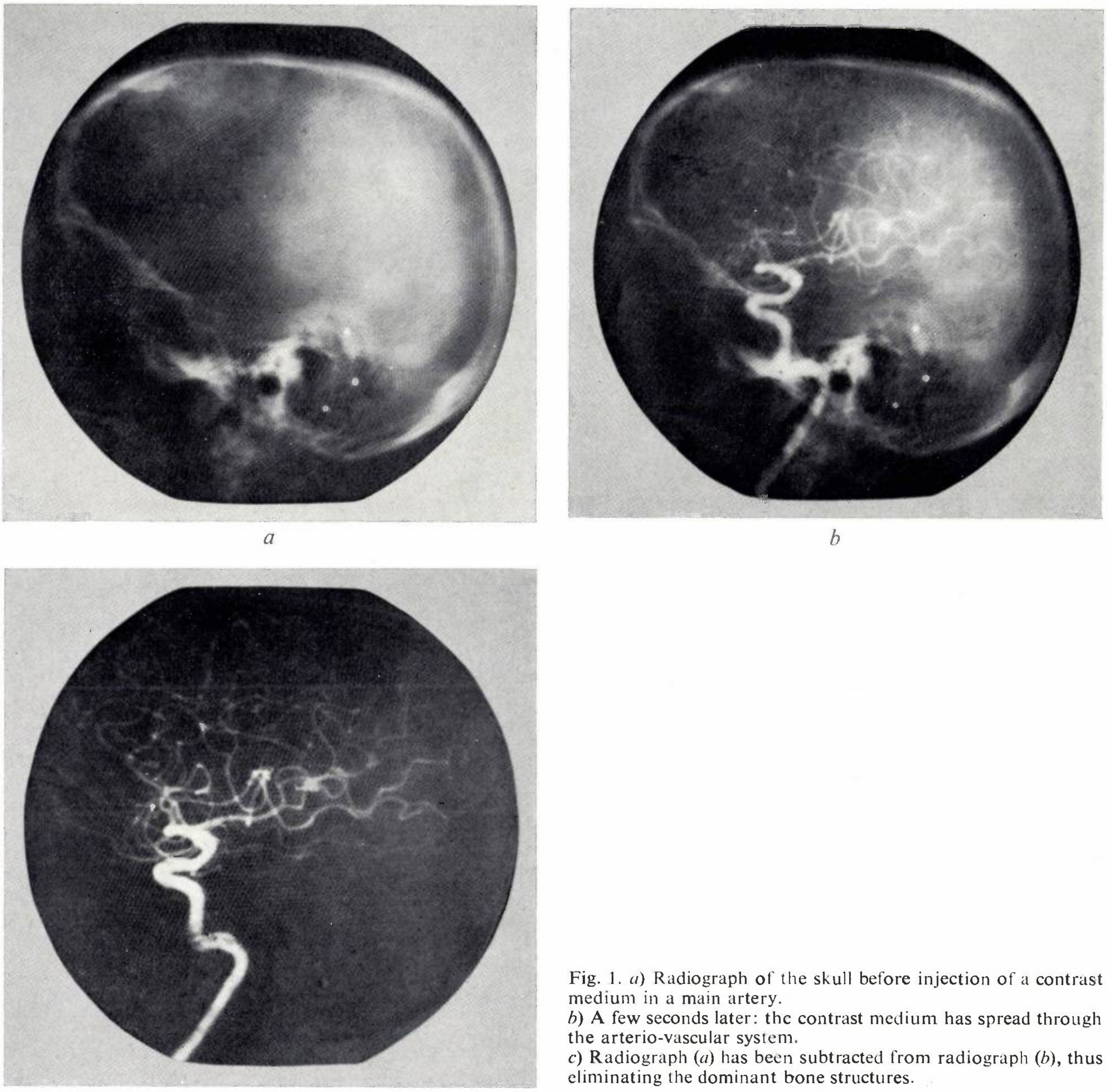
Prof. Dr. Ir. W. J. Oosterkamp, A. P. M. van 't Hof, and W. J. L. Scheren are with Philips Research Laboratories, Eindhoven. Prof. Oosterkamp is also a Professor Extraordinary in Applied Nuclear Physics at the Technical University of Eindhoven.

[1] Of the numerous publications on this subject, we shall mention only:
R. Janker, Die praktische und wissenschaftliche Verwendung der elektronischen Bildverstärkung und des Röntgenfernsehens, Fortschr. Röntgenstr. **88**, 377-385, 1958.
J. Feddema and J. E. Marquerinck, X-ray television with special regard to a newly developed vidicon: the "Plumbicon", Medicamundi **10**, 2-9, 1964 (see also Medicamundi **10**, 21, 1964).

[2] G. J. van Hoytema, W. J. Oosterkamp, A. M. C. van den Broek and A. Druppers, La neuroradiologie avec soustraction utilisant la télévision et une mémoire d'image magnétique, Neurochirurgie, in the press.
W. J. Oosterkamp, Th. G. Schut and A. Druppers, Röntgenbeeldsubtractie met behulp van een magnetisch beeldgeheugen, Ned. T. Geneesk. **108**, 2051-2054, 1964.

[3] B. G. Ziedses des Plantes, Planigrafie en subtractie, thesis, Utrecht, 1934.
B. G. Ziedses des Plantes, Subtraktion, Thieme, Stuttgart 1961.

[4] J. H. Wessels, A magnetic wheel store for recording television signals, Philips tech. Rev. **22**, 1-10, 1960/61. A slightly different version of a video store has been described by: O. Schott, Der Folienspeicher für Röntgenfernseh-Diagnostik, Röntgenblätter **16**, 65-72, 1963.



a

b

c

Fig. 1. *a*) Radiograph of the skull before injection of a contrast medium in a main artery.
b) A few seconds later: the contrast medium has spread through the arterio-vascular system.
c) Radiograph (*a*) has been subtracted from radiograph (*b*), thus eliminating the dominant bone structures.

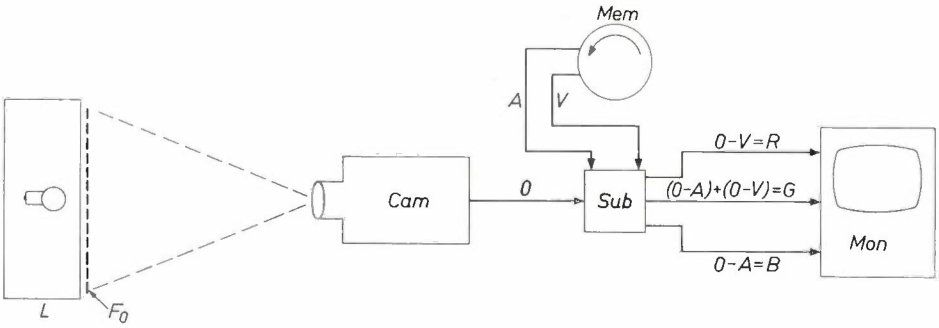


Fig. 2. System for image subtraction using closed-circuit colour television. The radiograph F_0 taken without contrast medium is placed in the viewing box L and viewed by a television camera *Cam*, after radiographs F_A and F_V, taken with the contrast medium in the arterial and in the venous systems, have been televised in the same way and recorded in the magnetic wheel store *Mem*. *Sub* is the subtraction circuit that supplies the colour difference signals R, G and B indicated in the figure which are applied to the red, green and blue channels of the colour monitor *Mon* respectively.

we obtain a colour picture with a light background (*fig. 3*) [5]. By using other combinations of signals one can also obtain a colour picture with a very dark background.

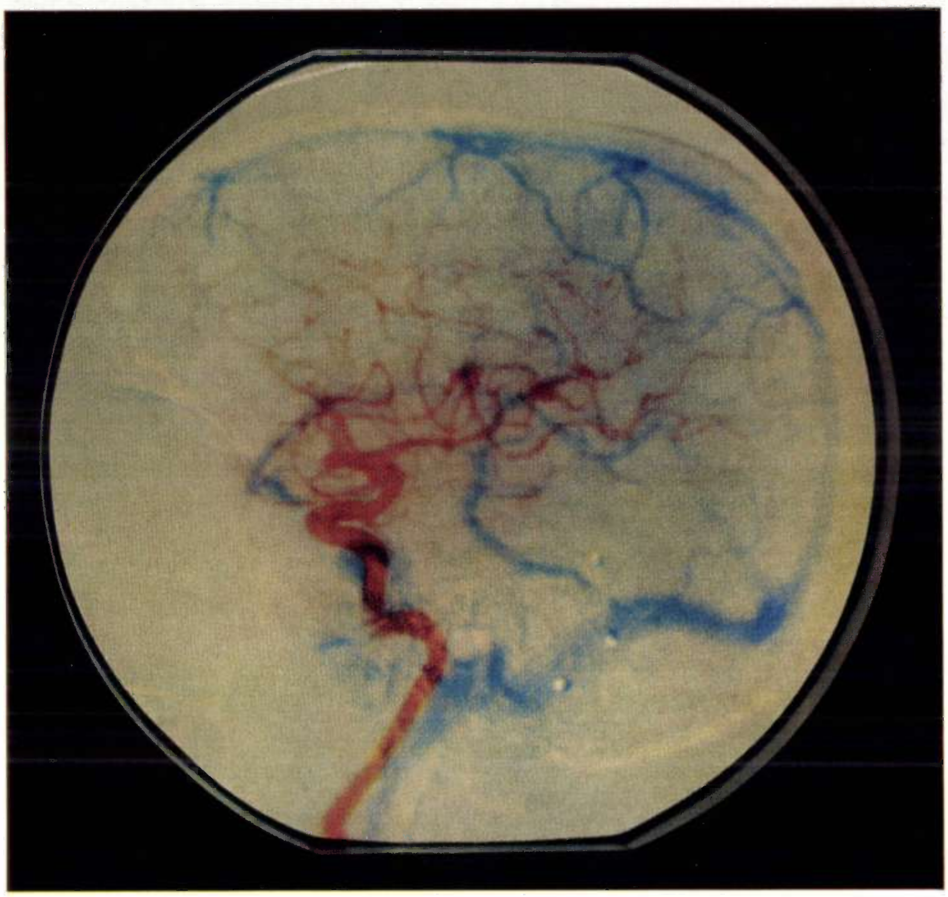
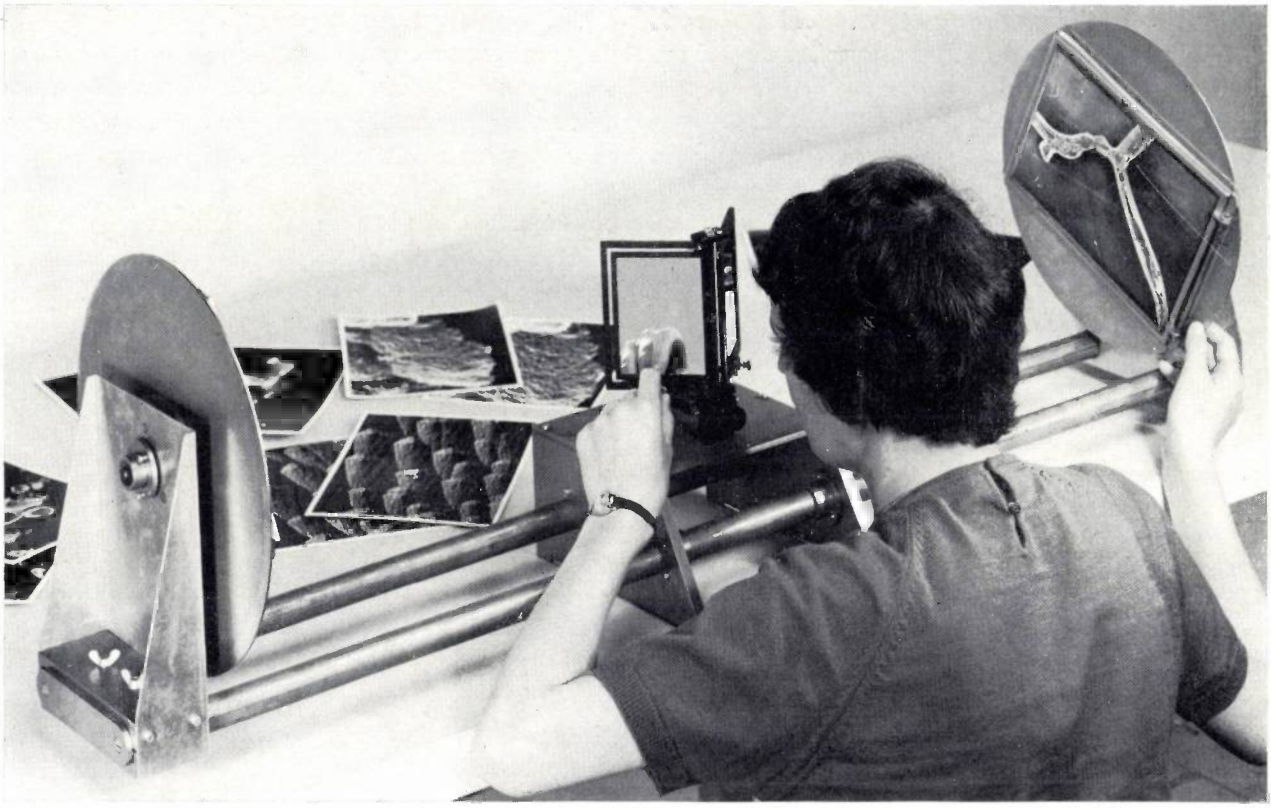


Fig. 3. The result of the image subtraction when the colour difference signals indicated in fig. 2 are applied to the colour monitor.

Apart from this application of X-ray colour television, where the difference between the individual radiographs is due to the movement of the contrast medium with the blood, many other applications are possible. Differences between images of the same object may also be due to the movement of the organ itself, to different angles of projection, to the use of X-rays of different hardness, to the successive introduction of contrast media in neighbouring organs, and so on. Colour display, which is relatively easy by the method described, will certainly be of instructional value, and may prove to be of practical interest through an increased facility of observation. The future will show to what extent this system will be able to assist the radiologist in diagnosis.

[5] The three radiographs used as starting material were placed at our disposal by Prof. B. G. Ziedses des Plantes of Amsterdam.



Stereophotography with the electron microscope

H. B. Haanstra

621.385.833.2:778.4

In order to be able to perceive "depth" when viewing a pair of stereophotographs taken with the electron microscope, the photographs have to be in a certain position relative to the viewer's eyes. A modified version of Wheatstone's stereoscope is discussed here which makes this readily possible. The great advantage which these stereophotographs have over two-dimensional pictures is demonstrated by a number of examples.

Introduction

When we look at an object we see it in three dimensions. We owe this ability to binocular vision, that is to say we observe the object with two eyes simultaneously. Each eye receives a picture which is a two-dimensional projection of the object. In this process parallax occurs — i.e. the eyes see the object from different angles — and as a result the two images are slightly different. The human brain is able to combine these two images to give a three-dimensional picture.

To perceive an object three-dimensionally it is not necessary to look at the object itself; the impression can also be obtained from two *photographs* that show the same images which the eyes would see if they were looking at the object directly. When viewing the photographs it is therefore necessary to ensure that each eye sees only the image intended for it, and various kinds of stereoscope have been devised for this.

In the most obvious method of making stereophotographs (thinking first purely of photography with visible light) the object is photographed by two cameras whose spacing is approximately the same as that between the eyes. This is illustrated schematically in *fig. 1*. L and R are the positions from which the photographs for the left and the right eye are taken. The object is indicated by the arrow PQ (for simplicity this is situ-

H. B. Haanstra is with Philips Research Laboratories, Eindhoven.

ated symmetrically with respect to L and R); $P'_lQ'_l$ and $P'_rQ'_r$ are the projections of the object that can be seen in the photographs. If the object is rigid and stationary it is obviously not necessary to take the two photographs simultaneously; one camera can then be used and displaced appropriately between the two exposures. In the following we shall be concerned only with stationary objects of this nature.

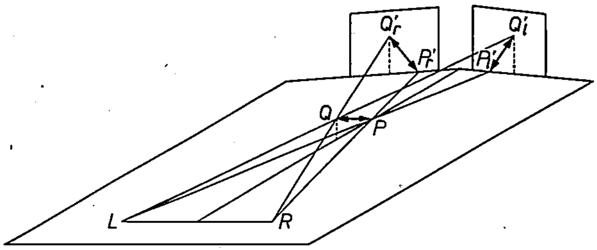


Fig. 1. Diagrammatic representation of a method of obtaining stereophotographs. The two photographs are taken from points L and R with two cameras, or with one camera which is displaced sideways between the exposures. $P'_lQ'_l$ and $P'_rQ'_r$ are the projections of the object PQ as seen on the photographs from L and R respectively.

There is another method of taking stereophotographs of such an object, again using only one camera. The parallax is obtained in this method by tilting the *object* through a small angle between the two exposures, while the camera remains fixed. *Fig. 2* shows the situation for this method. The camera is located at C. The dot-dashed line indicates the axis k about which the object is tilted. When the photographs are taken the object is situated at the positions P_lQ_l and P_rQ_r respectively; here again, $P'_lQ'_l$ and $P_r'Q_r'$ are the images on the two photographs, and k' is the image of the axis of tilt. If the tilting angle φ is approximately equal to the angle LPR in fig. 1, these photographs differ in the same way as those taken by the first method from positions L and R.

Stereophotographs have to be placed in a very definite position in the stereoscope to give a good three-dimensional impression. Photographs obtained by the first method must of course occupy the same positions in the stereoscope relative to the viewer's eyes as the projections had relative to the points L and R when the photographs were taken. In fig. 1 it can be seen that two corresponding image points on these photographs (e.g. Q'_r and Q'_l) lie in the same plane as LR when the photographs were taken. When viewing the photographs this same situation must be reproduced with respect to the eyes; in a stereoscope in which the photographs are placed side by side, the lines connecting the corresponding image points must therefore be parallel to the line connecting the two eyes.

This requirement must also be fulfilled when viewing photographs taken by the tilting method. It can be seen from fig. 2 that with these photographs the line connecting corresponding image points is always perpendicular to projection of the the axis of tilt k'. These photographs must therefore be positioned in the stereoscope in such a way that the direction of the axis of tilt in each photograph is perpendicular to the line connecting the observer's eyes. The axis of tilt need not be in the photographs; only its direction has to be known.

After the object has been tilted, its distance from the camera will usually have altered, so that in the two photographs the magnification of the object may differ; a correction must be made for this when the photographs are printed or when they are viewed in the stereoscope. (This complication is avoided in fig. 2 by making the two positions of the object symmetrical with respect to the line CK.) The position of the axis of tilt with respect to the object is of little significance; the tilting of the object around a particular axis may always be treated as rotation about another axis combined with a displacement which only affects the magnification. A three-dimensional impression does not depend critically on the angle through which the object is tilted: in direct observation as well we can experience three-dimensional vision of an object at very different distances, i.e. at very different values of the angle LPR in fig. 1.

Stereophotographs not only give a qualitative three-dimensional impression, but can also provide quantitative information about "depth". The depths in the picture can be calculated from the differences in the

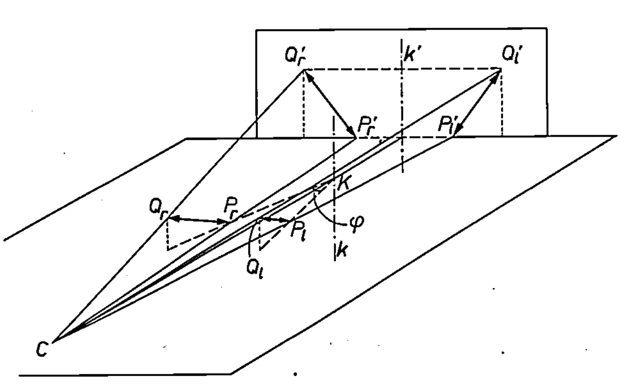


Fig. 2. Diagrammatic illustration of how stereophotographs can be made by tilting the object. The two photographs are both taken from C and the object is tilted through an angle φ about the axis k. P_rQ_r and P_lQ_l give the two positions of the object from which the photographs are taken. $P'_rQ'_r$ and $P'_lQ'_l$ are the images on the photographs, and k' is the image of the axis of tilt. The two positions of the object in this drawing are chosen symmetrical with respect to the line CK, so that the two images are identical in size, and the same photographs are obtained as with the method in fig. 1. As a rule, there will be some difference in optical magnification between the photographs, and correction must be made for this when viewing them.

positions of corresponding points on the two photographs; this is the principle underlying methods using a stereoscope in which markers are made to coincide with points in the image, allowing the depth to be read off directly. At the end of this article a system of this kind will be described. These methods are widely used in aerial and survey work.

Stereophotography with the electron microscope

A three-dimensional image may also be desired when photographing a specimen with the aid of an electron microscope [1]. In accordance with the foregoing, it is then necessary to pass the electron beam through the specimen from two different angles. The method first described cannot be considered here: in an electron microscope the direction of the beam cannot be changed; nor is it possible, owing to the very narrow angle of the electron beam, to achieve the effect by displacing the specimen. The tilting method, however, is particularly suitable here; owing to the small aperture of its objective lens the electron microscope has a very considerable depth of focus, so that a specimen making a small angle with the theoretical object plane can still be kept completely in focus. Most electron microscopes are therefore fitted (or can be fitted) with a device which enables the specimen to be tilted a few degrees each way with respect to the object plane.

In order to study such photographs in a stereoscope they must again be positioned in such a way that the direction of the axis of tilt on the two photographs is perpendicular to the connecting line between the observer's eyes. The difficulty now, however, is that the direction of the angle of tilt on these photographs is not known. This comes about because the electromagnetic lenses fitted in nearly all modern electron microscopes cause a *rotation* of the image relative to the object around the optical axis. This rotation also depends on the focal length of the lens, and since this length has to be varied to adjust the magnification and the sharpness of the image, the rotation is not constant. As it is usually necessary to re-focus after tilting the object, the rotation is moreover not as a rule identical for the two photographs. The result is that there is no simple way of deriving the direction of the representation of this axis in the image plane of the microscope (and hence on the photographs) from the direction of the axis of tilt in the microscope.

In order to be able to view the photographs three-dimensionally, in spite of the uncertainty of the location of the axis of tilt, they have to be mounted so that they can be rotated in the stereoscope; in addition there has to be some method of determining their correct position.

When an electron microscope with electrostatic lenses is used, or with photographs taken with an optical lens, one should still bear in mind that the photographs often have to be rotated through a certain angle in the stereoscope. Although the image rotation in these cases is zero (or 180°), so that the angle of tilt has the same direction on the photographs as in reality, this direction does not have to be parallel to one of the sides of the photographs placed in the stereoscope.

The image rotation in electromagnetic lenses has a second adverse effect in stereophotography. The image rotation increases with the third power of the distance from the object to the axis of the magnetic lens. The images therefore show some distortion, and, since the tilting alters the distance between the part of the object depicted and the axis of the microscope, the distortion is not identical in the two photographs. As a result, there is three-dimensional distortion when the photographs are viewed stereoscopically. If observation is limited to the neighbourhood of the microscope axis, and the part of the object to be depicted is placed as near as possible to the axis of tilt, the resultant error is negligible. In the most unfavourable situation, at the edge of the picture, the error can increase to a few per cent. If it is desired to use the photographs for accurate measurements, this effect must be taken into account [2]

The stereoscope which we use for studying these photographs is a modification of the oldest known stereoscope, designed in 1833 by Wheatstone [3]. This type of stereoscope has the great advantage that the only optical aids it needs are two plane mirrors — there are no lenses — so that the details of the photographs can be observed without any loss of definition. Furthermore, photographs of large format (e.g. 18 × 24 cm) can be viewed, enabling full use to be made of the optical magnification of the photographs. *Fig. 3* shows the instrument employed by us. The two stereophotographs, with the picture sides facing each other, are fixed on vertically mounted discs located about one metre apart. In the centre, between the two discs, there are two plane mirrors. The mirrors are set at an angle such that, when the observer brings his head close to them, his left eye sees only the reflection of the left-hand photograph, and the right eye sees only the reflection of the right-hand photograph (see also the title photograph). The observer can vary the angle between the mirrors to suit the spacing of his eyes, and he can also tilt one of the mirrors about a horizontal axis to bring

[1] J. G. Helmcke, Theorie und Praxis der elektronenmikroskopischen Stereoaufnahmen, Optik **11**, 201-225, 1954, and **12**, 253-273, 1955.

[2] J. G. Helmcke, Determination of the third dimension of objects by stereoscopy, Quantitative electron microscopy (Proc. Symp. Washington April 1964), p. 195-200, publ. Int. Academy of Pathology; also published in Laboratory Investigation **14**, 933-938, 1965 (No. 6^{II}).

[3] H. Mayo, Outlines of human physiology, page 288, Burgess Publ. Co., Minneapolis U.S.A., 1833.

the two images to the same height. This enables him to bring the two reflected images into register easily.

We have modified the original Wheatstone system by making it possible to rotate the discs, with the photographs fixed to them, independently of each other in their own plane. We have shown that the correct position is reached when corresponding image points on the two photographs always lie in the same plane as the line between the observer's eyes. This gives a basis for finding the correct position by means of the following method, which, after some practice, works fairly quickly. The photographs are placed in the stereoscope and, at any arbitrary angular position of the discs, the viewer adjusts the two images until they roughly coincide. He now shuts left eye and right eye momentarily in quick succession one after the other, observing while doing so a pair of points lying close together in the picture. Provided these points do not happen to lie at the same depth, he sees them shift in relation to one another because of parallax. Keeping his head upright, the viewer now has to rotate the discs to a setting at which he observes this displacement in the horizontal direction. In this position a three-dimensional image can easily be observed. Visual adaptation also makes it possible to see a three-dimensional image even when the photographs are not in perfect alignment, but this is not so easy, and prolonged viewing can be very tiring.

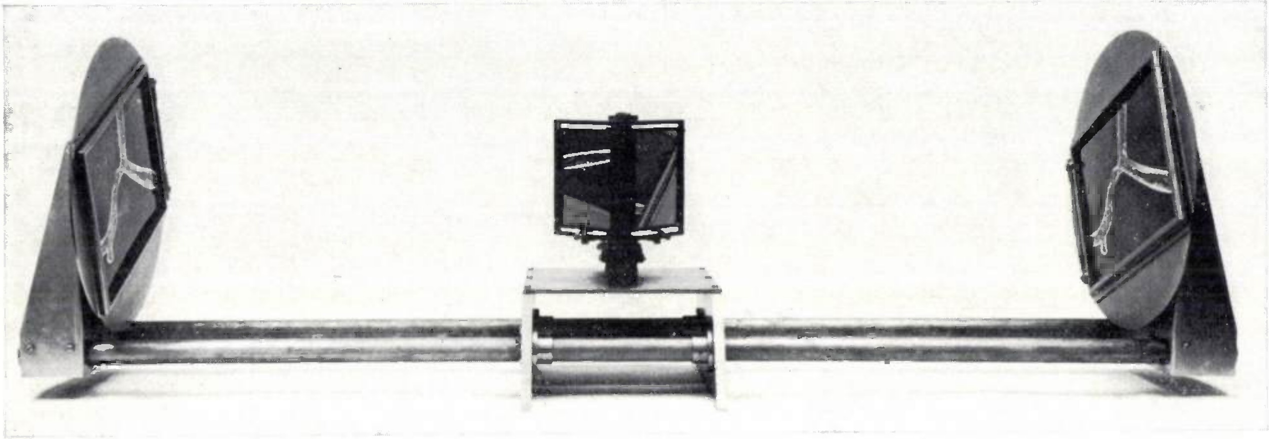


Fig. 3. Stereoscope specially designed for viewing stereophotographs taken with the electron microscope. On the left and at the right the photographs are attached to discs which can be rotated in their own plane. At the centre there are two mirrors mounted at an angle to each other, such that the viewer sees only the left-hand photograph with his left eye, and only the right-hand one with his right eye (see also the title photograph). Compensation can be made for a difference in optical magnification between the two photographs by adjusting the distance between one of the discs and the mirrors.

The smallest difference in depth observable by the method described here can be calculated with the aid of *fig. 4*. An object is tilted through an angle φ, usually chosen at 12°, about an axis passing through K (fig. 4*a*) and perpendicular to the plane of the drawing. The camera is located at C; for simplicity we assume that the tilting is symmetrical with respect to CK. A point V on the object which, seen from the camera, is in front of K, is in the positions V_r and V_l when the photographs are taken. The photographs now show two projections of the line element KV: these are $K_{rl}'V_l'$ and $K_{rl}'V_r'$. Fig. 4*b* illustrates diagrammatically the situation when the photographs are viewed in the stereoscope (the mirrors are omitted here and the angle between

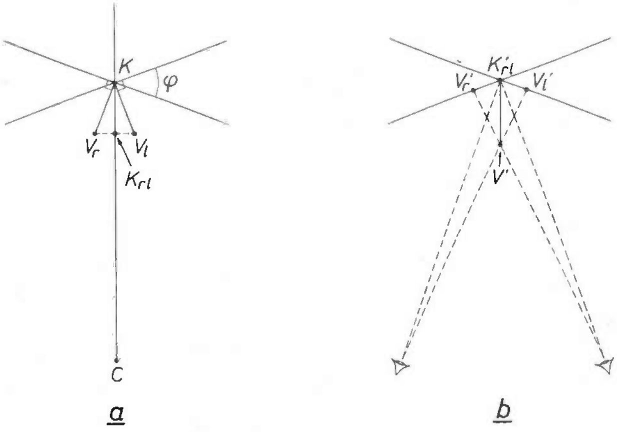


Fig. 4. To illustrate the calculation of the smallest difference in depth that can be observed on stereophotographs obtained by the tilting method.

the axis of the eyes is exaggerated). If the images are viewed at a sufficient (but not uselessly high) magnification, the structures only just resolved by the electron microscope (approx. 1 nm) are seen by the eyes as only just separated (approx. 0.5 mm at a viewing distance of 50 cm). The "true" size of $K_{r1}V_1$ must then be 1 nm in order for it to be just perceptible, and it can be shown from fig. 4*b* that the corresponding magnitude of $K'V'$, i.e. the smallest visible difference in depth, is about 10 nm.

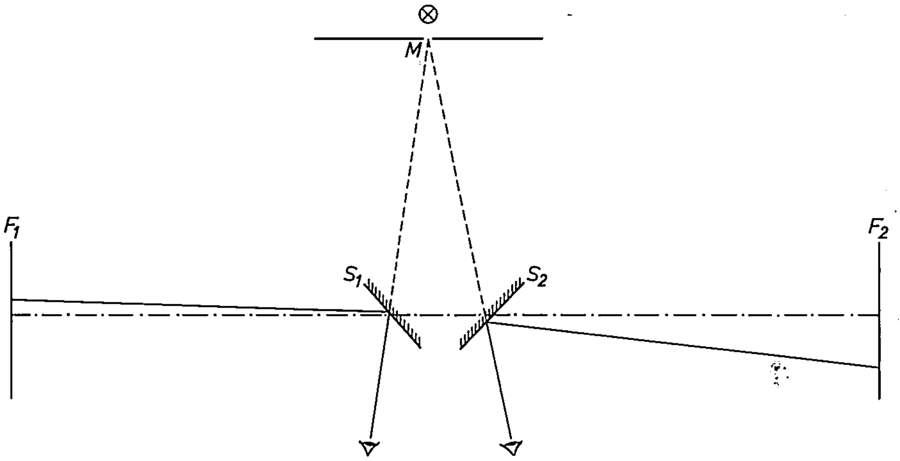


Fig. 5. Arrangement for determining depth measurements of an object from stereophotographs. S_1 and S_2 are the stereoscope mirrors shown in fig. 3, F_1 and F_2 are the discs with the photographs. In this arrangement the mirrors are half-silvered; behind the mirrors there is a marker M (e.g. an illuminated spot) which can be shifted in any direction. The viewer can make the marker coincide successively with various points of the three-dimensional image he sees in the stereoscope; the dimensions of the object can be calculated from the displacements of the marker.

If it is desired not only to view the photographs three-dimensionally but also to derive from them quantitative data about depths, then we must remember that the perspective image produced by an electron microscope corresponds to that produced by a telephoto lens. When a telephotograph is viewed at normal reading distance, the depth dimensions seem to be much smaller than they really are. We cannot therefore draw any direct conclusions about the dimensions of the object from the depth seen when viewing the photographs in the stereoscope. If however quantitative data are required, then, as noted above, we have to make a direct measurement of the differences in position of corresponding image points. The stereoscope described here can be adapted fairly simply for such measurements. This may be done by using half-silvered mirrors, with an illuminated marker on a support behind them which can be shifted in any direction; see *fig. 5*. While viewing in the stereoscope the observer can now make the illuminated marker coincide with various points of the three-dimensional image he sees. At a given enlargement of the photographs the dimensions of the object in depth can be calculated from the measured displacements of the support in the viewing direction.

This method takes no account of the image distortion caused by rotation of the electron beam. This effect sets a limit to the accuracy that can be achieved.

Some examples

When the correct position of a pair of stereophotographs has been found, the photographs can be viewed stereoscopically in other ways than with the instrument described, and can for example be reproduced for publication. If small photographs are sufficient (width no greater than the eye spacing = 6.5 cm) then the two photographs can be reproduced side by side. A stereo-pair of this kind can be viewed with a stereoscope or with two separate positive lenses (after practice some people are able to see the two images stereoscopically without optical aids). If photographs of a larger format are required, they can be viewed very suitably by reproducing them as an anaglyph.

By way of illustration we show here a few stereophotographs made with the Philips EM 100 electron microscope. The anaglyphs can be viewed with the aid of the red-blue spectacles issued with an article which appeared in this journal a few years ago, or with the transparent coloured sheets enclosed with this issue [4].

[4] The spectacles were issued with the article by A. E. Jenkinson, Projection topographs of dislocations, Philips tech. Rev. **23**, 82-88, 1961/62.
If the transparent sheets are used, hold the red one in front of the left eye and the blue one in front of the right eye. Some viewers may find that the stereo effect is enhanced if a double thickness of the blue sheet is used.

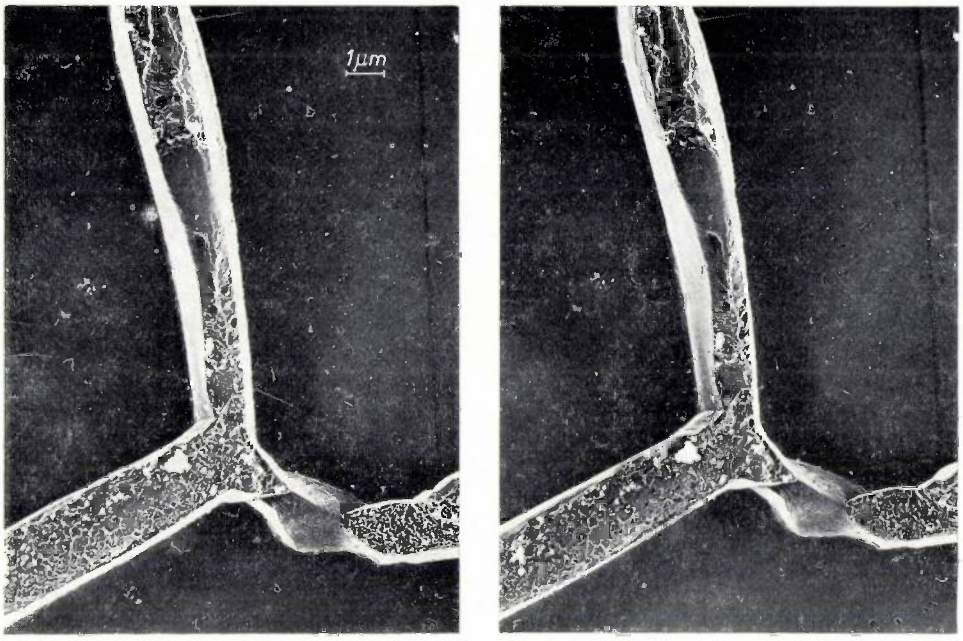


Fig. 6. Pair of stereophotographs of a carbon replica of sintered tantalum carbide after polishing and etching. The replica has been shadowed with platinum.

Fig. 6 and *fig. 7* show a photograph of tantalum carbide (Ta_4C_5), sintered after the addition of 1 per cent of magnesium by weight [5]. After polishing and etching a carbon replica was made and shadowed with platinum. Since it was mainly the crystal boundaries that were etched, a deep cleft can be seen in the photograph. The depth of this cleft is roughly 750 nm — calculated from the difference between the projections of the height of the side wall visible on the photographs.

Fig. 8 shows a carbon replica of exposed silver-bromide grains from a photographic plate. This replica was made in the following way. AgBr grains were

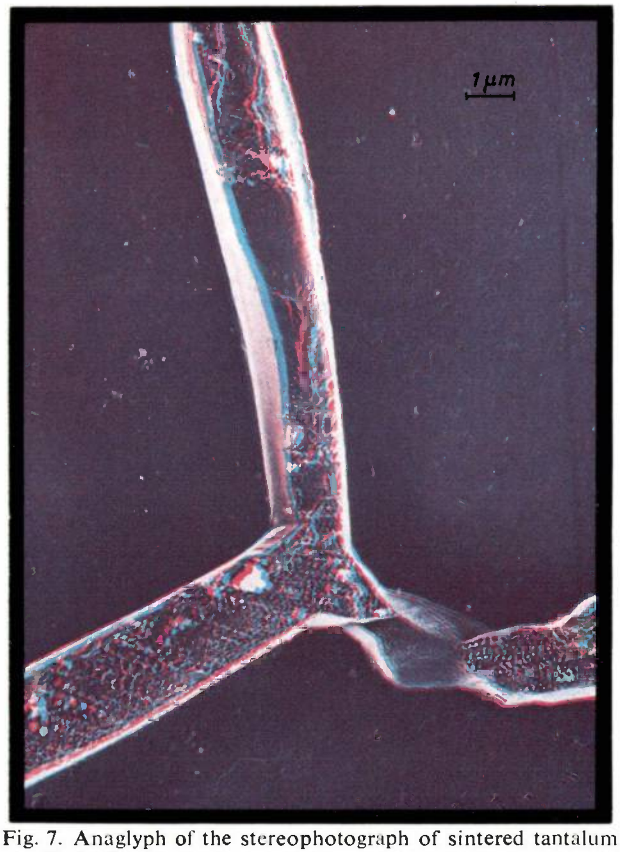


Fig. 7. Anaglyph of the stereophotograph of sintered tantalum carbide in fig. 6.

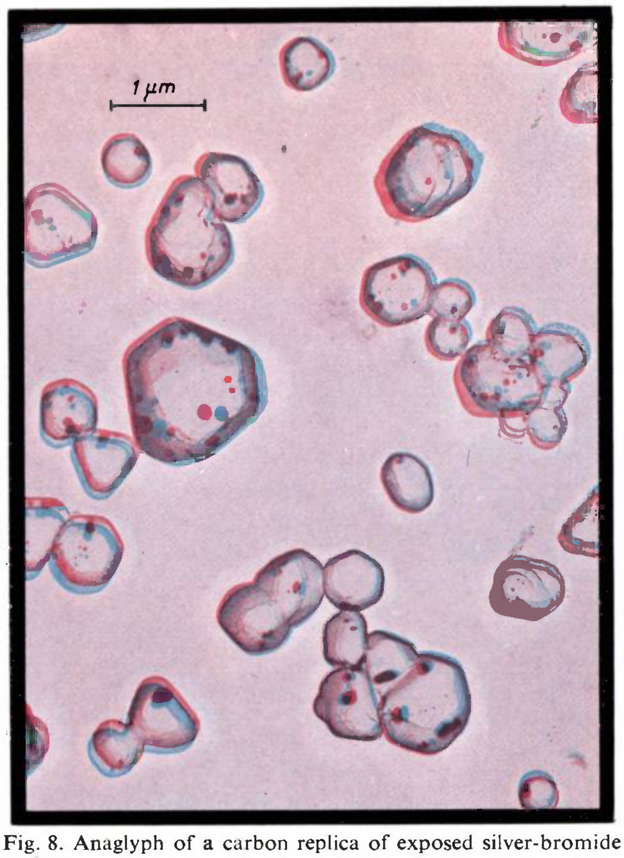


Fig. 8. Anaglyph of a carbon replica of exposed silver-bromide grains.

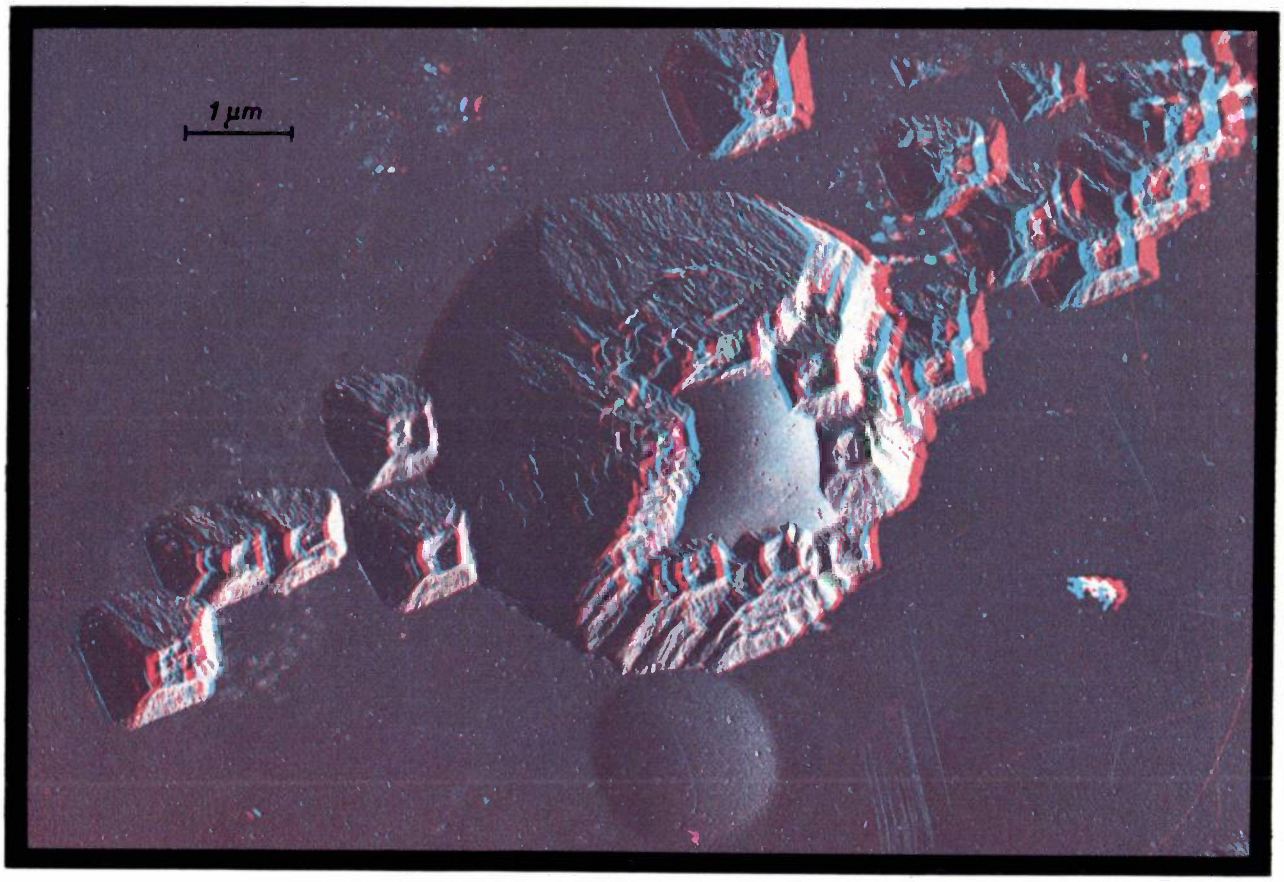


Fig. 9. Anaglyph of a carbon replica of etch pits in a tungsten single-crystal. The replica has been shadowed with platinum.

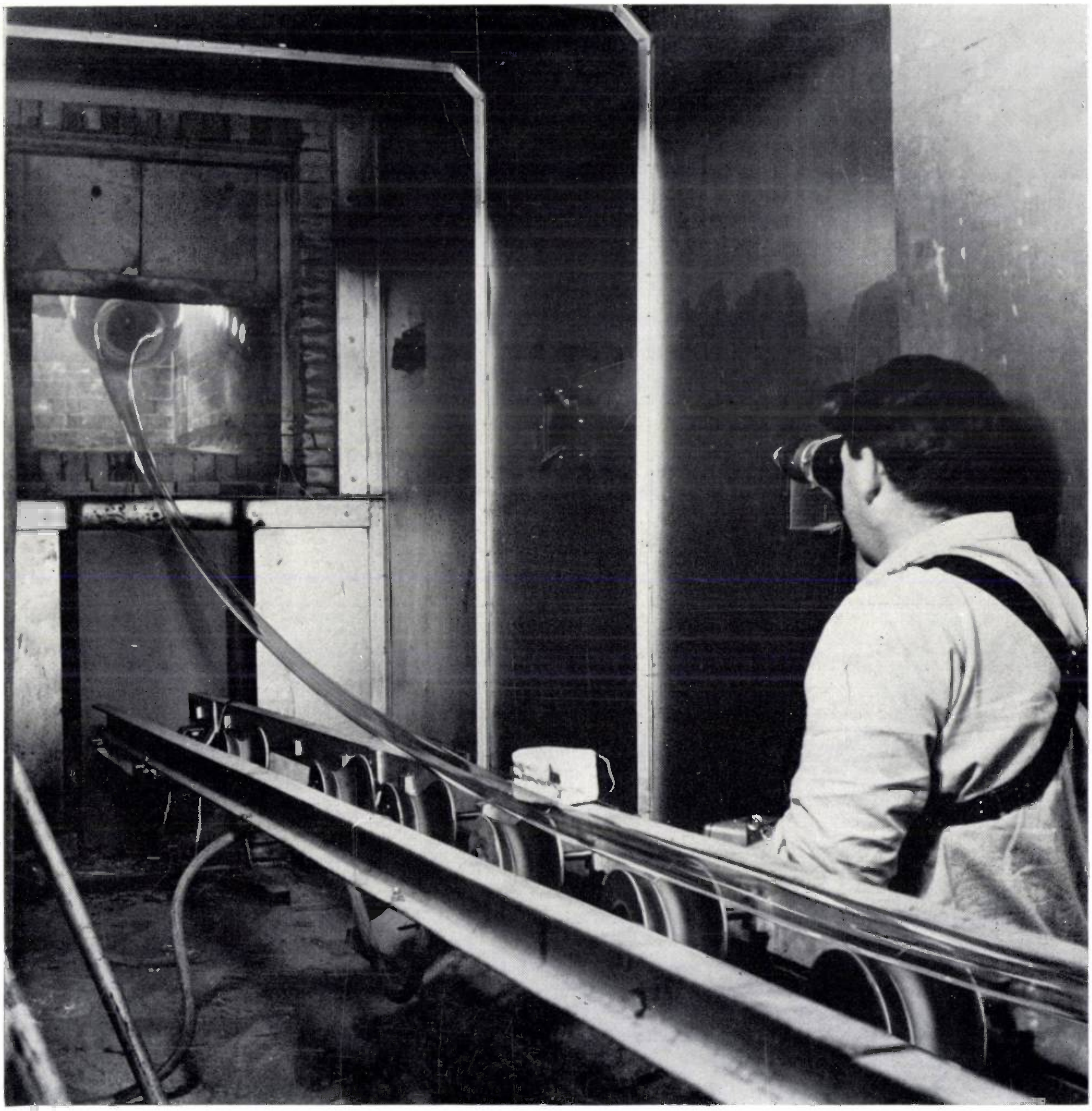
obtained from an unexposed photographic plate by removing the gelatine with a solvent. The separate grains were placed on a glass slide, upon which a layer of carbon was then deposited by evaporation. This was done by means of an arc discharge between two carbon electrodes. Due to the action of the incident light, part of the AgBr is decomposed in many of the grains. After the replica has been made, the non-dissociated AgBr is dissolved in sodium thiosulphate. The silver produced by exposure then remains visible as black patches in the grain walls, which have been made visible by the carbon film.

Fig. 9 shows some etch pits in a single-crystal of tungsten which has been electrolytically polished and then etched. The photograph was again made from a carbon replica shadowed with platinum.

Summary. Stereophotographs are taken with the aid of the electron microscope by slightly tilting the object between the two exposures (through an angle of about 12°). In order to obtain a three-dimensional image when viewing such a pair of stereophotographs in a stereoscope, the two photographs have to be given an angular position such that the direction of the axis of tilt, in the photographs, is perpendicular to the line connecting the viewer's eyes. This condition is not immediately fulfilled because electromagnetic lenses cause a variable rotation of the image relative to the object. This article describes a stereoscope — a modified version of the oldest known stereoscope, designed by Wheatstone — in which large photographs can be fitted and rotated in such a way as to give them the correct position for stereoscopic viewing. By a slight modification of the instrument, the depth of each detail in the pair of stereophotographs can be quantitatively determined. Some examples of stereophotographs produced in this way are shown.

[5] E. Roeder and M. Klerk, Untersuchungen mit dem Elektronenstrahl-Mikroanalysator an druckgesintertem Tantalkarbid mit geringem Mangan- und Nickelzusatz, Z. Metallk. **54**, 462-470, 1963.

Continuous drawing of glass tubing for fluorescent lamps



In the last twenty years or so, the picturesque process in which glass tubing is drawn by two glass-blowers has been almost completely superseded by a continuous mechanical process. The molten glass is run on to an inclined, rotating cylindrical mandrel (temperature 400-1000 °C). The glass flows off like a sleeve from the lower end of the mandrel (just visible top left in the photo). A mechanism for drawing the tubing away (not visible in the photo) is situated some dozens of yards from the mandrel. It can be seen how the tubing becomes thinner immediately after leaving the mandrel: a few yards away the glass has set. The continuous tube travels over rollers (in the foreground) towards the drawing gear. At some distance the temperature has decreased sufficiently for the tubing to be cut to the required lengths. The operator in the foreground is checking the temperature of the mandrel with a pyrometer.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands	*E*
Mullard Research Laboratories, Redhill (Surrey), England	*M*
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France	*L*
Philips Zentrallaboratorium GmbH, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany	*A*
Philips Zentrallaboratorium GmbH, Hamburg laboratory, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany	*H*
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium.	*B*

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

G. Bergmann: Untersuchungen über die Alterungserscheinungen des Ge-dotierten Bi_2Te_3.
Phys. Stat. sol. **14**, 311-318, 1966 (No. 2). *A*

G. Blasse: New compounds of the AB_2O_6 type.
J. inorg. nucl. Chem. **28**, 1122-1124, 1966 (No. 4). *E*

G. Blasse: Polymorphism of Bi_2MoO_6.
J. inorg. nucl. Chem. **28**, 1124-1125, 1966 (No. 4). *E*

G. Blasse: Dilanthanide molybdates and tungstates Ln_2MO_6.
J. inorg. nucl. Chem. **28**, 1488-1489, 1966 (No. 6/7). *E*

P. F. Bongers and **U. Enz**: Metamagnetism of $NaNiO_2$.
Solid State Comm. **4**, 153-157, 1966 (No. 4). *E*

A. H. Boonstra and **J. van Ruler**: The adsorption of various gases on clean and oxidized Ge surfaces.
Surface Sci. **4**, 141-149, 1966 (No. 2). *E*

G. Bosch: Anisotropy of the electrical conduction of α-SiC single crystals.
J. Phys. Chem. Solids **27**, 795-796, 1966 (No. 4). *E*

A. J. Bosman and **C. Crevecoeur**: Mechanism of the electrical conduction in Li-doped NiO.
Phys. Rev. **144**, 763-770, 1966 (No. 2). *E*

J. C. Brice, J. A. Cundall and **A. P. King**: Easy direction coercive force associated with domain wall motion in nickel-iron films.
J. Materials Sci. **1**, 170-185, 1966 (No. 2). *M*

J. C. Brice, P. A. C. Whiffin and **E. J. Millett**: Variation of distribution coefficients with melt composition.
Brit. J. appl. Phys. **17**, 563-564, 1966 (No. 4). *M*

A. Broese van Groenou: The compensation of positive and negative anisotropies in ZnNi and ZnLi ferrite.
Philips Res. Repts. **21**, 180-187, 1966 (No. 3). *E*

K. Bulthuis: The effect of local pressure on germanium *p-n* and *p-s-n* structures.
Philips Res. Repts. **21**, 85-103, 1966 (No. 2). *E*

H. B. G. Casimir: On Einstein-Bose condensation.
Proc. Kon. Ned. Akad. Wetensch. **B 69**, 223-229, 1966 (No. 2). *E*

J. R. Chamberlain and **R. W. Cooper**: Paramagnetic resonance in yttrium gallium garnet: Co^{2+} and Mn^{2+}.
Proc. Phys. Soc. **87**, 967-970, 1966 (No. 4). *M*

C. Z. van Doorn: Recombination mechanisms for the "edge" emission in cadmium sulphide.
Philips Res. Repts. **21**, 163-179, 1966 (No. 3). *E*

W. F. Druyvesteyn: Magnetization curves and resistance transitions of superconducting lead alloys.
Thesis, Eindhoven, July 1965. *E*

J. J. Engelsman, A. Meyer and **J. Visser**: Microdetermination of oxygen, carbon and water in inorganic materials using a carrier-gas technique.
Talanta **13**, 409-420, 1966 (No. 3). *E*

P. J. Flanders and **R. F. Pearson**: Measuring magnetization in anisotropic magnetic samples using a torque method.
Brit. J. appl. Phys. **17**, 521-524, 1966 (No. 4). *M*

G. E. G. Hardeman and **G. B. Gerritsen**: Displacement phenomena of boron acceptors in 6H SiC.
Physics Letters **20**, 623-624, 1966 (No. 6). *E*

J. Hasker: The influence of initial velocities on the beam-current characteristic of electron guns.
Philips Res. Repts. **21**, 122-150, 1966 (No. 2). *E*

J. Hasker: Calculation of lens action and lens defects of the post-acceleration region in oscilloscope tubes.
Philips Res. Repts. **21**, 196-209, 1966 (No. 3). *E*

J. C. M. Henning: ^{14}N hyperfine structure in ESR spectra of heterocyclic anions.
J. chem. Phys. **44**, 2139-2155, 1966 (No. 5). *E*

J. C. M. Henning and **P. F. Bongers:** Electron spin resonance of Mn^{2+} in Cs_3ZnCl_5.
J. Phys. Chem. Solids **27**, 745-747, 1966 (No. 4). *E*

J. Israël: A study of the thermoelectric properties of Pt/Ru alloys.
J. nucl. Mat. **18**, 272-277, 1966 (No. 3). *E*

G. H. Jonker: Magnetic and semiconducting properties of perovskites containing manganese and cobalt.
J. appl. Phys. **37**, 1424-1430, 1966 (No. 3). *E*

A. Kats, G. Piesslinger and **S. Rademaker:** Kiemvormings- en kristallisatieprocessen in glaskeramiek.
Chem. Weekblad **62**, 181-185, 1966 (No. 15).

S. R. de Kloet and **P. J. Strijkert:** Selective inhibition of ribosomal RNA synthesis in Saccharomyces carlsbergensis by 5-fluorouracil.
Biochem. biophys. Res. Comm. **23**, 49-55, 1966 (No. 1). *E*

W. F. Knippenberg and **G. Verspui:** The preparation of large single crystals of SiC polytypes by precipitation from solutions.
Philips Res. Repts. **21**, 113-121, 1966 (No. 2). *E*

J. E. Knowles and **A. Broese van Groenou:** A new manifestation of magnetic after-effect.
Phys. Stat. sol. **14**, 91-96, 1966 (No. 1). *M*

E. Kooi: Influence of heat treatments and ionizing irradiations on the charge distribution and the number of surface states in the $Si\text{-}SiO_2$ system.
IEEE Trans. on electron devices **ED-13**, 238-245, 1966 (No. 2). *E*

J. G. M. de Lau and **A. L. Stuijts:** Chemical composition and high-frequency properties of Ni-Zn-Co ferrites.
Philips Res. Repts. **21**, 104-112, 1966 (No. 2). *E*

F. A. Lootsma: Network planning with stochastic activity durations, an evaluation of PERT.
Statistica neerl. **20**, 43-69, 1966 (No. 1). *E*

M. H. van Maaren and **G. M. Schaeffer:** Superconductivity in group V^a dichalcogenides.
Physics Letters **20**, 131, 1966 (No. 2). *E*

J.-P. Mathieu: On the theory of perturbational optimal guidance of space vehicles.
Peaceful uses of automation in outer space, editor J. A. Aseltine, pp. 567-570, Plenum Press, New York 1966. *B*

R. Memming and **G. Schwandt:** Anodic dissolution of silicon in hydrofluoric acid solutions.
Surface Sci. **4**, 109-124, 1966 (No. 2). *H*

B. J. Mulder and **J. de Jonge:** Exciton diffusion and the photoconductivity spectrum of anthracene, pyrene and perylene.
Philips Res. Repts. **21**, 188-195, 1966 (No. 3). *E*

B. J. Mulder, J. de Jonge and **G. Vermeulen:** The photocurrent in anthracene crystals under illumination of the negative electrode.
Rec. Trav. chim. Pays-Bas **85**, 31-34, 1966 (No. 1). *E*

J. W. Orton, A. S. Fruin and **J. C. Walling:** Spin-lattice relaxation of Cr^{3+} in single crystals of zinc tungstate.
Proc. Phys. Soc. **87**, 703-716, 1966 (No. 3). *M*

C. J. M. Rooymans: Invloed van de druk op het mechanisme van de vastestofreactie.
Chem. Weekblad **62**, 189-194, 1966 (No. 16). *E*

H.-J. Schmitt and **G. Buchta:** Approximate theory of the transversely magnetized reciprocal phase shifter.
Proc. IEEE **54**, 308-310, 1966 (No. 2). *H*

D. A. Schreuder: Het vooruit bepalen van de gemiddelde wegdekluminantie bij openbare verlichting.
Electrotechniek **44**, 143-147, 1966 (No. 6).

L. A. Æ. Sluyterman: Amperometric argentometric titration of thiol groups in imidazole buffer.
Anal. Biochemistry **14**, 317-319, 1966 (No. 2). *E*

L. A. Æ. Sluyterman: Substrate binding by non-activated papain.
Biochim. biophys. Acta **113**, 577-586, 1966 (No. 3). *E*

F. A. Staas and **W. F. Druyvesteyn:** Flux flow in type-II superconductors with different field orientations.
Philips Res. Repts. **21**, 153-162, 1966 (No. 3). *E*

A. L. Stuijts: Keramische technologie.
Klei en Keramiek **16**, 35-39, 1966 (No. 2). *E*

R. Thees: Kleine Elektromotoren.
Elektrotechn. Z. A **87**, 171-175, 1966 (No. 5). *A*

B. Tuck and **R. W. G. Clarke:** Anomalous double-peaking effect from a monochromator.
J. sci. Instr. **43**, 196, 1966 (No. 3). *M*

J. S. C. Wessels: Studies with small fragments prepared by digitonin treatment of spinach chloroplasts.
Currents in Photosynthesis, Proc. 2nd W.-Europe Conf., Woudschoten-Zeist 1965, pp. 129-139, publ. Donker, Rotterdam 1966. *E*

W. J. Witteman: Theory of mode interaction in the case of small mode spacing.
Phys. Rev. **143**, 316-322, 1966 (No. 1). *E*

W. J. Witteman: Inversion mechanisms, population densities and coupling-out of a high-power molecular laser.
Philips Res. Repts. **21**, 73-84, 1966 (No. 2). *E*

Volume 27, 1966, No. 8 pages 213-240 Published 25th November 1966

PHILIPS TECHNICAL REVIEW

VOLUME 27, 1966, No. 9/10

Photosynthesis

A survey of the present state of research

J. S. C. Wessels and M. van Koten-Hertogs

581.132.1

The ability of plants to absorb, transform, and utilize energy in the form of light is the subject today of ingenious investigation. A possible practical application of this hard-won basic knowledge is in a planned intervention in plant metabolism. The creation of better circumstances for photosynthesis could lead to better harvests and an improvement in the world's food supply. One way of achieving this could be through an increase in CO_2 concentration. A directed synthesis by the plant, leading for example to more protein or carbohydrate production, may also become feasible. Moreover, the knowledge of photosynthesis which has been gained already enables a more informed action to be taken against undesirable plants. The chemical control of weeds will become an ever more vital necessity for mankind.

Introduction

All living organisms require energy for the synthesis of new molecules (chemical energy), for muscular contraction (mechanical energy), for conduction in nerve tissue (electrical energy), and for the movement of substances against a concentration gradient (osmotic energy). The cell, the unit of the organism, is able to make use of this energy through the combustion (oxidation) of nutritives, mainly sugars and fats. The ultimate supplier of this fuel is the plant. Plant cells have the power of using the energy from sunlight for building up high-energy organic compounds from low-energy inorganic compounds like carbon dioxide and water. Since light plays an essential role in this synthesis the process is known as photosynthesis. Needless to say, the process of photosynthesis is of vital importance for man and animals, as it is of course for the plant itself.

Since plants obtain the requisite energy directly from sunlight, so that in this respect they are fully self-supporting, they are said to be autotrophic. Man and animals, on the other hand, depend on the plant for their energy supply and are therefore said to be heterotrophic.

The energy liberated during the oxidation of nutritives (respiration) is not transformed into heat but is stored in the form of small "packets" of 7 kcal in molecules of high-energy phosphates. The most important high-energy phosphate compound is adenosine triphosphate (ATP). As to its chemical formula, suffice it to say that adenosine is built up from adenine (ad), an organic base, and ribose (rib), a sugar. Three phosphate groups are linked to the ribose; the bonds between the phosphate groups themselves have a high energy and are usually indicated by ~ P. In the hydrolysis of ATP a phosphate group is split off and adenosine diphosphate (ADP) is formed, with the liberation of about 7 kcal energy per mole:

$$\underset{\text{(ATP)}}{\text{ad-rib-P}\sim\text{P}\sim\text{P}} + \text{H}_2\text{O} \rightarrow \underset{\text{(ADP)}}{\text{ad-rib-P}\sim\text{P}} + \text{H}_3\text{PO}_4,$$
$$\Delta F^0 = -7 \text{ kcal.}$$

The ATP can be compared with a charged battery and the ADP to a battery in the discharged state. The energy liberated by the splitting of ATP can be used to supply the various energy requirements of the organism (*fig. 1*). Some idea of the importance of this formation and decomposition of ATP can be

Dr. J. S. C. Wessels and Drs. M. van Koten-Hertogs are with Philips Research Laboratories, Eindhoven. — Dr. Wessels was awarded the gold medal of the Royal Dutch Chemical Society in September 1965 for his work on photosynthesis.

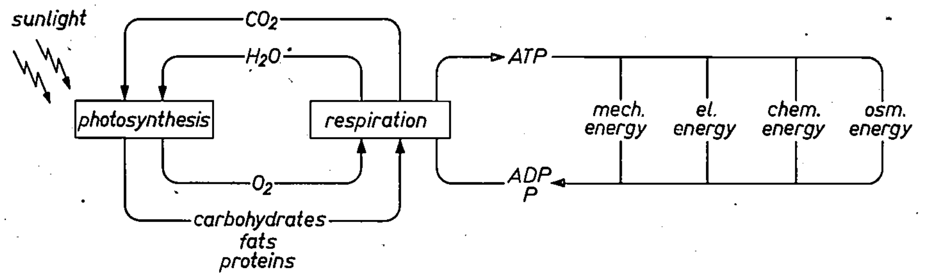


Fig. 1. The energy required by living organisms — mechanical, electrical, chemical, and osmotic energy — is supplied by the ATP that is formed in the oxidation of nutritives (respiration). The nutritives are built up by the *plant* in sunlight (photosynthesis).

seen from the fact that in 24 hours the human body produces and breaks down again some 70 kg of ATP, i.e. its own weight.

Historical development

About 200 years ago Joseph Priestley, an English Nonconformist minister, and Jan Ingenhousz, a Dutch physician, discovered that illuminated plants give off a gas that will accelerate combustion and in which animals can live. This was the beginning of systematic observations of the light-induced exchange of gases in green plants, and the first step that led to the discovery of the element oxygen by Lavoisier. It was soon found that plants would only give off oxygen if another gas, toxic to animals, was present. Since then we have learned that plants take up CO_2 from the air and, under the influence of light, "combine" it with water from the soil with the help of their green pigment. In this process organic substances, mainly carbohydrates (starch and sugars), are formed, while oxygen is liberated at the same time:

$$nCO_2 + nH_2O \overset{\text{light}}{\rightarrow} C_nH_{2n}O_n + nO_2. \qquad (1)$$

Agreement about the important role played by the light in this process was reached very early, but not about the true function of the light. For a long time it was believed that this function consisted of the splitting of CO_2, in which O_2 was liberated and "C" remained for incorporation into carbohydrates (hence also the name carbohydrates: hydrates of carbon). About 1880, however, two discoveries were made which did not accord with this idea. Winogradsky found that the "chemosynthetic" bacteria, which have no green pigment, were able to convert CO_2 into organic material *in the dark*. On the other hand Engelman found that purple bacteria do have the power of fixing CO_2 under the influence of light but do not give off O_2 in the process. While on the basis of these observations the theory of the splitting of CO_2 was gradually abandoned, the following began to be established:

1) The energy required for the formation of carbohydrates need not itself be supplied by light. Chemosynthetic bacteria can obtain the requisite energy by chemical means, by oxidation of certain chemical compounds. For example, *Nitrosomonas* oxidizes ammonia to nitrite, *Nitrobacter* oxidizes nitrite to nitrate. In short, the power to fix CO_2 is not restricted to the green cell.

2) In 1930 Van Niel suggested the following formula for the photosynthesis of purple bacteria:

$$nCO_2 + 2nAH_2 \overset{\text{light}}{\rightarrow} C_nH_{2n}O_n + 2nA + nH_2O, \quad (2)$$

in which AH_2 represented a hydrogen donor, e.g. H_2S, as is the case with sulphur bacteria. During illumination of these bacteria *sulphur* is formed instead of oxygen.

If it is now assumed that photosynthesis in the plant takes place in a manner analogous to that in the purple bacteria, then this means that *water*, not CO_2, is decomposed under the influence of light, giving O_2. (For the plant this then means that $AH_2 = H_2O$.) Experiments with labelled water ($H_2{}^{18}O$) did in fact provide clear evidence, by the evolution of $^{18}O_2$, that H_2O and not CO_2 was decomposed.

In agreement with this, Hill demonstrated in 1939 that oxygen could be generated even in the absence of carbon dioxide [1]. He illuminated certain preparations from plants (isolated chloroplasts, see below) in the presence of ferric ions and observed that O_2 was given off accompanied by a simultaneous reduction of the ferric ions to ferrous ions:

$$4Fe^{3+} + 2H_2O \xrightarrow[\text{chloroplasts}]{\text{light}} 4Fe^{2+} + 4H^+ + O_2. \quad (3)$$

In this "Hill reaction" ferric ions are reduced instead of the carbon dioxide.

Photosynthesis in outline

The view is generally accepted nowadays that photosynthesis is an oxidation-reduction process, with water as the reducing agent which reduces CO_2 to carbohydrate thereby becoming itself oxidized to oxygen.

This process requires the supply of energy, as could have been concluded directly above from the fact that the reverse reaction, the combustion of carbohydrates, liberates energy.

Closer examination of the process of photosynthesis reveals that a substance, called $NADPH_2$ for short, is formed as an intermediate which has a considerably greater reducing power than water and

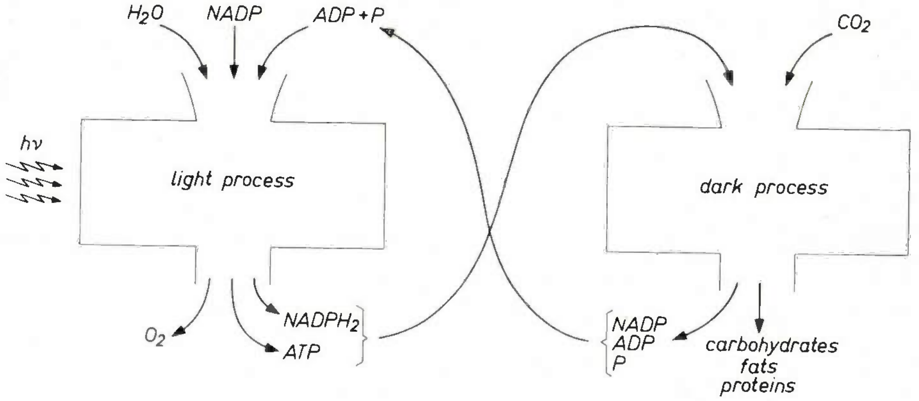


Fig. 2. The photosynthesis process takes place partly in the light, partly in the dark. In the light the splitting of water takes place, during which oxygen is given off. At the same time the products $NADPH_2$ and ATP are formed, which are necessary for the fixation of CO_2 in the dark and the formation of carbohydrates, fats and proteins.

which should be considered as the real reducing agent in photosynthesis. This intermediate stage takes up energy, but this forming of $NADPH_2$ may be seen as a temporary storing of energy.

The photosynthesis process consists of two parts, one of which takes place in the light and the other in the dark (*fig. 2*). In the *light process* decomposition, or photolysis, of the water takes place, during which the above strongly reducing compound $NADPH_2$ and the high-energy compound ATP are formed as well as oxygen. By means of the light process, all the chemical energy needed for carbohydrate formation proper is as it were stored up; such formation then takes place wholly in the dark. In this *dark process* the fixation of CO_2 and the synthesis of carbohydrates takes place as well as the synthesis of fats and protein, all these reactions being energized by the products ATP and $NADPH_2$ from the light process.

The above gives no more than a rough outline of what takes place in photosynthesis. In the following sections, in which we shall go into more detail, a very complicated mechanism will become apparent. After a section devoted to morphology the light and the dark reactions will be discussed separately. Our own work on photosynthesis is concerned exclusively with the light reactions. An important part of the dark reactions was clarified some time ago by Calvin, in a research for which he was awarded the Nobel prize. The frontier of the research has since shifted in the direction of the light reactions. This may perhaps justify to some extent the greater attention given to the light reactions in the present article [2].

The chloroplast

Photosynthesis takes place in the green parts of the plant. If we examine a part of a leaf under the microscope, then at sufficient magnification we can discern the separate cells from which the tissue is built up. In these cells there are lens-shaped structured particles arranged along the cell wall (*fig. 3*) [3]. The dimensions of these particles, which are called chloroplasts, are about 5×1 μm; there are between 20 and 100 chloroplasts per cell. The chloroplasts contain chlorophyll, the pigment in green leaves. Chlorophyll is therefore not, as might at first sight have been expected, distributed over the whole leaf. It is in these chloroplasts that photosynthesis takes place.

The chloroplasts can be removed from the plant by destruction of the cell walls, for example by pulverizing with quartz sand and then subjecting the mixture to fractional centrifuging. If this isolation is performed correctly, photosynthesis can still take place in the chloroplasts. A considerable part of the research on

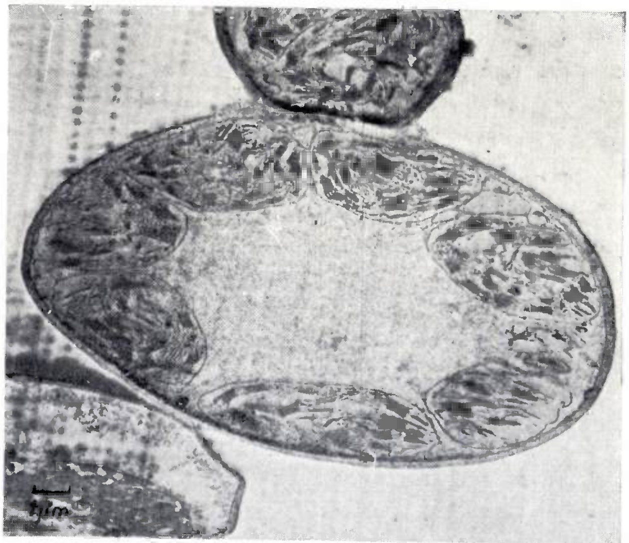


Fig. 3. Cell with chloroplasts along the cell wall. In these chloroplasts the whole process of photosynthesis is found to be localized. Fixation with $KMnO_4$. All photographs in this article have been made with an EM 200 electron microscope.

[1] R. Hill, Nature **139**, 881, 1937; R. Hill, Proc. Roy. Soc. **B 127**, 192, 1939.

[2] A small selection from the extensive literature on photosynthesis is given at the end of the paper for the use of the interested reader who is not an expert in this field but wishes to extend his knowledge of it. The literature referred to in the paper itself is concerned more with detail results, particularly with respect to investigations made in the last few years.

[3] The electron microphotographs appearing in this paper were made by H. J. van de Berg, Mrs. A. Dorsman, and A. J. Luitingh.

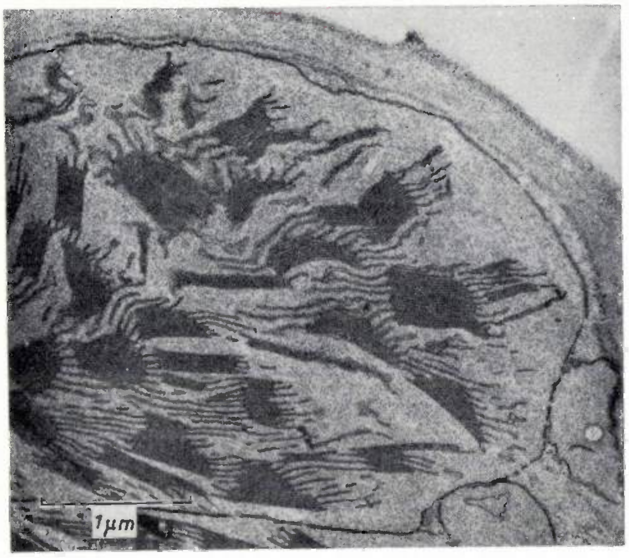


Fig. 4. Chloroplast with magnification about 5 times greater than in fig. 3. Fixation with $KMnO_4$. The light part is called the stroma, the dark parts are stacks of what are called lamellae or thylacoids (like stacks of coins seen from the direction of their edges). The light reactions take place in the lamellae, the dark reactions in the stroma.

photosynthesis is performed with isolated chloroplasts. One of the advantages of this is that there is no interference from other processes, such as respiration (oxidation), which take place in other organs of the cell, such as the mitochondria.

Light and dark spots can be distinguished in the chloroplast. With stronger magnification the dark parts of the chloroplasts are found to consist of stacks of lamellae (also known as thylacoids): see *fig. 4*. Such a stack of lamellae, looking rather like carelessly stacked coins, is called a granum. The grana are generally connected to each other by large lamellae. The light part of the chloroplast is called the stroma; the large lamellae are therefore also known as stroma lamellae or stroma thylacoids. The stroma accommodates most of the enzymes necessary for the photosynthetic dark reactions.

The lamellae are made up of about 50% protein and 50% lipid (fatty material). It is generally assumed that the membrane of a lamella is built up of a layer of lipid, 4 nm thick and covered on either side by protein (the "unit membrane") (*fig. 5*). After fixation of the material of a chloroplast with OsO_4 or $KMnO_4$ the protein layers can be seen in the electron microscope as dark bands with a thickness of about 2 nm. The lipid fraction includes the sunlight-absorbing photosynthesis pigments chlorophyll a and chlorophyll b. Not surprisingly therefore the light reactions of photosynthesis are found to take place in the lamellae.

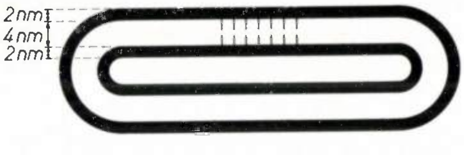


Fig. 5. Representation of the building-up of a lamella from protein layers (represented by the dark lines) and lipid layers. Lipids are fatty compounds consisting of a polar part, which is oriented towards the protein layer, and a non-polar chain oriented in the opposite direction (indicated by the thin lines).

Fig. 6 shows the structural formula of chlorophyll. The molecule consists of a porphyrin ring system with magnesium (four pyrrole rings joined by =CH groups) and a long side chain, the phytyl residue. Because chlorophyll possesses both a long non-polar side chain and also a polar porphyrin ring it will orient itself at the boundary between a polar and a non-polar solvent in such a way that the porphyrin ring is in the polar medium and the phytyl side chain is in the non-polar medium. This makes it probable that in the chloroplast lamellae as well the chlorophyll is oriented so that the porphyrin ring lies in the (polar) protein layer and the phytyl side chain in the (non-polar) lipid layer.

It seems likely that the picture given here of the arrangement of a chloroplast lamella is, however,

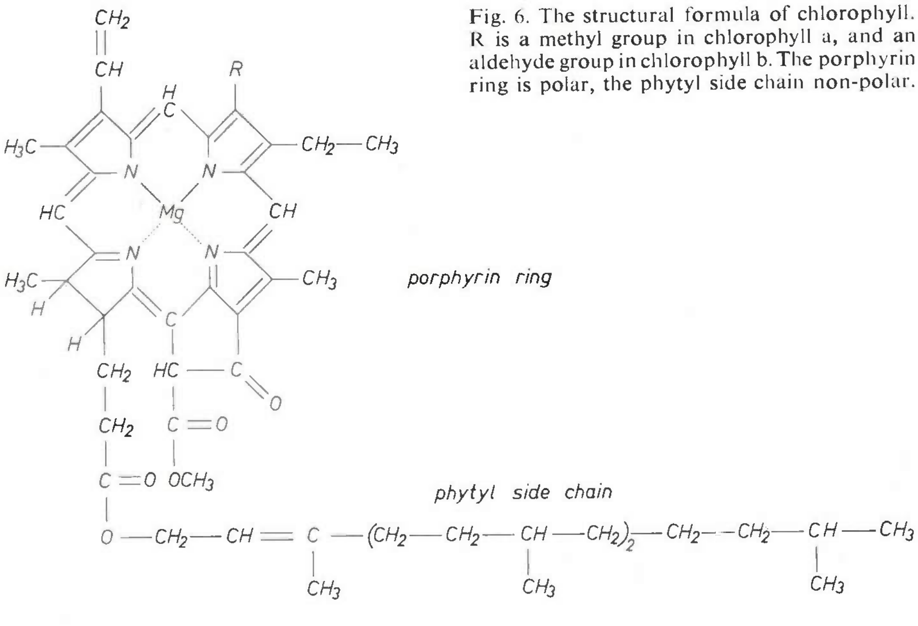


Fig. 6. The structural formula of chlorophyll. R is a methyl group in chlorophyll a, and an aldehyde group in chlorophyll b. The porphyrin ring is polar, the phytyl side chain non-polar.

still too simple. The use of new electron-microscopy techniques (negative-staining, freeze-etching) has made it possible to show that there are particles on the lamella membrane which have a diameter of about 10 nm and are thought to be built up of four sub-units (*fig. 7*). It is believed that these particles consist of protein. The fact that the protein layer, after fixation with OsO_4 or with $KMnO_4$ turns out to be only 2 nm thick would be explicable on the assumption that the fixation leads to denaturing of the protein and therefore to unrolling and spreading of protein molecules on the lipid layer [4] (*fig. 8*). For the present however this must remain a hypothesis; nothing can yet be said with certainty about either the composition or the function of the particles with a diameter of 10 nm. Another point that ought to be mentioned is that the cell organs responsible for oxidation processes in the cell, the mitochondria, show similar structures when examined with the electron microscope.

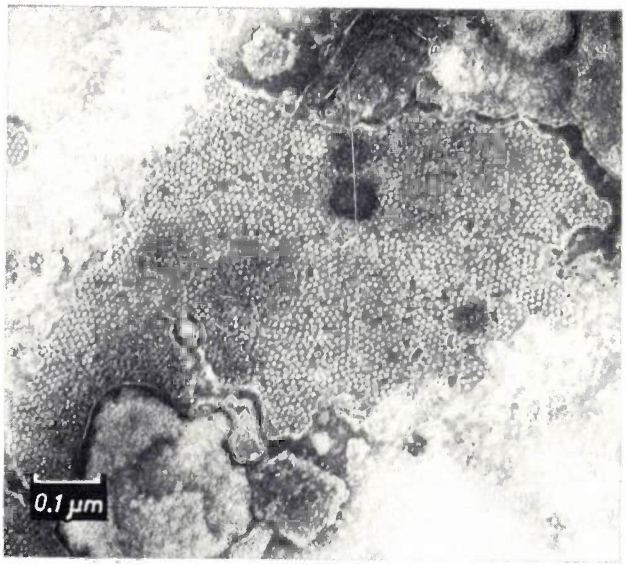


Fig. 7. A chloroplast lamella. Photograph made with the electron microscope by applying the negative-staining technique. The lamellae are here seen not on edge, as in fig. 4, but from above. In the photograph particles with a diameter of about 10 nm can be seen on the membrane of the lamella. These particles are thought to consist of four sub-units.

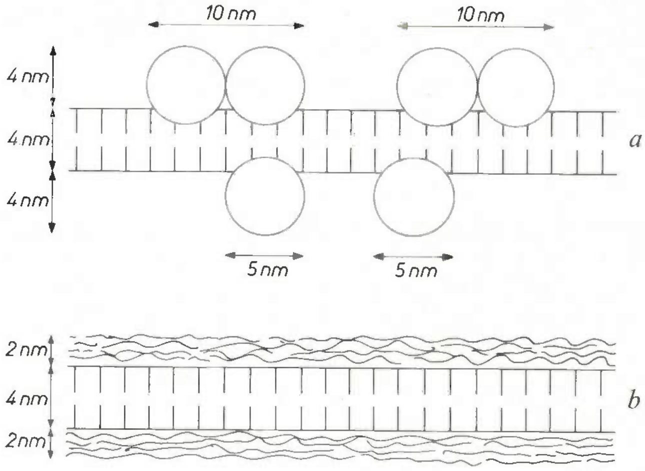


Fig. 8. It is assumed that the protein particles consisting of four sub-units, made visible by the negative-staining of freeze-etching techniques (*a*), unroll because of denaturing of the protein when fixed with $KMnO_4$ or OsO_4 (*b*).

The light reactions

Chlorophyll in the excited state

The sunlight is absorbed in the plant by the chlorophyll. This pigment, which exhibits an absorption maximum in the red as well as in the blue wavelength region is converted to the excited state by the sunlight. At room temperature the excited chlorophyll molecule is unstable; its life is 10^{-8} s. The excitation energy can be given up in the following ways.

a) Transfer of the excitation energy to neighbouring chlorophyll molecules.

b) Fluorescence, i.e. the emission of light. Since this light emission always takes place from the lowest vibration level of the excited state, the emitted quantum of light will have a longer wavelength than the absorbed quantum of light (cf. fig. 9). The fluorescent spectrum of chlorophyll is therefore shifted a little towards the infra-red compared with the absorption spectrum.

c) Loss of energy, without radiation, in the form of heat, e.g. by collision with other molecules.

d) The excitation energy can be used to set up a chemical reaction, bringing about a more or less permanent change in the molecule. It is understandable that this last possibility is the most important one for photosynthesis.

As we have said, chlorophyll has two absorption maxima in the visible wavelength region, at 664 and 431 nm in alcoholic solution; the molecule can be excited in principle by either a red or a blue light quantum. In the case of absorption of a blue quantum, however, part of this excitation energy is lost within 10^{-11} s in the form of heat, and the molecule then attains the same energy state as that resulting from the absorption of a red quantum. This excited state, the S_1 or first singlet state (*fig. 9*), has as we have already seen, a life of 10^{-8} s. Through the loss of energy unaccompanied by the emission of radiation, the excited electron can fall back from the singlet state to a third energy level T_1, the triplet state, which has a relatively

[4] K. Mühlethaler, H. Moor and J. W. Szarkowski, Planta **67**, 305, 1965.

long life of 10^{-3} s. Because of the long life of the triplet state it was for a long time believed that it was the excitation energy of this state that was used for chemical conversion. Recently, however, evidence has come to hand that various photochemical reactions originate in the singlet state. Energy can also be given up from the triplet state as light; in this case, because of the delay in the re-emission of the light, the phenomenon is called phosphorescence.

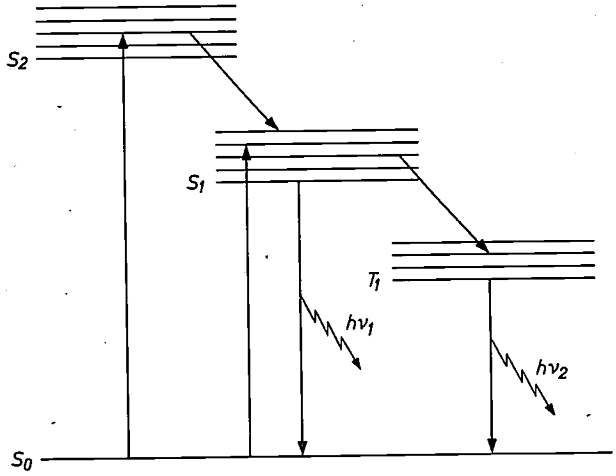


Fig. 9. The energy levels S_0, S_1, S_2, and T_1 of chlorophyll. The molecule can be excited from the ground state S_0 to the second singlet state S_2 and to the first singlet state S_1 by excitation with a blue and a red light quantum respectively. The life of S_2 is very short, and within 10^{-11} s the molecule attains the state S_1. From here it can reach the first triplet state T_1. Apart from the bringing about of a chemical reaction, the excitation energy can also be used for emission of fluorescent and phosphorescent radiation (quanta $h\nu_1$ and $h\nu_2$, respectively).

Photochemical reactions of chlorophyll

Photochemical reactions with chlorophyll in pyridine solution have been investigated in great detail by a group of workers in Russia led by Krasnovsky [5]. They demonstrated that chlorophyll under the influence of light could catalyse various oxidation-reduction reactions. Basically, oxidation and reduction are respectively the removing of electrons from an electron donor and their addition to an electron acceptor. Since these workers also observed a reduction and oxidation of the chlorophyll itself, they indicated the following scheme for the oxidation-reduction reactions catalysed by the chlorophyll:

1) Chl. $\xrightarrow{h\nu}$ Chl.* (absorption of light gives excited chlorophyll),
2) Chl.* acceptor → Chl.$^+$ + reduced acceptor,
3) Chl.$^+$ + donor → Chl. + oxidized donor.

Sum: acceptor + donor → reduced acceptor + oxidized donor, i.e. transfer of an electron from donor to acceptor. The catalytic function of the chlorophyll in this reaction is thought to be as follows: in the excited chlorophyll molecule there is an electron at a higher energy level, and this is therefore less strongly bound; this means that this electron can be easily transferred to an electron acceptor, so that the chlorophyll becomes positively charged (Chl.$^+$) and can in turn be reduced by an electron donor.

Another possibility is that the excited chlorophyll (Chl.*) first reacts with the donor, with the formation of a negatively charged chlorophyll (Chl.$^-$) and that the acceptor then oxidizes the Chl.$^-$ again to Chl. Which mechanism will be preferred depends on the concentration and the type of donor and acceptor and also on the medium in which the reaction takes place.

In connection with the structure of the chloroplasts and the orientation of chlorophyll in the protein-lipid boundary layer the experiments in organic solvents alone cannot be considered as fully characteristic of the role of chlorophyll *in vivo*. Better correspondence with the situation in the chloroplast is to be expected with experiments in which the chlorophyll is located at a lipid-water boundary. We have therefore investigated a number of photochemical reactions, catalysed by chlorophyll, in aqueous solutions of detergents (soaps) [6]. These are compounds built up of a polar and a non-polar part, e.g. cetyl-trimethyl-ammonium bromide, $CH_3(CH_2)_{15}N^+(CH_3)_3Br^-$, and sodium lauryl sulphate, $CH_3(CH_2)_{11}OSO_3^-Na^+$. If the concentration of the detergent in water exceeds a certain limit, the "critical micelle concentration", the detergent molecules coalesce to spherical aggregates called micelles (*fig. 10*). Here the polar part of the molecules, e.g. the -$N^+(CH_3)_3$ group, lies at the surface, i.e. it is oriented towards the aqueous phase, while the non-polar part, generally a long hydrocarbon chain, faces inwards. Since chlorophyll, which is insoluble in water, is also partly polar and partly non-polar, it can be taken up by such a micelle of detergent and thus "solubilized". In this process the porphyrin ring will orient itself on the surface of the micelle.

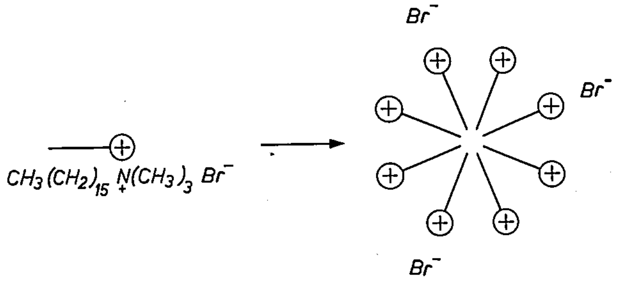


Fig. 10. Detergent or soap molecules have the tendency to unite into spherical micelles when above a certain concentration, the critical micelle concentration. A molecule of detergent consists of a non-polar part (thin line and a polar part ⊕, each of which tries to find a similar medium. In the above sketch it is assumed that the solvent is polar, e.g. water.

It turned out that chlorophyll-containing micelles are also able, on being illuminated, to catalyse many oxidation-reduction reactions. These systems can thus be considered as simple models for the functioning of photosynthesis. We can obtain a better understanding of the mechanism of the relevant photochemical reactions by varying the detergent, the donor, the acceptor, the *p*H, etc. and by studying the kinetics of these reactions. We must however realize that the structures occurring in the chloroplast are more complicated than these model systems. Thus we know, for example, that chlorophyll *in vivo* is linked to a protein and that the red absorption maximum of chlorophyll is, as a result of this, shifted about 15 nm towards the longer wavelengths (678.5 nm).

Although, as we have seen, many oxidation-reduction reactions can be made to take place *in vitro* with the help of chlorophyll, it has not yet been found possible to do this with the same substances as those that occur in the living cell. The cause of this failure must presumably be attributed to the fact that chlorophyll *in vivo* is bound to proteins.

The photosynthetic unit

In vivo the arrangements for photosynthesis have a special structure permitting reasonably rapid photosynthesis even at very low light intensities. The reason for this appears to be that the excitation energy can be transferred from one chlorophyll molecule to another, until finally a certain chlorophyll molecule specially suitable for bringing about a chemical reaction is excited. This special chlorophyll molecule, which may differ from the other chlorophyll molecules, for example, by being bound to an electron acceptor or an electron donor, is known as the reaction centre (*fig. 11*). The energy quanta intercepted by the chlorophyll molecules are thus transferred continuously while the centre is engaged in processing them one by one.

The reaction centre plus the chlorophyll molecules able to transfer their excitation energy to it is known as a photosynthetic unit. The existence of a photosynthetic unit was postulated as long as 30 years ago, after it had been found that on illumination with a very strong, brief flash of light, causing the excitation of all the chlorophyll molecules in the plant, the number of carbon dioxide molecules fixed amounted to only 1/2000 of the number of chlorophyll molecules present. In the course of investigations on the herbicide DCMU (dichloro-phenyl-dimethyl urea), nowadays used quite widely in research on photosynthesis as a specific inhibitor of light reactions, we found that this substance was active in a molar concentration of about 1/300 of the chlorophyll concentration [7]. From this we can conclude that only one in a few hundred chlorophyll molecules is able to catalyse a chemical reaction. The function of the other pigment molecules is then exclusively the absorption and transfer of light energy to the reaction centre. Though a correlation between the photosynthetic unit and the particles with a diameter of 10 nm on the membranes of the lamellae has been suggested, it is certainly not proved.

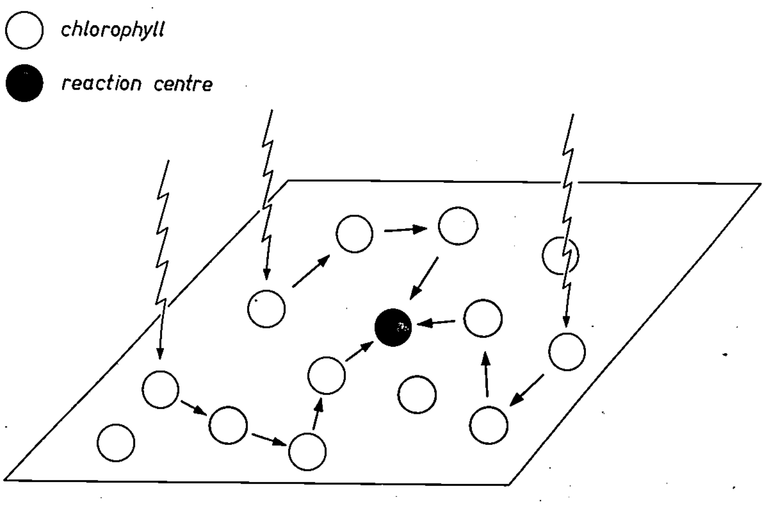


Fig. 11. Diagrammatic representation of the primary transfer of energy in photosynthesis. Excited chlorophyll molecules pass the excitation energy on to neighbouring molecules. One in about three hundred chlorophyll molecules, the "reaction centre", is able, very likely on account of its special bonding to donor and acceptor molecules, to utilize the excitation energy for an oxidation-reduction reaction.

To function as an efficient energy trap, the excited reaction centre has to have a lower energy than the other excited chlorophyll molecules. This implies that the reaction centre must absorb at a longer wavelength than the other chlorophyll molecules. Indications have in fact been obtained from absorption measurements which point to the existence of a pigment absorbing at 700 nm and therefore called P 700 for short. These measurements also give a ratio of 1 to 300 between the number of reaction centres and the number of chlorophyll molecules [8].

Calculation of the number of light quanta required for photosynthesis

In the light process in photosynthesis a compound NADP is reduced by H_2O to $NADPH_2$. During this process, as we have already seen, light energy is con-

[5] A. A. Krasnovsky, Ann. Rev. Plant Physiol. **11**, 363, 1960.

[6] M. Hertogs and J. S. C. Wessels, Biochim. biophys. Acta **109**, 610, 1965.

[7] J. S. C. Wessels and R. van der Veen, Biochim. biophys. Acta **19**, 548, 1956.

[8] B. Kok, Biochim. biophys. Acta **22**, 399, 1956 and **48**, 527, 1961.

verted into chemical energy — in this reaction the electrons are, as it were, pumped "upwards". Compounds occurring both in an oxidized and in a reduced form (e.g. NADP-NADPH$_2$) have a certain oxidation-reduction or redox potential E$_0$, which is characteristic for the particular redox system. Strongly oxidizing substances, such as O$_2$ (redox system O$_2$-H$_2$O) and FeCl$_3$ (redox system Fe^{+++}—Fe^{++}) possess a high redox potential, while strongly reducing substances, such as H$_2$ (redox system H$_2$O-H$_2$) and NADPH$_2$ (redox system NADP-NADPH$_2$) have a low redox potential. If two redox systems are combined, the system with a high redox potential will tend to oxidize the system with a lower redox potential; for example, O$_2$ oxidizes NADPH$_2$. The change in the free energy ΔF^0 in this reaction is proportional to the difference in redox potential of the two redox systems and is given, under standard conditions, by:

$$\Delta F^0 = -nN_{\mathrm{A}}e\Delta E_0, \quad . \; . \; . \; . \; . \; . \; . \quad (4)$$

where n is the number of electrons per molecule transferred in the reaction, N_{A} is the number of molecules per gram-equivalent (Avogadro's number), and e is the charge of the electron.

Given that at pH = 7 the system O$_2$-H$_2$O has the redox potential 0.81 V, and the system NADP-NADPH$_2$ the redox potential —0.32 V, the above formula makes it possible to calculate that in the reaction

$$\mathrm{NADPH_2} + \tfrac{1}{2}\mathrm{O_2} \rightarrow \mathrm{NADP} + \mathrm{H_2O} \quad . \; . \; . \quad (5)$$

the free energy changes by $\Delta F^0 = -52$ kcal. Since in photosynthesis the reverse process takes place, the light will thus have to supply 52 kcal for reducing 1 mole of NADP. From the fact that the energy of a red light quantum absorbed by chlorophyll amounts to only 42 kcal (calculated per mole) we can conclude that for the reduction of NADP by H$_2$O at least two light quanta are required. We shall now look at the way in which the energy of this number of two or more light quanta is used to make this reaction take place.

The splitting of the light process into two separate processes

About six years ago Emerson obtained the following results when measuring the quantum efficiency of photosynthesis over the whole of the visible region of the spectrum [9].

a) Monochromatic red light at wavelengths greater than 685 nm was found to be very ineffective for photosynthesis, and light at wavelengths greater than 700 nm gave no photosynthesis at all, even though chlorophyll still absorbs above 700 nm.

b) The addition of light at a wavelength below 685 nm increased the efficiency markedly, while the wavelength region in which photosynthesis takes place was extended further into the red.

These results can be explained if we assume that there are two pigment systems in the plant with different absorption spectra and that both must be excited for complete photosynthesis. Thus, if illumination takes place with red monochromatic light (e.g. 710 nm), which is absorbed by a "long-wave" pigment system, which we shall call P.S. I, but not by a "short-wave" pigment system (P.S. II), then complete photosynthesis cannot take place. The addition of light of shorter wavelength which P.S. II does absorb will then greatly increase the efficiency of the photosynthesis. Partly on the basis of these experiments, the following mechanism was postulated for the transfer of electrons from H$_2$O to NADP in the photosynthetic light reactions (*fig. 12*). The energy absorbed by P.S. II is used for splitting water, so that O$_2$ is liberated and a substance X is reduced, which has a redox potential between that of water and that of NADP. The energy absorbed by P.S. I is then used for transferring electrons from XH$_2$ to NADP. The energy necessary for the reduction of NADP is thus supplied by two light reactions in series, catalysed by the pigment systems I and II.

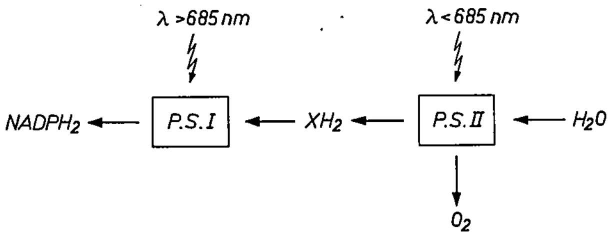


Fig. 12. For the reduction of NADP by H$_2$O, at least two light quanta are needed, which are absorbed by two different pigment systems P.S. I and P.S. II. The reduction takes place in at least two stages: the reduction of NADP by an intermediate substance, provisionally indicated by XH$_2$, and the reduction of this substance by H$_2$O.

Since the plant contains two green pigments, chlorophyll a and chlorophyll b, and since chlorophyll b absorbs at a shorter wavelength (650 nm) than chlorophyll a (678.5 nm), the obvious assumption was that P.S. I contained chlorophyll a and P.S. II contained chlorophyll b. From action spectra, however, it was found that this was not correct. Both pigment systems contain chlorophyll a and chlorophyll b, and the reason for the difference in the absorption maxima of P.S. I and P.S. II is not known at present.

The tracing of the electron carriers in the light process

We have seen above that chlorophyll can catalyse oxidation-reduction reactions and that in the photosynthetic light process electrons are transferred from

H_2O (via a number of intermediate steps) to NADP, when finally $NADPH_2$ and O_2 are formed. In principle there are two methods of investigating which compounds act in this process as intermediate electron carriers.

1) The first method is to try first to destroy the NADP-photoreducing activity of chloroplasts, by extraction with water, detergent solution, or petroleum ether, etc. and then to try to isolate from the extract a redox component which specifically restores the activity.

2) The demonstration of oxidation or reduction of certain components of the chloroplasts by measurement of changes in absorption during illumination.

Let us first look in a little more detail at these absorption measurements, since they have contributed considerably to our present knowledge of the process of photosynthesis. Since most substances have a different absorption spectrum in the reduced state from that in the oxidized state, we can in principle determine what substances are oxidized or reduced under the influence of light by measuring the changes in absorption occurring during the illumination of plant material (e.g. algae) or chloroplasts. A condition is, of course, that we know the absorption spectrum of the relevant compounds occurring in the chloroplast. Knowledge of the chemical composition of the chloroplast is therefore a prime requirement for characterizing the changes in absorption.

Absorption measurements are performed with the help of the equipment represented diagramatically in *fig. 13*. This consists basically of an ordinary spectrophotometer arranged so that the chloroplasts also can be illuminated from the side to bring about excited states, chemical conversions, etc., so that the accompanying changes in absorption can be measured. An objection with this method is that compounds *not* localized in the chain of electron transport but in equilibrium with it may also give rise to changes in absorption on being illuminated. As the former (extraction) method has the objection that one is no longer dealing with intact organs, a combination of both methods is generally required in order to prove with certainty that a substance is acting as electron carrier in the photoreduction of NADP.

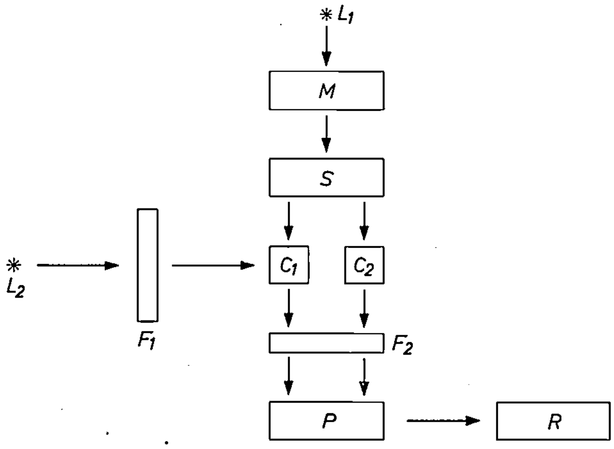


Fig. 13. A spectrophotometer with special provisions enabling measurement of absorption changes resulting from excitation. The excitation of the molecules in cuvette C_1 is achieved by light of high intensity emitted by light source L_2, from which by means of the filter F_1 a certain wavelength region is filtered out. Cuvette C_2 contains an unilluminated specimen serving as control. The measured quantity is the difference in light absorption between the two cuvettes; the two light signals are converted into electrical signals in the photomultiplier tube P and then subtracted from each other. L_1 light source for the measuring beam. M monochromator. S system for splitting the measuring beam. R recording device. The filter F_2 absorbs the scattered light from the exciting beam from L_2.

Up to now the following compounds have been identified as links in the chain of electron carriers: plastoquinone, cytochrome-f, plastocyanin, ferredoxin, and ferredoxin-NADP-reductase. In addition, of course, the reaction centres of the two pigment systems also act as electron carriers, As yet, however, only the reaction centre of pigment system I, the P 700 mentioned above, is known, and even then exclusively on the basis of absorption measurements.

During illumination of algae, decreased absorption was in fact observed at 700 nm, and this was put down to oxidation of the reaction centre P 700. The effect appears to be oxidation as a similar decrease in absorption also occurs in the dark on addition of an oxidizing agent such as potassium ferricyanide. As well as the changes in absorption at 700 nm, there are also changes in absorption at shorter wavelengths, and these changes could largely be put down to oxidation or reduction of known components of chloroplasts.

Fig. 14 shows a scheme for photosynthetic transport of electrons; the various electron carriers found are given with their redox potentials.

By means of the light reaction catalysed by pigment system II oxygen is liberated from water, and electrons (or hydrogen atoms) are carried to the plastoquinone, which therefore corresponds to the X of fig. 12. The function of the plastoquinone was explained a few years ago after it had been found that when chloroplasts were extracted with petroleum ether they lost the power to reduce ferric ions: this could however be restored by the petroleum ether extract. The active compound in this was isolated and identified as plastoquinone.

The following step is the transfer of electrons from plastoquinone to cytochrome-f, which this time is accompanied by a *rise* in the redox potential and therefore by a gain in energy. We shall return shortly to this interesting fact.

Pigment system I next catalyses the electron transfer

[9] R. Emerson, Science **127**, 1059, 1958; R. Emerson and E. Rabinowitch, Plant Physiol. **35**, 477, 1960.

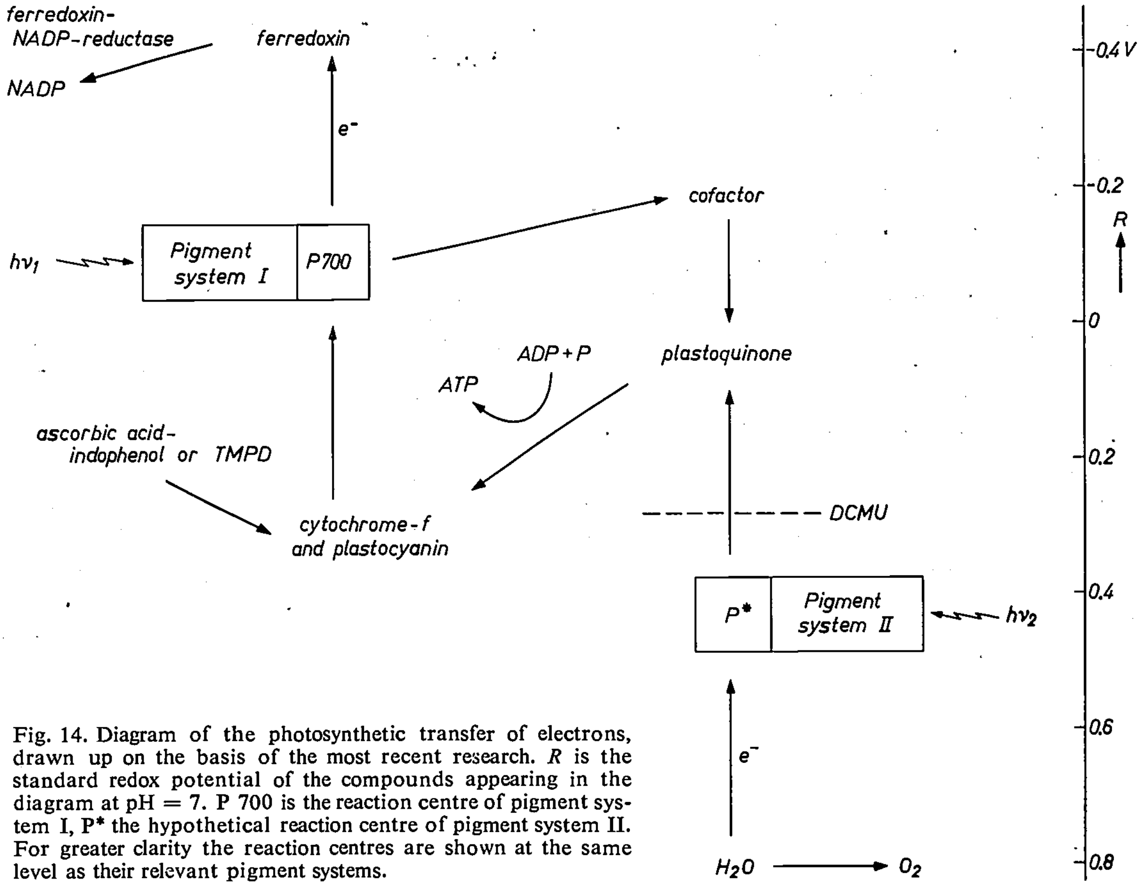


Fig. 14. Diagram of the photosynthetic transfer of electrons, drawn up on the basis of the most recent research. *R* is the standard redox potential of the compounds appearing in the diagram at pH = 7. P 700 is the reaction centre of pigment system I, P* the hypothetical reaction centre of pigment system II. For greater clarity the reaction centres are shown at the same level as their relevant pigment systems.

from cytochrome-f, very probably via plastocyanin, to ferredoxin, which has the lowest redox potential of the components concerned in the electron transport (—0.43 V). The reduction of NADP by reduced ferredoxin is finally catalysed by ferredoxin-NADP-reductase.

The formation of ATP

So far we have said nothing about the way in which ATP is formed in the light process. Energy is required for the formation of ATP from ADP and phosphate. We have just seen how light energy is used in the plant for the reduction of NADP to $NADPH_2$. As demonstrated in 1957 by Arnon this photoreduction of NADP is accompanied at the same time by the formation of ATP, and in fact 1 mole of $NADPH_2$ is formed for 1 mole of ATP:

$$\mathrm{NADP} + \mathrm{H_2O} + \mathrm{ADP} + \mathrm{P} \xrightarrow{\text{light}} \mathrm{NADPH_2} + \tfrac{1}{2}\mathrm{O_2} + \mathrm{ATP}. \quad . \ (6)$$

This process is known as *non-cyclic photophosphorylation.* Since in the transfer of electrons from plastoquinone to cytochrome-f sufficient energy is *set free* for the synthesis of an ATP molecule it seems obvious to assume that the phosphorylation of ADP accompanying the photoreduction of NADP takes place here. This has indeed been established, partly on the basis of our own investigations, which we shall come back to later.

We have already described the above photoreduction of NADP as a light-energized oxidation-reduction process, in which the redox system H_2O-O_2 is oxidized and the redox system NADP-$NADPH_2$ is reduced. The energy supplied by the light makes it possible for the redox system with the lowest redox potential to be reduced, a reaction that without light would run in the reverse direction.

With this picture in mind we can understand a number of different ways in which photophosphorylation can take place both *in vivo* and *in vitro.*

Chloroplasts washed with water and which have lost both ferredoxin and ferredoxin-NADP-reductase, so that they can no longer reduce NADP, are found on addition of a suitable redox system to be still able to give off O_2 and also to produce ATP. In this reaction the added redox system is immediately reduced by pigment system I instead of the ferredoxin. As an example we can take the reduction of ferric salts such as $K_3Fe(CN)_6$:

$$\mathrm{H_2O} + 2\mathrm{Fe^{3+}} + \mathrm{ADP} + \mathrm{P} \rightarrow 2\mathrm{Fe^{2+}} + 2\mathrm{H^+} + \tfrac{1}{2}\mathrm{O_2} + \mathrm{ATP}. \quad . \ . \ (7)$$

This is none other than the Hill reaction mentioned on page 242.

The redox system can also be chosen so that no oxygen is produced and only ATP is formed. The following chain of reactions then takes place: reduction of the redox compound and a spontaneous reoxidation, i.e. taking place without light, of the reduced component by plastoquinone. The electrons transported in this way can then be returned again by pigment system I under the influence of light, via cytochrome-f and plastocyanine, to the redox compound which — on illumination — is therefore alternately reduced and oxidized (see the curved arrows in fig. 14). Since we are dealing here with a cyclic process, in which one ATP molecule is formed per cycle (in the electron transfer from plastoquinone to cytochrome-f) we refer to *cyclic photophosphorylation.* The redox compounds able to catalyse this reaction are called cofactors of the cyclic photophosphorylation. The net result of the process is represented by the equation:

$$\mathrm{ADP} + \mathrm{P} \xrightarrow[\text{light}]{\text{cofactor}} \mathrm{ATP}.$$

The light energy is therefore used here exclusively for the synthesis of ATP. The cofactors are generally redox compounds with a redox potential less than 0.1 V. This is understandable, since only these redox substances can spontaneously reduce the plastoquinone, which has a redox potential of about 0.1 V. A few examples of cofactors of the cyclic photophosphorylation are: vitamin K_3, flavin-mononucleotide (FMN for short) and the phenazine derivatives phenazine methosulphate and pyocyanin. It is not yet clear which substance acts as cofactor of the cyclic photophosphorylation *in vivo*, i.e. in the intact plant. It is possible that ferredoxin plays this role and that the ratio of NADP to $\mathrm{NADPH_2}$ determines whether the electrons are transported from ferredoxin to NADP (non-cyclic process) or back to plastoquinone (cyclic process). The fact that the ferredoxin is washed out during the isolation of chloroplasts in an aqueous environment might then be the reason why a cofactor has to be added to obtain cyclic photophosphorylation *in vitro.*

From an energy standpoint, formation of a second ATP molecule in the cyclic photophosphorylation is possible. Although there are some indications that the transfer of electrons — either from pigment system I to the cofactor or from the cofactor to plastoquinone — is connected with the synthesis of a second molecule of ATP, we are at present not yet certain of this.

We have found that the cofactors of the cyclic photophosphorylation can also be reoxidized by *oxygen* [10]. Taking FMN as an example we have the following mechanism for the reaction:

$$\left.\begin{array}{l} \mathrm{H_2O + FMN + ADP + P} \\ \qquad\qquad \rightarrow \mathrm{FMNH_2} + \tfrac{1}{2}\mathrm{O_2} + \mathrm{ATP} \\ \quad \mathrm{FMNH_2 + O_2} \rightarrow \mathrm{FMN + H_2O_2} \\ \qquad\qquad \mathrm{H_2O_2} \rightarrow \mathrm{H_2O} + \tfrac{1}{2}\mathrm{O_2} \\ \hline \text{Sum}: \mathrm{ADP + P} \xrightarrow[\text{light}]{\text{FMN}} \mathrm{ATP} \end{array}\right\} . \quad (8)$$

The net result is thus the same as in cyclic photophosphorylation, though here there is no question of a truly cyclic process. We therefore call this process *pseudocyclic photophosphorylation.* Whether a cofactor will react with oxygen or with plastoquinone also depends, apart from the presence of O_2, on the concentration of the cofactor and on its affinity for oxygen. If we want to make certain that only the cyclic process occurs then O_2 has to be excluded completely.

Separation of the light reactions

Important corroboration of this scheme for electron transport is offered by the possibility of separating light reactions I and II.

a) From our investigations of the herbicide DCMU, referred to above, we found that this substance was a specific inhibitor of the photosynthetic transport of electrons. It blocks this reaction at a concentration as low as 10^{-7} M; the point of attack is the light reaction catalysed by pigment system II. The result is that in the presence of DCMU no reduction of NADP by H_2O can take place. Photoreduction of NADP can however be made to occur if we substitute ascorbic acid + indophenol for water as the electron donor. This is because the ascorbic acid + indophenol system reduces the cytochrome-f immediately, so that the reduction of NADP can be catalysed by the pigment system I on its own, i.e. it has become independent of pigment system II (see fig. 14). The indophenol serves only to facilitate electron transfer from ascorbic acid to the cytochrome-f tightly bound in the chloroplasts. Tetramethyl-phenylene-diamine (TMPD) can also be used for this. The net reaction can be represented by the equation:

$$\mathrm{AH_2 + NADP} \xrightarrow{\text{light}} \mathrm{A + NADPH_2},$$

in which $\mathrm{AH_2}$ and A represent ascorbic acid and dehydroascorbic acid respectively.

We have shown that, in accordance with the scheme of electron transfer given in fig. 14, the cyclic photophosphorylation as well can still take place in the presence of DCMU [11]. The pseudocyclic phosphorylation, however, which according to eq. (8) depends

[10] J. S. C. Wessels, Biochim. biophys. Acta **29**, 113, 1958.

[11] J. S. C. Wessels, Biochim. biophys. Acta **65**, 561, 1962; J. S. C. Wessels, Proc. Roy. Soc. **B 157**, 345, 1963.

on the photolysis of water and therefore on pigment system II, was completely blocked by DCMU. This inhibitor thus also enables us to distinguish between a cyclic and pseudocyclic photophosphorylation.

The experimentally established fact that the photoreduction of NADP by ascorbic acid-TMPD is *not* accompanied by formation of ATP was one of the important arguments for localizing the phosphorylation between plastoquinone and cytochrome-f [12].

b) The same effect as with DCMU can be obtained if chloroplasts are illuminated with monochromatic light of, for example, 710 nm which is still absorbed by pigment system I but not by pigment system II. With light of 710 nm cyclic photophosphorylation thus still occurs and NADP can be reduced, provided that ascorbic acid-indophenol is added as an electron donor.

c) By treatment with the detergent digitonin, a steroid, chloroplasts can be broken up and solubilized. The lamellae swell, become detached from each other, and constrictions then give rise to vesicles with a diameter of 50 to 200 nm (*fig. 15*). The surface of these vesicles is covered with particles of 8 to 10 nm diameter which are strongly reminiscent of the particles observed on the membranes of the lamellae, but need not be identical with them. We found that the fragments of chloroplasts obtained in this way had completely lost the power to split water and that they were able only to perform the light reaction catalysed by pigment system I [11]. Treatment of chloroplasts with high concentrations of digitonin gave fragments that were still able to reduce NADP with ascorbic acid-indophenol as an electron donor but were no longer able to synthesize ATP by the cyclic photophosphorylation mechanism. Apparently pigment system I is still intact, but the link with the phosphorylation has been lost. Pictures obtained with the electron microscope show that the surface of these fragments was also no longer covered with particles of 8 to 10 nm (*fig. 16*). Whether these particles are concerned with the mechanism of phosphorylation or have another function is at present being investigated.

In the above experiments fragments of chloroplasts precipitated in the ultracentrifuge at 80 000 *g* were always used. However, the supernatant was still found to be coloured green also, and since the chlorophyll-containing particles in it were extremely small — under the electron microscope they no longer exhibited any clear structure — we tried to find out whether it would be possible to bring about a physical separation in it of the two pigment systems I and II. The supernatant was therefore centrifuged through a sucrose gradient, i.e. a centrifuge tube filled with a sucrose solution whose concentration increases linearly with

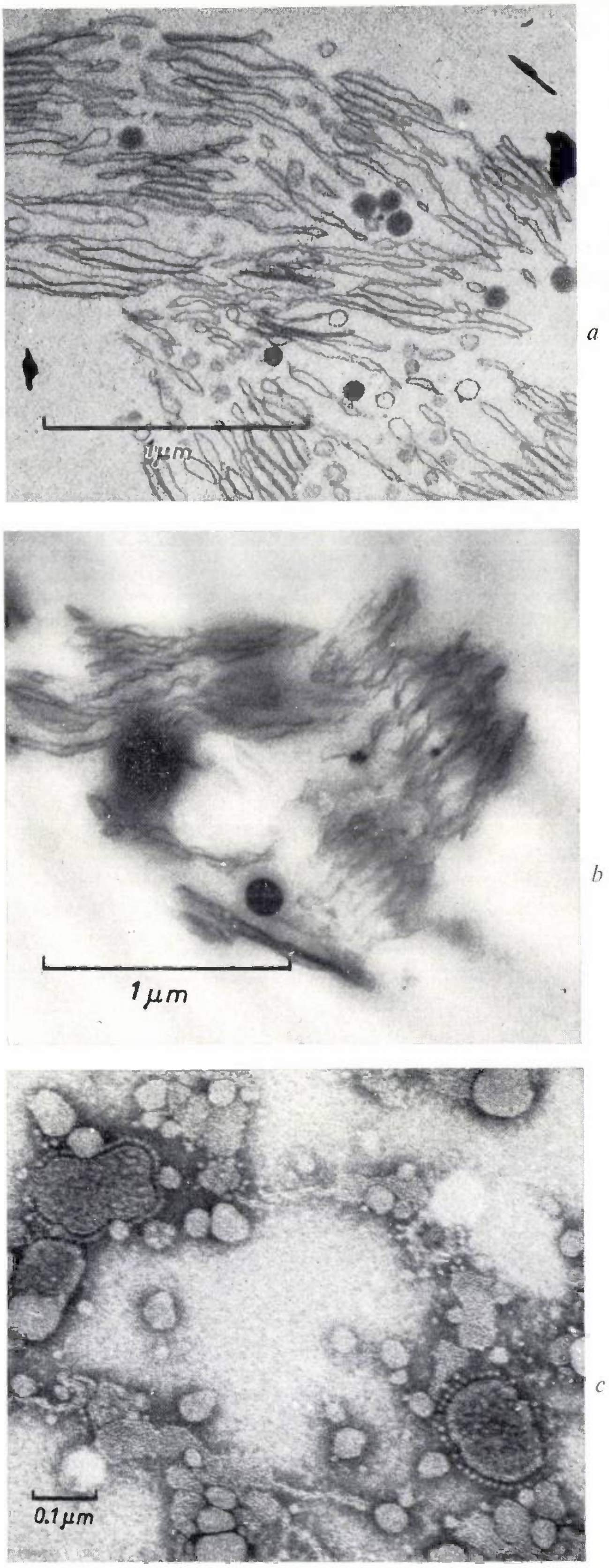


Fig. 15. Electron-micrographs of chloroplasts treated with digitonin, in different stages of treatment. First the lamellae become detached and swell up (*a*), which is followed by a constriction (*b*). Finally separate fragments are formed, on the surface of which particles can be discerned with a diameter of 8-10 nm (*c*). For (*a*) and (*b*) OsO_4 was used for fixing, for (*c*) the negative-staining technique was applied.

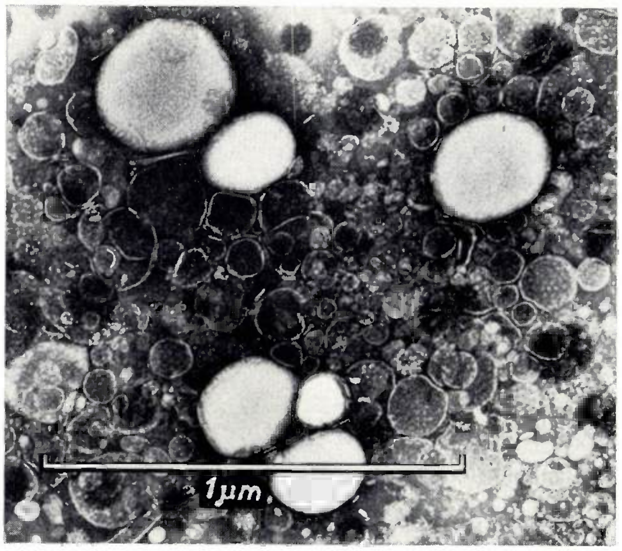


Fig. 16. The separate chloroplast fragments obtained by treatment with much more highly concentrated digitonin. The surface is now not covered by particles such as can be seen in fig. 15*c*.

depth, e.g. from 10 to 30%. If now a small amount of the solution under investigation is carefully added on top of this sucrose solution and the tube is centrifuged for some time (say 40 hours) the various particles of the test solution will finally occupy places where the density of the sucrose solution corresponds to that of the different particles. The result was that after centrifuging two green bands were found which corresponded, so far as their red absorption maximum was concerned, with the two pigment systems I and II [13]. The lower green band absorbed at 678 nm, the upper at 671 nm.

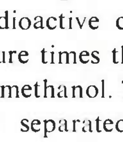
Although the former band now corresponded, in its absorption, with pigment system I, it was found that under illumination this fraction would not reduce NADP with ascorbic acid-indophenol as the electron donor. This was the more remarkable because the solution from which it was separated had in fact been able to do this. To find out whether this fraction had, in centrifuging through the gradient, lost a component essential for the reduction of NADP, the various fractions of the gradient were examined in more detail. This was done by puncturing the centrifuge tube at the bottom with an injection needle, collecting the contents in fractions and analysing them. It was now found that at the top of the gradient there was a component which could completely restore the NADP-reducing activity of the lowest green fraction. This component was isolated and purified and was found to be identical with the copper-containing protein plastocyanin [14]. The activity was not restored by the cytochrome-containing zone which we had found between the two green bands, nor by purified cytochrome-f.

We had thus proved that plastocyanin is essential for the photoreduction of NADP and very probably acts as an electron donor for pigment system I. We are at present engaged on further purifying and analysing the active fraction of the gradient, absorbing at 678 nm, in order to find out what other components apart from chlorophyll are required for the electron transfer from plastocyanin to ferredoxin. We also hope to learn something from this about the chlorophyll-(lipo)protein interaction, which causes a shift of the absorption maximum to 679 nm, and about the nature of the P 700 reaction centre.

The dark reactions

The investigations described in the previous section have enabled us to understand the way in which the light processes yield the products ATP and $NADPH_2$ required to bring the mechanism of the dark reactions into action. This mechanism of the reductive conversion of carbon dioxide into carbohydrates has already been more or less completely explained by Calvin and his co-workers with the help of radioactive tracers. In this research Chlorella, a green alga, was illuminated in the presence of radioactive carbon dioxide, $^{14}CO_2$. After various exposure times the cells were killed and extracted with 80% methanol. The compounds formed from $^{14}CO_2$ were separated by two-dimensional paper chromatography of the methanol extract and then identified. The position of the compounds containing ^{14}C was determined with X-ray film laid over the paper chromatogram. At the places where there were ^{14}C-compounds the X-ray film was blackened by the β-rays emitted by the ^{14}C. In addition the radioactivity of the different ^{14}C- compounds on the paper chromatogram was then measured by means of a Geiger counter. The following aspects were now studied:

a) The appearance of ^{14}C in each compound as a function of the time of photosynthesis with $^{14}CO_2$. Thus it was found that after 30 seconds large ^{14}C-molecules, such as sucrose, were formed, while only smaller molecules, mainly phosphoglyceric acid



had become radioactive after 5 seconds.

b) The distribution of ^{14}C over the various C-positions of the radioactive compounds. This showed that after

[12] J. S. C. Wessels, Biochim. biophys. Acta **79**, 640, 1964.

[13] J. S. C. Wessels, in: Currents in photosynthesis (Proc. 2nd W.-Europe Conf. on photosynthesis, Woudschoten, Zeist 1965), edited by J. B. Thomas and J. C. Goedheer, p. 129, published by Donker, Rotterdam 1966.

[14] J. S. C. Wessels, Biochim. biophys. Acta **109**, 614, 1965.

a short $^{14}CO_2$ fixation time the only atom that was labelled was the carbon atom in the carboxyl group (-COOH) of the phosphoglyceric acid.

The identity of the radioactive compounds can be determined by comparison of their place on the chromatogram with the position occupied by a known substance in the same circumstances. It is also possible to cut the spot out of the paper and to analyse the

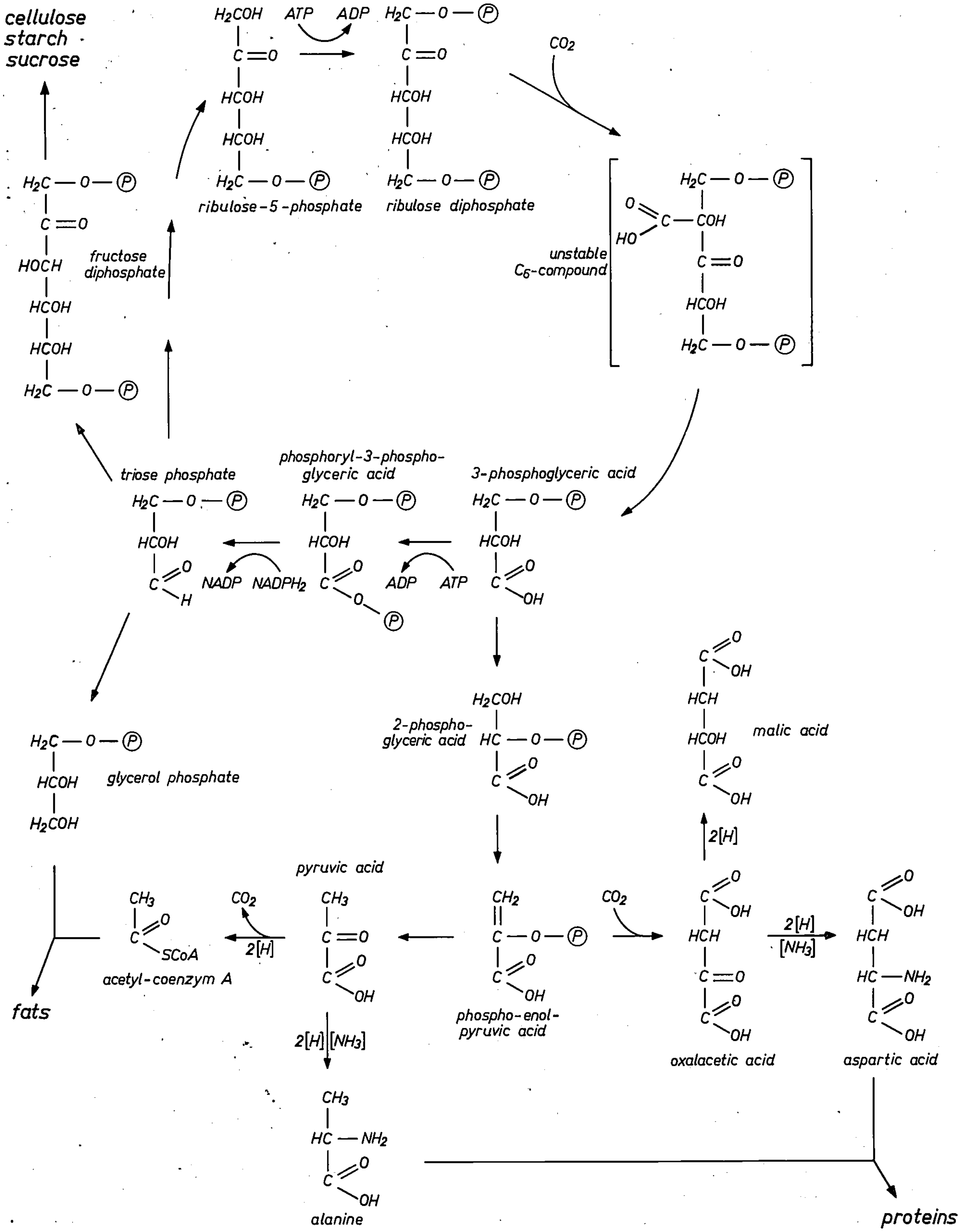


Fig. 17. Simplified Calvin cycle and the related biosynthesis of sugars, starch, fats, proteins, etc.

compound after eluting it with a solvent. It will be readily understood that this was no simple problem, and the many years of this research were rightly rewarded in 1961 by the Nobel prize.

The results of these investigations have led to the formulation of the "Calvin cycle" of photosynthesis which is shown in a simplified form in *fig. 17*. The CO_2 is attached to ribulose diphosphate, a sugar with five carbon atoms, so that an unstable C_6-compound is produced which is immediately split into two molecules of phosphoglyceric acid. Phosphoglyceric acid is reduced, via phosphoryl-phosphoglyceric acid, to a triose phosphate (phosphoglyceraldehyde) by the ATP and $NADPH_2$ formed in the light reactions. Subsequently five of these triose phosphate molecules undergo a series of reactions in which finally three molecules of ribulose-5-phosphate are formed (*fig. 18*). The ribulose-5-phosphate is phosphorylated by ATP to ribulose diphosphate, so that the CO_2-acceptor is re-formed again and a carbon dioxide molecule can once more be fixed. Since each of the three molecules of ribulose diphosphate thus formed can add a CO_2-molecule, so that six C_3-fragments are formed in total, the net result of the Calvin cycle is that three molecules of CO_2 are used per cycle for forming one molecule of an organic compound with three carbon atoms. This C_3-compound can then be used for the synthesis of specialized end products. Two molecules of triose phosphate can, for example, be condensed to a C_6-sugar, and two C_6-sugars to sucrose (ordinary cane or beet sugar). Another possibility is that of linking a number of glucose molecules to form a complex polysaccharide, such as starch or cellulose. The presence of starch in the chloroplasts after illuminating a plant can be clearly shown (*fig. 19*). All these different reactions are catalysed by specific enzymes which are thought to be located in the stroma part of the chloroplast.

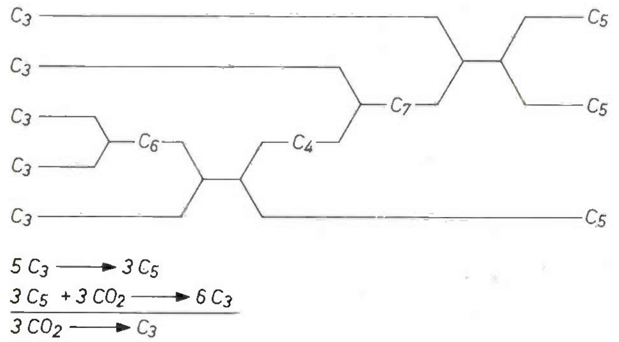


Fig. 18. The net result of the Calvin cycle is that three molecules of CO_2 are used per cycle for the formation of one molecule of a C_3 compound.

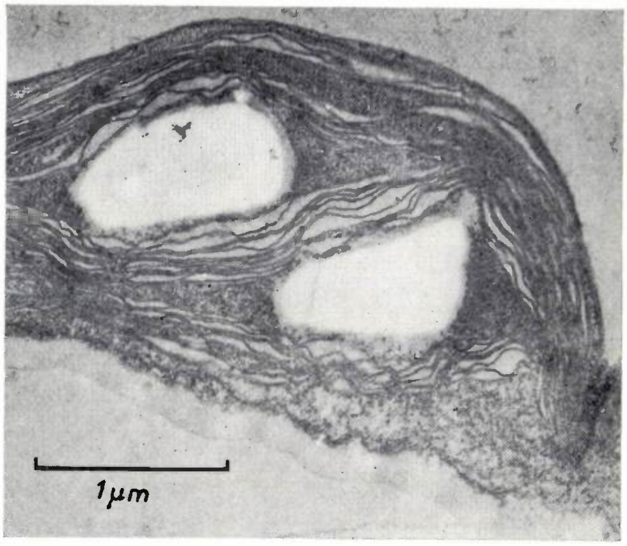


Fig. 19. Chloroplast with two starch grains. Fixation with OsO_4.

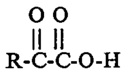
A problem which will have to be dealt with is the extension of the Calvin cycle in order to explain all the experimental results. For example, in certain cases the distribution of the radioactivity over the different carbon atoms of the intermediate compounds discovered cannot be explained by the scheme given. Presumably there is already a partial reduction of the unstable C_6-compound, so that one molecule of phosphoglyceric acid + one molecule of triose phosphate are produced; not two molecules of phosphoglyceric acid. Furthermore the experiments with $^{14}CO_2$ by Calvin and his co-workers have been carried out at much higher concentrations of carbon dioxide than that in air. In particular, if CO_2 concentrations comparable with those in the atmosphere are used, then glycolic acid ($CH_2OH.COOH$), which does not figure at all in the Calvin scheme, is found to appear as one of the first radioactive products of photosynthesis. Further investigations will undoubtedly make the scheme shown in fig. 17 even more complicated.

As well as the building up of carbohydrates other biosyntheses are linked to the Calvin cycle, namely the building up of dicarboxylic acids (e.g. malic acid), fatty acids, fats, amino acids, and proteins.

In these processes the 3-phosphoglyceric acid is converted, via 2-phosphoglyceric acid, into phospho-enol-pyruvic acid, which can now undergo reactions such as:

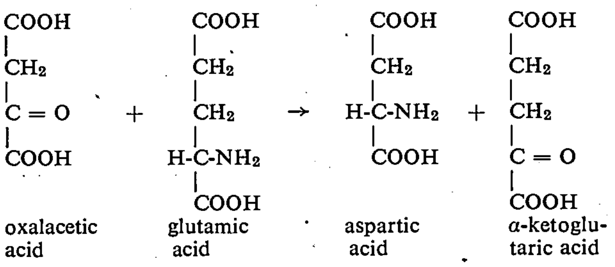
a) Carboxylation to oxalacetic acid, from which malic acid, and the amino acid aspartic acid can be produced.
b) Formation of pyruvic acid, which can be converted into the amino acid alanine or into acetyl-coenzyme A, which serves as a basis for the synthesis of fatty acids. The fats are built up from fatty acids and the glycerol phosphate formed by the reduction of triose phosphate.

The conversion of a keto acid

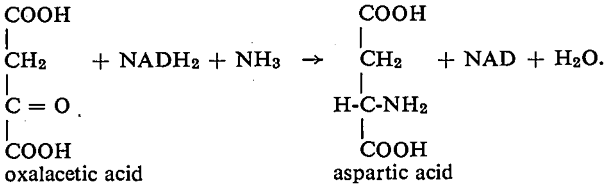


into an amino acid, e.g. the conversion of oxalacetic acid into aspartic acid, can be achieved in the following way:

a) by trans-amination; this is a reaction between a keto acid and an amino acid in which the keto acid is converted into the corresponding amino acid and the amino acid into the corresponding keto acid:



b) by reductive amination; this is a reaction with NH_3 and simultaneous reduction:



The function of the hydrogen donor is performed here by $NADH_2$, which differs from $NADPH_2$ only by containing one phosphate group less.

The nitrogen required for the building up of amino acids is supplied ultimately by the nitrate of the soil, which is reduced by $NADPH_2$ with the help of FMN and the enzymes ferredoxin-NADP-reductase, ferredoxin, nitrate and nitrite reductase to ammonia, nitrite being an intermediate step. No light is required for this reaction. Losada has recently shown that in *illuminated* chloroplasts, the nitrate can also be reduced directly by FMN and ferredoxin, i.e. not via $NADPH_2$ [15]. *Fig. 20* shows the transport of electrons in the light and in the dark to nitrate and nitrite. An analogous mechanism was demonstrated by us for the reduction of organic nitro-compounds [16].

This survey, which is by no means complete, of the possibilities of synthesis shows that the chloroplast is an autonomous system that is able to build up all the amino acids, enzymes, fats and other materials which are necessary for it to function. In recent years it has moreover been found that chloroplasts contain ribonucleic acid (RNA) and deoxyribonucleic acid (DNA), which are different from the RNA and DNA of the cell nucleus. We can conclude from this that the chloroplasts also possess their own system of genetic information.

Photosynthesis research and its value in practical application

The unravelling of photosynthesis, one of the most important processes in living matter, and for this reason an objective of prime importance in itself, also deserves special attention on account of the practical by-products resulting from the work in this field.

For the control of undesired plants it is preferable to use agents which are harmless to humans and animals. It is obvious that substances which act on a process characteristic of the plant should be used. This condition is of course complied with by specific photosynthesis inhibitors, and many of the herbicides at present in use do in fact block this process to a large extent. Examples are CMU (monuron), simazine and DCMU (mentioned above), all of which operate upon the light reaction catalysed by pigment system II. These substances therefore block the photosynthetic transfer of electrons from H_2O to NADP. We are also familiar with compounds which "uncouple" the photosynthetic phosphorylation, which means to say that transfer of electrons does take place from water to NADP but that the formation of ATP accompanying this is now interrupted. As ATP, as well as $NADPH_2$ is required for the dark reactions, it will be clear that in this case also the plant will be killed. Some specific photosynthesis uncouplers are: simple amines such as methylamine (CH_3NH_2), ethylamine ($C_2H_5NH_2$), etc.

On the other hand there are also substances which favour the cyclic photophosphorylation at the expense of the reduction of NADP. Paraquat is thought to act in this way.

Indirect interference with photosynthesis is also a possibility, for example by preventing the building up of chlorophyll or the formation of the chloroplasts. One of the substances which have this action is aminotriazole (amitrole). Recent research has shown that this herbicide primarily attacks an enzyme of the biosynthesis of the amino acid histidine.

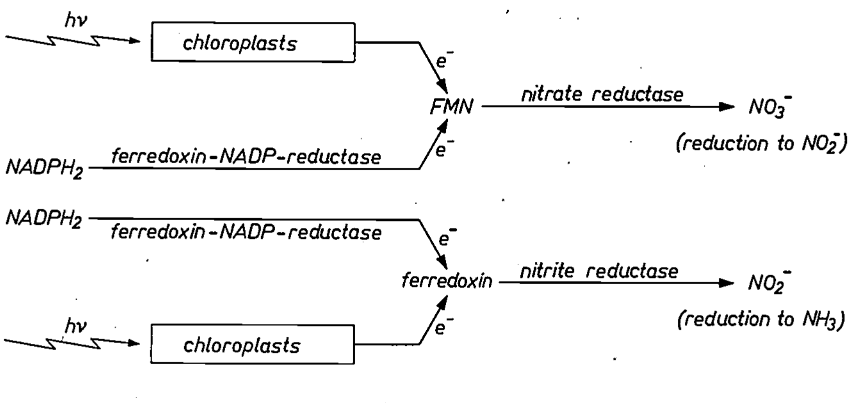


Fig. 20. Diagram of the electron transfer for nitrate and nitrate formation, in the light and in the dark.

Inhibitors of the dark synthesis of carbohydrates, of proteins, and of fats, will in general have little specificity, since analogous processes take place in animal organisms. Specificity can only be expected if the substance acts on a dark process characteristic of the plant, such as reduction of nitrate or synthesis of vitamins. An example here is 2,2-dichloro-propionic acid (dalapon), which inhibits the formation of the vitamin panthothenic acid.

These herbicides will in principle be able to attack every type of plant. It is found, however, that certain plants are less susceptible than others, for example because the particular inhibitor is less easily taken up or transported, or because these plants possess a mechanism for the chemical conversion of the inhibitor into an inactive substance. All these effects together determine the "availability" of the substance. It is through the differences in availability that many herbicides possess a certain degree of selectivity, so that we can use them for the control of certain weeds without damaging the crop itself. We should be able to achieve even greater selectivity if we could utilize a difference in metabolism between the weed and the crop to be protected. For the time being however the difference in availability is the chief factor determining the selectivity of a weed killer.

At present, in conjunction with the Philips-Duphar laboratories, we are examining many substances for their effect on photosynthesis and respiration. Specific agents must inhibit photosynthesis but not respiration. Needless to say, specific inhibitors of photosynthesis thus found are in turn of great value for fundamental research.

In quite a different branch of science, space travel, there is a practical interest in investigations of photosynthesis, as this process could well perform a function in planetary exploration. On long journeys through space the metabolic excretion products of astronauts, CO_2 and water, could be converted by plants (algae) into oxygen and edible vegetable material.

Another interesting discovery — possibly of value in the provision of food in the future — is that the CO_2 concentration in the air (0.03 %) is not optimum and that we can obtain better yields from photosynthesis by increasing the CO_2 concentration.

It is moreover not improbable that man will succeed in finding substances able to direct the photosynthesis of the plant in a certain direction, so that a measure of control can be exercised on the proportion of carbohydrates, proteins, or fats synthesized by the plant.

For the present, however, it looks as though those working on photosynthesis will have plenty to do for many years in removing the many question-marks so liberally distributed around this process. Questions which come to mind are: the detailed mechanism of the photochemical reactions, the difference in the way that chlorophyll acts in pigment systems I and II, the nature of the reaction centres, the mechanism of the light reaction catalysed by the pigment system II and the evolution of oxygen accompanying it, whether there are more locations at which ATP is formed, the mechanism of the link between electron transfer and ATP formation, and of the significance of the morphological structure, etc.

As more experimental results become available the scheme of the photosynthetic reactions of electron transfer here given will no doubt be extended further or modified, but the general principles will probably remain unchanged.

Recommended literature

Historical article: D. I. Arnon, The chloroplast as a complete photosynthetic unit, Science **122**, 9-16, 1955.

Research of Calvin: M. Calvin, The path of carbon in photosynthesis, Science **135**, 879-889, 1962 (Nobel prize lecture).

Article on the photosynthetic phosphorylation: D. I. Arnon, Cell-free photosynthesis and the energy conversion process, in: Light and Life, edited by W. D. McElroy and B. Glass, The Johns Hopkins Press, Baltimore, 1961, pp. 489-569.

Extensive article on photosynthesis in a handbook: H. Gaffron, Energy storage: photosynthesis, in: Plant physiology, edited by F. C. Steward, Part IB, Academic Press, New York, 1960, pp. 3-277.

More profound summary article: A. T. Jagendorf, Biochemistry of energy transformations during photosynthesis, in: Survey of biological progress, Part IV, Academic Press, New York, 1962, pp. 181-345.

The physical aspects of the process of photosynthesis are treated in: L. N. M. Duysens, Photosynthesis, Progress in biophysics **14**, 1-104, 1964.

[15] M. Losada, J. M. Ramírez, A. Paneque and F. F. Del Campo, Biochim. biophys. Acta **109**, 86, 1965.

[16] J. S. C. Wessels, Biochim. biophys. Acta **109**, 357, 1966.

Summary. The process of photosynthesis takes place in the chloroplasts of plants, partly in the light and partly in the dark. In the light an oxidation-reduction process, i.e. a transfer of electrons, takes place with the reduction of NADP to $NADPH_2$ by H_2O as the final result, the H_2O itself being oxidized to O_2. At the same time ATP is formed. The $NADPH_2$ and ATP thus formed are used for the fixation of CO_2 in the dark and the formation of carbohydrates, fats, and proteins. It is shown that two pigment systems play a role in the light reactions. A number of electron carriers, such as plastoquinone, cytochrome-f, plastocyanin, and ferredoxin could be identified by means of the extraction method and by measurement of changes in absorption during illumination. The formation of ATP (photophosphorylation) could be localized. Important evidence in favour of the given scheme of electron transfer has been given by the possibility of splitting the chloroplast into various systems in which component reactions of the light process take place separately. Separation was achieved with the help of inhibitors such as DCMU, with detergents, such as digitonin, and by centrifuging through a sucrose gradient.

The dark reactions have largely been explained by the research of Calvin, which won him the Nobel prize. A survey of these is given and a picture is drawn of the way in which the Calvin cycle is thought to link up with the formation of carbohydrates, fats, and proteins. Finally, the value of photosynthesis research for practical application, as in herbicide development, is dealt with.

Behind-the-ear hearing aids

B. de Boer

534.773.2

Advantage has been taken of the miniaturization of active and passive electronic components to make hearing aids as unobtrusive as possible. One form which is quite popular is the spectacle hearing aid. An even less conspicuous version is the one in which the aid is worn behind the ear in a banana-shaped, flesh-tinted case.

Since the electronic hearing aid was introduced in the thirties, its development has been largely determined by the reduction of physical dimensions. The transition from miniature thermionic valves to transistors some ten years ago marked an important stage in this development [1]. Now another important change is in progress — the transition from circuits with separate transistors to integrated circuits [2]. The first stage made it possible to make aids which do not have to be carried in the pocket but can be worn behind the ear (or built into spectacles). In the near future the second stage will enable the realization of an aid which can be completely concealed *inside* the ear itself.

The replacement of the pocket hearing aid by much smaller aids of this type has, of course, only been made possible because changes in the active elements of the circuit have been accompanied by a significant reduction in the size of the passive elements such as resistors, capacitors, microphones, earphones and batteries, while mechanical and technological developments have contributed to current advances in miniaturization. We are referring here principally to the technique of printed wiring which has made possible the construction of very small switches.

The development which has taken place in the past ten years is illustrated by *fig. 1* in which a pocket hearing aid of 1954 is compared with the behind-the-ear aid which Philips have been producing since May, 1964. Weight and volume have both been reduced by a factor of 15, and yet the two types are completely equivalent in acoustic performance and adjustability. *Fig. 2* illustrates side by side the chief components of both aids. The volume control has undergone the greatest reduction in size — a factor of 30 — while the resistors have only been reduced in size by a factor of 2, as the resistors of 1954 were already quite small.

B. de Boer is with the Hearing Aid Department, Philips Radio, Gramophone and Television Division, Eindhoven.

In this article we shall describe the behind-the-ear hearing aid type KL 6710/01 and also discuss briefly the changes which have been introduced in this aid by the application of an integrated circuit type OM 200.

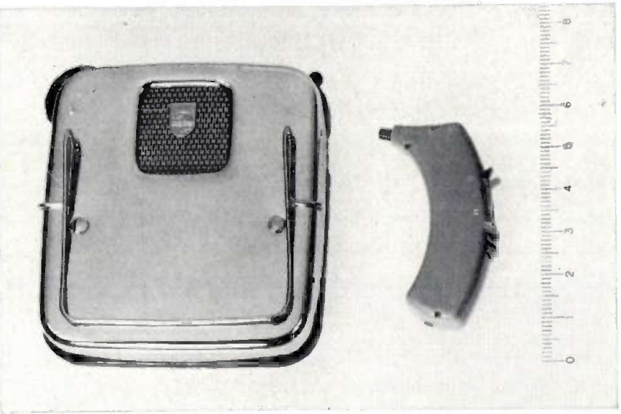


Fig. 1. Hearing aid from the 1954 series (type KL 5500) and the behind-the-ear aid of 1964 (type 6710); these aids are completely identical in the scope of their adjustments and in their acoustic characteristics.

Description of the KL 6710/01 hearing aid

The small dimensions and light weight — only 10 grams — of the hearing aid in question make it eminently suitable for wear behind the ear, at present certainly the most inconspicuous position for a hearing aid. It is adapted for this kind of wear by being housed in a slightly curved, flesh-coloured case.

The aid is suitable for acoustic gains up to 55 dB. If greater amplification is desired a pocket aid must be used. This is necessary if, for example, sound cannot be transmitted via the ear-drums and bone conduction is prescribed.

In contrast to the usual pocket aid method of placing the earphone in the ear, here it is incorporated in the hearing aid case and the output signal is led along an acoustic path to the opening of the auditory canal. The acoustic path is formed by an "elbow" fixed to the top of the aid, a flexible polyvinyl-chloride tube and an earpiece placed in the opening of the auditory canal (see *fig. 3*). This arrangement helps to make the device unobtrusive as the earpiece can be made much smaller than an earphone which is placed completely in the ear. Furthermore, this arrangement offers the possibility of modifying the total acoustic frequency characteristic of the hearing aid, if desired, by a judicious choice of the internal dimensions of the elbow. Three versions of the elbow are therefore available with constrictions at different places (*fig. 4*). The use of different elbows is equivalent to the use of different types of earphone with pocket hearing aids. Various "ready-made" forms of earpiece are available but it is better to use a "made-to-measure" earpiece which is cast after taking an impression. Since this latter earpiece fits exactly into the auditory canal there is no acoustic leak — an important condition in preventing howl. Some earpieces are illustrated in *fig. 5* including one "made-to-measure" example.

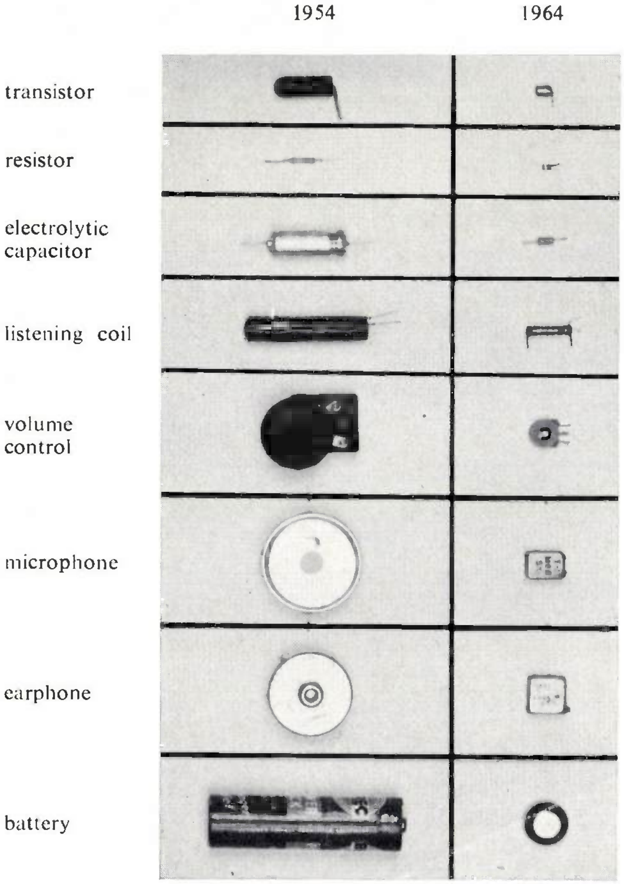


Fig. 2. The chief components of the 1954 and 1964 aids illustrated in fig. 1 are shown in the left and right hand columns.

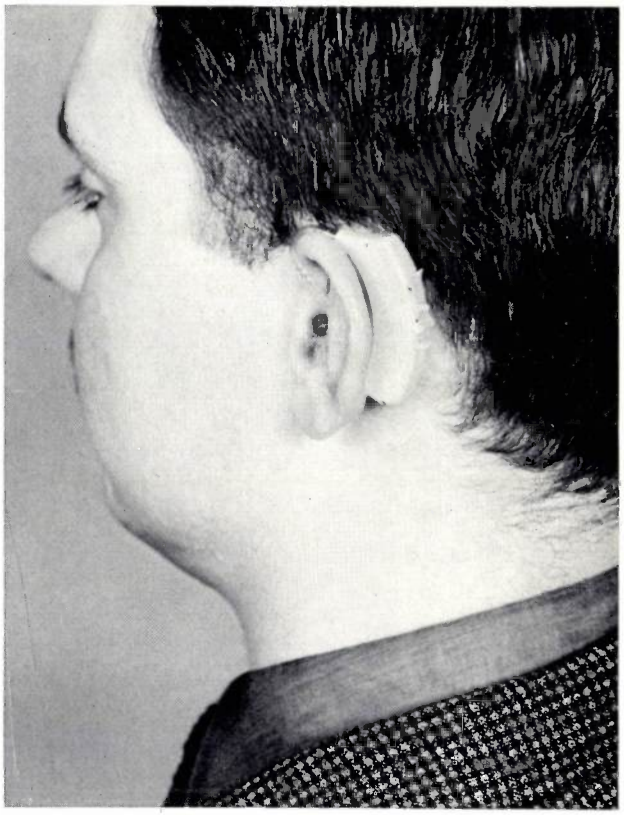


Fig. 3. The hearing aid, type KL 6710 as worn behind the ear. The lever for switching on the battery and listening coil, and the knurled knob for volume control are on the right.

The microphone is in the base of the aid. The acoustic input signal is led to the microphone through a rubber tube via an opening in the case. The microphone and earphone are therefore situated as far as possible from each other in the case and, in addition, are flexibly mounted — measures which have again been taken to prevent feedback and resultant howl.

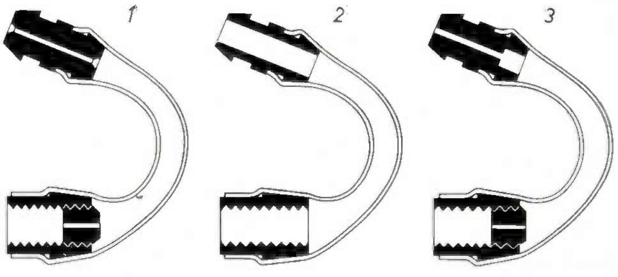


Fig. 4. Sectional drawings of the three optional elbow pieces, approximately twice actual size.

[1] See P. Blom and P. Boxman, A transistor hearing aid, Philips tech. Rev. **19**, 130-139, 1957/58.

[2] These circuits have been described in a series of articles in Philips tech. Rev. **27**, 180-206, 1966 (No. 7).

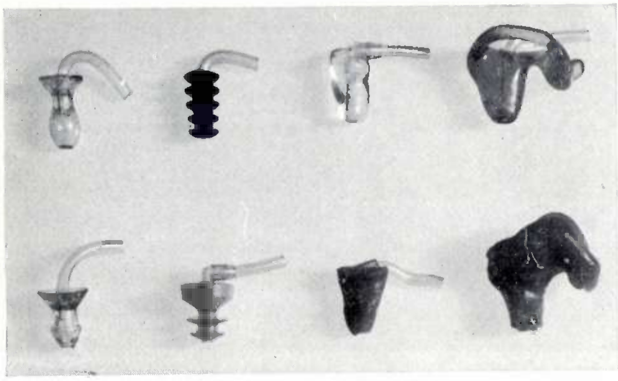


Fig. 5. Some earpieces which can be used with the behind-the-ear aid. An example of a "made-to-measure" aid can be seen bottom right.

Some details of the aid

The circuit diagram is illustrated in *fig. 6*. The aid includes a four stage amplifier with four *P-N-P* germanium transistors. The first three stages (OC 58 transistors) are directly coupled, the output stage (OC 60 power transistor) has capacitor coupling.

The electrical power amplification at battery voltage of 1.3 V is 85 dB; this amplification is, of course, considerably larger than the resulting acoustic amplification (see above) owing to the limited efficiency of the double electro-acoustic conversion and the transmission losses in the acoustic links of the system. The current consumption is 2.5 to 3 mA, depending on the battery voltage. Either a mercury cell, a nickel cadmium accumulator or a silver oxide cell can be used as the battery, as desired. Silver oxide cells give the greatest amplification, have the most even discharge characteristic and are less temperature-dependent, but are somewhat more expensive in use than the other cells. The nickel cadmium accumulator has the lowest capacity. This is no obstacle, however, since it can be charged regularly and a special charging device has been developed for this purpose.

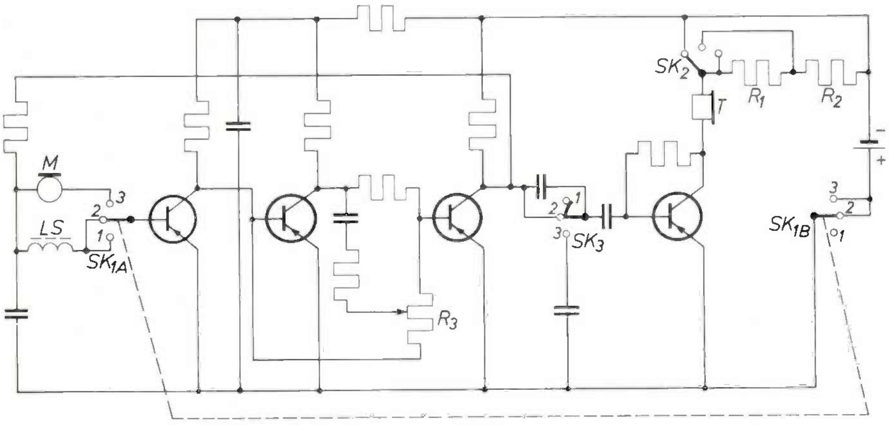


Fig. 6. Circuit of the KL 6710/01 hearing aid. *M* microphone. *LS* listening coil. *T* earphone. SK_{1A} and SK_{1B} form a combined switch: at position *1* the aid is switched off, at position *2* the listening coil is switched on, and at position *3* the microphone is switched on. SK_2 is an attenuation control, SK_3 a tone control: position *1*, low cut-off; position *2*, normal; position *3*, high cut-off. R_3 volume control.

The electromagnetic earphone is connected directly into the collector circuit of the output transistor. By means of a switch (SK_2) the resistors R_2 or $R_2 + R_1$ can be connected into the collector circuit of the output transistor as desired. This reduces the maximum power output of the output stage, which is 0.6 mW at a battery voltage of 1.3 V, by 6 or 12 dB respectively. This makes it possible to avoid exceeding the threshold of pain — this is especially important for sufferers from regression deafness, who have a low threshold of pain [3]. The audiologist, doctor or dealer can adjust the switch with a miniature screwdriver through an opening in the case of the hearing aid.

The electromagnetic microphone is connected directly into the input circuit of the first transistor. Using the switch SK_1 it is possible to switch to a listening coil (*LS*) which can be used when telephoning (elimination of external noises) and for inductive sound transmission in halls by means of loop circuits. The user can operate the switch himself with the small lever which can be seen in fig. 3. It has three positions and operates two electrically separate switches (designated SK_{1A} and SK_{1B} in fig. 6) which are also used for switching off the batteries. By making use of printed wiring in the switch its total thickness is cept down to 1.2 mm.

The tone control SK_3, which can be adjusted in the same way as SK_2, makes it possible to choose from three tone settings: low cut-off, normal, and high cut-off. In conjunction with the three elbows, nine different frequency characteristics in all can therefore be obtained.

The circuit with direct coupling of the three pre-amplifier stages has been specially developed for use in hearing aids and has the advantage of economizing on capacitors and resistors normally used for coupling amplifier stages. At the same time this has the effect of reducing the effect of temperature on the amplification, as *fig. 7* shows. In direct coupling each transistor obtains its d.c. base voltage from the collector voltage of the previous one, while the d.c. base voltage of the first transistor in the pre-amplifier comes from the collector voltage of the last one.

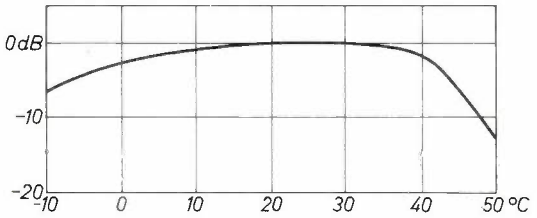


Fig. 7. Relative amplification as a function of temperature, with respect to the amplification at 25 °C, for the KL 6710/01 hearing aid.

It was not possible to use the normal potentiometer circuit for volume control here since the d.c. operating point of the transistors would then have been affected by the volume setting. A rather more complicated circuit was therefore used in which the potentiometer R_3 functions as a volume control in an *RC* network between the second and third stages. Besides causing an alteration of the impedance between the second and third transistors, any variation of R_3 also causes an alteration in the degree of feedback between the collector and base of the second transistor. These two effects reinforce one another and make volume control very effective. The knurled knob of the volume control can be seen in fig. 3. Over three-quarters of the angle of rotation the maximum value of the amplification falls uniformly by 30 dB, while it falls by a further 30 dB over the last part of its travel.

The amplifier is mounted on two resin-bonded laminated paper panels with printed wiring. The microphone, earphone and battery compartment together with the volume control are also mounted on one of the panels (see *fig. 8*). The case consists of two halves screwed together, and the battery holder is fitted between the two halves on a hinged mounting.

The microphone has a balanced magnetic system. A magnetic "Wheatstone bridge" is used for this, its "voltage source" being formed by a small permanent magnet and two of the "resistance arms" by air gaps. The size of these air gaps depends on the position of the microphone diaphragm, so the vibrations of the diaphragm throw the bridge out of balance. This causes an alternating magnetic flux in the "galvanometer" arm of the bridge which is formed by the armature of the microphone. The alternating magnetic flux induces an a.c. voltage in a small coil fitted around the armature. In spite of the small dimensions the sensitivity is still 0.22 mV/μbar [4] at 1000 c/s, while a satisfactory characteristic is obtained (see *fig. 9*). The microphone case is made of mu-metal which provides good protection against interference from external magnetic fields.

The earphone has a similar balanced magnetic system. Its dimensions are somewhat larger than those of the microphone, while the case is again made of mu-metal. At an input signal of 1 mVA at 1000 c/s the earphone produces a sound pressure signal of 123 dB above the normal reference level of 2×10^{-4} μbar (approximately the threshold of hearing, while 123 dB at 1000 c/s lies near to the threshold of pain). To measure this the earphone is connected by a 75 mm long tube to a standardized artificial ear with an air content of 2 cm³, whose acoustic load characteristics approximate closely to those of the human ear in the speech frequency range. The frequency characteristic of the earphone is illustrated in *fig. 10*.

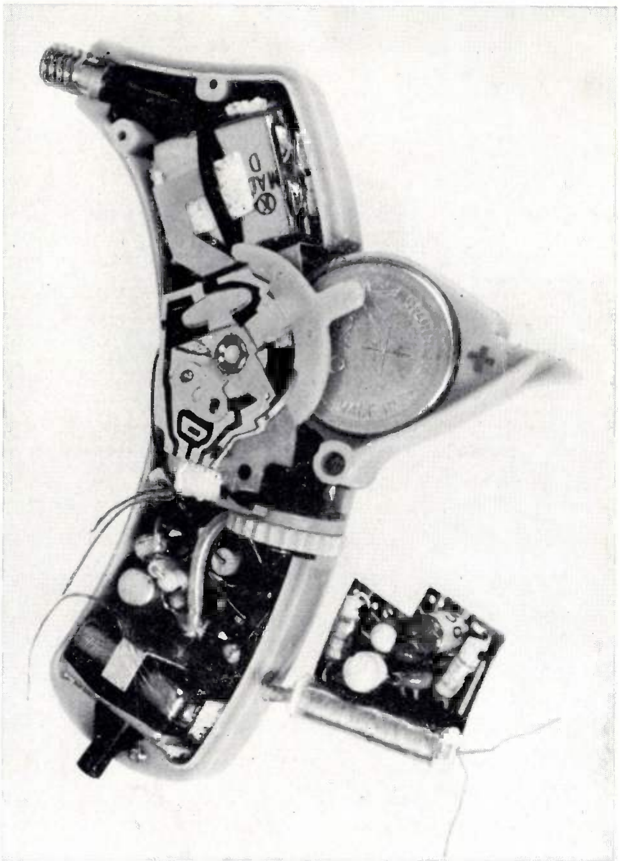


Fig. 8. The interior of the KL 6710/01 hearing aid.

[3] See P. Blom, An electronic hearing aid, Philips tech. Rev. **15**, 37-48, 1953/54.

[4] 1 μbar corresponds to 10^{-1} N/m² (that is approximately 10^{-6} kgf/cm²).

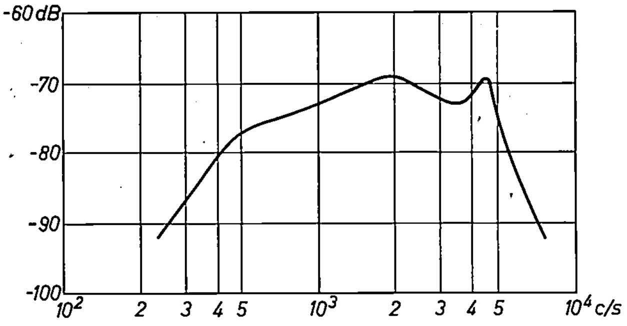


Fig. 9. The frequency characteristic of the microphone with a 5000 Ω load. The level plotted is that of the output signal (in dB with respect to a reference level of 1 volt) which is obtained with an input signal of 1 μbar.

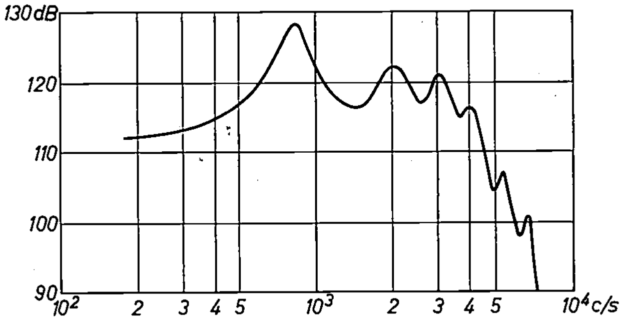


Fig. 10. The frequency characteristic of the earphone acoustically loaded by an artificial ear with an air content of 2 cm³. The level plotted is that of the output signal (in dB with respect to the usual sound pressure reference level of 2×10^{-4} μbar) which is obtained with a 1.5 mA a.c. input signal.

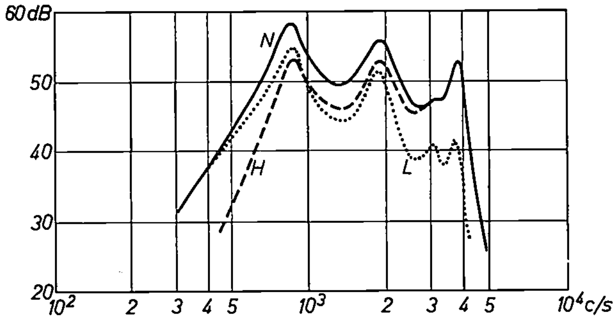


Fig. 11. The three overall frequency characteristics of the KL 6710/01 hearing aid, using the elbow No. 2 and with the tone control SK_3 at the positions "normal" *N*, "high" *H* and "low" *L*. The acoustic amplification is plotted for an input signal level 50 dB above the reference level (2×10^{-4} μbar).

Since the frequency characteristic of the amplifier is almost flat, the total acoustic characteristic of a complete hearing aid is very largely determined by those of the earphone and microphone and is therefore approximately equal to the sum of the two. *Fig. 11* illustrates the three overall characteristics corresponding to the three positions of the tone switch, for elbow number 2.

It must be pointed out that the threshold of pain of the user's ear must never be exceeded. *Fig. 12* gives the maximum output signal of the aid as a function of the frequency. This curve is recorded by increasing the input signal at various frequencies until the output signal stops rising. The magnitude of the output signal (in dB with respect to the reference level of 2×10^{-4} μbar) gives the points through which the curve is drawn. Comparison of this curve with the one showing the user's threshold of pain indicates whether it is necessary to switch in attenuation by means of switch SK_2 in fig. 6.

The use of an integrated circuit

Besides the aid described above, a similar aid has now been put into production (type KL 6710/11) in which the first three amplifier stages with *P-N-P* germanium transistors have been replaced by an integrated circuit on silicon (type OM 200).

Mechanically speaking the two aids are almost identical and electrically they are also very similar, as can be seen by comparing the diagrams in fig. 6 and *fig. 13*. On the silicon substrate of the OM 200 circuit, measuring 0.75×0.75 mm, there are three *N-P-N* transistors and two resistors, which are connected to a 3-stage amplifier, again directly coupled. The gain is controlled in a different way from that in fig. 6, however, as the use of the integrated circuit has made the second amplifier stage inaccessible. Furthermore, the use of an integrated circuit increases the tendency to instability as a result of capacitive coupling; this is suppressed by screening.

The silicon crystal is enclosed in a plastic encapsulation measuring 2.8×2.8×1.1 mm and fitted with four connecting strips.

Besides the three transistors, the integrated circuit also replaces three resistors while an electrolytic capacitor has been added [5] so that the number of components has been reduced by four. This has resulted in a decrease of about 90 mm³ in the volume of the components. This is not a very spectacular gain in space considering that the total volume of the hearing aid is 6500 mm³. At the moment, therefore, the great advant-

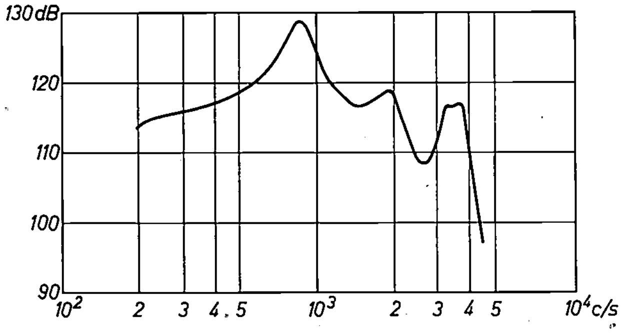


Fig. 12. The maximum output signal of the KL 6710/01 hearing aid as a function of frequency using elbow No. 2 and with the tone control at "normal". The input signal is raised until there is no further alteration in the output signal, shown by the curve in dB with respect to the reference level of 2×10^{-4} μbar.

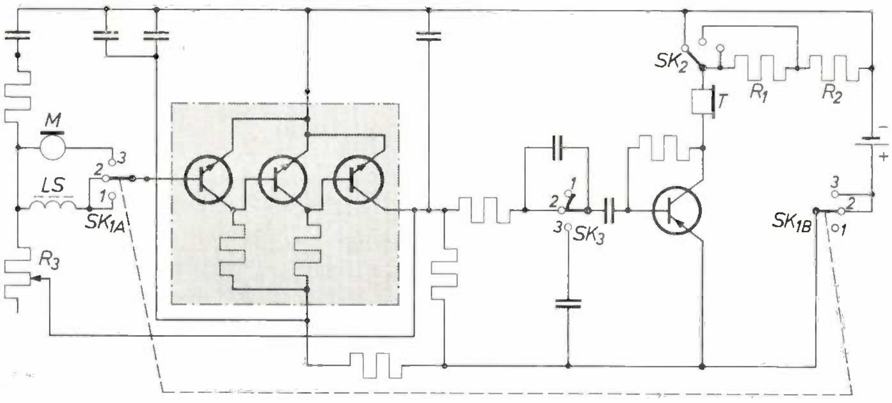


Fig. 13. Circuit of the KL 6710/11 hearing aid. For explanation of the symbols see fig. 6. Here the three transistors of the pre-amplifier form part of an integrated circuit (OM 200), which is shown in grey.

age of using integrated circuits in hearing aids is not so much the gain in space, but the reduction of the number of soldered joints — which gives greater reliability and longer life (with some extra gain in space as a bonus). Assembly is faster and easier, fault detection simpler, and better repairs are possible. Moreover, the use of silicon transistors in place of germanium transistors brings with it the attendant advantages of reduced noise, 13% lower battery consumption and less temperature-variation in the amplification (see *fig. 14*).

Final conclusions

The application of integrated circuits marks a new phase in the progressive miniaturization of hearing aids. Their use is a stimulus to a further reduction in size for the other components of the hearing aid such as microphone, earphone and battery. When dimensions have been sufficiently reduced and an earphone smaller in diameter than the auditory canal is available, an "in-the-ear" hearing aid can be constructed which will be even less conspicuous when worn. This will, however, involve making some concessions with regard to amplification (from 4 stages to 3 stages), battery costs, (smaller and therefore more expensive batteries) and the possibilities of variation (tone controls and attenuation and the listening coil will have to be dispensed with). Broadly speaking, at the moment the in-the-ear hearing aids will be suitable for acoustic amplification up to 40 dB, the behind-the-ear aids to 60 dB and pocket aids for amplification beyond 60 dB. The degree of the wearer's deafness will therefore determine, to a greater extent than in the past, how inconspicuous and comfortable his aid can be.

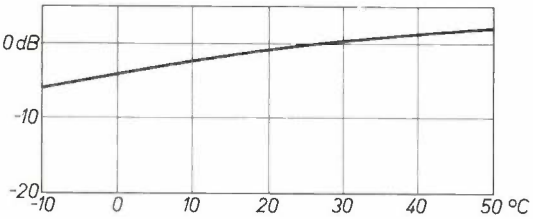


Fig. 14. Amplification of the KL 6710/11 hearing aid as a function of temperature, shown relative to the amplification at 25 °C.

[5] In fact two capacitors are connected in parallel to obtain the correct capacitance value. In principle, however, one capacitor would have sufficed.

Summary. The transition from valves to transistors and the reduction in size of other electronic components have made it possible to develop hearing aids considerably smaller than the pocket aids previously in use. The KL 6710/01 hearing aid is designed so that it can be worn behind the ear, making the aid as inconspicuous as at present possible. The circuit and characteristics of this aid are described. At the same time a survey is given of the advantages obtained by using an integrated circuit in place of the first three amplifying stages with conventional transistors. Until the stage of an "in-the-ear" hearing aid has been reached — for which the earphone and other components will have to be made still smaller — increased reliability and not the space gained is the main argument in favour of this replacement. It may be expected that in-the-ear hearing aids will become available in the near future; due to the anticipated relatively low acoustic amplification of such aids, (e.g. 40 dB) the behind-the-ear aids and pocket aids will also continue to find application in cases of more severe deafness.

Tunable integrated circuits

621.3.049.7:621.37

The use of integrated filter and oscillator circuits, particularly for low frequencies, is greatly limited by the fact that it is difficult to incorporate inductors in them. For some years now attempts have been made to overcome this difficulty by using transistorized circuits in which there are no inductors and in which a suitable feed back ensures that the desired characteristics are obtained — i.e. active *RC* circuits are used consisting exclusively of resistors, capacitors, diodes and transistors. Two problems associated with this are that the characteristics of such circuits depend very much on the temperature and the supply voltage. For example the mutual conductance of a transistor and the differential resistance of a diode are respectively proportional and inversely proportional to the quotient I/T of the current and the temperature. This means that even small variations in the supply voltage or the temperature could make such a circuit impracticable. In particular, the amplification in the feedback loops must be very constant.

We have discovered a combination of technical tricks which enables these difficulties to be mastered and is especially suitable for circuits which have to be tunable. The first trick consists in using a transistorized circuit (*fig. 1*) in which the current I is independent of the voltage and proportional to the temperature as a source of current for the amplifying circuit. The variation in the mutual conductance of the amplifier transistor etc. which occurs due to an alteration in temperature is therefore exactly compensated by the change in current. This characteristic of the current source is obtained by giving the base (b) of the transistor used in it a constant and well-defined bias V_b. The current can be set to the desired level by the variable resistor R_e.

The second trick ensures that variation of I has no effect on the loop gain. We have used a series of diodes instead of a resistor as the load resistance of the transistor in the amplifying circuit (*fig. 2*). These diodes have the same current-voltage characteristic and hence the same differential resistance r_0 as the emitter-base *P-N* junction of the transistor. For the a.c. component i of the current flowing through this *P-N* junction and the load diodes, the whole circuit may thus be considered as a kind of potentiometer circuit; the voltage gain is therefore independent of the current and simply equal to the number of diodes.

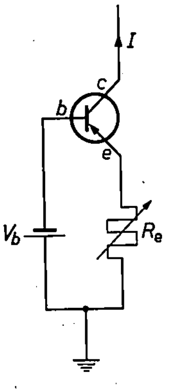


Fig. 1

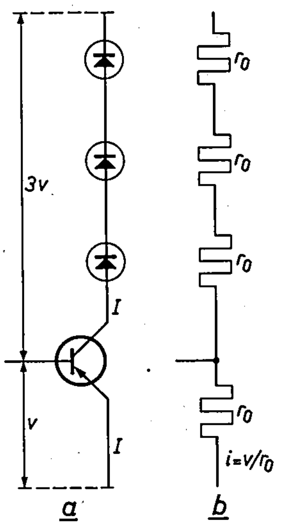


Fig. 2

The resistance of the diodes (including the *P-N* junction in question) is of course dependent on the current — their characteristic is exponential. It is therefore possible to adjust an *RC* circuit in which the circuit of fig. 2 is employed, whilst retaining the correct gain in the feedback loop, simply by altering the current I by means of R_e (fig. 1).

Very stable active filters and oscillators can be made by means of the two tricks described. As an example of the circuits we have already developed, we will refer to a filter for the medium waveband which is tunable between 0.5 and 1.5 Mc/s (quality factor Q between 70 and 80) and an oscillator which can be tuned electrically from 20 to 20 000 c/s. If desired, several circuits can be tuned simultaneously using only one variable resistor R_e.

Filters of the type described above generate much more noise than say an *LC* circuit. In addition — at least as far as the tunable filters are concerned — the maximum permissible signal strength is relatively low, since the characteristics of diodes and transistors are very non-linear. Consequently the extreme limits of the range within which the signal strength can vary are relatively close to one another.

All things considered, application of these filters will for the time being be chiefly confined to cases in which the rather less satisfactory noise and dynamic range characteristics do not cause too much inconvenience and in which Q need not be too high, but where the possibility of electrical tuning is of great importance. On the other hand, there are many possible applications for the oscillators, and they have already been realized in experimental form, for example as generators for telephony and for special-purpose measuring-oscillators. The oscillators are also suitable for applications in frequency modulation and telemetering, etc.

A. J. W. M. van Overbeek
W. A. J. M. Zwijsen

A. J. W. M. van Overbeek and W. A. J. M. Zwijsen are with Philips Research Laboratories, Eindhoven.

Dimming fluorescent lamps

J. C. Moerkens

621.327.534.15:621.382.233:621.387.132.22

Recent developments in the equipment for controlling the luminous intensity of fluorescent lamps are largely responsible for the present increasing application of these lamps in situations where a continuously variable light level is desired — a field in which formerly incandescent lamps were used almost exclusively. Application of this type of dimmer is not confined to indoor projects; lighting installations with fluorescent lamps are also used at present in road tunnels, where the light level required in the day-time depends upon the outside light level.

Introduction

A continuously variable light level is required for many kinds of lighting projects. This is the case in theatres, cinemas, lecture halls, etc. This requirement is not so easily met with fluorescent lamps as it is with incandescent lamps. As early as 16 years ago there was an article in this journal which dealt with the problems arising in continuously variable control of the luminous intensity of fluorescent lamps [1]. The equipment then described used thyratrons (relay valves). Although this system allows gradual control of the luminous intensity over a wide range, it has not been employed very extensively. The main reasons were the fairly large dimensions and the high cost of the devices necessary for control, as well as the fact that these dissipated considerable power. The availability of silicon controlled rectifiers (thyristors) has made possible the development of simpler, smaller and cheaper control equipment. In addition, recent improvements in the circuit have removed a serious difficulty which was present in the older equipment. We refer here to the tendency to uneven operation (flicker) which appeared when fluorescent lamps were adjusted to give a low light output. The causes of this, and the measures taken to prevent it, will be discussed in the following paragraphs. First, however, we shall briefly recapitulate the principle on which the control of the luminous intensity of fluorescent lamps is based, and also the operation of thyristors.

Controlling the luminous intensity of fluorescent lamps

The control methods which can be used for varying the luminous intensity of incandescent lamps, i.e. series resistors or variable transformers, cannot be used with gas-discharge lamps such as fluorescent lamps. The main reason is that the lamp goes out completely at a voltage lower than the re-ignition voltage. Moreover, where a number of lamps are controlled simultaneously, a very undesirable effect can occur, as the re-ignition voltages of different lamps of the same type may vary considerably, with the result that not all the lamps go out at the same time. A method of control much more favourable in this respect is one in which the voltage is not reduced, but is applied to the lamp for only a fraction of each half-cycle. Variation of this fraction controls the mean lamp current, and hence the luminous intensity. As we mentioned above, in modern equipment thyristors are used for "cutting out" a part of each half-cycle.

Thyristors

The operation of a thyristor [2] corresponds in many respects to that of a thyratron (relay valve). When the anode has a negative voltage with respect to the cathode, the anode current is virtually zero. If a positive voltage is applied, there is still no current flow until the rectifier has been "switched on" by an appropriate signal at the control electrode. Once this has happened, the conducting state continues, irrespective of the voltage applied to the control electrode. Only when the anode current has dropped below a certain

J. C. Moerkens is with Philips Lighting Division, Eindhoven.

[1] K.W. Hess and F. H. de Jong, Controlling the luminous intensity of fluorescent lamps with the aid of relay valves, Philips tech. Rev. **12**, 83-93, 1950/51.

[2] For the construction of such rectifiers reference may be made to J. J. Wilting, DC/AC converters using silicon controlled rectifiers for fluorescent lighting, Philips tech. Rev. **23**, 272-278, 1961/62.

very low value, the holding current, does the non-conducting state occur again.

In the conducting state the voltage drop across a thyristor is much smaller than for a thyratron. As a result the power loss is much smaller when thyristors are employed, and therefore less heat is generated in the equipment. This simplification in the problem of cooling, and the much smaller dimensions of the thyristors, allow substantial reductions in the dimensions of the control equipment.

Control with the aid of thyristors

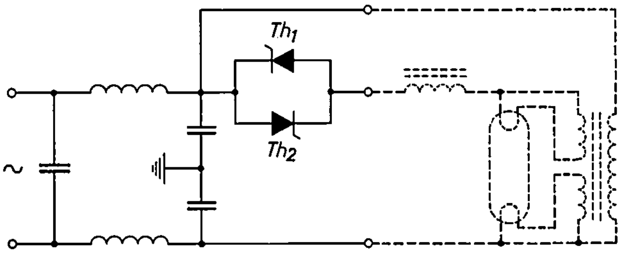


Fig. 1. Anti-parallel connection of two thyristors, Th_1 and Th_2, for dimming fluorescent lamps. The inductors and capacitors on the left form an interference filter. Each fluorescent lamp is provided with a filament-supply transformer with two separate secondary windings.

Since a thyristor conducts only in one direction, alternating current can be transmitted only if two thyristors are connected in parallel in opposite senses (anti-parallel). *Fig. 1* shows such a circuit, with the mains voltage connected to the input and one or more fluorescent lamps connected to the output. Each lamp has its own ballast choke. For normal circuits with no control the cathodes are kept at the correct temperature by the discharge-current itself (ion bombardment). If the mean current through the lamp is decreased in order to reduce the luminous intensity, the cathodes would have too low a temperature, and this is detrimental to the lamp life. For this reason each fluorescent lamp is provided with a filament-supply transformer with two separate secondary windings.

Signals 180° out of phase with each other are applied to the control electrodes of the thyristors, with the result that the thyristors conduct alternately during a fraction of each successive half-cycle. The current curve during this time can be calculated to a good approximation. For the purposes of the calculation we consider the thyristors as ideal switches, which are closed from the instant at which the control signal is applied until the instant at which the anode current is zero again. The maintaining voltage of the lamp, which is virtually independent of the current, is taken to be constant [3] during the conducting time. In *fig. 2*, which shows an equivalent circuit applicable during conduction, the lamp has for simplicity been replaced by a battery of voltage V_l. The thyristor is represented here by the switch *Th*. If we represent the instantaneous value of the mains voltage by $v_n = V_n \sin \omega t$, then the current i which flows after the switch has been closed follows from the equation:

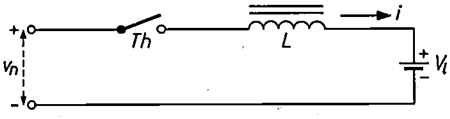


Fig. 2. Equivalent circuit for a fluorescent lamp with ballast choke L and a thyristor, connected to the mains voltage v_n. The lamp is replaced by a battery of constant voltage V_l. The thyristor is represented as a switch *Th*.

$$V_n \sin \omega t = L\frac{di}{dt} + V_l\,.$$

Integration of this equation gives the following expression for i:

$$i = \frac{V_n}{\omega L}(\cos \omega t_1 - \cos \omega t + \frac{V_l}{V_n}\,\omega t_1 - \frac{V_l}{V_n}\,\omega t)\,. \qquad (1)$$

Here t_1 is the instant at which the thyristor starts to conduct (i.e. at which in fig. 2 the switch is closed): at t_1 the instantaneous value of the current is zero.

In *fig. 3* half-cycles of the mains voltage and of the current have been drawn. v_n/V_n and $i/(V_n/\omega L)$ are shown as functions of ωt. Three current curves are shown for which the control signal at the thyristor is applied at three different instants of time, determined by $\omega t_1 = 45°$, $\omega t_1 = 90°$ and $\omega t_1 = 135°$. In preparing fig. 3 the value 0.35 has been assumed for V_l/V_n, which corresponds to a working voltage of 110 V and a mains voltage with an r.m.s. value of 220 V. (Hence $V_n = 220\sqrt{2}$ V.)

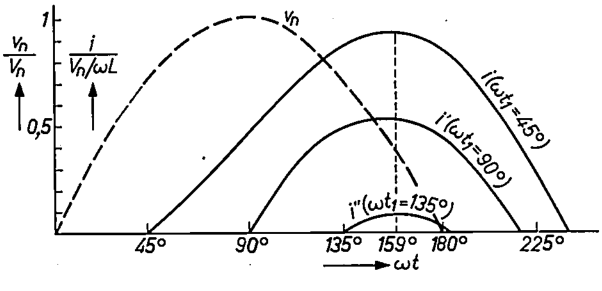


Fig. 3. Curves for a half-cycle of the mains voltage v_n and of the current in the circuit shown in fig. 2. The three current curves i, i' and i'' apply to the cases where the control signal is applied to the thyristor at the instants of time given by $\omega t_1 = 45°$, $\omega t_1 = 90°$ and $\omega t_1 = 135°$. The ratio between the maintaining voltage of the lamp V_l and the r.m.s. value of the mains voltage has been taken as 0.35.

It can be deduced from (1) that the time t_m at which the current reaches its maximum value is independent of the instant at which the thyristor starts to conduct. It is determined by the relation:

$$\sin \omega t_m = \frac{V_l}{V_n} . \quad \ldots\ldots \quad (2)$$

In our case $\omega t_m = 159°$. If the control signal at the thyristor is given at this instant or later, then there is no current at all, and the lamp remains out. If the control signal is applied at an earlier instant there is a current surge in the circuit, whose duration and peak value increase as the control signal is advanced. When the current has passed through the peak value and becomes zero again, the thyristor is in the non-conducting state. No current then flows in the lamp until the other thyristor conducts in the next half-cycle.

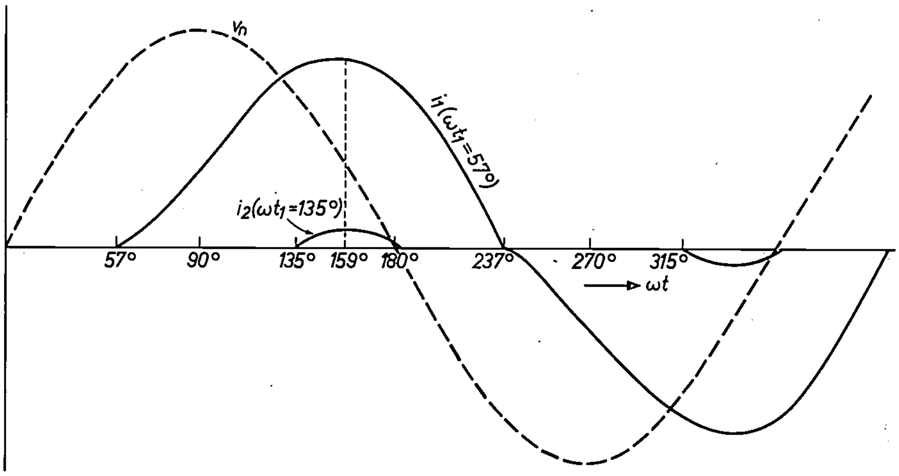


Fig. 4. Curves for a cycle of the mains voltage v_n and the current in a circuit for dimming fluorescent lamps. Two current curves have been drawn: i_1 for maximum luminous intensity ($\omega t_1 = 57°$) and i_2 for a very low light level ($\omega t_1 = 135°$).

The highest r.m.s. value of the current, and therefore the highest luminous intensity, is obtained when the control signal is applied to both thyristors at such an instant that the current is again zero exactly a half-cycle later. This is therefore the same instant at which the other thyristor starts to operate. From (1) it can be deduced that for this to occur the following condition must be met:

$$\cos \omega t_1 = \frac{V_l}{V_n} \times \frac{\pi}{2} . \quad \ldots\ldots \quad (3)$$

For $V_l/V_n = 0.35$ it follows that $\omega t_1 = 57°$. For this case the lamp current curve during a full cycle (in which, therefore, both thyristors operate) has been drawn in *fig. 4*. This diagram also shows the current curve for ignition at a completely different time, corresponding to a very low light level ($\omega t_1 = 135°$).

Even if the control signal is given before the instant of time corresponding to $\omega t_1 = 57°$, no higher luminous intensity is obtained, for in the circuit employed neither of the two thyristors can conduct until the other one is in the non-conducting state. As a result neither can conduct for longer than a half-cycle. Bringing the control signal further forward does not therefore cause a further increase in the mean lamp current.

So that the luminous intensity can be varied from the maximum value to zero, it has to be possible to adjust the phase of the signal voltages applied to the control electrodes, in such a way that ωt_1 changes from 57° to 159° and from 237° to 339°.

The shape of the control signal

The obvious course would be to use a sinusoidal alternating voltage of variable phase as the control signal. This has been found to be not so suitable however, for the following reason. The control signals required for different thyristors of the same type show some spread. This may mean that there is a difference in the times at which the two thyristors start to conduct, leading to an inequality of the two parts of the current cycle. The result is that the light from the lamp reveals a noticeable ripple at a frequency of 50 c/s. To ensure that each thyristor starts to conduct at the correct instant, a square-wave voltage is used instead of a sinusoidal voltage as the control signal. The sharp voltage front which occurs every

[3] In fact the maintaining voltage increases if the conduction time becomes very short. For very low light levels, the theory given here is therefore a fairly rough approximation.

time the polarity changes ensures that if the voltage is high enough each thyristor will begin to conduct, irrespective of a spread in the required control voltage.

Since a thyristor remains conducting after receiving a control signal until the current has fallen below the fairly low holding current level, a series of pulses could be used as a control signal; after the thyristor begins to conduct no further signal at the control electrode is necessary. However, with this scheme as well, a 50-c/s ripple in the light from the fluorescent lamp may occur, particularly on setting to a low light level. This comes about because there is also a certain spread in the holding currents of different thyristors of the same type. As a result the two thyristors would not conduct for the same length of time, and the current waveform would again become unsymmetrical. If a *square-wave voltage* is employed as control signal, there is always a voltage on the control electrode during a half-cycle. The non-conducting state does not occur again until the anode current has become zero; the value of the holding current is then no longer of significance. In this case the two thyristors operate entirely symmetrically, and there is no undesirable 50-c/s ripple in the light output.

Principle of the control circuit

A square-wave voltage whose phase is adjustable with respect to the mains voltage can be obtained with the circuit shown in *fig. 5.* A series arrangement of a choke and two Zener diodes is connected to the autotransformer Tf_1, directly at one end and by the capacitor C and the variable resistor R at the other. The phase of the voltage at point B can be set with the aid of R. In a Zener diode, current begins to flow in the reverse direction once the voltage exceeds a certain value, the Zener voltage. An increase in the current through the diode then causes hardly any change in the voltage. Such a diode therefore corresponds in this respect to a gas-discharge lamp: we note however that a Zener diode shows this effect at only one polarity of the voltage; at the other polarity the diode conducts for all values of the voltage. Coupling two Zener diodes of opposite polarity in series ensures that in both directions the voltage cannot exceed a certain value. A sinusoidal current then gives rise to a square-wave voltage at the diodes. *Fig. 6a* shows the curve of the current in the Zener diodes and that of the voltage on these diodes, together with the voltage applied between A and B.

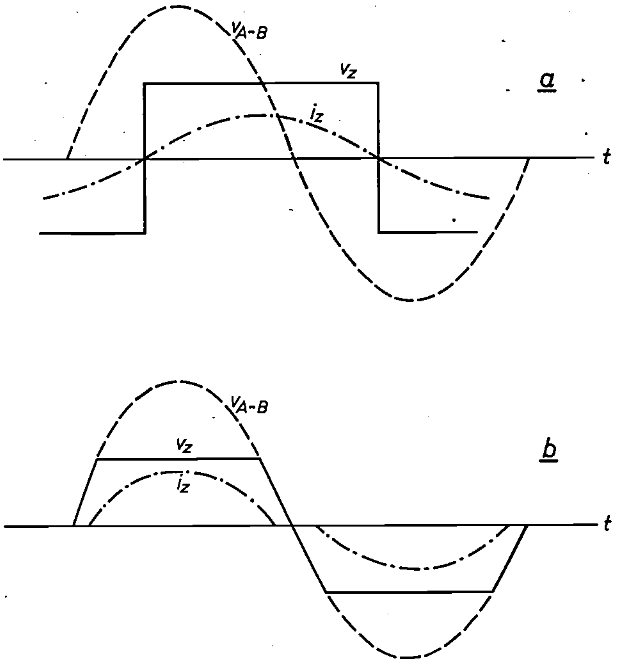


Fig. 6. Curves for the voltage $v_{A\text{-}B}$ between points A and B in fig. 5, the voltage v_z at the Zener diodes and the current i_z in these diodes, as a function of the time:
a) for the circuit of fig. 5;
b) using a resistor instead of the choke L.

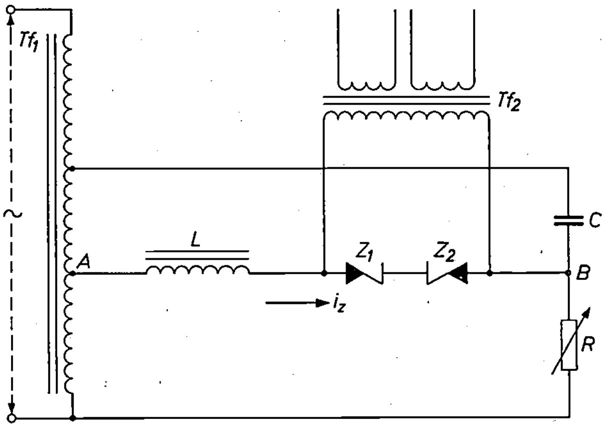


Fig. 5. Circuit for generating the control signals for the thyristors. Z_1 and Z_2 are Zener diodes.

Two square-wave voltages in phase opposition, as required for controlling the two thyristors, are obtained with the aid of the transformer Tf_2, which has two secondary windings.

It is also possible to connect a resistor instead of a choke in series with the Zener diodes. This however would give a square-wave with much less steep edges. The steep edges of the voltage applied to the Zener diodes occur because the phase of the current i in the circuit shown in fig. 5 is behind that of the voltage $v_{A\text{-}B}$. If the current passes through zero in a certain direction, then this voltage has already exceeded the Zener voltage in the other direction, and a voltage of opposite polarity immediately appears at the Zener diodes. This would not be so if a resistor was used instead of a choke. Current and voltage between A and B are then in phase. The voltage v_z then has the shape of a "flattened sine curve", and there are interruptions in the current. The curves concerned are reproduced in fig. 6*b*.

The effect of mains-voltage fluctuations

If the current flowing through a fluorescent lamp is varied as described above, then at low current values the light level is found to be highly sensitive to small mains-voltage fluctuations, with the result that the light can give an impression of unsteadiness. To illustrate this phenomenon, the percentage decrease in the mean current (the light level obtained) which occurs for a 10% reduction in the mains voltage is shown in *fig. 7* as a function of ωt_l (the instant of ignition, hence the set light level). The high degree of sensitivity of the light level to mains-voltage fluctuations when the lamp is heavily dimmed is clearly seen in this diagram.

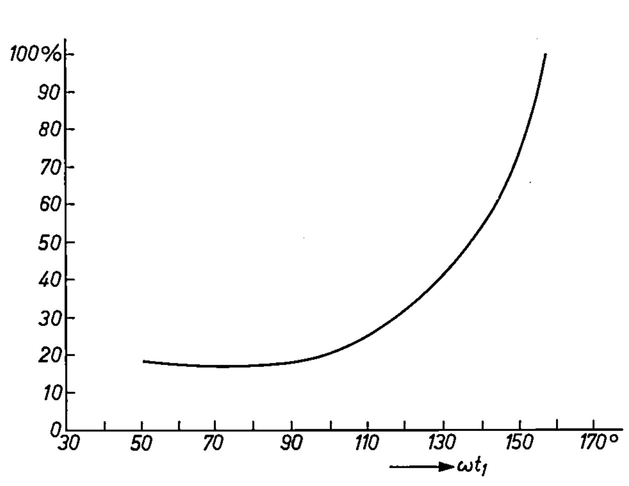


Fig. 7. Percentage current decrease in a dimmed fluorescent lamp for a 10% reduction in the mains voltage, as a function of the instant of ignition.

The cause of this perhaps rather astonishing sensitivity to mains-voltage fluctuations is to be found in the fluorescent lamp itself. Because of the constant value of the maintaining voltage, such a lamp does not have a constant resistance; a decrease in the current increases the equivalent resistance of the lamp. When a lamp of this kind is connected in series with a choke, a reduction in the applied voltage will therefore bring about a reduction in the phase lag of the current with respect to the voltage. If it is now desired to obtain a current curve at the lower voltage like that at the higher voltage, a reduction of the mains voltage will have to be accompanied by an advance of the control signal. If this is not the case, then at the lower mains voltage the control signal comes "too late". This results in a sharp decrease in the mean lamp current, particularly when the lamp is adjusted to a low light level. These effects are illustrated in *fig. 8*, which shows the current curve over a half-cycle for the nominal mains voltage and for a 10% lower value of this voltage. In both cases it has been assumed that the thyristor starts to conduct at $\omega t_l = 115°$. The much lower mean value of i' in relation to i means that there is a much smaller light output from the lamp. If however at the lower mains voltage the control signal is sufficiently advanced that the thyristor starts to conduct at $\omega t_l = 95°$, the current follows the curve i'', whose mean value differs only slightly from that of i. A reduction in the mains voltage then has only a slight effect on the light level.

If the circuit shown in fig. 5 is used to derive the control signals, an automatic advancement of the control signals is obtained for a decreasing mains voltage. This comes about as the voltage is also independent of the current in Zener diodes, and therefore these also have resistances which increase for decreasing current. If the mains voltage falls, the phase of voltage v with respect to voltage $v_{A\text{-}B}$ accordingly decreases. The control signals at the thyristors then occur earlier, and choosing suitable values throughout has ensured that the lamp current is not affected by mains-voltage fluctuations within fairly wide limits. This arrangements therefore gives stable lamp operation at low luminous intensities.

Dimmer for limited power

Although in principle it is possible to connect the control electrodes of the two anti-parallel-connected thyristors to the secondary windings of transformer Tf_2 in fig. 5, a difficulty arises. The control current of a thyristor is so high (e.g. 100 mA) that it cannot be neglected with respect to the current in the Zener diodes. A result would be that the square-wave voltage controlling the thyristors would have much less steep edges. (As already explained, this may give rise to an unsymmetrical lamp current curve, which would give an undesirable 50-c/s ripple in the luminous intensity.) To maintain sufficiently steep edges in the control signal, the current taken by a circuit in parallel with the Zener

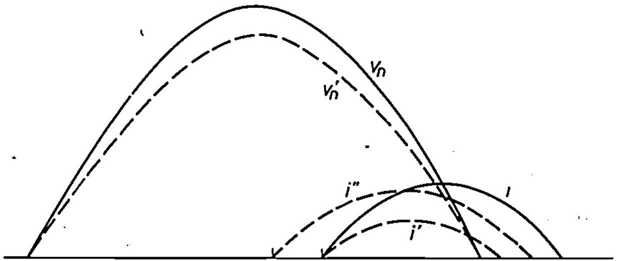


Fig. 8. Curves for the mains voltage and the lamp current when the thyristor starts to conduct at $\omega t_l = 115°$. The curves v_n and i relate to the nominal mains voltage, the curves v_n' and i' to a voltage 10% lower. The curve i'' is the current which results if it is arranged that the thyristor starts to conduct at $\omega t_l = 95°$ at the lower mains voltage.

diodes must not exceed about 5% of the current in the Zener diodes. So that this condition can be met, but nevertheless a sufficiently high control current for the thyristors can be obtained, the two square-wave voltages supplied by the circuit in fig. 5 are applied to the thyristor control electrodes via current amplifiers. The complete circuit is shown in *fig. 9*, and *fig. 10* shows a dimmer in which this circuit has been used. A device of this kind can be used to control a number of fluorescent lamps with a maximum total power of 2 kW. This means that fifty 40-W fluorescent lamps could for example be connected to it. The resistor R, by means of which the luminous intensity of the entire installation is adjusted, does not have to be mounted in the dimmer, and may be fitted in any convenient position.

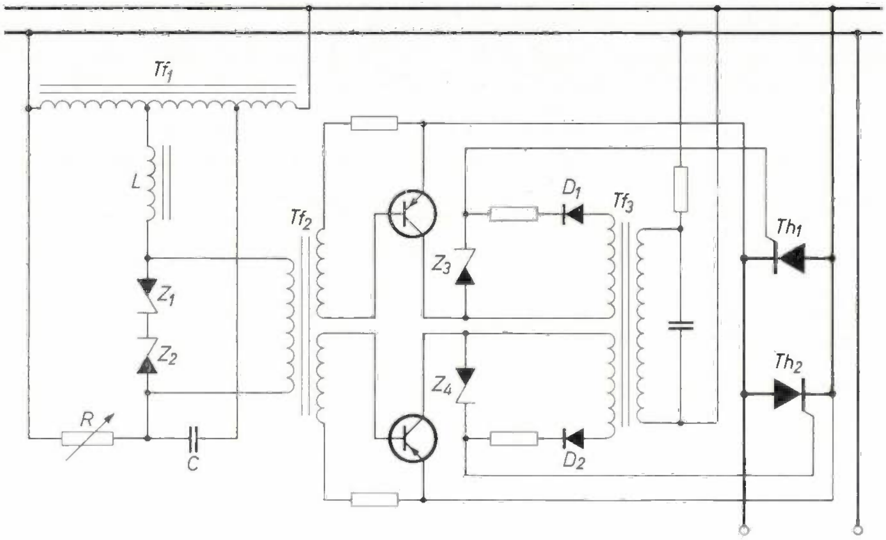


Fig. 9. Circuit of a dimmer for fluorescent lamps. A number of lamps with a total power of 2 kW can be connected to this device. The left-hand part of the circuit corresponds to fig. 5. The two transistors are current amplifiers for the control signals applied to the thyristors Th_1 and Th_2. The transformer Tf_3, the diodes D_1 and D_2 and the Zener diodes Z_3 and Z_4 deliver the supply voltages for the transistors.

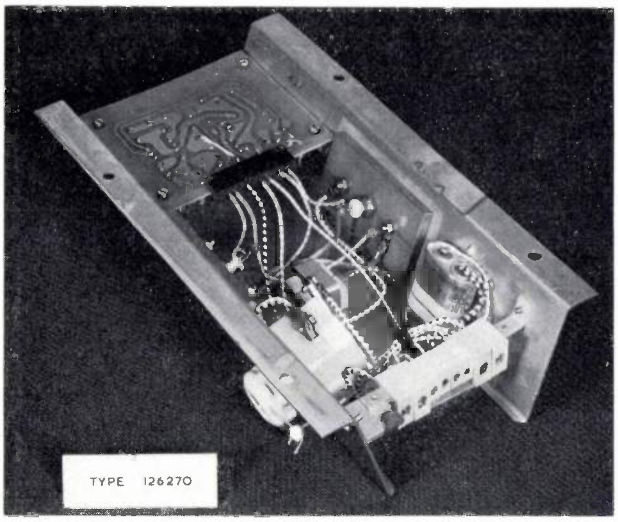


Fig. 10. A dimmer which uses the circuit of fig. 9, shown open. When closed, the dimensions are 21 × 35 × 8 cm.

Control of high powers

Where the number of fluorescent lamps to be dimmed simultaneously is so large that at full luminous intensity the total consumption exceeds 2 kW, a number of dimmers of the type illustrated in fig. 10 can be employed. These are usually connected to different phases of a three-phase mains supply. To give one-knob control for the luminous intensity of the whole installation, the sliding contacts of the variable resistors can be fixed to one spindle, and this is in fact done for a small number of dimmers (e.g. two or three). Uniform control of a large number of dimmers is, however, difficult to achieve by this method, because very accurate tracking of many variable resistors and a very small degree of spread between the different dimmers are then required. For this reason a special device was constructed for controlling high powers; its circuit is reproduced in *fig. 11*. Here the phases of the control signals are adjusted by means of a saturable reactor (*Td* in the diagram), instead of a variable resistor. One of these circuits is connected to each phase of the three-phase supply. The complete circuit is accommodated in what is termed a "central control box" (*fig. 12*). The control windings of the three saturable reactors are connected in series, and a d.c. current is fed through all three windings. Variation of this current enables the individual phases of all the control signals to be adjusted with respect to the corresponding supply phase. The signals from the central control box are fed to dimmers, each of which can control a number of lamps with a total power of 2 kW. control box. Each phase of a central control box can have 50 dimmers connected to it, so that a number of lamps up to a total power of 300 kW can be dimmed.

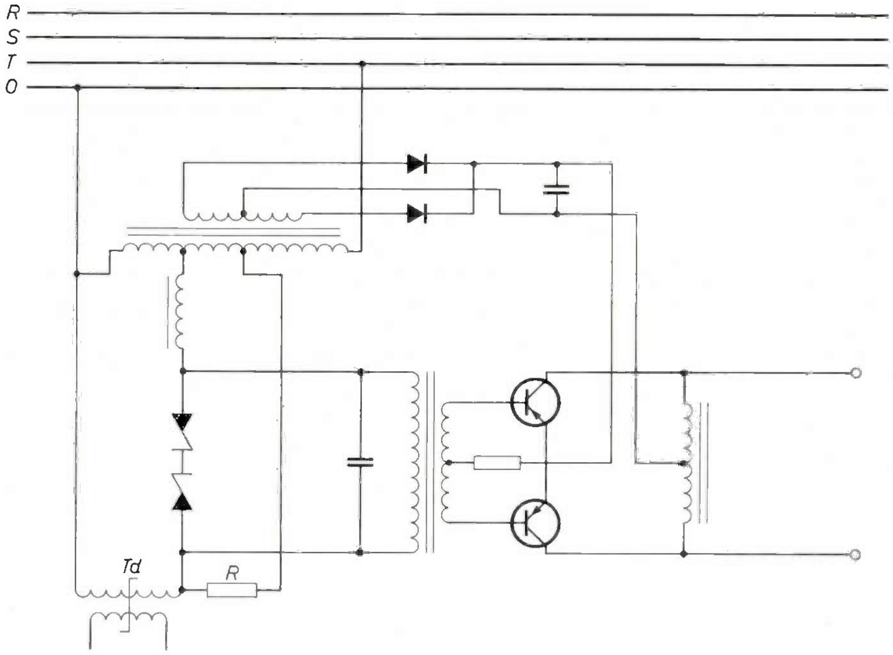


Fig. 11. Circuit of one of the three sections of a central control box. One of these circuits is connected to each phase of the three-phase supply. The phases of the control signals are adjusted with the aid of a variable d.c. current in the series-connected control windings of the saturable reactors *Td*. A total of 3 × 50 dimmers can be connected to each control box, and each dimmer can control a number of fluorescent lamps with a total power of 2 kW.

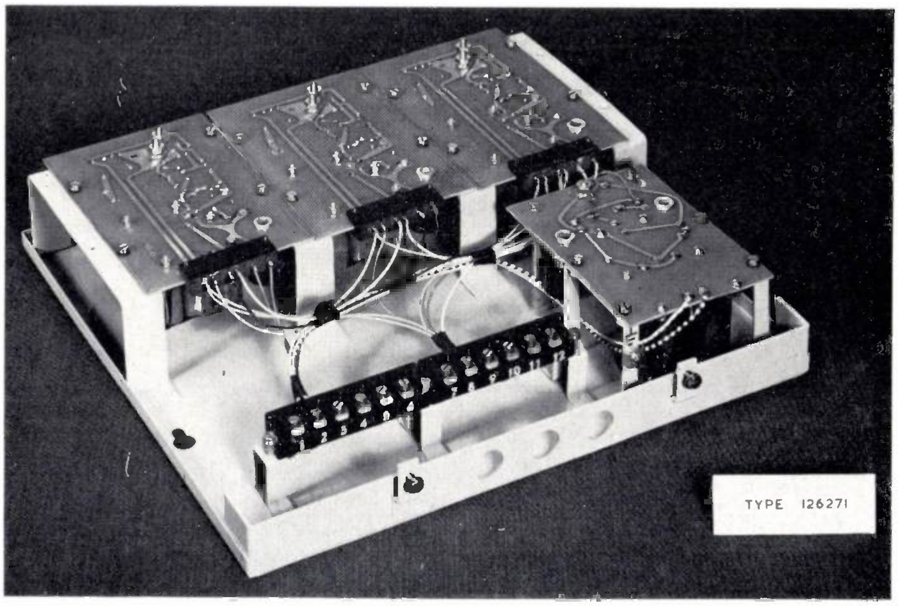


Fig. 12. Central control box, to which a maximum of 3 × 50 dimmers can be connected. When closed, the dimensions are 35.6 × 35 × 8 cm.

These differ from the dimmer of figs. 9 and 10 in that only the amplifier section and the thyristors are required. The left-hand part of the circuit of fig. 9, which generates the variable-phase control signals, is not required here, as these signals are supplied by the central control box.

Outdoor lighting

The growing interest in variable outdoor lighting with fluorescent lamps, particularly for road tunnels, has made it necessary to consider some questions which are of little significance in indoor lighting.

First of all, the fairly wide temperature variations which occur out of doors should be taken into account. This is important because the luminous flux of a fluorescent lamp depends to a rather marked extent on the temperature of the glass wall. In *fig. 13* curve Φ shows this dependence, on a relative scale, for 65-W fluorescent lamps, for constant mains voltage. The r.m.s. values of the lamp voltage and the current, as well as the power consumed, are also shown in the diagram.

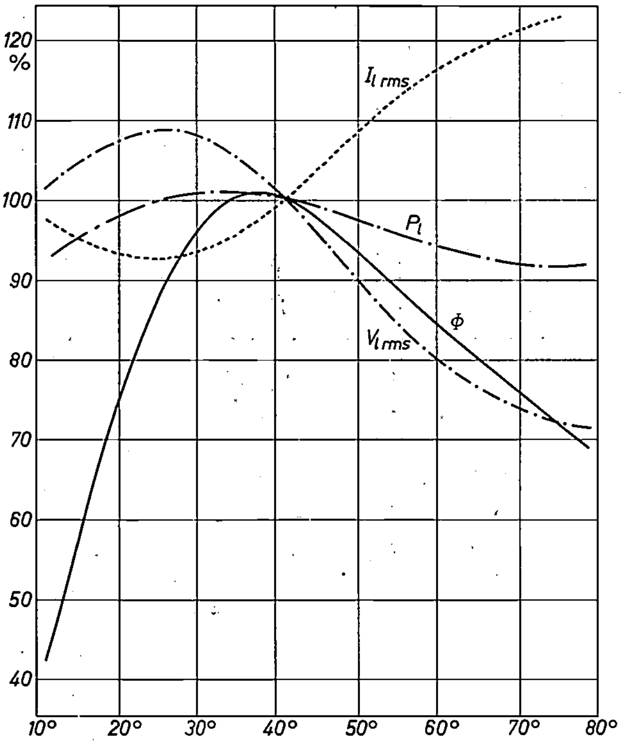


Fig. 13. Luminous flux Φ, r.m.s. value of the lamp voltage $v_{l\,\mathrm{r.m.s.}}$, r.m.s. value of the lamp current $i_{l\,\mathrm{r.m.s.}}$ and power P_l for a fluorescent lamp, as a function of the temperature of the glass wall. The lamp is connected to a constant-voltage mains supply by means of a ballast choke.

From fig. 13 we see that the maximum luminous flux is produced at a wall temperature of about 40 °C. This is obtained if the lamp is operating at room temperature (22 °C). At higher temperatures, and even more so at lower temperatures, the light output is smaller. Lower temperatures are more usually applicable for outdoor lighting; in most countries the average temperature is much lower than 22 °C.

To increase the light output of fluorescent lamps at low ambient temperatures the wall temperature must therefore be raised. This may be simply achieved by placing the lamps in a closed fitting. This method can give wall temperatures of about 30 °C above the ambient temperature for lamps operating at full luminous intensity. If an average temperature of 10 °C is assumed for the year, the optimum wall temperature is thus reached again. Owing to the high-heat-absorption capacity, such a unit also affords protection against variations in luminous flux when there are rapid temperature fluctuations.

The temperature increase in a closed fitting of course occurs only when the lamps are producing the maximum luminous flux. When the lamps are operated dimmed, less heat is generated and hence the increase in the wall temperature is smaller. As a result the light output falls much more sharply with decreasing current than it would at a constant wall temperature. Therefore to achieve a certain relative decrease in the light level, the lamp current does not have to be reduced as much at a low ambient temperature as it does at higher temperatures. This is a fortunate situation, as it has been shown that the method described for controlling the mean lamp current can be employed only to a limited extent at low temperatures. Whereas at room temperature the current, and hence also the light level, can be made to decrease continuously to zero, unstable operation (flicker) of the lamp occurs fairly early if the lamp current is reduced at low temperatures. The cause lies in the fact that at a low wall temperature the mercury vapour-pressure in the lamp is too low to allow stable operation. At a wall temperature of 0 °C, for example, the lamp starts to flicker if the current is adjusted to 45% of the nominal value; at —10 °C this effect sets in at a current as high as 70% of the nominal value.

Despite the limitation mentioned in the permissible *current*-control range at low temperatures, the effect we have mentioned can nevertheless be made use of to realize a sufficiently large variation in the *light* level. It has been shown that at the most unfavourable temperature 65-W fluorescent lamps can be dimmed to a luminous flux of 9% of the maximum value without flicker. For 40-W fluorescent lamps this value can be as low as 4%.

Automatic control

In a road tunnel, the light level has to be arranged to suit the level outside the tunnel [4]. It is quite clear that in this application automatic control is desirable, both because the intensity of the outdoor light can show fairly large unexpected variations, and also because setting to a certain lamp current does not determine the light level in the tunnel: this level is also dependent on the temperature. An automatic control system should therefore regulate the *light* level inside the tunnel; it is not sufficient to adjust the *current* in the lamps.

A circuit with which good control is obtained has been drawn in *fig. 14*. A number of light-sensitive resistors are mounted both inside and outside the tunnel. These are incorporated in the arms R_1 and R_2 of a

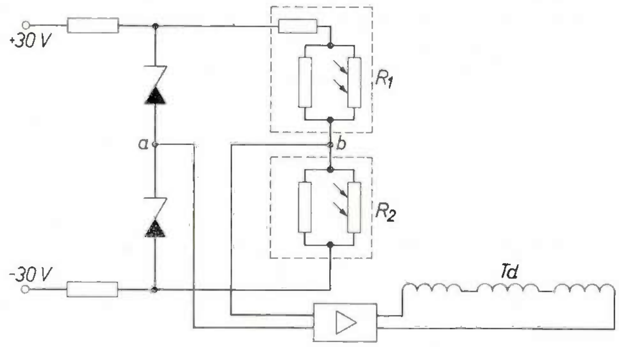


Fig. 14. Circuit for automatic control of the light level in a road tunnel as a function of the luminous intensity outside the tunnel. A number of light-sensitive resistors are incorporated in each of the resistor combinations R_1 and R_2. (Only one of these has been drawn.) R_1 and R_2 are placed outside and inside the tunnel, respectively. *Td* indicates the control windings of the three saturable reactors of a central control box (see fig. 11).

bridge circuit, whose other arms are formed by two Zener diodes. The bridge voltage is supplied to an amplifier, which supplies the current through the saturable reactors of a central control box. The intention is to keep the bridge very nearly balanced, and for the input voltage of the amplifier therefore to be low. If for example the outside light level increases, R_1 decreases. The voltage difference between *a* and *b* increases, and the amplifier supplies a higher current to the saturable reactors. The light level in the tunnel therefore also increases, with the result that R_2 decreases and the bridge returns very nearly to balance. A simplified approximation is that the ratio between inside and outside levels always adjusts itself in such a way that R_1 equals R_2. By suitable choice of the various resistors the desired relation between the two light levels can be obtained. *Fig. 15* shows R_1 and R_2 values from a practical example as functions of the light level. By mounting a resistor in parallel as well as one in series with the light-sensitive resistors outside the tunnel a range of R_1 from a maximum of 36 kΩ to a minimum of 14 kΩ can be obtained. R_2, which consists of the parallel arrangement of several light-sensitive resistors and a normal resistor, can vary over a greater range. As however the inside light level is always adjusted such that R_1 equals R_2 then R_2 will also have a maximum of 36 kΩ and a minimum of 14 kΩ. The inside light level will therefore be controllable between 20 and 200 lux.

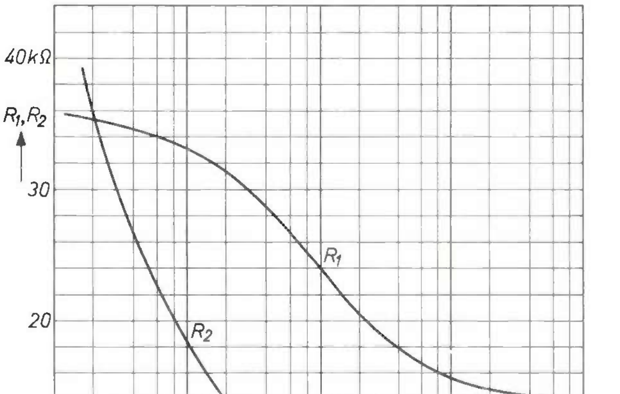


Fig. 15. Values of the resistor combinations R_1 and R_2 of the circuit of fig. 14, as functions of the light levels outside and inside the tunnel. Depending on the outside light, the light level in the tunnel varies between 20 and 200 lux.

The resistor combination R_1, placed outside the tunnel, is accommodated in a watertight closed fitting (*fig. 16*). To prevent it from being obscured by snow, which would result in an undesirable decrease in the light level in the tunnel, the fitting is heated. A thermostat keeps the temperature at a constant value.

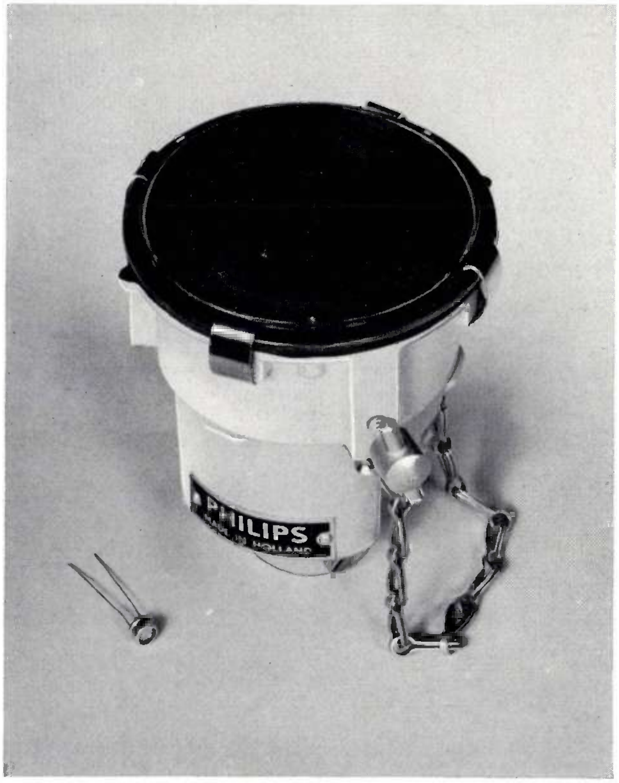


Fig. 16. Fitting in which the resistor combination R_1 (see fig. 14) is placed. To prevent it from being obscured by snow, the fitting is provided with heating controlled by a thermostat. A light-sensitive resistor is shown in the foreground.

The resistor combination R_2 is divided over several lamp fittings (*fig. 17*). In front of the light-sensitive re-

[4] For further information see D.A. Schreuder, Physiological aspects of the lighting of tunnel entrances, Philips tech. Rev. **27**, 76-86, 1966 (No. 3/4).

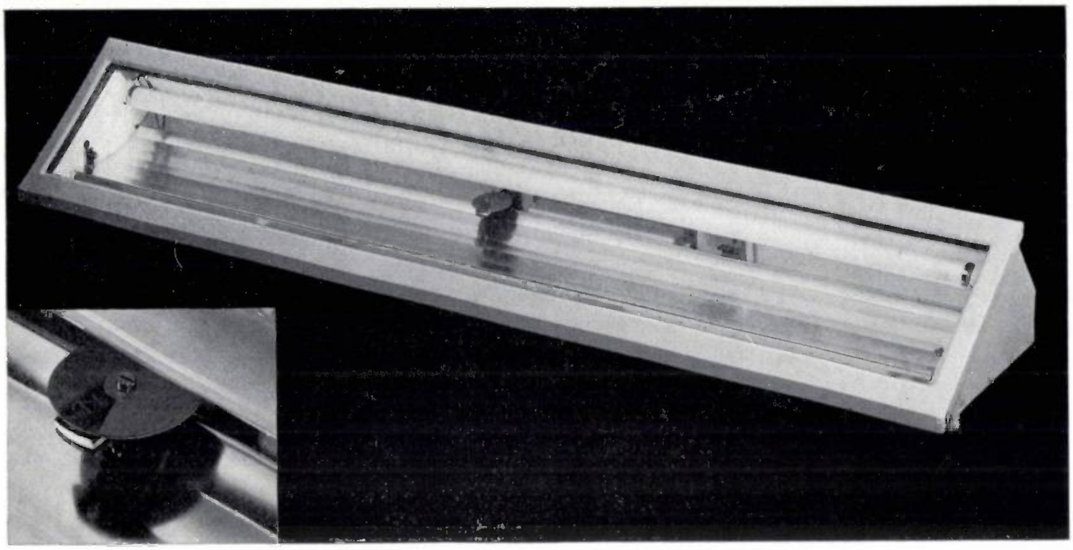


Fig. 17. The resistor combination R_2 (see fig. 14) is divided over several fluorescent lamp fittings. The light-transmission of the small rotatable discs varies with angle. The discs can be used to vary the amount of light received from the lamp by the light-sensitive resistors.

sistors small rotatable discs are mounted here, whose light-transmission varies with angle. By setting these discs, a desired relation between the light level and R_2, and hence also between outside light and inside light, can be obtained.

Installations using the principle described are already being employed in some road tunnels, notably the Leoben Tunnel in Austria and the Coen Tunnel in the Netherlands. *Fig. 18* shows a number of dimmers in use in the Coen Tunnel. The Y Tunnel in Amsterdam will also be equipped with a similar installation.

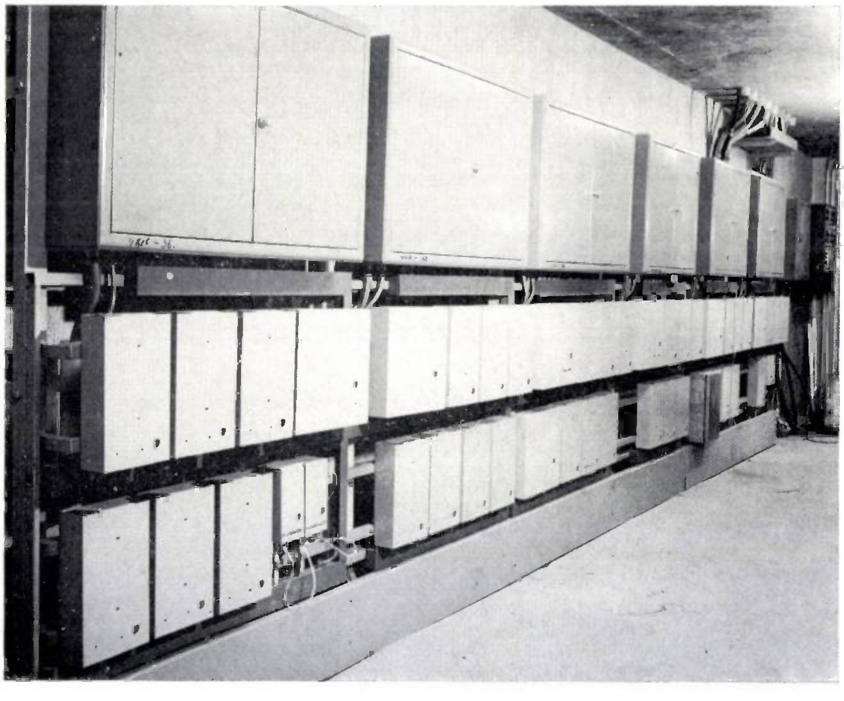


Fig. 18. A number of dimmers and central control boxes in use for automatic control of the light level in the southern half of the Coen Tunnel. The upper row of cabinets contain fuses. A similar number of apparatus have been installed for the northern half of the tunnel.

Summary. In modern dimmers for fluorescent lamps thyristors offer considerable advantages over the thyratrons formerly employed. After an introduction to methods of dimming fluorescent lamps, several devices used for this purpose are discussed. These are a dimmer, which can handle a number of lamps with a total power of up to 2 kW, and a central control box, with which, if desired, 150 such dimmers can be operated at the same time. For outdoor use of fluorescent lamps the possibility of considerable variations in ambient temperature has to be taken into account; on average the ambient temperature is below room temperature. The average light output can be increased if lamps for outdoor use are placed in a closed fitting. Despite the fact that at low temperatures the current in a fluorescent lamp can be reduced to only a limited extent without flicker, a variation in the light level sufficient for practical purposes can nevertheless be achieved. Finally a circuit is discussed which can be used to control the light level in a road tunnel as a function of the outside light.

Beam-plasma amplifier tubes

M. T. Vlaardingerbroek and K. R. U. Weimer

533.9:621.375:621.385

During the last twenty or thirty years "plasmas" (ionized gases) and their interaction with electromagnetic waves and beams of charge carriers have attracted a great deal of attention. To some extent this has been stimulated by the prospect of being able to control fusion reactions between atomic nuclei, but this was by no means the only motive. There was, for example, a purely scientific interest in plasmas on the part of astrophysicists, while on the other hand engineers were interested in the possibility of using the interaction between a plasma and an electron beam to obtain microwave amplification. In the following article amplifier tubes are described whose operation depends on this interaction. Although these tubes have a very high gain and a wide bandwidth, the prospects for their practical application are not very favourable due to the presence of low-frequency instabilities of a fundamental nature. Their investigation has, however, provided a number of results of considerable general interest in plasma physics.

Discovery and investigation of plasma oscillations

In about 1925 Langmuir and Mott-Smith [1] discovered that electrons can occur in a sustained low-pressure gas discharge which have a higher velocity than that corresponding to the local potential. As in their experiments the average mean free path of the electrons was far larger than the dimensions of the discharge chamber, these high speeds could not be the result of collisions.

It occurred at the time that this could be accounted for by the presence of charge carrier oscillations in the discharge, but this could not at first be demonstrated.

By coupling an open-wire line with a matched detector crystal to a gas discharge, Penning [2] was able to show that oscillations can indeed occur in an ionized gas in which the concentrations of the positive and the negative charge carriers are equal. The frequency of these oscillations was found to be very sharply defined. Langmuir repeated Penning's measurements, and in 1929 he and Tonks published their classic article on this subject [3]. In this article, they showed that the angular frequency ω_p of the plasma oscillations of the electron gas depends exclusively on the charge density ϱ_0 of this gas in accordance with the equation:

$$\omega_p^2 = \frac{e\varrho_0}{m\varepsilon_0}. \quad \ldots\ldots \quad (1a)$$

Here e and m are the charge and mass of the electron and ε_0 is the dielectric constant of free space. If the charge density is expressed in n_0, the number of electrons/cm³, it follows from (1a) that the frequency f_p of the plasma oscillations is given by:

$$f_p = 8980\ \sqrt{n_0}. \quad \ldots\ldots \quad (1b)$$

It follows from Langmuir's theory that the oscillations do not propagate independently under normal circumstances: a group of electrons oscillating about its equilibrium position must be regarded as an isolated harmonic oscillator. Of course, it is conceivable and quite possible, as we shall see later, that various groups of electrons can be brought into oscillation in such a way that the resultant effect assumes the appearance of a wave phenomenon.

For a description of the behaviour of the electrons in such a plasma, we may make use of the following picture. Let us begin by assuming that the ions, which are much heavier than the electrons, are immobile. Let us also assume that the ions are uniformly distributed over the chamber at a charge density of n_0e. Within this "lattice" of ions the electrons are likewise uniformly distributed (charge density $-n_0e$), but these, however, are mobile. Now assume that in an infinite plasma the electrons in a layer of thickness d travel a distance ζ in the direction vertical to the surfaces of the layer, i.e. in the z direction of *fig. 1*. On one side, a layer of thickness ζ then arises which only contains ions, and on the other side at a distance d there arises a layer of equal thickness in which the electron concentration is doubled. As the system is infinite, the field lines between the two layers are parallel to the z axis.

Dr. Ir. M. T. Vlaardingerbroek is with Philips Research Laboratories, Eindhoven. Dr. K. R. U. Weimer was with Philips Research Laboratories until his untimely death in June 1966.

[1] I. Langmuir, Phys. Rev. 26, 585, 1925.

[2] F. M. Penning, Physica (N.T.N.) 6, 241, 1926.

[3] L. Tonks and I. Langmuir, Phys. Rev. 33, 195, 1929.

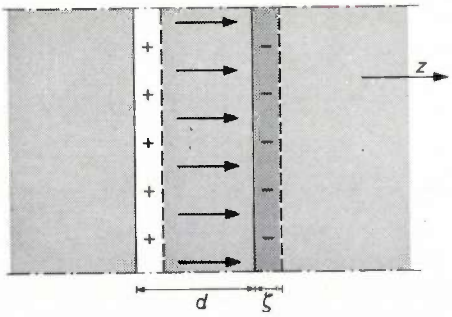


Fig. 1. Explaining the nature of plasma oscillations.

The field strength F of this homogeneous electric field is given by:

$$F = \frac{n_0 e \zeta}{\varepsilon_0}. \qquad (2)$$

It is immediately clear from this equation that the displaced electrons will perform a simple harmonic oscillation on their release, as the force $-Fe$ tending to return the electrons is proportional to the displacement ζ. The equation of motion of the electrons is as follows:

$$m\partial^2 z/\partial t^2 = - Fe = - \frac{n_0 e^2 \zeta}{\varepsilon_0}, \qquad (3a)$$

or:

$$m\partial^2 z/\partial t^2 + \frac{n_0 e^2 \zeta}{\varepsilon_0} = 0 . \qquad (3b)$$

It follows from this that the frequency is given by [4]:

$$f_p = \frac{1}{2\pi}\sqrt{n_0 e^2/m\varepsilon_0} = 8980\sqrt{n_0}. \qquad (1b)$$

The experiments of Langmuir et al. and Penning still did not solve the question of how plasma oscillations occur naturally and how they can be induced artificially. About 15 years ago it became clear that plasma oscillations can occur when a beam of electrons moves through a plasma, and that this brings about a gradually increasing periodic variation in the charge density of the beam [5]. This discovery led quite naturally to attempts to design microwave amplifiers based on these phenomena. Little success was at first achieved, but in 1957 Boyd, Field and Gould achieved good results [6] with the arrangement shown diagrammatically in *fig. 2*. Using a helix coupled capacitively to an input waveguide the beam was modulated at a trequency close to the plasma frequency. The modulafed beam was then passed through a mercury vapour discharge (pressure about 2×10^{-3} Torr) and finally through a second helix coupled to an output waveguide.

Both helices act as delay lines. The pitch and diameter are so chosen that the phase velocity of the electromagnetic waves along the axis is approximately equal to that of the velocity of the electrons in the beam. This is essential for good energy exchange between the helix and the beam. As the velocities are approximately equal, the electrons experience the effect of the same field for a considerable period of time. This cannot be achieved with a waveguide alone. We shall return to this important point later.

Theoretical and experimental investigations in our laboratory during recent years have shown that it is not necessary to modulate the beam before it enters the plasma. The input helix may therefore enclose the plasma: the same applies for the output helix.

Moreover, it has been found that in principle separate generation of the plasma by means of a gas discharge is unnecessary. It is sufficient if the electron beam is passed through a gas of suitable pressure. The beam ionizes a number of gas atoms along its path and it appears that these can perform quite adequately the function of the plasma. An arrangement for this type of operation is shown diagrammatically in *fig. 3*. As can be seen, the tube portrayed in this figure shows a certain similarity to a travelling-wave tube [7]. To prevent divergence of the beam, the entire structure is located in a longitudinal magnetic field.

Investigation of the behaviour of plasmas and of the interaction between a plasma and a beam of charged particles is not only of interest because it might in principle assist in the development of new microwave amplifier tubes. Plasma physics is also of basic importance in *astrophysics*. For example, the fact that the emission of radio waves by the sun during periods of sun spot activity is greater than at other times has been explained as due to the interaction between the cloud of ionized gas thrown off during a solar eruption and the plasma of the corona [8].

Knowledge of the behaviour of plasmas is equally important for research into the possibility of controlled fusion of atomic nuclei. One of the main problems here is the generation of the very high temperatures (10^7 to 10^8 °C) required for nuclear fusion. Research into

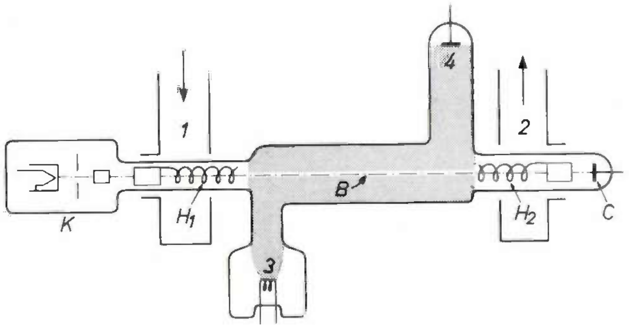


Fig. 2. The amplifier tube due to Boyd, Field and Gould [6]. The beam is modulated and the output signal is taken off by means of a helix. The plasma is obtained by setting up a gas discharge (shaded) between two auxiliary electrodes. *K* gun. *B* electron beam. *C* collector. H_1 and H_2 input and output helices. *1* input waveguide. *2* output waveguide. *3* auxiliary cathode for gas discharge. *4* auxiliary anode for gas discharge.

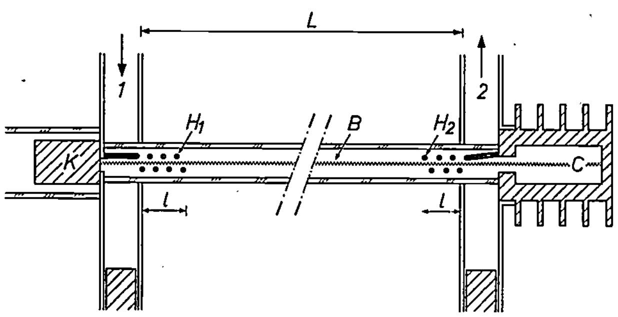


Fig. 3. Diagrammatic representation of a tube in which the beam itself generates the plasma. The letters and figures have the same significance as in fig. 2. L is the length of the helices-beam-plasma system, and l the length of the helices.

beam-plasma systems may perhaps contribute towards solving this problem in that we may learn something about the conversion of the "directed" energy of the beam electrons into "non-directed" (thermal) energy of the plasma particles.

In this article we shall confine ourselves to the field of microwave amplification by beam-plasma tubes [9]. We shall see that extremely high gain is possible with these tubes, and that they can have a considerable bandwidth. Their practical application, on the other hand, has hitherto been prevented by inherent disadvantages: the noise level is very high, and moreover, troublesome low-frequency oscillations arise when the input power is increased.

The interaction between beam, plasma and helix

Space charge interaction between beam and plasma

Let us assume that we have a beam all of whose electrons are moving at the same velocity and that this beam, just as in a klystron, passes through two grids positioned close to one another and between which a small high-frequency a.c. voltage is applied. In passing through these modulation grids an electron is either accelerated or delayed according to the phase of the field, and the beam is thus subjected to a certain periodical velocity modulation. As a result, the accelerated electrons gradually catch up with the delayed ones so that bunching takes place in the beam. The bunches move at approximately the speed of the beam.

The bunching is of course linked with the occurrence of forces between the electrons. As like charges repel, these space charge forces oppose this overtaking effect and reduce it eventually to zero. There is thus a return to the original state: a velocity-modulated beam of homogeneous density. The whole procedure is then repeated, etc. The final effect of the modulation grids is thus that space charge waves are set up in the beam, their character being completely identical with that of the space charge oscillations in the plasma: these waves appear to be stationary to an observer moving along with the beam at the speed of the electrons. The distance along the beam axis at which the same modulation pattern is repeated in amplitude and phase is called the plasma-wavelength of the beam. The maximum value of the charge density depends of course on the amplitude of the high-frequency voltage at the modulation grids [10].

If we now take a look at what happens at a given fixed point of the line along which the beam electrons move, it is clear that if f is the frequency of the modulation signal, f bunches per second will pass that point. The electrical fields mentioned earlier naturally travel at the same speed as the corresponding bunches, so that at this fixed point the direction of the field alternates between the forward and backward directions at a frequency f. A stationary electron at this point will therefore be brought into oscillation by the effect of the passing bunches.

In order to see what happens if, instead of meeting a single stationary electron in its path, the beam moves through a plasma, let us briefly recapitulate how a system with a certain eigenfrequency behaves when it is exposed to a periodically varying external force. Unlike an isolated electron in a field-free space, a group of electrons in a plasma should of course be considered as such a system. (The situation is then mathematically described by an equation such as (3b) in which the zero on the right-hand side is replaced by an expression which describes the periodic external force.) *Fig. 4* gives a graphical representation of the behaviour of such a system. If the frequency f of the external force is very high, the amplitude A of the system is nearly equal to zero. With decreasing f the amplitude increases, until f reaches the eigenfrequency f_{res}. At this frequency, in theory, A becomes infinite. In practice its size is limited, partly because at high amplitude nonlinear effects come into

[4] A more general derivation of this equation can be found in reference [3].

[5] The first indications of this were found by A. V. Haeff (Phys. Rev. **74**, 1532, 1948; see also Proc. I.R.E. **37**, 4, 1949) in a study aimed at finding the effects causing radio emission of the sun. Independently of Haeff, J. R. Pierce and W. B. Hebenstreit (Bell Syst. tech. J. **28**, 33, 1949) obtained the same results in a study of the propagation of noise fluctuations in a non-homogeneous electron beam.

[6] G. D. Boyd, L. M. Field and R. W. Gould, Phys. Rev. **109**, 1393, 1958.

[7] See for instance B. B. van Iperen, Philips tech. Rev. **11**, 221, 1949/50.

[8] A. V. Haeff, Phys. Rev. **75**, 1546, 1949.

[9] A considerable contribution to the design of these tubes and to the measurements obtained with them was made by H. Bodt and J. A. L. Potgiesser of this laboratory.

[10] A detailed consideration of space charge waves in electron beams is to be found in H. Groendijk, T. Ned. Radiogenootschap **26**, 51, 1961. See also A. H. W. Beck, Space-charge waves, Pergamon Press, London 1958.

play whose effect is negligible when A is small. If f decreases still further, the amplitude A is again reduced, but always remains greater than zero. For $f > f_{res}$ the phase difference φ between the driving force and the oscillation of the system is equal to 180°, for $f < f_{res}$ it is zero.

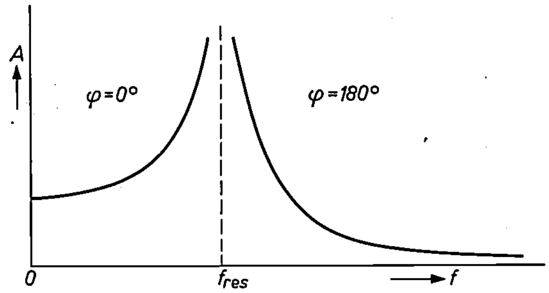


Fig. 4. Graph of the amplitude A of a system whose eigenfrequency is f_{res}, at the frequency f of a periodically varying force to which it is subjected. When $f = f_{res}$, the amplitude is very large (resonance). If f becomes much greater than f_{res}, A approaches zero. If however f decreases below f_{res}, A approaches asymptotically to a value differing from zero. The phase difference φ between the driving force and the oscillation of the system is 0° in one frequency range and 180° in the other.

The interaction between a plasma and a modulated electron beam can be described as follows. If the modulation frequency f of the beam is much greater than the plasma frequency f_p, then there will be very little reaction from the plasma-electrons. The beam is then hardly affected by the plasma and passes through it without any noticeable interaction taking place. If the signal frequency f is now allowed to decrease and to approach the plasma frequency f_p from above, then the plasma-electrons will gradually oscillate more strongly, but in the opposite phase to the field they are subjected to ($\varphi = 180°$; see fig. 4). Bunching occurs in the plasma with exactly the same distribution as the bunching in the beam. This bunching in the plasma is exactly in step with that in the beam, but as we see, it is due to electrons which oscillate about a fixed point and whose position, when time-averaged, is thus stationary.

As the bunches occur together ($\varphi = 180°$), the oscillating plasma electrons amplify the electrical field caused by the differences in beam density. The bunches in the beam therefore diverge more rapidly — and new ones are created more quickly — than would be the case without plasma, so that the plasma wavelength of the beam decreases. If f approaches the value f_p, this wavelength approaches zero.

If, on the other hand, the signal frequency f is less than the plasma frequency, then the electrons oscillate *in phase* with the space-charge field. The bunches of the plasma now occur precisely *between* those of the beam and *attenuate* the electrical field. In fact, the field is more than compensated, so that the resulting field is in the opposite sense and drives the electrons in the beam bunches further towards one another. This effect increases with the path length that a given bunch has traversed in the plasma, i.e. the density modulation gradually increases along the path of the beam. This is therefore a case in which the interaction between beam and plasma causes amplification. It is no longer true to speak of a plasma wavelength: because of the increasing wave amplitude there is no space charge pattern which can be described as recurring in amplitude and phase.

All in all, it is therefore clear that for $f < f_p$ the modulation is amplified, that for $f > f_p$ modulation is not amplified but the plasma wavelength of the beam is reduced and that for $f \gg f_p$ there is no noticeable interaction between beam and plasma.

It should also be mentioned that in a plasma, in addition to the *longitudinal* oscillations mentioned above, *transverse* oscillations can also occur. The principal types among these are the cyclotron oscillations. These are characterized by the cyclotron frequency f_c, which is equal to the number of turns traversed per second by an electron following the helical or circular path obtained when a magnetic field is applied. The occurrence of cyclotron oscillations can be avoided in tubes by a suitable choice of magnetic field.

The beam-plasma-helix system

Let us consider the case in which the modulation is applied by means of a helix inside or around the plasma, as in tubes of the type shown in fig. 3. In addition to the modulation of the beam by the helix and the interaction between beam and plasma, there is now a direct interaction as well between helix and plasma. The result of this is that the characteristics of such a system differ to some extent from those of a system in which the plasma is affected exclusively by the beam, as discussed above.

If a suitable mathematical model is chosen for the situation in a helix-beam-plasma system its characteristics can then be calculated. We have made such a calculation [11], whose main approximation is the assumption that only longitudinal oscillations are possible in the dirction of the magnetic field. The main characteristics of the helices-beam-plasma system which were revealed by our calculations can be briefly summarized as follows.

1) To obtain coupling, care must be taken that the velocity u_0 of the beam electrons is higher than the phase velocity v_f of the wave on the helix.

2) The coupling is tightest and the gain highest for frequencies which are immediately above the plasma frequency f_p.
3) Coupling and gain increase as the difference between u_0 and v_f increases. An increase in this difference leads however to decreased bandwidth.

If these characteristics are compared with those of a free beam-plasma system, it is found that the addition of the helix has decreased the gain but increased the bandwidth. The difference is less at greater values of $u_0 - v_f$. Furthermore, with the free beam-plasma system f had to be somewhat smaller than f_p, whereas here, on the other hand, it has to be somewhat greater.

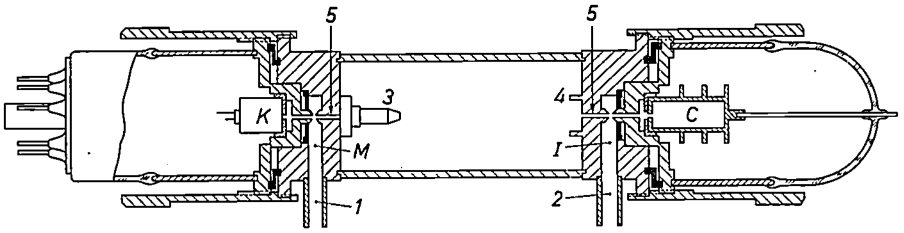


Fig. 5. Simplified diagrammatic section of a beam-plasma tube used by us, with signal input and output by means of resonant cavities. The plasma is obtained here by means of an auxiliary discharge. *K* gun. *M* input cavity. *I* output cavity. *C* collector. *1* and *2* input and output waveguides. *3* plasma cathode. *4* plasma anode. *5* opening for electron beam.

To find out whether we were justified in not taking transverse waves into account in the theory of our tubes — and also to obtain a better insight into the behaviour of beam-plasma systems in general — we have also calculated the characteristics of a beam-plasma system in a cylindrical waveguide. As the problem here is a very complicated one, we have solved it in three successively less simplified steps [12] [13] [14]. These calculations showed that the approximations used are permissible when the plasma and cyclotron frequencies differ considerably. If this is not so, then relatively strong cyclotron oscillations also occur — which cannot of course be detected with a helix — and the increment of the plasma oscillations is slight. This agrees exactly with the results obtained with our tubes. When the magnetic field strength of a tube which initially gave a good amplification was increased to such a level that f_c became equal to f_p, the output signal decreased to zero. At the normal magnetic field f_c was no greater than $\frac{1}{2}f_p$.

Construction and characteristics of some beam-plasma amplifiers

The first type of tube with which we obtained gain by beam-plasma interaction may be considered as a klystron in which two auxiliary electrodes excite a non-independent low-pressure mercury-vapour discharge in the space between the two resonant cavities. An example of this type is shown in *fig. 5*. This tube gives a gain of more than 50 dB at a frequency of 10 000 Mc/s, i.e. at $\lambda = 3$ cm. (There is no gain without plasma.) The maximum output power is 0.2 mW. The use of resonant cavities gives tubes of this type a rather narrow bandwidth (about 2 Mc/s for the tube shown in fig. 5). Another undesirable feature is that the cavities become detuned when plasma penetrates into the coupling gaps.

In many respects tubes with helical coupling (fig. 3) are better. Helical coupling is not only efficient but also gives a good bandwidth. Detuning is, of course, impossible here. We have designed and tested two kinds of helical tube, one with and one without auxiliary discharge. A photograph of a tube with no auxiliary discharge is shown in *fig. 6*: in these tubes the electron beam makes its own plasma. The length L (cf. fig. 3) is 20 cm in the tube shown in the photograph; in other tubes of this type which we have made it is 10 cm. The length l of the helices was 4, 5, 9, 18 and 40 mm in various tubes: in the tube shown in fig. 10 $l = 18$ mm. The relation between the input power P_i and the output power P_o is shown in *fig. 7* for a tube 20 cm long with 18 mm helices. As can be seen, when P_i is low the output signal over a certain range does not rise above the noise level (range I), then there is a range (II) in which P_o increases linearly with P_i, while when P_i is high (range III) the output power decreases again, sometimes quite markedly. All beam-plasma tubes give this type of curve; tubes of the klystron type (fig. 5) as well as helix tubes hav-

[11] M. T. Vlaardingerbroek and K. R. U. Weimer, T. Ned. Elektronica- en Radiogenootschap **29**, 73, 1964.

[12] M. T. Vlaardingerbroek, K. R. U. Weimer and H. J. C. A. Nunnink, Philips Res. Repts. **17**, 344, 1962.

[13] M. T. Vlaardingerbroek and K. R. U. Weimer, Philips Res. Repts. **18**, 95, 1963.

[14] H. Groendijk, M. T. Vlaardingerbroek and K. R. U. Weimer, Philips Res. Repts. **20**, 485, 1965.

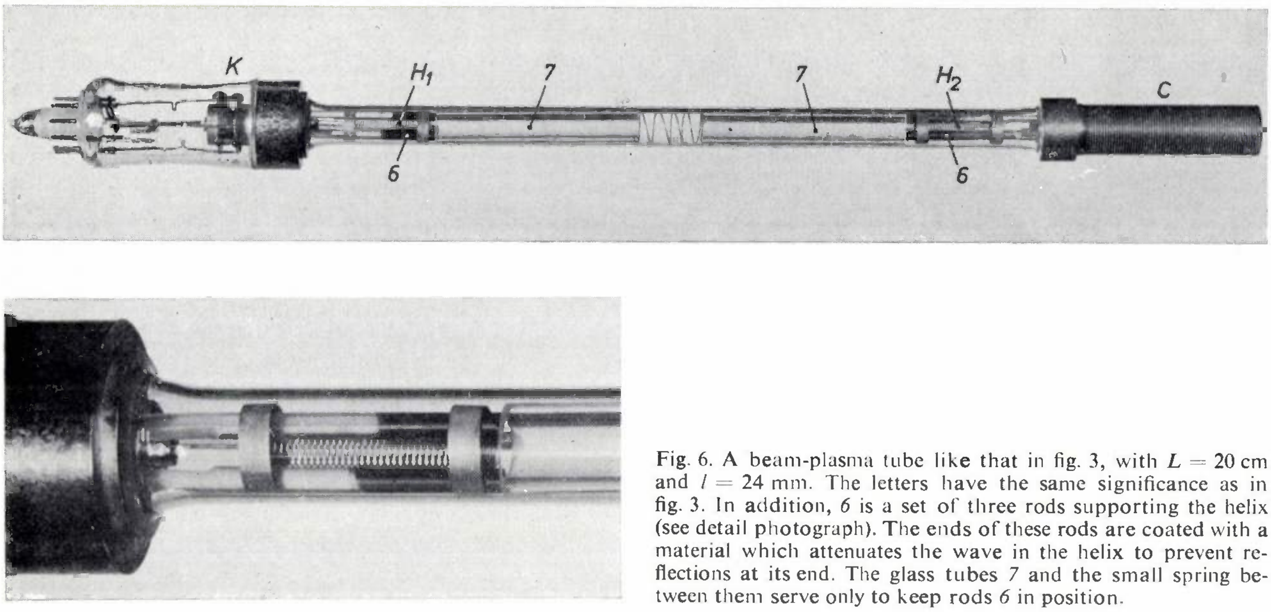


Fig. 6. A beam-plasma tube like that in fig. 3, with $L = 20$ cm and $l = 24$ mm. The letters have the same significance as in fig. 3. In addition, *6* is a set of three rods supporting the helix (see detail photograph). The ends of these rods are coated with a material which attenuates the wave in the helix to prevent reflections at its end. The glass tubes *7* and the small spring between them serve only to keep rods *6* in position.

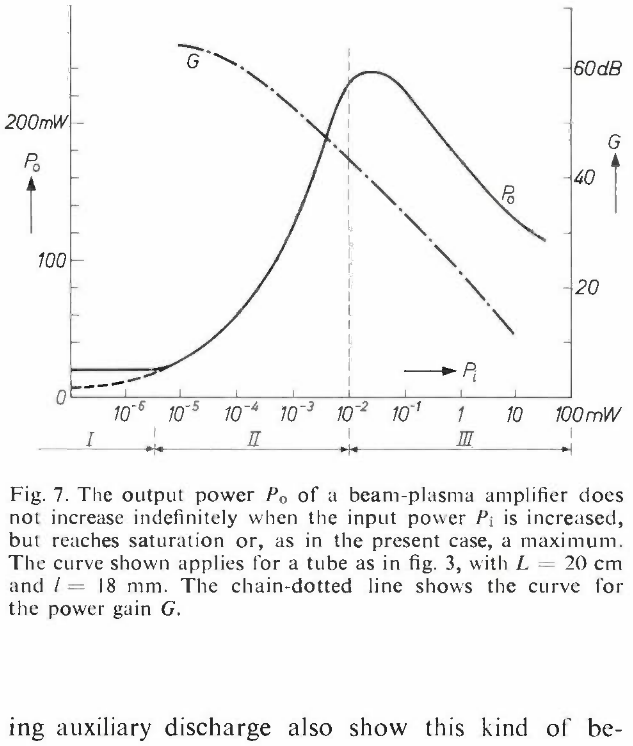


Fig. 7. The output power P_o of a beam-plasma amplifier does not increase indefinitely when the input power P_i is increased, but reaches saturation or, as in the present case, a maximum. The curve shown applies for a tube as in fig. 3, with $L = 20$ cm and $l = 18$ mm. The chain-dotted line shows the curve for the power gain G.

ing auxiliary discharge also show this kind of behaviour. The reason why the output power decreases again when P_o increases will be discussed later.

The variation of the gain G in range II as a function of frequency is shown in *fig. 8* for a particular tube of the type in question. The four curves apply for different values of the beam current. At 4.5 Gc/s and 8 mA the gain is about 52 dB. It can be seen that the bandwidth of these tubes is quite large. The highest gain achieved is 65 dB.

With tubes of this type, however, noise level is quite high; noise figures of the order of 55 dB are encountered. (The noise level is the ratio (expressed in dB) of the output noise power to G times the available noise power in the input signal. If the tube itself makes no noise contribution, this ratio is unity, i.e. the noise figure is 0 dB.) This very considerable noise is found to be partly due to the fact that in tubes having no auxiliary discharge there is a plasma over the whole length of the beam so that the beam noise is already considerably amplified at the input helix. For tubes which had a much lower beam current and an auxiliary discharge between the helices to obtain the required plasma density — i.e. tubes of the second type — the noise level was in fact lower. The plasma density for the path between the cathode and the input helix of this tube was so low that no noise gain could occur at about 4000 Mc/s.

A far greater part of the noise was found to be related to the type of electron gun used. In helix tubes of

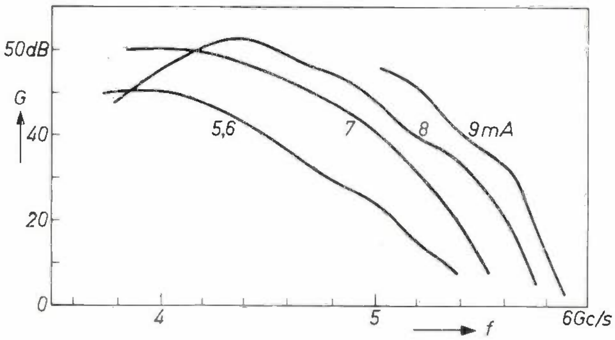


Fig. 8. The power gain G for the helix tube of fig. 6 as a function of the frequency f of the input signal for four values of the beam current. The curves apply only for relatively low input powers.

the second type (*fig. 9*) in which not only is the length of the plasma reduced to the necessary minimum, but a different type of gun is used, the noise level only amounts to 25 dB. About 5 dB of the difference between this value and the noise figure for tubes of the first type is accounted for by the plasma gain and a further 20 dB by the gun.

The gun which has given this improvement is the same as the one used in low-noise [15] travelling-wave tubes. Unlike the other gun, it lies entirely inside the magnetic field used for focusing the beam, so that the cross-section of the beam is approximately constant right from the cathode. Various electron movements which occur in beams of diminishing section do not occur here.

All the tubes discussed above operate in the frequency range from 4000 to 10 000 Mc/s (wavelength 7 to 3 cm). By using a higher gas pressure and thus increasing the plasma density and plasma frequency, tubes can also be made which amplify in the mm range. Sub-millimetre operation is however difficult to achieve: it is by no means easy to form plasmas of the density required ($n_0 > 10^{15}$ cm^{-3}) and penetration of an electron beam through such plasmas will also present problems.

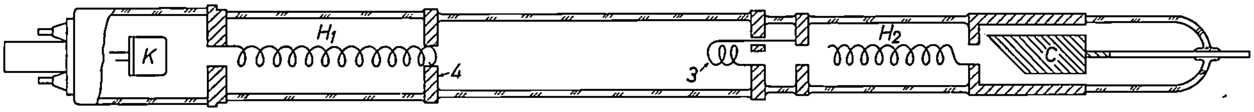


Fig. 9. Helix tube with auxiliary electrodes for gas discharge. An extremely low beam current is chosen for this tube to avoid noise. The letters and figures have the same significance as in figs. 3 and 5.

Increasing the frequency also affects signal input and output: in particular, resonant cavities and helices suitable for use in the millimetre range are difficult to make We found recently that the output signal could be derived from the conical electromagnetic waves which are set up in a medium when an electron moves through it (or a narrow tunnel in it) at a speed greater than that of light in the medium. A tube in which the output signal was taken off in this way gave a gain of 20 dB at 10 Gc/s. The inverse effect can be used at the input. So far, however, these tubes have only given an extremely low maximum output power [16].

The factors determining the upper limit of the output power

We have just seen (fig. 7) that the output power does not continue to increase with increasing input power but shows saturation, or may even eventually decrease again. In our view this must be due to the fact that the oscillating plasma-electrons can absorb so much energy that they are able to ionize the gas molecules (or atoms). The charge density of the plasma increases when this happens, and the plasma frequency also increases (equation 1b). If the plasma frequency rises above the signal frequency the gain falls, which results in saturation of the output power.

This explanation is supported by various experiments. Measurements of the output power, the Q (quality factor) and the parallel resistance of the output cavity of the two-cavity tube shown in fig. 5 can be used to derive a good estimate of the a.c. component of the beam current at the height of the output cavity gap. From this in turn one may determine the oscillation velocities which the plasma electrons can have. It now appears that the highest levels found for the kinetic energy of these electrons are in fact in the neighbourhood of the ionization energy of the gas. On the other hand, in the tube in question, at the input signal power at which saturation occurs, the modulation is still small enough for non-linear phenomena to be of very little importance.

Further support is obtained from the result of experiments in which measurement is carried out simultaneously at two frequencies, one being well above the plasma frequency. If the input power of the signal at the lower frequency is increased, it is found not only that the output power shows saturation at that frequency, but also that the gain increases at the highest frequency, see *fig. 10*. This is precisely what is to be expected with the assumed increase of the plasma frequency, and it proves that the saturation phenomenon is not due to beam disintegration. A further indication that the density of the plasma rises is to be found in the increase in the light emission. This light

[15] See W. Kuypers and M. T. Vlaardingerbroek, Philips Res. Repts. 20, 349, 1965.

[16] Some theoretical considerations on this method of signal input and output as well as a description of the construction and characteristics of our tubes are given in Electronics Letters 2, 368-370, 1966.

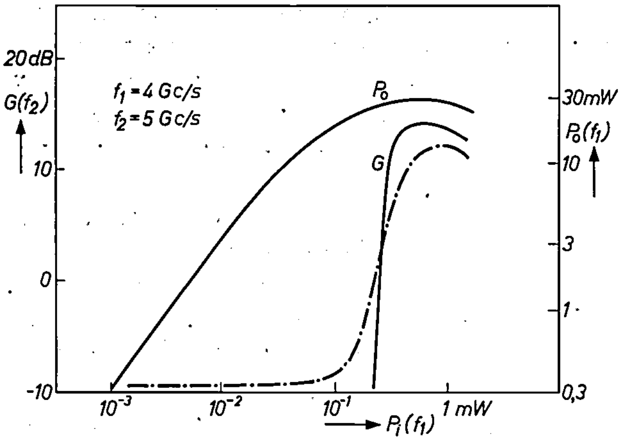


Fig. 10. Result of experiments using two input signals. One signal had a frequency f_1 (4 Gc/s) at which considerable amplification was possible: the frequency f_2 of the other (5 Gc/s) was well above the plasma frequency. If the power P_i of the first signal is increased, then at the value of P_i, at which the output power P_0 reaches its maximum, the gain G of the other signal increases considerably. This shows that the plasma frequency has approached close to f_2, i.e. the density of the plasma has increased. In agreement with this, the light emission (chain-dotted line; ordinate scale in arbitrary units) shows a similar increase.

emission is caused by the return of ionized atoms to the ground level and is therefore proportional to their density.

During the very short time after the sudden application of a strong input signal, it should indeed be possible for a considerable amplification to take place before the plasma density reaches its new value. We have in fact succeeded in demonstrating this experimentally. When a powerful input signal was suddenly applied to a helix tube the output signal was considerably stronger during the first few microseconds than afterwards. As the light emission showed, the plasma density in that period had not yet reached its equilibrium value (*fig. 11*).

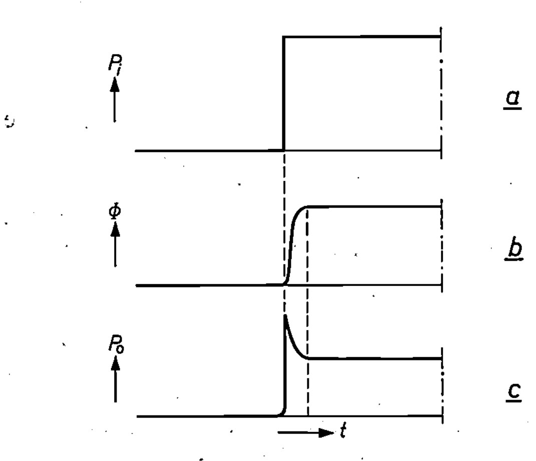


Fig. 11. The phenomena occurring when a powerful input signal (power P_i) is suddenly applied (*a*) — in all three graphs the time t is shown horizontally — also indicates that the amplification is limited by ionization. During the first few milliseconds in which, as appears from the curve of the light emission Φ, the plasma density has not yet reached its equilibrium value (*b*), the output signal (power P_0) is much greater than its final value (*c*).

A more detailed experiment using a series of input signals increasing in power, showed that P_0 at first increased with P_i; the initial value of P_0 then remained constant while the equilibrium value obtained after a few μs decreased steadily (*fig. 12*). This corresponds completely with the form of the curve of fig. 7. The constancy of the initial value shows that at the relevant P_i value the kinetic energy of the plasma electrons had exceeded the value E_i required for ionization. The output power could therefore not increase further, for electrons whose kinetic energy has exceeded E_i can only retain this energy until they are involved in a collision leading to ionization. At the gas pressure and degree of ionization applicable in our tube this time was however extremely short (order of magnitude 10^{-8}s), so that the higher kinetic energy of the plasma-electrons was mostly used up for ionization and could only negligibly contribute towards the output power.

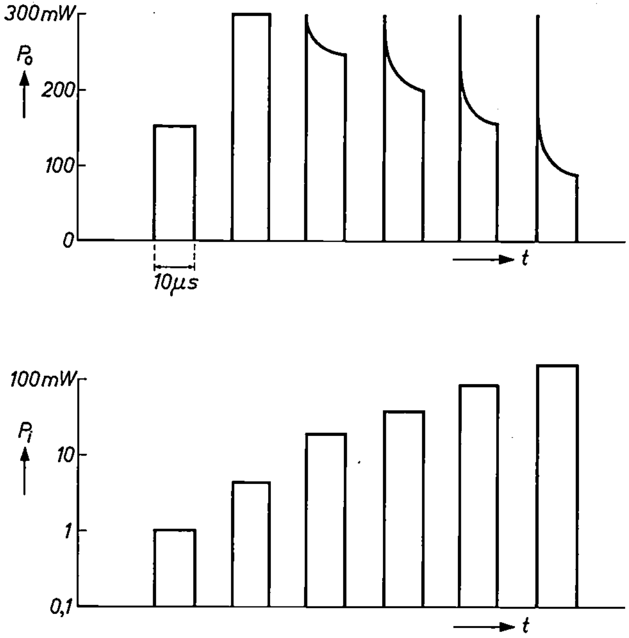


Fig. 12. As in fig. 11, this time for six successively increasing values of P_i. When P_0 had reached its equilibrium value — which takes place within 10 microseconds — the signal was removed. As soon as the effect shown in fig. 11 occurred, the left-hand side of the curves for P_0 remained at a constant height and the height of the right-hand side steadily decreased.

The conclusion to be drawn from the above is that a higher c.w. output power can only be achieved by working with a fully ionized plasma; the charge density then cannot increase further and therefore the plasma frequency f_p cannot change, thus reducing the gain. This in effect limits the choice of gas to hydrogen or deuterium, as otherwise double ionizations may be encountered. The charge density for an f_p of say 6000 Mc/s, as in our tubes, is of course obtained at a lower gas pressure when ionization is complete than with incomplete ionization; a pressure of 2.5×10^{-5} torr is sufficiently great. At so low a pressure the chance of a collision between beam electrons and gas atoms is slight, however, so that an extremely powerful beam must be used to obtain the complete ionization desired. This method of obtaining a large pulsed output power has been used by Allen and Chorney [17] in a tube for 3000 Mc/s.

Low-frequency oscillations

The ionization which occurs because the oscillating plasma-electrons attain too high a kinetic energy at high input power causes yet another unwanted effect besides the saturation of the output power just mentioned. On increasing P_o to the saturation value oscillation occurred in each of our tubes at a frequency between 0.5 and 2.5 Mc/s. This oscillation revealed itself through: 1) the presence of sidebands in the output signal, 2) modulation of the collector current, 3) modulation of the light emission; see *fig. 13*. This oscillation, whose character appears to be related to the fluctuations in the microwave emission of large beam-plasma systems, can best be explained by assuming that there are local density fluctuations in the plasma. So far nothing has been discovered to contradict this hypothesis, in contrast with others which may be put forward. The hypothesis that interaction with the ion gas takes place could be rejected, as the ion plasma frequency of mercury vapour is about 10 Mc/s in a tube for 6000 Mc/s. Nor can the low-frequency oscillation be due to any transverse oscillation: the cyclotron frequency in our tube is about 2000 Mc/s for electrons, and only 5.5 kc/s for ions.

We believe that the plasma density fluctuations in these tubes arise as follows. As soon as the plasma electrons acquire a sufficiently large oscillation energy to cause ionization, the plasma density at the back of the tube becomes greater than elsewhere. Due to ambipolar diffusion (i.e., diffusion of both kinds of charge carriers together) the bunch spreads out in the direction of the electron gun. In a considerable part of the tube, because of the increase in the charge density, the plasma frequency then rises so far above the frequency of the input signal that the gain is reduced. As a result the plasma electrons at the back of the tube cannot absorb as much energy as before nor can they reach a high enough velocity to cause ionization. This state continues until the plasma density has decreased to its original level, and the process then repeats.

The power necessary for ionizing collisions of the plasma electrons, is of course taken up by these from the d.c. power of the beam through the beam-plasma interaction. As soon as low-frequency oscillations take place, the power given up by the beam will thus vary

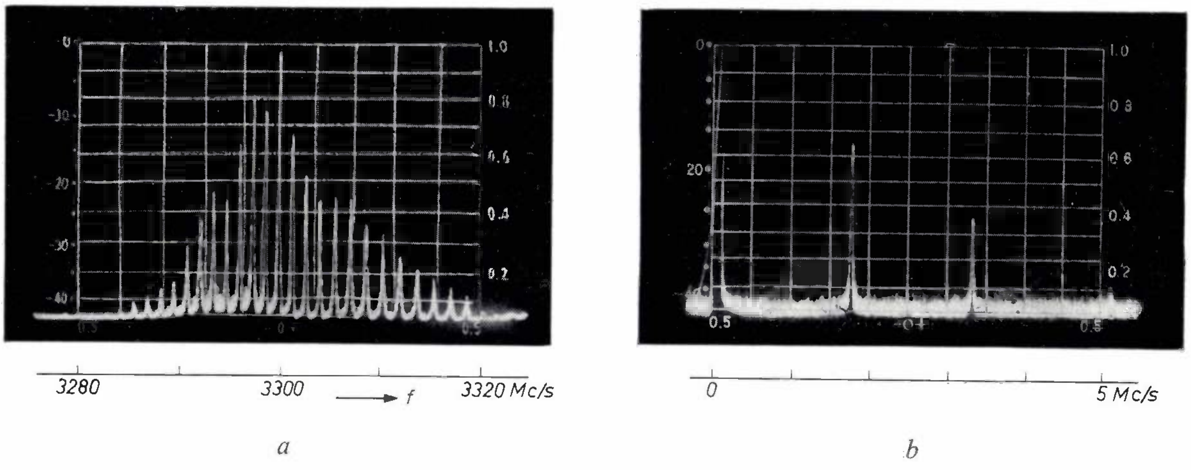


Fig. 13. *a*) Spectral energy distribution of the output signal of a heavily loaded beam-plasma amplifier. In addition to a peak at the input signal frequency (3300 Mc/s), there are a number of associated peaks. The frequency difference Δf between all neighbouring peaks is the same (1.5 to 2 Mc/s), which indicates that low-frequency oscillations at a frequency Δf occur in the tube.
b) If the light emission from such a tube is detected and the periodicity of its variations is analysed, it is then found that the frequencies which occur are equal to $n\Delta f$ ($n = 1, 2, \ldots$).

[17] M. A. Allen and P. Chorney, Int. Conf. on the microwave behaviour of ferrimagnetics and plasmas, London 1965.

with the periodicity of the oscillations and its average level will moreover be greater than before. This implies that the average velocity of the beam electrons must vary in the same way. Attempts to measure the velocity distribution in the beam give results which indicate that this is in fact the case.

On looking through the foregoing it can be readily understood that beam-plasma tubes are not particularly attractive as microwave amplifier tubes. The noise is relatively high and moreover, as we have just seen, there is the unwanted effect of low-frequency oscillation at high input signal levels. One further purely practical point which weighs against the beam-plasma tube is the fact that its cathode life is relatively short as a result of ion bombardment.

On the other hand however, the study of these tubes provides a contribution to plasma physics in general, both directly and as far as methods of approach are concerned. This also appears to apply to a branch of plasma physics not discussed above, the study of plasmas in solids. Research in this field may be expected to give an insight into the nature of the current instabilities that have been observed in a number of semiconductors.

Summary. In a neutral plasma, electron oscillations can occur which do not spread out and whose frequency f_p, the plasma frequency, is dependent on the charge density of the electron gas. They can be excited by a beam of charged particles, e.g. electrons. Through interaction of the beam and plasma a density modulation of the beam can be amplified provided its frequency f is lower than the plasma frequency f_p. This effect is the fundamental principle of beam-plasma microwave amplifier tubes. In these tubes the beam is modulated and an output signal is taken off by means of a resonant cavity or a helix. Calculations show that the amplification in helix tubes is greatest when f is a little higher than f_p; moreover, the beam velocity should be a little higher than the phase velocity of the signals in the helices. The experiments confirmed this. Helix tubes have a greater bandwidth than free beam-plasma systems or tubes with resonant cavities. A tube 10 cm long gave a gain of 60 dB at 5000 Mc/s. The noise figure could not be reduced below 25 dB. The output power reaches its limit when the plasma electrons themselves begin to give ionization, causing f_p to decrease. At high input power unwanted low-frequency oscillations occur (at about 2 Mc/s). This last effect presents the greatest obstacle to the practical application of beam-plasma. tubes as microwave amplifiers.

Volume 27, 1966, No. 9/10 pages 241-284 Published 17th January 1967

PHILIPS TECHNICAL REVIEW

VOLUME 27, 1966, No. 11

The "COBRA", a small digital computer for numerical control of machine tools

R. Ch. van Ommering and G. C. M. Schoenaker

621.9-523.8

In the last few years, machine tool exhibitions have shown an increasing emphasis on automatic control. The latest development in this field is the application of a small universal computer for the actual control process; this arrangement, which has already been taken up by several control equipment suppliers, offers a number of exceptional advantages. In the present article the authors discuss the general properties of such systems and describe the prototype of a computer developed specially at Philips Research Laboratories for machine control. An account of early practical experience obtained with the system is given.

Numerical control of machine tools gradually appeared in engineering workshops and other metal-working industries some ten years ago. Many problems arose in the development of this form of automation; the first of these was with the reliability of the electronic equipment. As well as this it soon appeared that the machine tools current at the time were not particularly suitable for numerical control; the same was true of the mechanical systems for tool and workpiece movement and the systems that measured these displacements. It was also found that the practical utility of numerical control depended to a considerable extent on the programming possibilities, i.e. on the way in which workpiece data could be fed to the automatic control apparatus.

After only a few years these problems have been largely solved; we have now reached a situation in which electronic and mechanical systems no longer give rise to fundamental difficulties, and it has also been possible to find suitable methods of programming.

The solution of the old problems however has brought new ones in their turn — problems determined primarily by the increasing demand for numerical control. These relate to the total design of the control system and its adaptation to the various types of machine tools, taking this term to include machines such as drawing and engraving machines. We shall deal particularly with these problems in this article, which presents the desirability and possibility of a more efficient and more universal system of numerical control than is usually used at present. But first we shall recapitulate a few general aspects of numerical control: an extensive examination of this subject appeared earlier in this journal [1] [2].

Numerical control

"Numerical control" usually refers to a system in which all data relevant to the making of a workpiece are fed to the machine tool in the form of numbers: this operation is performed by means of a "command unit". These data relate not only to the dimensions of the workpiece, but also to spindle and cutting speeds of the machine tool, and to tool changing, if this is done automatically. The command unit interprets the data and feeds signals to the units which are to be controlled. Generally, for example with slide movements, the signal indicates a desired displacement which is then compared with the actual displacement, after which a control system ensures that the difference between the two values is kept smaller than a preset value. *Fig. 1* shows this arrangement in the form of a block diagram.

The numerical data can be fed to the command unit in various ways. The simplest method is manual input, in which the numerical values are set by means of

Ir. R. Ch. van Ommering is with the Philips Development Coordination Office, Eindhoven, and Ir. G. C. M. Schoenaker is with Philips Research Laboratories, Eindhoven.

[1] T. J. Viersma, Some considerations on the numerical control of machine tools, Philips tech. Rev. 24, 171-179, 1962/63.

[2] J. A. Haringx, R. Ch. van Ommering, G. C. M. Schoenaker and T. J. Viersma, A numerically controlled contour milling machine, Philips tech. Rev. 24, 299-331, 1962/63.

switches and buttons arranged on the command unit. This method is only used if the amount of information is limited, e.g. for the control of very simple jig boring machines. If more data for the workpiece or machine are needed, punched tape or magnetic tape has to be used. A good example of this is a contour milling machine, for which manual input of the data is out of the question.

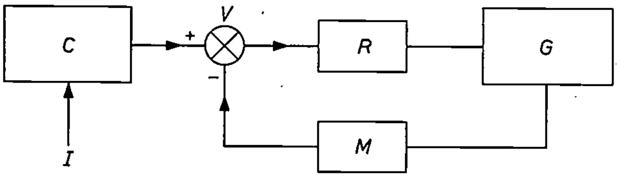


Fig. 1. Block diagram of the numerical control of a machine tool G with the help of a command unit C. The data concerning the workpiece, I, are fed to the command unit, and this calculates from these what actions have to be performed and gives instructions to the machine tool. For the performance of the necessary displacements such as slide movements, there is a control system R with a measuring system M and a comparator unit V.

The contour is described by a series of discrete points on the tape. If a punched tape were to be made for a complicated workpiece with very many points the tape would be unmanageably long. To get over this difficulty, the command unit is made to perform a computation, an *interpolation* between two given points; the tape then carries only as many points as are necessary to ensure that, after interpolation, the desired piece of work is produced with the accuracy required.

For simple workpieces, e.g. with only a limited number of plane surfaces, punched tape has the advantage, in comparison with magnetic tape, that it can be made by a programmer by hand (using a typewriter with a punched tape output). If the workpiece is more complicated, then the determination of the points that have to be fed to the command unit often demands so much calculation that the help of a computer is necessary. A computer is able to transfer the workpiece data as easily to magnetic tape as to punched tape, but it must be borne in mind that for magnetic tape an expensive output unit is also required; on the other hand the actual control of the machine is simpler. Both kinds of tape are used in practice. Punched tape has however, apart from the possibility already mentioned of being prepared by hand, a number of other advantages. It is easy to handle, to prepare, and to duplicate. Furthermore punched tape is already in common use in many workshops.

Once the workpiece dimensions have been properly recorded on the tape, the production of the workpiece on the machine requires a minimum of time. The cutting process can in fact be performed without intermediate checking of the dimensions and without having to set up the workpiece again. Besides the great saving of time that can thus be obtained, particularly for complicated workpieces, there is the further advantage that fewer auxiliary tools (jigs, gauges etc.) are needed than with conventional methods of manufacture.

The above features of numerical control show that this method of processing will provide special advantages in the production of a few pieces or small batches of workpieces where strict requirements have to be met.

The desirability of a universal control system

As a result of the growing demand for numerical control and the variety of the applications, many different control systems have arisen which are suitable for only one particular machine tool or one particular type of machine tool.

The making of a specialized machine does not at first sight present much difficulty, at least if long runs of one type can be made. However, in Europe the numbers in any one type are still small, and for most types it looks as though this situation will not change very much in the next five years. For this reason efforts have been made to find a universal control system, i.e. a system easily and quickly adaptable to many different machine tools. Such a system can be built in large production runs. It is then possible to supply a control system quickly, even for exceptional applications, and maintenance and the supply of spare parts can be made easier.

There is a further reason why a universal control system is attractive. With the existing control systems the processes of computation (e.g. interpolation) and the instructions ("start", "stop" etc.) are built into the wiring; this is called "wired logic". The utility of such a control system is then not only limited to one type of machine tool, but moreover only those computing processes and instructions can be performed with it which were envisaged during the design of the system. Subsequent changes can be made only with difficulty. The possibility that changes (e.g. new computing processes) may be needed is far from imaginary. Numerical control is developing quickly, and future requirements are difficult to predict. This implies that even during a period of five to seven years, in which the machine has to be written off, the need may arise for modifications to the machine. A universal system in which this can be done easily will have the advantage that it can always be adapted to include the latest facilities, so that the combination of machine tool and control system, as far as useful value is concerned, will be a match for the most modern equipment appearing on the market. A system that cannot be adapted will fall in value from year to year.

Application of a digital computer as a control system

A small digital computer of the type used for process control is quite suitable for use in a universal control system. Such a computer consists of an arithmetic unit, a store, a unit for internal control which ensures that the computing processes are carried out correctly, and a unit for the input and output of information. A computer can be adapted to the machine tool to be controlled by feeding the data relating to the machine tool (e.g. co-ordinate system and measuring accuracy) and the information necessary for a correct interpretation of the workpiece data into the store by means of a punched tape. This simple adaptation procedure takes very little time, once one has the punched tape, and thus complies with the requirements we established above for a universal control system. The method in which the data are put into a store is known as "stored logic". The use of a computer as an automatic control unit is really rather obvious, since all actions to be performed by an automatic control unit can be reduced to simple arithmetic operations; this is also the case in automatic control units with "wired logic". The processes which the design of a piece of work undergoes between the drawing office and the machine tool are also almost exclusively computing processes, and these are often performed with a large computer. The dividing line between the computing processes performed by the automatic control system and the processes performed by the large computer can be drawn at various different places.

We have stated that a control computer is adapted to a machine tool by putting a series of computing processes in the store. It follows from this that the size of the store limits the possibilities of adaptation. The control of a contour milling machine, for example, calls for a larger store than the control of a jig boring machine. There is no difficulty, however, in designing the store in such a way that it can be easily extended, for example by units of equal size. This does not make the control system any less flexible, while an efficient utilization of the storage capacity is always obtained.

Another important advantage of the use of a true computer is that the electronic components are used efficiently. This is because a large number of different calculations can be carried out successively with the same components owing to the extremely high computing speed possible with these components. A stored logic system will in general contain fewer components than the equivalent system with wired logic.

Finally the use of a computer has the advantage that it is possible to profit directly from the extensive research work and the rapid developments which are taking place in this field.

The COBRA

In the foregoing we have stated that a small digital computer is suitable for use as a universal control system. To investigate the correctness of this proposition a trial model, popularly known as the "COBRA", was built at Philips Research Laboratories; the design of this computer makes it suitable for two-dimensional contouring control.

We shall now briefly describe the various components of the COBRA (*fig. 2*) [3]. This will be followed by a treatment of the various computing processes which the COBRA can perform, and attention will then be given to the part programming i.e. the preparation of the punched tape knowing the workpiece data and the machining method.

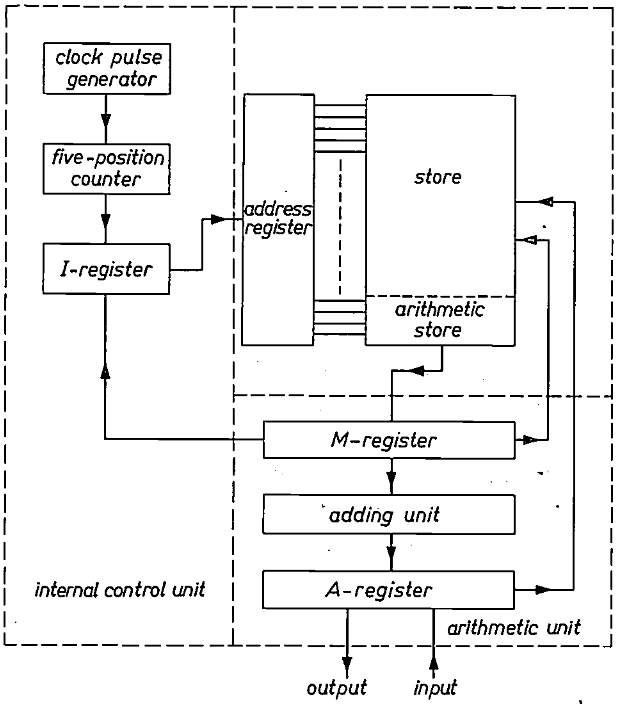


Fig. 2. Diagram of a small electronic computer. The functions of the three principal components, which are the arithmetic unit, the store, and the internal control unit, are clarified further in the text.

The arithmetic unit contains two registers and an adding unit. The *M*-register forms the connection with the store; during addition the *A*-register stores one of the two numbers being added. After the calculation this number is replaced by the sum of the two numbers. The addition is performed by a 24 bit parallel adding circuit.

[3] For the terms and concepts from digital computing techniques used here, see e.g. W. Nijenhuis, The PASCAL, a fast digital electronic computer for the Philips Computing Centre, Philips tech. Rev. 23, 1-18, 1961/62.

A magnetic core store working on the coincidence principle is used for data storage. This contains 1024 addresses, each having a word length of 24 bits: it is therefore possible to store numbers of about seven decades ($2^{24} = 16\,777\,216$). A section of the store, with 64 addresses, is used as a "computer store": special wiring enables this to be used to calculate the complement of a number ($A \rightarrow 2^{24} - A$) and to multiply a number by a power of 2 ($A \rightarrow A \times 2^k$, k being an integer). We shall not discuss this further here [4].

The data held in the store consist of numbers relating to the slide positions, results of calculations etc. and of instructions indicating the actions to be performed by the computer. An instruction consists of three parts (*fig. 3*). The first part, the code part, indicates a certain process, such as addition, multiplication by 4, etc. The second part determines the place in the store at which the number to be processed is to be found (the operand address). Finally the third part indicates the location in the store where the next instruction can be found. This location is called the instruction address.

code	*operand address*	*instruction address*
6 bits	*8 bits*	*10 bits*

Fig. 3. An instruction consists of a code part of six bits for the indication of the desired operation (e.g. Add), an operand address of eight bits indicating the place in the store where the number to be processed can be found, and an instruction address of ten bits referring to the subsequent instruction.

Of the 24 bits available for an instruction, 10 are used to indicate the address of the subsequent instruction. These can be used to indicate each of the 1024 addresses ($2^{10} = 1024$). To indicate the operand address, only 8 bits are used, which at first sight does not seem sufficient. From the analysis of a normal control programme it is found, however, that the major part of the store is occupied by instructions and only a fairly small part by numbers. If we make sure that these numbers are accommodated at the beginning of the store we do not need more than 8 bits.

Finally, there are 6 bits left over for the code part.

The use of 10 bits for the instruction address and 8 bits for the working address makes it clear that the store cannot be extended without taking further measures. If an extension, e.g. a doubling, is necessary, then it is possible to indicate which of the two parts of the store contains the relevant numbers and instructions by means of a separate instruction beforehand. For efficient time use, the data for successive computing processes should obviously as far as possible be accommodated in the same part of the store.

The internal control unit ensures that the instructions are carried out. This unit consists basically of a clock pulse generator, a 5-position counter, and a register, the *I*-register, in which the instruction to be carried out is temporarily stored. The pulse generator supplies pulses with a frequency of up to about 300 000 c/s (the frequency is variable for reasons to be mentioned later). The pulses are fed to the 5-position counter which divides the time into five intervals 1-5; in each set of five successive time intervals an instruction is carried out in five stages. We shall attempt to clarify what takes place at each stage by means of an example: the adding of a certain number from the store to a number which is already in the *A*-register.

Let us start with the situation in which a previous instruction is still present in the *I*-register. Its instruction address indicates the location in the store of the adding instruction which we are considering for our example. In the first time interval (the first stage) the new instruction is transferred from the store to the *M*-register of the arithmetic unit. In the second stage this instruction is brought over to the *I*-register and is also returned at the same time to the original place in the store. In the following stage the number indicated in the working address of the instruction, now to be found in the *I*-register, is transferred form the store to the *M*-register. In the fourth stage this number is added to the number in the *A*-register. In the fifth stage, finally, the result of the calculation is put into the store at the location from which the first number was taken. The first two stages are the same for every instruction. The three other stages depend on the instruction that is being executed.

At the maximum frequency of the pulse generator some 60 000 instructions can be carried out per second. This speed is necessary because various simple calculations, each consisting of a number of instructions, must be carried out about 1000 times per second for proper functioning of the control.

As well as the simple calculations there are also a number of more complicated ones which need to be carried out only once or twice per second. These complicated calculations, which require a large number of successive instructions, have to be interrupted several times, again with the help of one or two instructions, in order to carry out quickly one of the simple calculations.

Finally we must mention the input and output units of the COBRA, which are connected to the arithmetic unit via the *A*-register. These units handle the supply of workpiece data and of information derived from the measuring systems, and supply the commands to the servomotors. These operations are again carried out with the help of instructions. The COBRA is equipped with four input and four output facilities. One input is reserved for 8-channel punched tape and a second input for switches for setting up numbers of up to five decades (we shall see later that separate setting up of the cutter diameter, for example, is required). A wide variety of measuring systems and servomotors can be connected to the other inputs and outputs with the help of different types of matching equipment. The digital circuits of the COBRA are made up from 800 circuit blocks [5].

The computing processes

Let us now investigate what computing processes the COBRA has to perform by considering the special case of a two-dimensional contour milling machine.

Interpolation

Let the full line in *fig. 4* represent part of the desired contour. The co-ordinates of the points *A*, *B* and *C* are given on the punched tape. (The co-ordinates of the points given on the punched tape are rounded off to an integral multiple of the smallest unit of measurement. *A*, *B* and *C* will therefore as a rule not lie exactly on the contour.) The COBRA must now be able to interpolate between the points *A* and *B*, *B* and *C* etc. in order to calculate the approximate contour (the broken line in fig. 4). This is done by parabolic interpolation. This computing process, which has to be carried out both for the movement in the *x*-direction and for the movement in the *y*-direction, has to be repeatable 1000 times per second in order to obtain a given accuracy for the workpiece at a certain cutting speed.

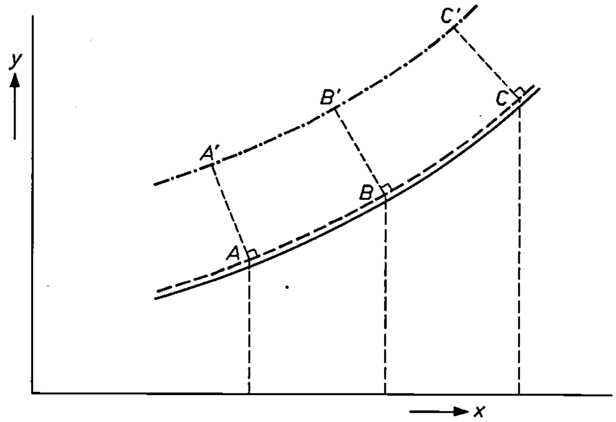


Fig. 4. The full line is an element of the desired workpiece contour. *A*, *B*, *C* are three points given on the punched tape. *A'*, *B'*, *C'* are points on the tool path which lie at a distance of half the milling cutter diameter from the workpiece contour.

The interpolation is carried out by means of an interpolation formula due to Bessel. For the determination of the value of *x* at the n^{th} point of *N* required points between two successive given values of *x* (x_k and x_{k+1}), the formula takes the following form:

$$x_n = x_k + \frac{n}{N}(x_{k+1} - x_k) + \frac{n(n-N)}{4N^2}(x_{k+2} - x_{k+1} - x_k + x_{k-1}) + \ldots .$$

Here x_n is the desired position of the tool in the *x*-direction. In the parabolic interpolation used here the terms of higher power are neglected.

An earlier article in this journal [2] has already described how a skilful choice of the data recorded on the punched tape enables the calculation of x_n to be reduced to between 3 and 6 additions. The amount of data to be recorded on the tape by this method — the method of "third differences" — is small, and this gives the advantage of a small tape length.

The cutter path

Knowledge of the contour of the workpiece is not always a sufficient condition. In a milling machine the cutter path is particularly important. In fig. 4 the chain-dashed line indicates the cutter path and the distances *AA'*, *BB'* and *CC'* are equal to half the diameter of the cutter. The co-ordinates of *A'*, *B'* and *C'* have again been rounded off to an integral multiple of the smallest unit of measurement. With the aid of the computer with which the punched tape is prepared it is possible to calculate the points *A'*, *B'* and *C'* of the cutter path instead of the points *A*, *B* and *C* on the workpiece contour, and to record these data on the punched tape. This method is in fact usually adopted, but then the difficulty arises that the diameter of the cutter which must be used for shaping the workpiece is as tightly specified as the workpiece. To overcome this objection the COBRA is provided with a computing process which makes it possible to calculate the cutter path for an arbitrary diameter of the cutter or to calculate the correction of the path for an arbitrary *change* in the diameter of the cutter.

The computing process consists essentially in setting up the perpendiculars *AA'*, *BB'* and *CC'* on the contour at points *A*, *B* and *C* (fig. 4). For the calculation of the small distances ($x_{A'} - x_A$) etc. use is here made of the formula:

$$x_{A'} - x_A = \frac{R_f F_y}{\sqrt{F_x^2 + F_y^2}} .$$

[4] H. J. Heijn, Representations of switching functions and their application to computers, thesis, Delft, 1960, also published in Philips Res. Repts. **15**, 305-341 and 448-491, 1960.

[5] These circuit blocks are of a type to be described in an article by E.J. van Barneveld, which will appear shortly in this journal.

Here $R_{\rm f}$ is half the diameter of the cutter (or half the correction of the cutter diameter). F_x and F_y are the x and y components of the cutting speed; their values are determined in a computation to be discussed below.

Once the co-ordinates of the points A', B', C', etc. are calculated, the intermediate points of the cutter path are determined by interpolation.

The cutting speed

In milling, the cutting speed must always remain approximately equal to a certain desired value which is given by means of the punched tape. The COBRA can control the cutting speed by changing the frequency of the pulse generator in the internal control unit. For this adjustment the x and y components of the cutting speed F used are calculated with an interpolation equation:

$$F_x = A \left\{ \frac{1}{N}(x_{k+1} - x_k) - \right.$$

$$\left. - \frac{2n-1}{6N^2}(x_{k+3} - 3x_{k+2} + 3x_{k+1} - x_k) + \ldots \right\},$$

and similarly for F_y. The constant A contains the frequency of the pulse generator. Knowing F_x and F_y we then have:

$$F = \sqrt{F_x^2 + F_y^2}.$$

The difference between F and the desired value supplies the correction signal used to control the frequency of the pulse generator.

The equations show that for calculating the cutting speed and for calculating the cutter path it is several times necessary to multiply, divide, and extract a root, and these calculations therefore require considerably more instructions and time than the calculation of the approximate contour. On the other hand the cutting speed and the points of the cutter path have to be calculated only a few times per second.

Various other calculations

Many workpieces have their shape determined by straight lines and arcs of circles. A special computing process sometimes required in practice is the determination of such workpiece contours from only a few co-ordinates. A simple hand-prepared punched tape normally suffices for these contours. The relevant computing process is rather complicated for arcs of circles, and the formulae to be used will not be treated here.

In addition the COBRA can carry out a number of checking calculations to investigate whether interpolation and the calculation of the cutter path are proceeding satisfactorily. In the interpolation between the points A and B it is possible, for example, to ascertain whether the end point of the calculated path element does indeed correspond to point B. A continuous check is also made to find out whether the tape-reader is able to read the information off the tape fast enough at the stated speed. These checking computations generally consist of a number of simple additions.

One further computation which the COBRA is sometimes called upon to undertake in the control of a contour milling machine is the comparison of the actual location with the desired location. This requires the transfer of the information from the measuring system to the COBRA. This comparative calculation, like the checking computations, is fairly simple but, especially if an incremental measuring system [2] is used, it has to be performed at a repetition frequency which is high with respect to the maximum number of displacements of the tool per second.

The programming of the COBRA

We shall now deal in a little more detail with the programming, which has already been briefly mentioned above.

Two programmes are supplied to the COBRA by means of punched tape. The first programme consists of the workpiece data and is therefore called the *workpiece programme*. The second programme consists of instructions and numbers required for a correct interpretation of the workpiece programme and for a correct control of the machine tool. This second programme, which is accommodated semi-permanently in the COBRA store, is the *control programme*. The workpiece programme for the COBRA is assembled in much the same way as for control systems with "wired logic". The drawing up of a control programme only occurs with systems which have "stored logic".

The workpiece programme

The data for a workpiece are subjected to four processes before machining can start. The first three processes supply the workpiece programme; the fourth process is carried out by the COBRA itself. The four processes are:

1) Description in "programming language" of the geometry of the workpiece and of the way in which the tool has to machine it.
2) Calculation of the ideal cutter path (or the workpiece contour) with the help of a large computer, in a computer centre.

[6] APT is an abbreviation of "Automatically Programmed Tools". An American research group (the Illinois Institute of Technology Research Institute—IITRI) is working on the further development of this language, supported by a large number of American firms and some ten European firms, Philips among them.

3) Determination, again with the help of a large computer, of the form and nature of the co-ordinates and other data which must be included on the punched tape so that the control system used will give the desired workpiece with the required accuracy.
4) Calculation, from the data of 3) above, of the actual cutter path. The COBRA does this itself, by interpolation between the points specially established, during the third process, for the COBRA and the interpolation method to be used.

The programming language chosen by Philips is the APT language [6], specially designed for programming numerically controlled machine tools.

Like many other programming languages, the APT language also consists of a large number of "statements" — short indications of a piece of information or of a mechanical action — which describe the workpiece and its production. The first step in programming thus consists of drawing up a list of statements.

In the second step the workpiece data, in APT language, are supplied with the help of punched cards to a computer in the computer centre. This computer must be equipped with a computing programme adapted to the APT language. For many large computers, special computing programmes of this kind, known as compiler programmes, have already been set up.

The results of the computations can be . stored, with further data, on magnetic tape.

The workpiece data have now been converted to a form which is still quite independent of the control system to be used. By means of a second special computing programme, the postprocessor programme, the typical features of the control system to be used (e.g. the method of interpolation) are therefore included in the programme during the third step of the programming process. This third step in the programming, which like the second step takes place in the computer centre and is often carried out immediately after the second step, results in the punched tape which will be supplied to the control system, in our case to the COBRA.

Each type of control system requires its own post-processor programme. The COBRA is no exception in this. But, as we shall see, the COBRA does have facilities for making certain parts of the post-processor programme itself.

The creation of a workpiece programme is shown schematically in *fig. 5*.

Fig. 6 shows a drawing of a simple workpiece which will serve as an example for considering the program-

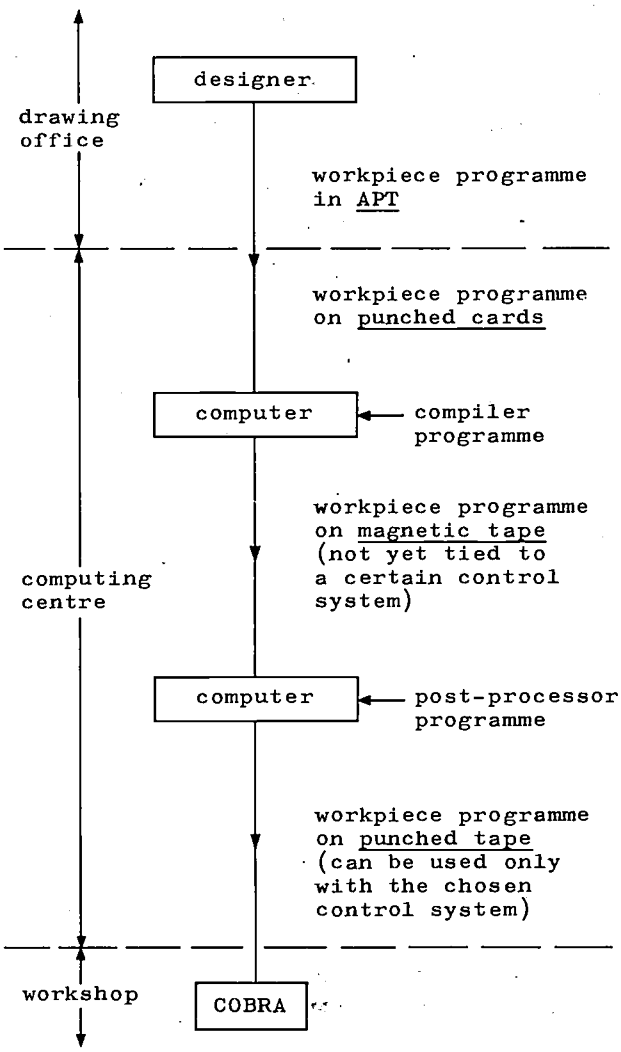

designer
drawing office
workpiece programme in APT
workpiece programme on punched cards
computer
compiler programme
workpiece programme on magnetic tape (not yet tied to a certain control system)
computing centre
computer
post-processor programme
workpiece programme on punched tape (can be used only with the chosen control system)
workshop
COBRA

Fig. 5. Diagram of the processes undergone by the workpiece programme between the drawing office and the workshop. The various processes and the special computing programmes required for them are explained in more detail in the text.

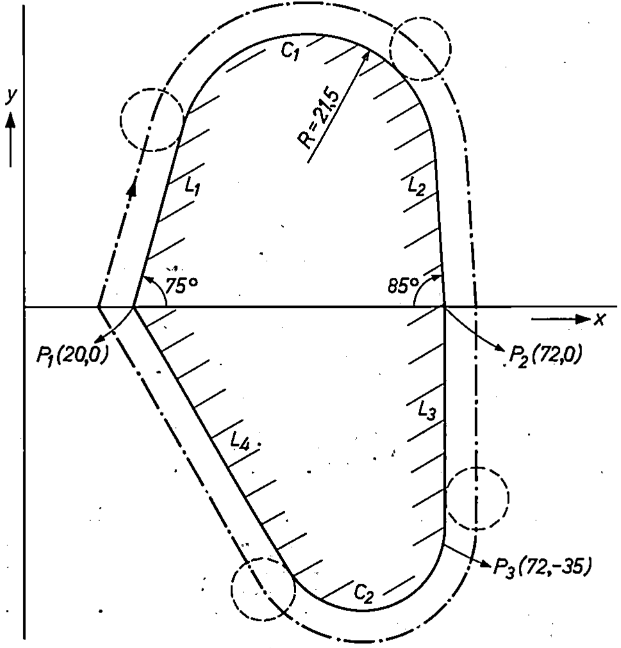
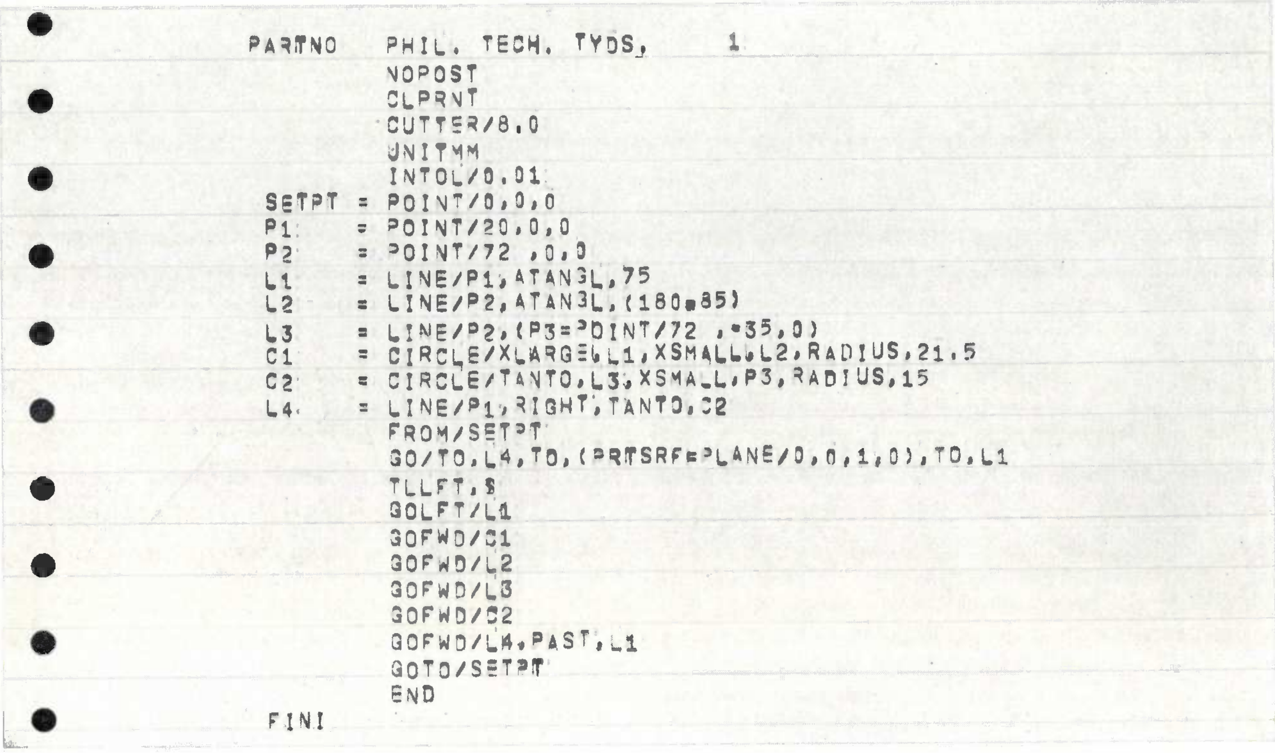


Fig. 6. A simple workpiece consisting of straight lines and arcs of circles. The chain-dashed line is the path of the centre of the milling cutter; the dashed-line circles represent the cutter (diameter 8 mm). The programme, in APT, for this workpiece is given in Table I.

Table I. The workpiece programme, in APT language, for the workpiece of fig. 6.



```
PARTNO   PHIL. TECH. TYDS,        1
         NOPOST
         CLPRNT
         CUTTER/8.0
         UNITMM
         INTOL/0.01
 SETPT = POINT/0,0,0
 P1    = POINT/20,0,0
 P2    = POINT/72 ,0,0
 L1    = LINE/P1,ATAN3L,75
 L2    = LINE/P2,ATAN3L,(180•85)
 L3    = LINE/P2,(P3=POINT/72 ,•35,0)
 C1    = CIRCLE/XLARGE,L1,XSMALL,L2,RADIUS,21.5
 C2    = CIRCLE/TANTO,L3,XSMALL,P3,RADIUS,15
 L4    = LINE/P1,RIGHT,TANTO,C2
         FROM/SETPT
         GO/TO,L4,TO,(PRTSRF=PLANE/0,0,1,0),TO,L1
         TLLFT,$
         GOLFT/L1
         GOFWD/C1
         GOFWD/L2
         GOFWD/L3
         GOFWD/C2
         GOFWD/L4,PAST,L1
         GOTO/SETPT
         END
 FINI
```

me in APT (*Table I*) and the data supplied to the COBRA by means of the punched tape (*Table II*).

The workpiece is a flat shape which is to be milled with a tolerance of 0.01 mm by a cutter of 8 mm diameter. The circumference of the workpiece is formed by four straight lines and two arcs of circles. The points of intersection P_1 and P_2 of L_1 and L_2 with the x-axis (see fig. 6) and the angles made by L_1 and L_2 with the x-axis are given.The radius of C_1 is given. A further piece of information is that C_1 must meet L_1 and L_2 tangentially. L_3 is a straight line through P_2 and P_3. The radius is also given of C_2; furthermore C_2 must meet tangentially L_3 at the point P_3. L_4 is a straight line through P_1 which is tangent to C_2.

In the APT programme (Table I) a number of general data are given in the first seven lines, including the desired milling tool diameter and the tolerance. The points, lines, and arcs of the circles are defined in the next eight lines. The remainder of the programme indicates that the cutter must move from the starting point to the point of intersection of L_1 and L_4 and then from there to the left of L_1 upwards, along C_1, etc. until the end point is reached again.

Table II. The workpiece data for the workpiece of fig. 6, as given on the punched tape fed to the COBRA.

F20G8X1571Y—54	M3G1U2378V2956	G2U7381V2518	
I2463J—660	U7600V17G9	Y—3517G1	G2U4046V—44
35I—1900	U1571V—54G9	M4U0V0G8	M2

The APT programme can be supplied to the computer centre, where the path of the centre of the cutting tool is calculated. This has been drawn in fig. 6 Only a very limited number of data describing the path of the centre of the cutting tool are finally put on the punched tape, which can then be fed to the COBRA.

These data are shown in Table II; U and V indicate the absolute x and y co-ordinates and X and Y the displacements in these directions. The other letters are code indications of processes.

Examples are:

F 20 → cutter speed 20 mm/s
G 8 → straight line with cutter stationary at start and finish
M 3 → cutter down
etc.

The data shown in Table II occupy 35 cm of tape.

The control programme

Just as each workpiece requires a special programme to be made, so too does each combination of the COBRA with a certain type of machine tool necessitate a separate control programme. This task corresponds in content to the design of the wiring for a control system with "wired logic"; we shall see, however, in an example that the making of a COBRA control programme is rather more speedily completed.

In compiling the control programme it is necessary to take into account the way in which the workpiece programme has to be interpreted. This means to say that the data included in the workpiece programme with the help of the post-processor programme must also be considered in the control programme. However, the control programme must also ensure proper adaptation of the automatic control system to the machine tool, which means that allowance must be made for the measuring systems used, the servomotors, the number of slide movements etc. This aspect also covers the degree of acceleration and deceleration of the slide movements which, in some applications is connected with the accuracy to be achieved.

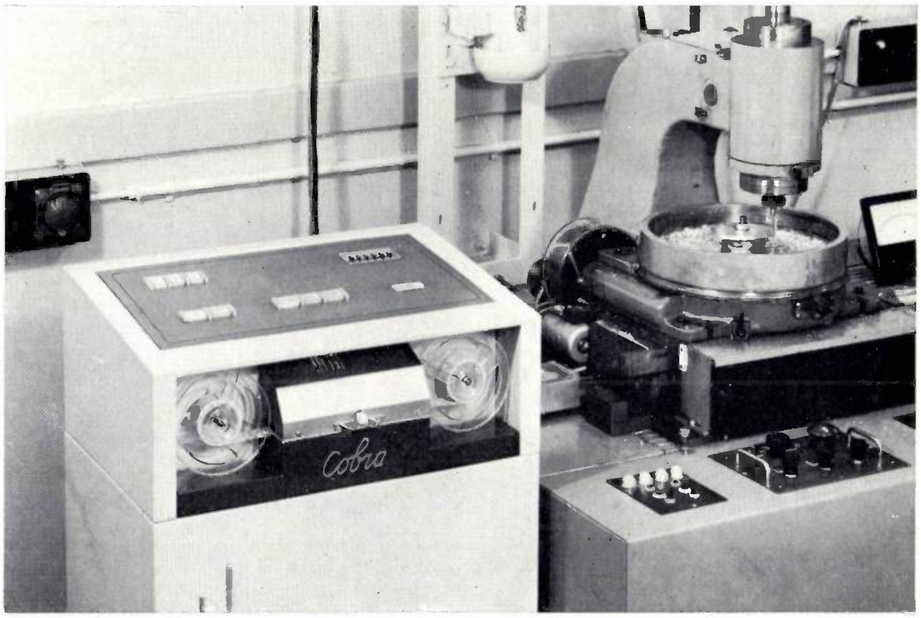


Fig. 7. The COBRA as an automatic control unit for a cam milling machine. The cam is fastened to a turn-table which can be moved backwards and forwards by means of a slide. The cutter cannot be displaced.

We have stated that with a universal control system a short delivery time should be possible, even for exceptional applications. In preparing the COBRA for a new application the drawing up of the control programme requires the most time. It is therefore necessary — and the possibility already exists in principle — that in future the same pattern should be followed for setting up the control programme as for setting up the workpiece programme. This means that a "language" must be developed which can be used to give a simple description of the control programme, so that with the help of the computer centre again, the punched tape with the control programme could be made automatically. The first steps towards the development of a suitable "language" have already been taken.

A special feature in the preparation of control programmes relates to the "mixed" running of complicated computations (e.g. cutter path or cutting speed) and simple ones (e.g. interpolation). This means that the times required for each of the stages of the computing processes must be accurately known. These times, the interruption points, and the storage of the intermediate results must be included in the programme of the computer, and this makes the preparation of a control programme specially complicated. This still holds if all instructions are made of equal length, a simplification introduced in the COBRA.

Although the preparation of the control programme for a particular machine tool only has to be done once, the facility of automatic interruption, which we call the "interrupt feature", will be incorporated in a future version of the COBRA. This can be done by including another pulse generator which, for a situation in which the rapid calculation has to be carried out 1000 times per second, would give pulses at a frequency of 1000 c/s; these pulses would always initiate a separate instruction which would interrupt the longer computation, irrespective of the stage in which it is at that moment. At the time of the interruption the intermediate result must also automatically be stored, and after the conclusion of the rapid computation i.e. a fraction of 1/1000 s later, the long calculation must be restarted.

Practical experience

The practical experience obtained with the COBRA has been with applications of widely different types, aptly illustrating the many possibilities of the system.

Control of a cam milling machine

The COBRA was tested first of all in combination with a simple cam milling machine. This had one slide on which a turntable was mounted (*fig. 7*), and the workpiece, the cam, could be fixed to the turn-table. Basically this milling machine corresponded to the machine described in reference [2]: the position of the rotary table is measured, and depending on this the position of the slide is adjusted. The setting of the slide is therefore a function of the angular position of the turn-table. In this case only a one-dimensional control is required.

The potentialities of the COBRA were not fully used with this one-dimensional control. As a consequence many of the functions of the electronic peripheral equipment, i.e. the equipment required for the coupling of the servomotor and the measuring system to the control, could be accommodated in the COBRA. (In the future it will probably be normal practice for the COBRA to perform these functions as well.)

The control programme of the COBRA was arranged so that the combination of COBRA and milling machine formed a system identical to that of an existing milling machine with "wired logic" control. Workpiece programmes for this existing control could therefore also be processed in the COBRA. A complication, by no means uncommon in workshop practice, was that the workpiece data for the existing machine were set out in metric units. In the test combination these had to be converted by the COBRA itself to measurements in inches. The tests proved that the COBRA, starting with the same punched tape, could without difficulty make the same pieces of work as the existing combination.

The Bessel interpolation method, which, with its use of the third difference, requires only a short length of tape, had not yet been applied in the existing wired-logic control system. For a second test a control programme was now prepared which differed from the first programme only in the method of interpolation. A new punched tape with workpiece data was made for this new control programme and a workpiece was milled. The resulting workpiece was found to be identical with that produced by the first control programme, while only about a third of the original tape length was needed.

In particular, this second test clearly illustrates the flexibility of the COBRA. With "wired logic", adaptation to the better method of interpolation requires so drastic a modification to the wiring that it would be ruled out in practice on account of the costs involved and because the milling machine would then be out of action for a month. On the other hand, a new control programme was drawn up for the COBRA within a fortnight, and when it was ready it could be read into the memory in two minutes. No changes were made in the system.

The COBRA calculates its own workpiece programme

In most control systems used for cam machining, the workpiece programme is prepared by a computer centre. The designer of a cam makes a table in which he puts data such as setting-up heights, setting-up angles, base radius of the curve to be followed by the centre of the cam roller, the diameter of the cam follower roller, and the cutter diameter to be used. The computer centre has a programme which can be used to calculate a cam profile from these data, by choosing a profile from forms such as linear, parabolic, sinusoidal, skew sinusoidal, or fifth, seventh, or ninth degree curves. The result is a table of the coordinates of the profile to be milled, on a punched tape which has to be supplied to the control system of the machine tool.

It turns out that the skew sinusoidal profile is widely used in practice. It therefore seemed interesting to try to draw up a "control" programme for the COBRA which could be used to compute this cam profile without calling in the computer centre. To simplify matters

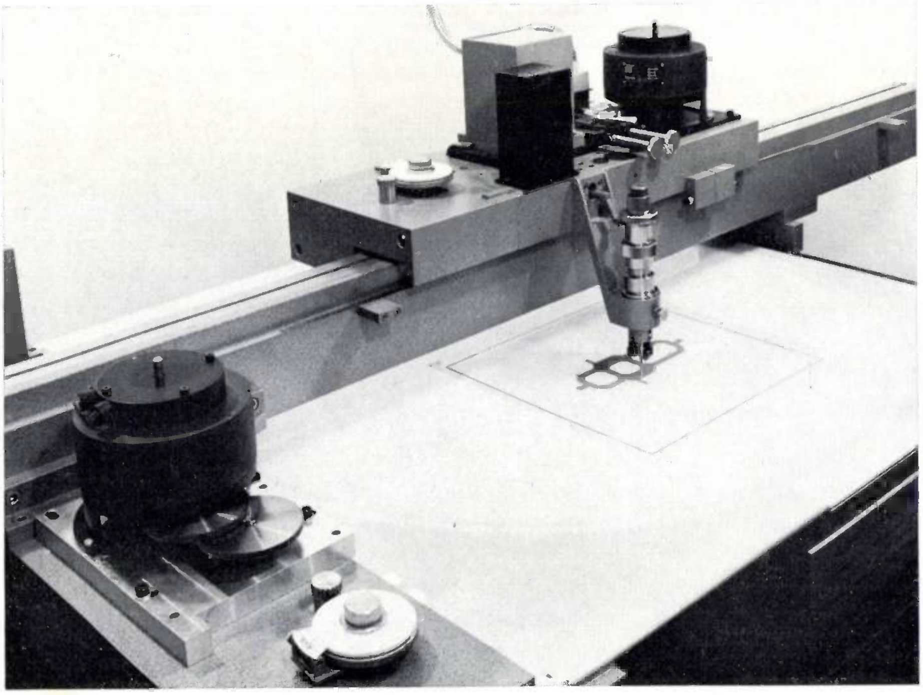


Fig. 8. An automatic drawing table used an as engraving machine. The upper support can be moved in the x-direction on a guide fixed to a second support, which can be moved along the y-direction (cf. fig. 7). The servomotors that can be seen in the photograph are used for the movement of the supports.

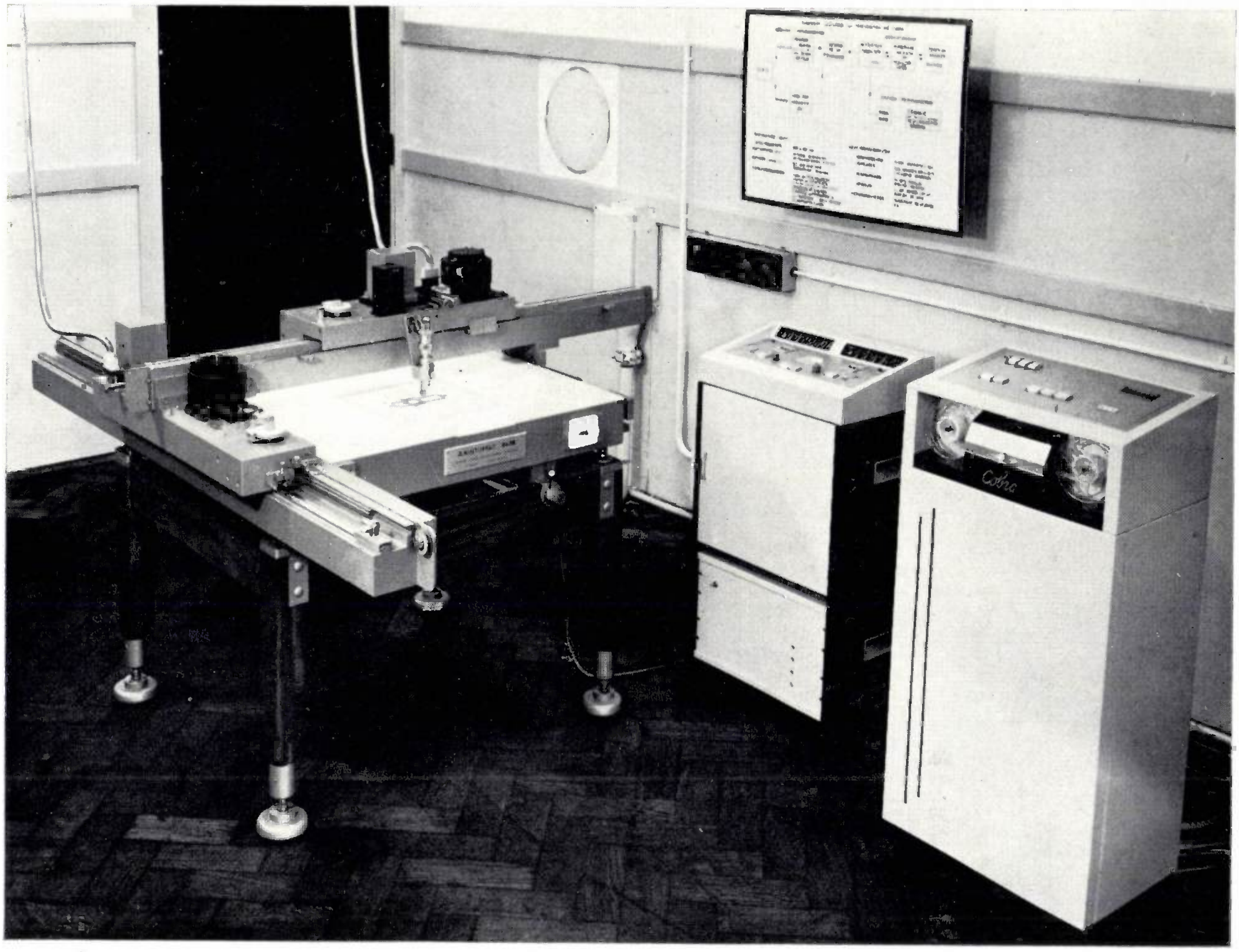


Fig. 9. The COBRA with the drawing table of fig. 8. The cabinet standing between the COBRA and the drawing table contains electronic apparatus for numerical display of the position of the drawing pen.

we assumed that the diameter of the cutter to be used was equal to the diameter of the cam follower roller, avoiding the difficult calculation of the path of the cutter centre. The parabolic method due to Bessel was chosen for interpolation.

This "control" programme could be put into practice without much difficulty. It was found that the COBRA, making use of about 500 store addresses, took one minute to supply a complete punched tape for a cam profile covering 360° with an arbitrary number of rises and falls. The programme could also be extended easily so that a choise could be made from four different profile forms.

This performance illustrates the computing potentialities of the COBRA. It is even more interesting if we consider that a control programme for the COBRA can be prepared which combines the calculation of the punched tape and the control just described for the cam milling machine (the control also requires about 500 addresses in the store). This means in fact that as well as the intermediate stage of the computer centre, the intermediate stage of the punched tape can be also dispensed with. The setting-up heights, setting-up angles, etc. can then be fed directly to the COBRA, for example by means of decimal switches, after which the required cam can be milled.

Control of a drawing and engraving machine

Numerical control of drawing and engraving machines will be used on a large scale in the near future for making photographic masks for integrated circuits, evaporated circuits, spiral groove bearings, etc. In anticipation of this development we have tried the COBRA as a unit for such a machine (*fig. 8*).

The electronic peripheral equipment, which in this experiment was kept completely outside the COBRA, was sufficiently extensive to permit manual operation of the machine if desired. The measuring systems, for example, provided signals not only for the control system but also for a position indicator, so that there was always a numerical indication of the slide positions (*fig. 9*).

The computing possibilities which the control programme possessed for this application were different from those in the examples described above. Linear and "circular" interpolation, for instance, were used instead of quadratic interpolation. To keep manual programming as simple as possible, the programme includes a gradual acceleration and deceleration which is of importance in starting and stopping and for making sharp turns (see Table II). It was also necessary to be able to work with scale factors, either with the x and y scales the same, or with different scales.

The test combination was used for making many "workpieces". *Figures 10, 11 and 12* show three examples. The first illustration (fig. 10) shows the contours (lines of equal height) of a mould for pressing the cone of a television picture tube. The data for these contours are calculated previously and put on punched tape, and the shape of the contours is then checked by drawing with the help of the COBRA and the drawing machine. At the same time this reveals possible errors in the punched tape.

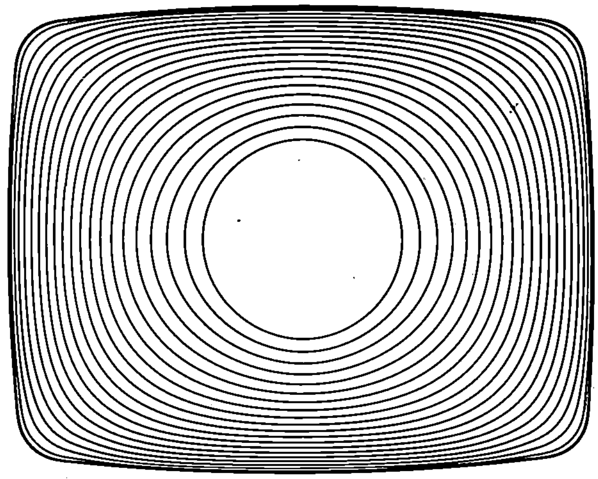


Fig. 10. A workpiece made with the drawing machine of figs. 8 and 9. The figure shows the contours (lines of equal height) of a mould for pressing the cone of a television picture tube.

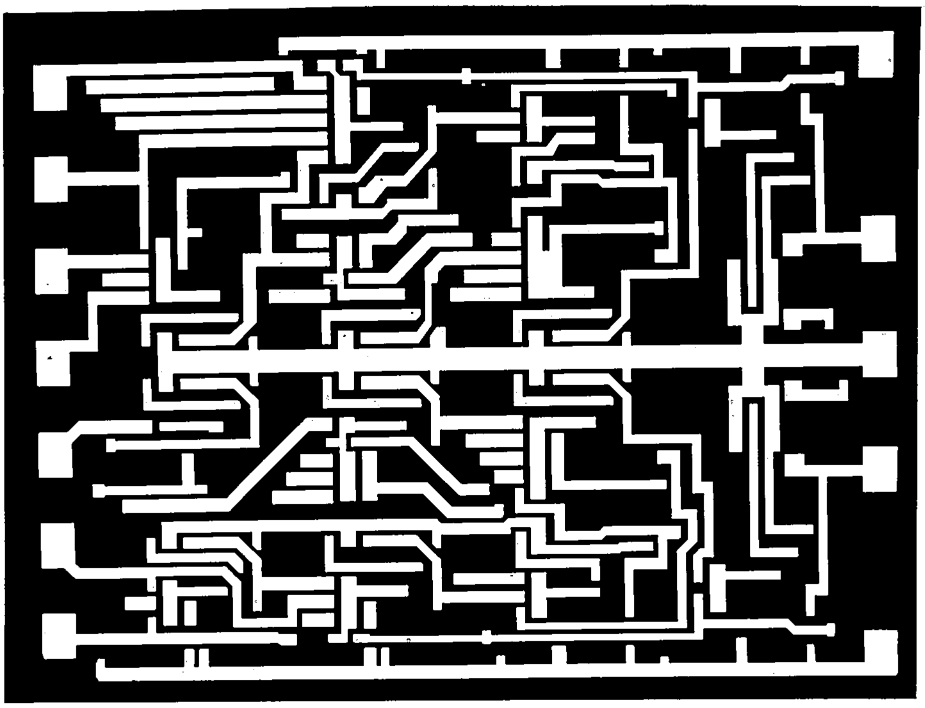


Fig. 11. A photographic mask for an integrated circuit, cut out by the COBRA-controlled drawing machine.

The second illustration (fig. 11) shows a photographic mask for an integrated circuit. Engraving a mask like this takes about a day with manual control while the COBRA-controlled engraving machine can do it in half an hour after five hours of programming. Apart from obtaining a more accurate product, there is also a saving of time. Even more time can be saved if it is necessary to make a change in the photographic mask. Instead of spending another day on engraving by manual control, all that now has to be done is to insert a change in the punched tape, which usually takes very little time, and run the machine for another half hour.

Fig. 12 shows a photographic mask of a spiral groove bearing, in which a particular pattern is repeated a number of times, the workpiece to be engraved being turned through a few degrees before each repeat. For work in which a simple pattern has to be repeated very frequently, the programming time is short in comparison with the processing time.

Future expectations

The COBRA can already show a better performance than any other existing control system for machine tools. Even so, its potentialities are as yet far from exhausted. The development of the COBRA is related to the rapid further development of digital electronic computers. As a result of this development we can expect two things in the near future; a more economic construction of small computers and an increase in the computing speed.

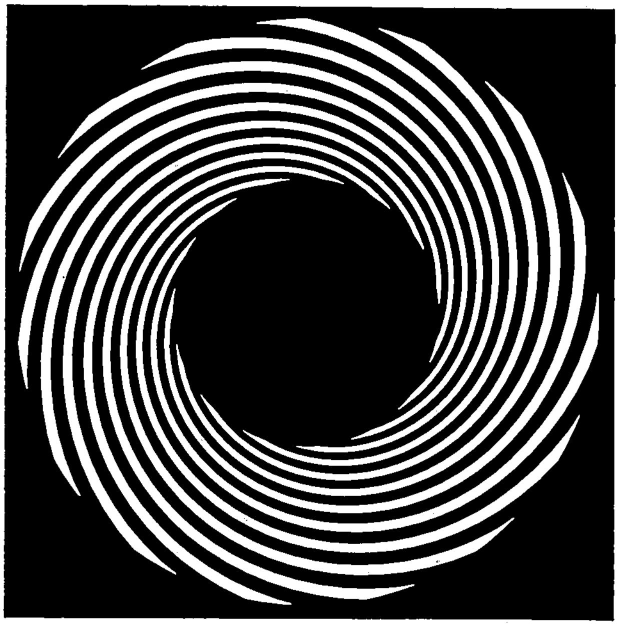


Fig. 12. A photographic mask for a spiral groove bearing, cut out by the COBRA-controlled drawing machine.

The importance of the first factor is self-evident, With respect to the second, the greater speed of calculation, coupled with the use of larger stores, will make it possible for the COBRA to calculate its own workpiece programme for a variety of applications. This has already been achieved for the simple cam milling machine. In one of the examples we have seen that the COBRA can take over many of the functions of the electronic peripheral equipment. Large stores and faster computing speeds will allow a wider use of this facility. A possibility likely to be achieved in the future is that of correcting a workpiece after carrying out an automotic series of measurements on it. It will also be possible for the COBRA itself to determine the optimum machining conditions.

Summary. The use of a small digital computer for the numerical control of machine tools gives rise to a universal system that can be matched to any machine tool by introducing the appropriate control programme into the store. The paper describes COBRA, a prototype of such a system, and also a number of experiments clearly demonstrating the utility and, in particular, the flexibility of the system. Workpiece programming for COBRA is found to run in exactly the same way as the programming for other systems, while hand programming is much simp er. The preparation of a control programme, however, is only r quired with systems like COBRA and this has to be performed once for every application of COBRA. Suitable means for this are in principle available, utilizing a simple programming language and the help of a computer centre. Partly because of the rapid general development of digital computers a considerable extension of the COBRA system can be expected in the future, both in numbers and in applications.

Calculation of traffic-light sequences

651.056

The derivation of an optimum sequence [1] for the traffic-lights at a complicated intersection is no easy task. A traffic-light sequence has to be able to process the incoming traffic properly, it has to prevent different streams of traffic from coming into conflict with one another, and to avoid unnecessary delays it should have as many "green waves" as possible. The sequence can be considered to be optimum if, with these conditions, the total time of the "green" periods of the lights is as great as possible.

We have developed a computer programme (VERKEER, i.e. "TRAFFIC") which can be used to give rapid solutions of this problem for very complicated intersections. The programme deals only with fixed, purely periodic, light sequences. The system considered operates in a fixed cycle, with one green phase and one red phase for each light (the amber phase is taken as part of the green). The times at which a green phase starts and ends are called a green point and a red point respectively. These times are the variables in the mathematical formulation of the problem. We shall clarify the method of formulation with the aid of a rather arbitrarily chosen example following the situation shown in *fig. 1.*

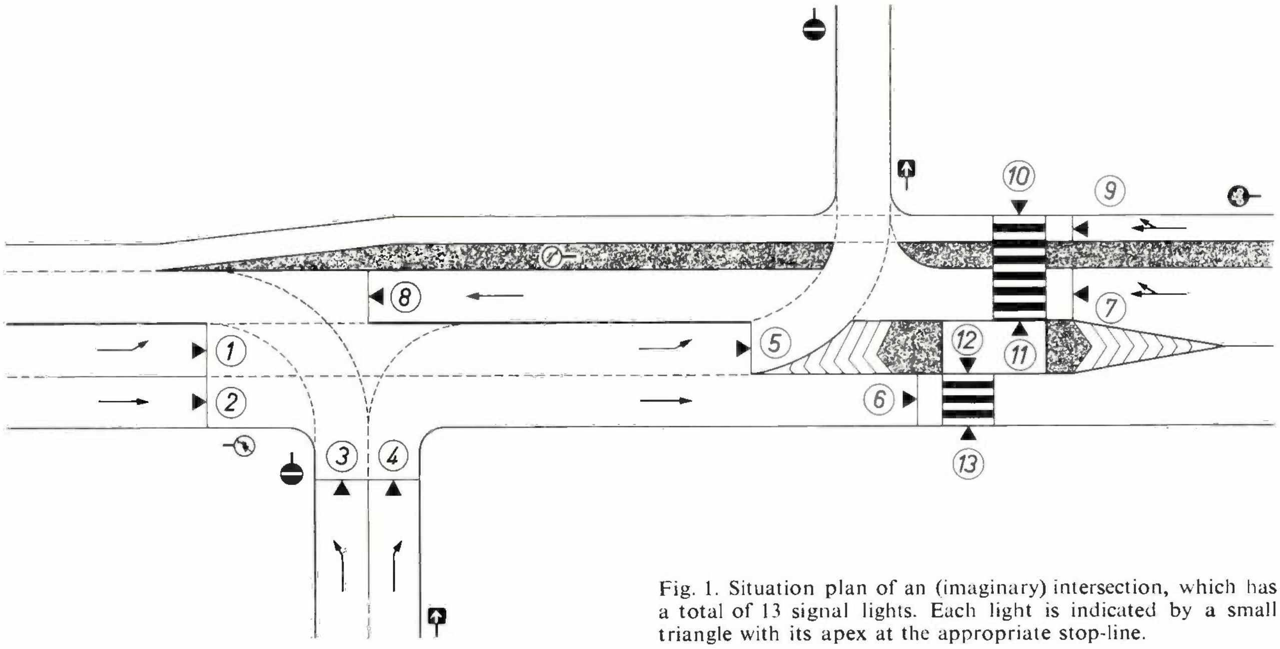


Fig. 1. Situation plan of an (imaginary) intersection, which has a total of 13 signal lights. Each light is indicated by a small triangle with its apex at the appropriate stop-line.

From the diagram, we can discern a number of *cross-linked pairs*, i.e. signal lights that control streams of traffic that cross one another, and a number of *wave-linked pairs*, i.e. signal lights which are "in cascade". The following relationships may then be set up.

1) The wave relationships. For each wave-linked pair (p, q) there are two linear inequalities, which express the fact that a green phase of q covers a green phase of p, displaced by the transit time. (1,5), (4,5), (4,6) are wave-linked pairs in fig. 1. The determination of the transit time depends of course on various considerations of traffic conditions, which we shall not deal with here.

2) The conflict relationships. For each cross-linked pair (j,k) there are two linear inequalities, which express the fact that streams let through by j and k do not come into conflict with one another; in other words a green phase of j lies inside a "displaced" red phase of k, and *vice versa.* (1,3), (2,3), (2,4), (5,7) are cross-linked pairs in fig. 1.

In order to be able to state the wave and conflict relationships unambiguously, it is necessary to make a rough preliminary survey to see how the streams of traffic alternate, connect, and merge with one another, in other words, to make what is called a phase-matching scheme. For example, if (p,q) is a wave-linked pair, it must be decided beforehand whether traffic let through by a given green phase of p arrives in the next green phase of q, or in the second, or the third etc. Similarly, multiples of the cycle time arising in the conflict relationships are established in the phase division.

3) An obvious condition is that each green period is shorter than the cycle time. This again gives a relationship for each light.

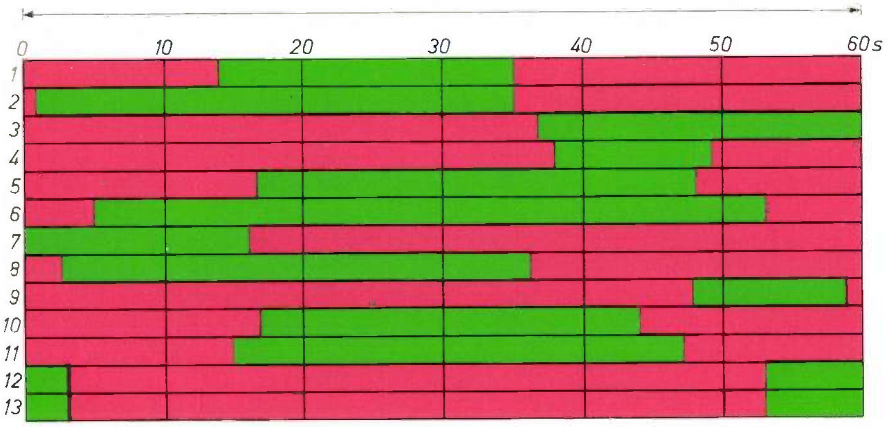


Fig. 2. Bar diagram of an optimum traffic-light sequence calculated under specified conditions for the intersection of fig. 1.

4) The green period for each light is no smaller than a certain minimum value appropriate to that light: these values are external data for the problem in hand. This relationship serves to ensure that the incoming traffic is properly handled. The minimum green period has to be calculated from the traffic density on the basis of known criteria.

In general there are many solutions which satisfy the above system of inequalities. The mathematical problem now reads: find, for the given conditions, the solution for which the sum of the green periods in a single cycle is a maximum. If required, the green periods in this sum can be weighted, so that the more important traffic routes exert a greater effect. The problem has now been brought back to one of linear programming: i.e. the maximization of a linear function subject to a number of linear constraints [2].

The TRAFFIC computer programme is based on this formulation of the problem. The information supplied to the computer consists of the data concerning the intersection, i.e. the structure (cross-linked and wave-linked pairs) and the phase-matching scheme, and also the transit times, the minimum green periods, the weights of the green periods, and the cycle time. There are two parts to the programme: the first uses the data to form the inequalities for the linear programming problem; the second solves this problem. The solution is given in the form of a bar diagram.

The optimum light sequence thus found for the traffic situation of fig. 1 is shown in *fig. 2.* There are 14 cross-linked pairs in the intersection of fig. 1, all of which have been taken into account, and 7 wave-linked pairs, two of which (10,12 and 13,11) have not been taken into consideration. The linear programming problem therefore has 26 variables in this case (viz, a green and a red point for each signal light) and 64 linear inequalities, which are:

13 lights	→	26	inequalities
14 cross-linked pairs	→	28	,,
5 wave-linked pairs	→	10	,,
		64	,,

It can readily be seen from this that even with relatively simple traffic situations the problems we have described may be quite complicated, particularly if an analysis of the consequences of different choices of cycle time and phase-matching schemes is required.

The TRAFFIC programme — which we will not describe in detail here — can be used to carry out this analysis very rapidly for complicated intersections (or systems of intersections). Up to 100 signal lights may be handled in the analysis. The solution shown in fig. 2 for the problem of the relatively simple junction of fig. 1 required about 20 seconds of computing time on the CD 3600 computer [3].

A. J. Dekkers
A. van Duuren
F. A. Lootsma
J. Vlietstra

[1] The term "traffic-light programme" is often used, but in this connection it appears less suitable.

[2] For linear programming, see S. I. Gass, Linear programming, methods and applications, McGraw-Hill, New York 1958. — Linear programming problems have been discussed several times in this journal: H. W. van den Meerendonk and J. H. Schouten, Trim-losses in the manufacture of corrugated cardboard, Philips tech. Rev. **24**, 121-129, 1962/63; W. F. Schalkwijk, Operations research, Philips tech. Rev. **25**, 105-113, 1963/64.

[3] A more extensive treatment of the linear programming problem dealt with here will shortly be published elsewhere.

A. J. Dekkers, Ir. A. van Duuren, Drs. F. A. Lootsma and J. Vlietstra, who are with Philips Research Laboratories, Eindhoven, are attached to the Philips Computing Centre.

An automatic X-ray spectrometer

S. A. Wytzes

543.422.8

Qualitative and quantitative spectrochemical analysis, using the emission spectra in the X-ray wavelength region, is nowadays widely employed. There has been considerable progress in the development of equipment for this technique, and spectrometers are now available that can analyse a number of samples entirely automatically. An automatic spectrometer of this kind is the subject of this article.

Spectrochemical analysis by means of X-rays is nowadays widely used for determining the concentration of one or more elements in a sample. The principles of this method of analysis were dealt with several years ago in this journal [1]. The present article will be concerned with the advances made since that time in the development of equipment. Fully automatic X-ray spectrometers are now being made which are so simple in operation that they can be used by persons with very little previous training. One such instrument is the type PW 1212 spectrometer (*fig. 1*) produced by Philips Industrial Equipment Division. The construction and various interesting features of this spectrometer will be discussed in this article.

. The instrument is employed in many branches of industry. Typical examples are cement works, foundries and steel works, where this rapid method of analysis is used both for the inspection of manufactured products and the investigation of waste products, such as slag.

Before describing the instrument we shall briefly recapitulate the principles underlying the operation of an X-ray spectrometer.

X-ray spectrochemical analysis

The light which a substance emits when it is heated or burned has long been used for identifying the elements of which the substance is composed. In an analogous technique, the *X-rays* emitted by a substance when it is bombarded with fast electrons or irradiated with an X-ray beam can be used instead of light. Nowadays the second method of excitation is in more general use. By analogy with the optical method, the term *fluorescence* is used in this connection, and this method of analysis is frequently referred to as *X-ray fluorescence analysis*.

Unlike the optical spectrum, the characteristic emission spectrum of an element in the X-ray wavelength region is extremely simple. For the lighter elements this consists of only two lines, the Kα line and the Kβ line. For the heavier elements there are the L lines as well, and, in some cases, the M lines.

Because of the simplicity of the X-ray spectrum it is a fairly simple matter to identify an element from it. The wavelength λ of the spectral lines depends on the atomic number Z of the element in accordance with Moseley's law, which states that for every kind of line, e.g. Kα, there is a linear relationship between $\lambda^{-1/2}$ and Z. The intensity of a given line is a measure of the concentration of the element emitting the line, and by measuring this intensity it is possible to determine the concentration of that element in a sample. For the lighter elements the Kα line is generally used; for analysis of the heavier elements it is usual to take an L line.

Apart from the simplicity of the spectra, X-ray spectrochemical analysis has various other advantages over the optical method. One of them is the fact that the substance to be analysed need not be heated or burned; the method is *non-destructive*. This is a great practical advantage as the reference standards, with which, both in the optical and X-ray methods, the specimens to be examined are compared, remain intact (we shall return to this later). Furthermore, X-ray analysis is on the whole more accurate than optical spectral analysis and also covers a wider range of concentrations. Finally, the difficulties encountered in determining the concentration of certain elements by the optical method are absent in the X-ray method. A limitation of X-ray analysis is that the determination of elements lighter than sodium is difficult or impossible.

Dr. S. A. Wytzes is with Philips Industrial Equipment Division, Scientific Instruments Design Group, Eindhoven.

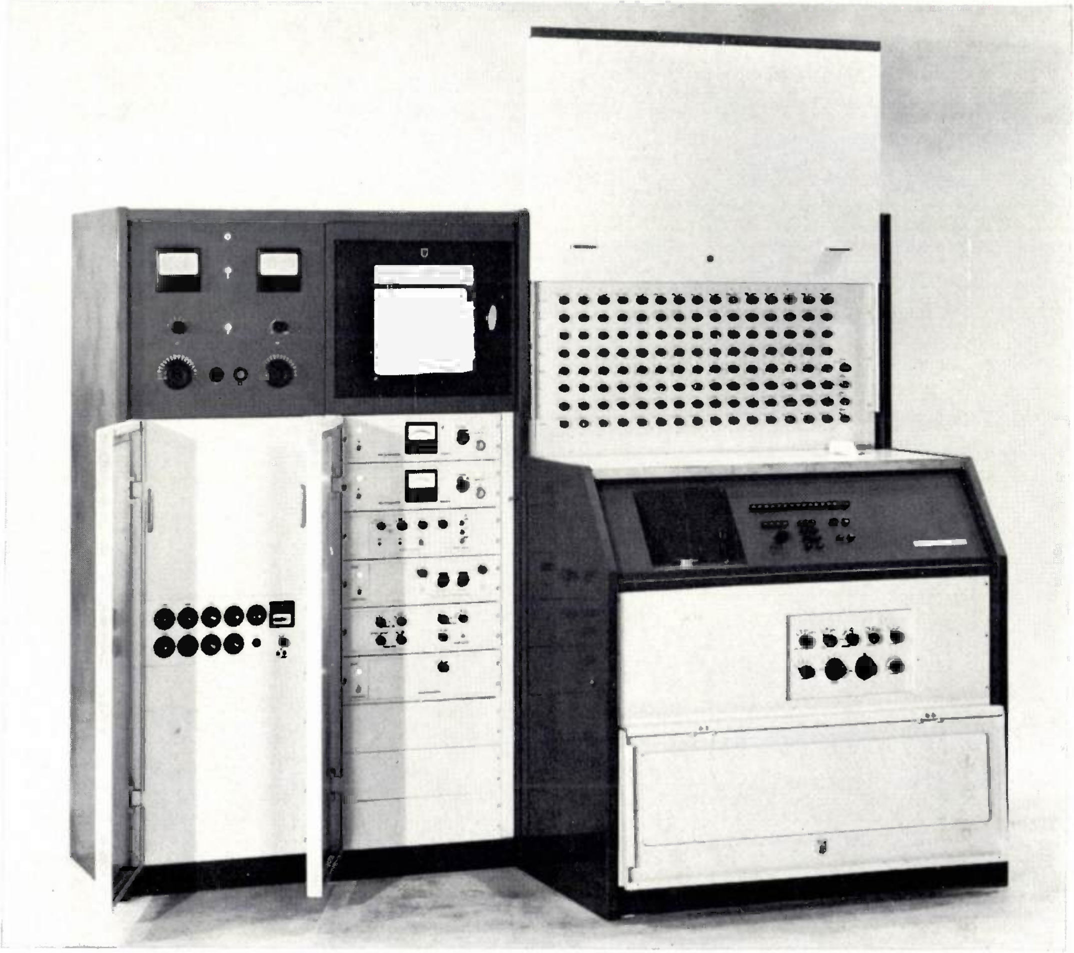


Fig. 1. Front view of the type PW 1212 automatic X-ray spectrometer. The cabinet on the left contains the circuits for stabilizing the X-ray tube voltage and current. The cabinet beside it contains the electronic devices for counting the X-ray quanta. The rear section of the cabinet on the right supplies the high voltage for the X-ray tube; it also contains the electronic circuits that control the automatic measuring programmes. The front section of this cabinet contains the actual analysing section, of which a number of details can be seen in fig. 3.

Principle of the spectrometer

The principle of an X-ray spectrometer is illustrated in *fig. 2*. The atoms of the specimen are excited by the radiation from the X-ray tube. The secondary X-radiation which the specimen then emits is analysed in terms of wavelength. A parallel beam for analysis is selected from the secondary rays, which are emitted in random directions, with the aid of a *collimator*. The beam is made to fall on the *analysing crystal*, a single crystal in which a particular strongly reflecting lattice plane is oriented parallel to the surface. The X-ray beam is, so to speak, reflected by this crystal, but only when the angle of incidence Θ obeys the Bragg equation:

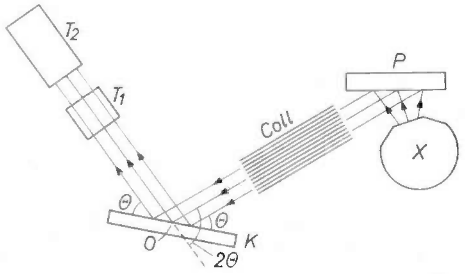


Fig. 2. Arrangement of an X-ray spectrometer. *X* X-ray tube head, *P* specimen for analysis, which, under irradiation, emits characteristic X-rays (fluorescence). *Coll* collimator. *K* analysing crystal. T_1 flow counter. T_2 scintillation counter. The crystal and the two counters are fixed to the arms of a goniometer, which can rotate about an axis *O*.

$$2d \sin\Theta = n\,\lambda. \qquad (1)$$

Here n is an integer (the order of the reflection) and d is the spacing of the lattice planes. At fixed values of n and d the values of Θ at which reflection occurs thus give a measure of the wavelength λ, so that these values indicate the presence of the relevant element in

[1] W. Parrish, X-ray spectrochemical analysis, Philips tech. Rev. **17**, 269-286, 1955/56.

the specimen. The intensity of the spectral radiation is likewise a measure of the concentration of that element. (All this is entirely independent of the chemical compound in which the element appears in the specimen.)

In order to be able to analyse radiation of a particular wavelength, the crystal is rotated until, at a certain angular position Θ, the Bragg equation is satisfied. The reflected radiation is measured with a detector, which is rotated about the same axis (angular setting 2Θ) and which counts the number of incident X-ray quanta as separate pulses. The pulse density or counting rate (the number of quanta counted per unit time) is now a measure of the concentration of the element that emits rays of the relevant wavelength.

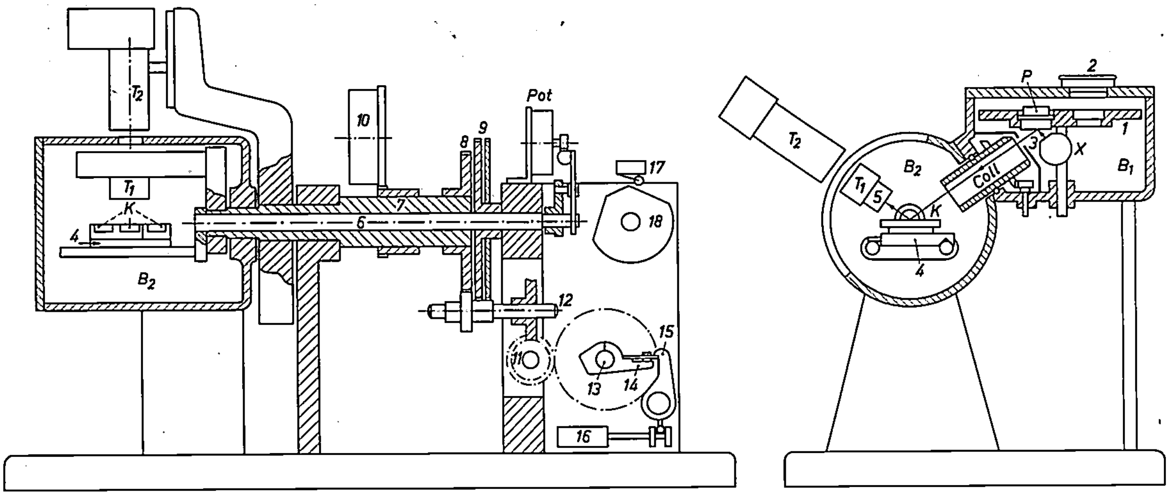


Fig. 3. Construction of the analysing section of the spectrometer. *X* X-ray tube head. *P* specimen under analysis. *Coll* collimator, consisting of a coarse and a fine section. *K* analysing crystal. T_1 flow counter. T_2 scintillation counter. B_1 specimen chamber. B_2 crystal chamber. *1* turntable accommodating four sample holders introduced through the aperture *2*. The diaphragm *3* admits the secondary X-ray beam to the fine or the coarse part of the collimator. *4* slide with three analysing crystals. *5* coarse collimator. The goniometer mechanism consists mainly of the central shaft *6* and the hollow shaft *7*, to which the slide for the analysing crystals and the counters are attached by arms. The shafts are moved by gearwheels *8* and *9*; the latter consists of two parts pressed away from each other by springs to eliminate backlash. The spring *10* attached to the hollow shaft serves the same purpose. The goniometer is driven from the worm shaft *11* by means of the shaft *12*. The mechanism which stops the goniometer at the spectral-line positions comprises a shaft *13* to which 24 braking arms (*14*) are attached; their position on the shaft can be adjusted as required. Opposite each braking arm is a magnet arm *15*, which can be moved by an electromagnet *16*. This magnet is energized when a corresponding switch *17* is closed by the controller drum *18*. *Pot* potentiometer by means of which the height of the pulses is multiplied by sin Θ (see also fig. 10 and fig. 11).

To be able to determine as many elements as possible, a detector is needed which is sensitive to a wide range of wavelengths. It has not proved possible to meet this requirement with a single detector. For this reason our instrument uses a combination of two detectors; a *flow counter*, i.e. a proportional counter through which a specific quantity of gas flows during the determination, and a *scintillation counter*. The first counter is sensitive to soft, long-wave radiation, while the second detects mainly hard rays. In addition to the normal entrance window, the flow counter has an exit window, behind which the scintillation counter is situated. The hard rays, which are not absorbed in the flow counter, thus impinge on the scintillation counter. Preselection makes it possible to use either one of the counters or both of them together. We shall return to this facility later.

The lightest elements, whose radiation is so soft that it is appreciably absorbed by the atmosphere, can only be analysed in a vacuum. Therefore the head of the X-ray tube, the specimen, the collimator, the analysing crystal and the flow counter are contained in an evacuated space. The scintillation counter, which detects the short-wave radiation, is situated outside this space.

The basic requirement to be met by the specimen is that the irradiated side should be flat. It may be an alloy, a compacted powder, a compressed tablet, a cooled melt or a solution.

Description of the equipment

The actual analysing section of the equipment is represented in *fig. 3*. The space that can be evacuated consists of two parts, the specimen chamber B_1 and the crystal chamber B_2. The former space contains the head of the X-ray tube and a turntable which accommodates four sample holders containing the specimens. The sample holders are placed on the turntable through an aperture which can be made vacuum-tight. *Fig. 4* shows a sample holder being loaded. The turntable is rotated by a motor (not drawn) so that the four specimens enter the X-ray beam one after the other. During the analysis the specimen rotates around its own axis, so that the measured average intensity of the fluorescence radiation is not affected by surface irregularities, such as grooves left after grinding the surface.

The rotating specimen table makes it possible to analyse four specimens before the loading aperture is opened and the space again evacuated. The successive introduction of the samples into the X-ray beam and their analysis take place automatically in accordance with a preselected programme.

The collimator holder is located between the vacuum spaces B_1 and B_2. The holder contains two collimators, each consisting of a series of thin parallel plates; the distance between the plates in one of the collimators is 0.15 mm, in the other it is 0.45 mm. By rotating the holder, one or the other collimator is introduced into the X-ray beam, which enters through a diaphragm *3*. The reasons for employing two collimators will be dealt with below.

The crystal chamber (*fig. 5*) contains a slide with three analysing crystals, which can be introduced selectively in the X-ray beam emerging from the collimator. The slide is also operated by a motor. The position of each crystal can be preset with respect to the slide, so that the reflecting lattice planes of the various crystals can be oriented exactly parallel with one another. The reasons for employing three exchangeable analysing crystals will also be discussed later.

The crystal chamber also contains the flow counter. A coarse collimator is placed in front of this counter. The wall of the crystal chamber is fitted with a mylar window, behind which the scintillation counter is situated.

The *goniometer* mechanism, which moves the analysing crystal and the detector in such a way that the angular displacements are in a ratio of 1 : 2, is shown in fig. 3 and explained in the caption.

One revolution of the worm shaft *11* causes a rotation of the detector by 1° (thus rotating the crystal by $\frac{1}{2}$°). This driving system moves the goniometer at a constant speed, which can be set to 4°, 1° or $\frac{1}{2}$°(2Θ) per minute with the aid of a small gearbox. In this way

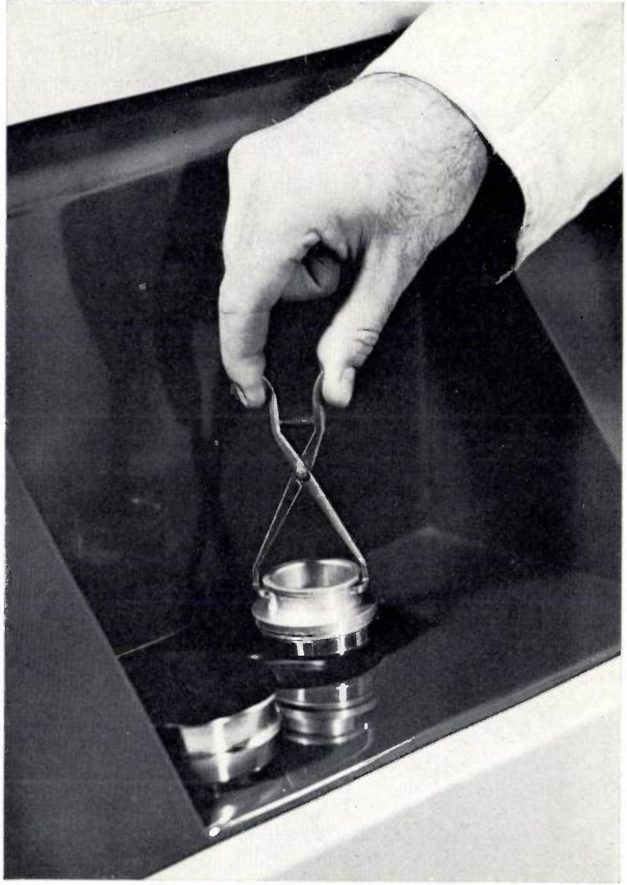


Fig. 4. Introducing a sample holder into the specimen chamber.

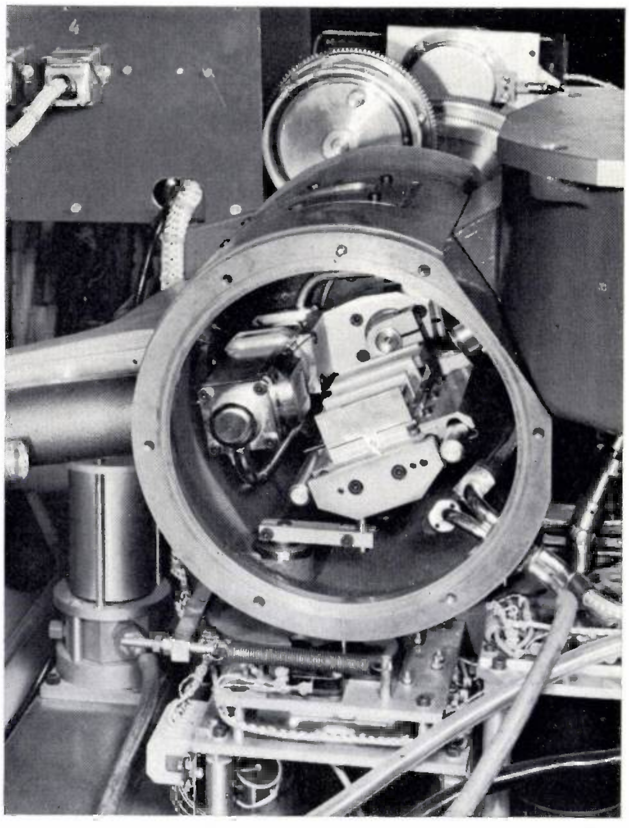


Fig. 5. The crystal chamber with side wall opened, showing the slide with the analysing crystals and the flow counter.

the whole spectrum of a specimen can be scanned fairly quickly in order to obtain some first ideas about the elements present in the specimen. For accurately determining the concentration of certain elements the procedure is different: the goniometer travels rapidly to a preset angular position corresponding to a particular spectral line. Here it stops and the analysis takes place. The goniometer then travels quickly to an angular position corresponding to another spectral line, stops here again, and so on.

For fast goniometer travel another motor is used; the speed is then 214°(2Θ) per minute. The mechanism that stops the goniometer at different angular positions is also shown in outline in fig. 3. The shaft *11* drives the shaft *13* in addition to the goniometer. On this shaft there are 24 braking arms, which can be clamped to the shaft in any desired angular position. (One of these arms, *14*, is shown in the drawing.) Opposite each braking arm is a magnet arm which can be moved by an electromagnet. This mechanism is shown in *fig. 6*. About 1°(2Θ) before the goniometer has to stop, a switch *17* is closed by the controller drum *18*. This reduces the speed to 52° per minute, and at the same time one of the magnet arms *15* is moved into a position in which the claw at the end of it stops the corresponding braking arm after a further rotation of 1°. The goniometer is then blocked. A friction coupling enables the motor to continue running. Once the relevant spectral line has been measured, the circuit of the magnet engaged is broken and the goniometer can move to the next measuring position, and so on. By means of this mechanism the goniometer is thus able to stop at 24 accurately predetermined positions, 15 of which can be included in an automatic programme. In this way each of the four specimens on the rotating table can be automatically analysed for the presence of 15 different elements.

The positions at which the goniometer must stop do not have to be selected in ascending order of angles. A special mechanism enables the goniometer to run back automatically. With this type of operation there could be the danger that the position in which the goniometer stops to measure a particular spectral line may not be the same in both directions of travel. The fact that backlash in the mechanism is eliminated by springs does not imply that there is no dead movement at all: this may also be caused by elastic deformation and by friction. The circuits are therefore so arranged that the goniometer always moves forward a little distance before stopping after a reverse movement.

For this purpose each of the 24 discs on the controller drum has a second disc attached to it, which closes another switch 2° past the point where the first switch is closed (that is 1° past the relevant measuring position). When *both* switches corresponding to a particular measuring position are closed, the goniometer travels in reverse. If the goniometer now has to stop at, say, 10°, coming from an angle greater than 11°, then when the position 11° has been passed in the return travel, one of the switches is opened and the mechanism is again switched to forward travel. This does not happen immediately, however, the presence of a delay circuit (a capacitor in parallel with the relevant relay) allows the goniometer to continue in reverse for 2° or 3°. At about 8° to 9° the motion then reverses, and the goniometer stops in the normal way at 10°.

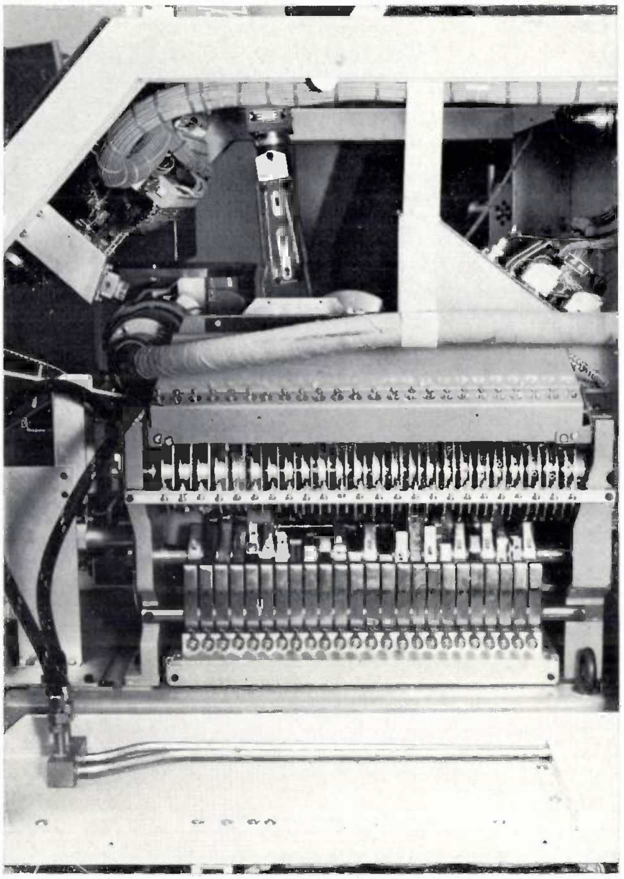


Fig. 6. The mechanism that stops the goniometer at 24 fixed positions. On the underside it contains 24 braking arms with associated magnet arms. Above this there is a controller drum (see fig. 3).

Accuracy of the goniometer

The goniometer is aligned manually by turning the worm shaft *11* until the peak intensity of the required spectral line is found. The braking arm *14* is then fixed in the position to which the goniometer has then turned. In this connection it is not of primary interest that the angle 2Θ read on the goniometer in this position should correspond exactly to the angle that can be found in a table for each crystal. The important thing is that in all subsequent determinations of the element concerned, the goniometer should return exactly to the angle thus set. An idea of the accuracy required can be obtained by measuring the profile of a normal spectral line. It is then found that a deviation of 0.005° from the crystal position for maximum reflection gives a reduction of 0.6% in the measured intensity [2]. In order, then, to prevent goniometer deviations from causing a greater error than 0.6% in automatic intensity measurements, each setting of the crystal must be reproducible to an accuracy of 0.005°. In ratio measurements (to be dealt with later) the goniometer setting is much less critical, and greater accuracy can be achieved. By using springs and stiff components, and by keeping friction as low as possible, the dead movement could be reduced to less than 0.003°. Since the goniometer always stops from the same direction, the actual inaccuracy in angular re-setting is even less. It should be added here that an error may also be caused by the re-setting inaccuracy in the position of the collimator and of the slide carrying the analysing crystals. Both these errors have been kept to less than 0.001°.

Table I. Reflection angle 2Θ (in degrees) for the Kα radiation of a series of elements, using seven different analysing crystals (with different lattice plane spacing *d*).

crystal (*d* in nm) / Element ($\lambda_{K\alpha}$ in nm)		gypsum	ADP	PE	EDDT	quartz	LiF	topaz
		0.7593	0.5324	0.4371	0.4404	0.3343	0.2014	0.1356
Na	1.191	103.3	—	—	—	—	—	—
Mg	0.990	81.3	136.6	—	—	—	—	—
Al	0.834	66.6	103.7	145.1	142.5	—	—	—
Si	0.713	56.0	84.0	109.2	108.0	—	—	—
P	0.616	47.8	70.6	89.5	88.7	134.0	—	—
S	0.537	41.4	60.6	75.9	75.2	106.9	—	—
Cl	0.473	36.3	52.7	65.5	64.9	90.0	—	—
K	0.374	28.5	41.2	50.7	50.1	68.3	136.7	—
Ca	0.336	25.6	36.8	45.2	44.8	60.3	113.1	—
Ti	0.275	20.0	29.9	36.7	36.4	48.6	86.1	—
Fe	0.194	14.6	21.0	25.6	25.4	33.7	57.5	91.2
Cu	0.154	11.6	16.6	20.3	20.1	26.7	45.0	69.3

Choice of crystal

For short-wave radiation, i.e. with relatively heavy elements, the analysing crystal is required to have a small lattice plane spacing *d*. Otherwise the angle Θ, and therefore $\mathrm{d}\Theta/\mathrm{d}\lambda$ too, would be very small, so that the spectral lines would be too close together for good resolution. For long-wave radiation (light elements), however, a large *d* is required because the reflection angle 2Θ could otherwise be greater than the angle over which the detector can be rotated. If the lattice plane spacing is less than a half wavelength, then according to Bragg's equation there will be no reflection at all.

The choice of the analysing crystal must therefore be adapted to the wavelength of the fluorescent radiation of the element under investigation.

The automatic X-ray spectrometer described here provides a choice of seven analysing crystals. They consist of gypsum, ammonium dihydrophosphate (ADP), penta-erythritol (PE), ethylene diamine-ditartrate (EDDT), quartz, lithium fluoride (LiF) and topaz. The analysing crystals are set out here in descending order of *d* values. These values are listed in *Table I*; the table also gives the wavelength of the Kα line for a number of elements, and the reflection angle 2Θ corresponding to the first order reflection for this line as obtained with the different analysing crystals.

It might be asked why such a relatively large number of crystals is provided, and whether the two extremes, gypsum and topaz, might not be sufficient. These would certainly be sufficient as far as the reflection angles are concerned, but it would mean that the spectral lines obtained would not have maximum intensity in all conditions, as the reflecting efficiency of the various crystals is widely different. The reflection efficiency of PE, for example, is much greater than that of EDDT. In choosing between these two crystals, which are in many respects equivalent in view of the virtually identical values of *d*, the first will be used if maximum sensitivity is desired, that is if the element to be detected is present in only a very small quantity in the sample. In some cases, however, the second crystal may be preferable for a different reason, namely thermal expansion. A change in temperature, by causing expansion in the direction perpendicular to the reflecting lattice planes, gives rise to a change in *d* and this results in a shift of the spectral lines. The linear coefficient of expansion in the direction perpendicular to the set of lattice planes used differs considerably, however, from one crystal to another. It is greatest in PE and least in topaz. Because of this PE cannot be used in all cases.

The shift of a spectral line resulting from a change in *d* also depends on the magnitude of the angle of incidence Θ. It follows from Bragg's equation that:

$$\Delta(2\Theta) = -2\frac{\Delta d}{d}\tan\Theta. \quad . \ . \ . \quad (2)$$

At greater values of Θ, therefore $\Delta(2\Theta)$ shows a greater dependence on Δd. Curves drawn with the aid of equation (2) and using the measured linear coefficient of expansion for the various crystals, are shown

[2] The line profile depends to some extent on the collimator and on the analysing crystal. The figures given here hold when the fine collimator and a lithium fluoride crystal are used.

in *fig: 7*. It can be seen that, for example, when aluminium is analysed with PE ($2\Theta = 145.1°$), a temperature change of 1° causes a deviation of as much as 0.045° in the reflection angle 2Θ. The higher sensitivity obtained with PE is thus accompanied by a reduction in accuracy.

Choice of collimator

An analysing crystal is not perfectly uniform; it behaves like a mosaic of small blocks, each at a slightly different orientation. Because of this the crystal not only reflects radiation incident at the angle Θ, but also radiation at a slightly different angle of incidence, $\Theta + \alpha$. The reflection efficiency I of a good crystal decreases rapidly with increasing value of $|\alpha|$. *Fig. 8a* gives a qualitative representation of I as a function of α.

If we assume the collimator to be placed in such a way that the parallel plates make an angle Θ with the surface of the crystal, then radiation with this angle of incidence is fully transmitted, apart from a shadow factor due to the thickness of the plates. For radiation at an angle of incidence $\Theta \pm \alpha$, the degree of transmission D decreases linearly with α. It becomes zero when $\alpha = \pm a/l$, where a is the spacing between the collimator plates and l their length. Fig. 8*a* also shows D as a function of α (isosceles triangle). The contribution to the reflected radiation made by the radiation with a deviation between α and $\alpha + \mathrm{d}\alpha$ from the Θ direction is proportional to the product of I and D. The total reflected radiation (the counting rate at the maximum of the spectral line) is therefore proportional to $\int ID \, \mathrm{d}\alpha$.

If the crystal is rotated through an angle β, the triangle then shifts with respect to the curve I (D'), as in this case the radiation in the direction $\Theta + \beta$ is completely transmitted. The total reflected radiation will decrease more quickly as the base of the triangle becomes smaller. In other words, the resolution increases as the triangle is made smaller, but the counting rate at the maximum of the spectral line decreases. The choice of collimator, then, is determined by the compromise that has to be made between the time taken by the measurements and the resolution required. For the heavy elements, the counting rate is usually high, and since the spectral lines are close together a high resolution is needed. The fine collimator is used for these elements, with $a = 0.15$ mm and $l = 95$ mm. For light elements the reflection angles of the spectral lines are farther apart, but the intensity of the radiation is lower. This means that with these elements a high counting rate is of greater importance than a high resolution, and the coarse collimator is therefore used, with $a = 0.45$ mm. The counting rate in this case is two or three times greater than when the fine collimator is used.

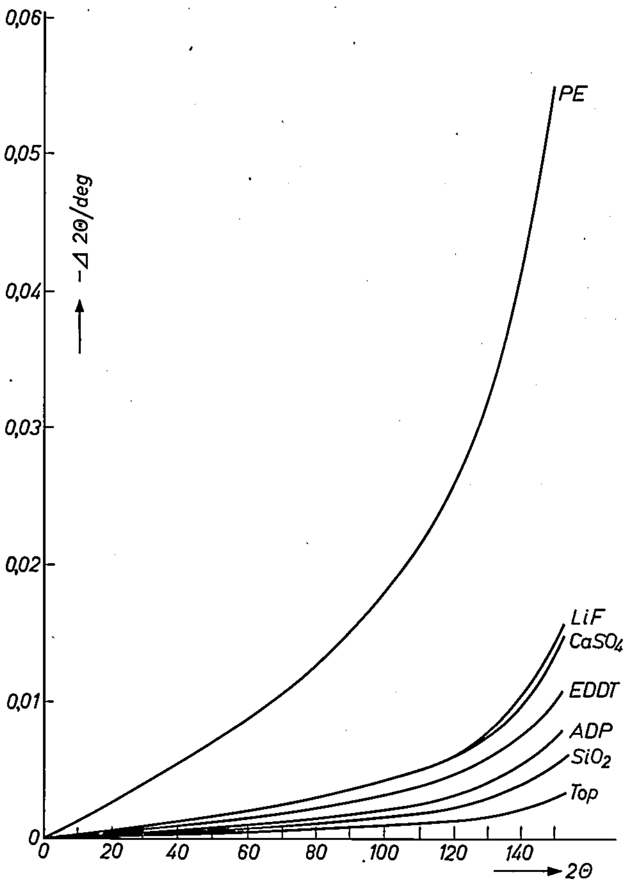


Fig. 7. Spectral line shift $\Delta 2\Theta$ for a 1° change in the temperature of the analysing crystal, as a function of the total reflection angle 2Θ, for the seven crystals available in the spectrometer.

The plate spacing in the collimator in front of the flow counter (*5* in fig. 3) is sufficiently large to have hardly any effect on the line profile measured. This collimator serves two purposes. On the one hand it helps to suppress unwanted radiation (background)

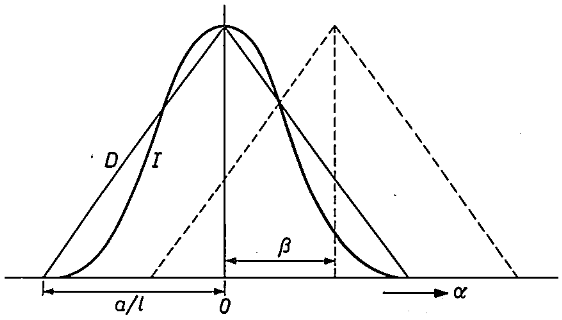


Fig. 8. Reflection coefficient I of a crystal as a function of the angle α by which the angle of incidence deviates from that at which optimum reflection occurs. The triangle D represents the transmission of a collimator as a function of the angle between the direction of the collimator plates and that of the radiation. a is the spacing and l the length of the plates. The product of I and D determines the line profile measured. If the crystal is rotated through an angle β, the triangle D is displaced with respect to the curve I (dotted curve).

and on the other hand it ensures that, at small values of Θ, part of the radiation coming from the first collimator does not go straight to the counters.

To obtain a very high resolution in special cases, a second fine collimator is placed in front of the scintillation counter. This reduces the counting rate, however, by a factor of 3 or 4. The use of this collimator also increases the effect dealt with earlier of a change in the lattice spacing *d* with temperature.

Choice of counter

Most of the detectors used for measuring X-radiation have already been described in this journal some considerable time ago [3]. We shall therefore confine ourselves here to a few salient points.

The *flow counter* is a proportional counter filled with argon gas (+ 10% methane), fitted with a side window made of 6 μm thick mylar for the transmission of soft X-rays [4]. Since this material permits the diffusion of oxygen and water vapour, the counter would soon become unserviceable if arrangements were not made to keep the gas filling constantly flowing through the tube to remove the contaminating gases. The flow need not be large, a few litres per hour being sufficient.

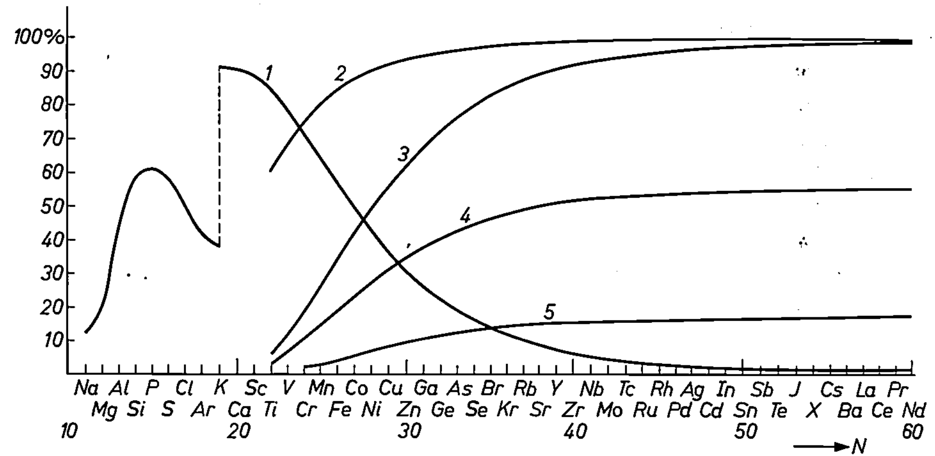


Fig. 9. Calculated counter efficiency for Kα radiation, as a function of atomic number *N*. Curve *1* relates to the flow counter, *2* to the scintillation counter. Curve *3* is arrived at by taking into account the attenuation caused by the flow counter placed in front of the scintillation counter and by the window in the crystal chamber. Curve *4* is found by additionally taking into account the difference in the distance between the counters and the crystal and the difference in the size of window. (Curve *5* relates to the scintillation counter when an extra collimator is placed in front of it to obtain a very high resolution.)

The quanta not absorbed in the flow counter are able to leave the tube through a beryllium exit window, pass through the window in the wall of the crystal chamber (60 μm thick mylar) and then impinge on the scintillation counter. This contains a sodium-iodide crystal, activated with 1% thallium, which is placed against the window of a photomultiplier tube. Every X-ray quantum absorbed in the NaI crystal produces a scintillation in it, which gives rise to a voltage pulse at the anode of the photomultiplier tube.

Which counter can best be used in any given case depends upon considerations of efficiency (the counted fraction of the quanta incident on the counter window).

The efficiency of a counter is given by the product of two fractions: the percentage of the incident quanta transmitted by the window, and the percentage of the transmitted quanta which is absorbed in the counter medium. *Fig. 9* gives the efficiencies of the counters for the Kα radiation of the various chemical elements [5]. Curve *1* relates to the flow counter; except for the very lightest elements, the radiation here always penetrates well into the counter medium, but the harder the radiation the more it passes right through the medium without being counted. With the scintillation counter (curve *2*) the situation is different: everything that gets through to the medium is absorbed and fully counted, but the entrance window is not so transparent to soft radiation. Furthermore, in our arrangement the radiation cannot reach this window until it has first passed through the flow counter and the window in the vacuum chamber. The efficiency is therefore reduced all along the line, to curve *3*. If we also take into account the fact that the window of the scintilla-

[3] P. H. Dowling, C. F. Hendee, T. R. Köhler and W. Parrish, Counters for X-ray analysis, Philips tech. Rev. **18**, 262-275, 1956/57. See also reference [1].

[4] The counter may also be fitted with a window only 1 μm thick, which considerably increases the counting rate for the lightest elements. The life of these windows, however, is limited.

[5] The *K* radiation has an absorption edge; particulars of this will be found in the articles mentioned under reference [3].

tion counter is smaller than that of the flow counter and is farther away from the crystal, then we must multiply all values on curve *3* by a factor of 0.55. This brings us to curve *4*.

A comparison of curves *1* and *4* shows that the flow counter has greater sensitivity for elements lighter than copper, whereas the sensitivity of the scintillation counter is greater for the heavier elements. For a few medium heavy elements the sensitivity can be considerably increased by using both counters together.

In fig. 9 curve *5* gives the efficiency of the scintillation counter when the second fine collimator referred to above is placed in front of it to improve the resolution. This curve was derived from curve *4* by taking another geometric factor into account.

The pulse-height discriminator

In the flow counter as well as the scintillation counter the average pulse height depends on the energy (wavelength) of the quanta producing the pulses, and when a particular spectral line is being measured the pulses counted are therefore all roughly of the same height. For this reason the effect of unwanted pulses (the background) can be substantially reduced by using a pulse-height discriminator, which passes only those pulses whose height lies between two close limits. A complication in this respect is that the limits would have to be separately selected for every wavelength, that is to say for every element. The automatic spectrometer contains a device, however, that makes this manipulation unnecessary.

The energy of a quantum is inversely proportional to the wavelength. Since the mean pulse height V_{m} is proportional to this energy, it may be expressed, with the aid of the Bragg equation, as:

$$V_{\mathrm{m}} = C \frac{n}{d} \frac{1}{\sin \Theta}, \quad . \; . \; . \; . \; . \quad (3)$$

where C is a constant that depends on the voltage on the counter. If the pulses are now multiplied by a factor $(d/n) \sin \Theta$ for every position of the goniometer and for every crystal, then $V_{\mathrm{m}}=C$. This means that we can choose two fixed limits for the discriminator within which the pulses must be counted.

The multiplication by $\sin \Theta$ is carried out by applying the pulses via a linear potentiometer which is coupled with the goniometer in such a way that its spindle rotation is proportional to $\sin \Theta$. In fig. 3 this potentiometer is indicated by *Pot*. *Fig. 10* shows a sketch of the mechanism employed.

The circuit containing this potentiometer is shown as a block diagram in *fig. 11*. The pulses from the counter or counters T go through an amplifier A, which is followed by a transistor in common collector configuration (emitter follower). In series with the potentiometer there is a resistor R. When the spectrometer changes to another crystal (different lattice spacing d), or to a reflection of another order (different n value) a different value of R is automatically selected, such that $R_{\mathrm{p}}'/(R_{\mathrm{p}} + R + R_{\mathrm{i}})$ is proportional to d/n. (Here R_{i} is the internal resistance of the emitter follower, R_{p} the resistance of the potentiometer and R_{p}' the part of the potentiometer resistance that would be tapped off by the sliding arm at $\sin \Theta = 1$.) The instrument is fitted with resistors R for six different crystals [6] and for first and second order reflections. (The resistance values are so chosen that R is zero for the first-order reflection from the crystal with the greatest value of d, i.e. the gypsum crystal.)

A difficulty arises from the fact that the pulse amplitude from the flow counter depends on the density of the gas filling; a 1% increase in density lowers the height of the pulses by 6%. Since the limits between which the pulses are counted are fixed, the chance of a counting error arises. To avoid this a regulator is employed which makes the density of the gas flowing through the counter independent of temperature and of atmospheric pressure. The construction of this regulator is shown in the diagram of *fig. 12*.

Absolute measurements and ratio measurements

The quantity of a certain element contained in a specimen can be determined in two ways. In an *absolute measurement* the number of pulses for a given spectral line is counted for a specific time. This time is

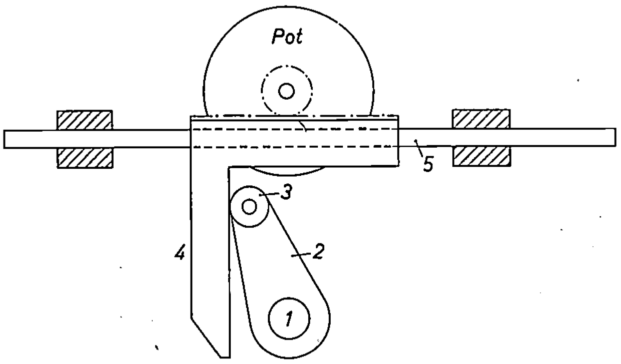


Fig. 10. Mechanism which makes the angular displacement of the potentiometer *Pot* in fig. 3 proportional to sin Θ. *1* shaft of crystal holder. *2* arm. *3* ball bearing. *4* bracket with gear rack, fixed to the sliding shaft *5*.

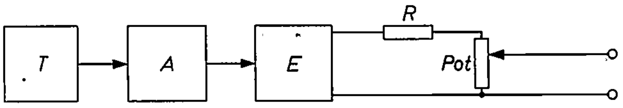


Fig. 11. Block diagram of the circuit for maintaining constant pulse-height. *T* counter. *A* amplifier. *E* emitter follower. *R* interchangeable resistor. *Pot* potentiometer (see fig. 3 and fig. 11).

made sufficiently long to minimize the statistical error. The time needed varies from about 10 seconds to 2 minutes. If the background is relatively high, a count is made when the goniometer is not on a spectral line. The difference between the two counts is then a measure of the concentration of the relevant element. For every element a calibration curve can be made in which the concentration is plotted against the number of pulses counted. The curve is usually almost a straight line.

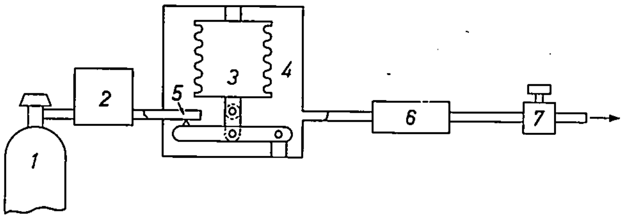


Fig. 12. Construction of the regulator for maintaining constant density of the gas in the flow counter. *1* gas cylinder. *2* control valve set at 1.5 atm. *3* closed metal bellows, in which the pressure is 1.1 atm. at a temperature of 15 °C. In the equilibrium state the pressure and temperature are the same in space *4* as in the bellows. When the latter expands, as a result of a change of temperature or a pressure drop in *4*, valve *5* opens. *6* flow counter. *7* needle valve. Because the expansion of *3* is very slight, the density of the gas inside it can be considered to be constant, and hence the density of the gas in space *4* is constant.

A difficulty with absolute measurements is that, to obtain reproducible results, very strict requirements have to be set for the accuracy of the goniometer and for the constancy of the radiation from the X-ray tube. These requirements are much less strict for *ratio measurements*. In this method a particular element is first measured in a standard specimen. The number of pulses to be counted is decided upon, and the time required is recorded in an electronic store. This element is then measured for the same period of time in the unknown specimen. The ratio of the numbers of counted pulses is equal to the ratio of the concentration of the relevant element in the standard to that in the unknown sample. Since the turntable has four places for the specimens, one of which is occupied by the reference standard, it is possible in this way, after a single measurement on the standard, to determine the content of the element concerned in the three unknown samples. The goniometer does not have to be reset between these measurements, so that there is no reset error to be taken into account. If, for example, the crystal is not exactly in the right position, this has the same relative effect on all the measurements, and therefore the ratio between the numbers of counted pulses is not affected. The same applies to any change that may take place in the course of time in the reflection efficiency of the analysing crystal or in the intensity of the X-ray beam. This means that greater accuracy is to be achieved with a ratio measurement than with an absolute measurement. The error that may arise from inaccuracies in the instrument can be put at 0.5 to 1% in absolute measurements and at 0.1 to 0.2% in ratio measurements.

Ratio measurements do however have the disadvantage that they take longer, because of the time needed for measuring the standard. Moreover, the statistical effect adds to the time required: if we assume that the statistical errors in the counts for standard and unknown specimen are identical, then in order to ensure that the resultant error in the calculated concentration is no greater than in a corresponding absolute measurement, it is necessary to take twice the number of counts for both specimens. Thus, if one unknown sample is under investigation, the total time needed for a ratio measurement is four times longer than for an absolute measurement. As, when two or three unknown samples are to be analysed, the reference standard is analysed only once, the loss of time per specimen is not so great. It is easily seen that the total time needed for measuring two samples is increased by a factor of 3, and for three samples by a factor of 8/3.

Changes in the intensity of the X-ray beam which occur so quickly that they fall within the measuring period cannot, of course, be compensated in ratio measurements either. The high voltage on the X-ray tube and the tube current must therefore be well stabilized for both methods of measurement. The circuits used for this purpose need not be discussed here.

Programming

As we have seen, the instrument can analyse four specimens entirely automatically; in the ratio method one of these is the reference standard. The analysis can relate at the most to 15 elements per specimen. In normal applications of the instrument the choice of the 15 elements (for which 15 of the braking arms shown in fig. 3 have to be set at the appropriate angles Θ) remains unchanged for a long period. In the analysis of a copper alloy one may be interested, for example, in the content of aluminium, silicon, phosphor, manganese, iron, nickel, zinc, arsenic, tin and lead.

The programming of the automatic analysis is arranged as follows. For each of the 15 elements the control panel, seen in the upper right of fig. 1, has a row of 8 knobs with which the following can be selected: the collimator, the crystal and the reflection order, the counter, the time (for absolute measurements), the number of pulses (for ratio measurements), the current and the high voltage on the X-ray tube. The last con-

[6] The same resistors can be used for both PE and EDDT.

trol knob can be used for indicating whether the element in question is to be analysed or passed over. Once this programme has been set for a particular sample, all that has to be done is to introduce the sample and push the starting button. The evacuation and the analysis programme are then carried out automatically. The results of the analysis are printed out on a strip of paper.

If the instrument is regularly called upon to analyse samples of different types for which it has not normally been preset, but which involve the analysis of no other elements apart from the 24 to which the braking arms have been set, a programme for each type can be set up on an interchangeable board with plug-leads and sockets. This dispenses with the necessity of setting a fairly large number of controls.

Summary. After a brief review of the principles underlying the operation of an X-ray spectrometer, an instrument is described which automatically analyses a number of samples for their content of 15 different elements. The number of samples per measuring cycle is four in absolute measurements, or three in ratio measurements. Seven different analysing crystals are available for use in the measurements, and there are two collimators, and two detectors, one of which is a flow counter and the other a scintillation counter. These components are automatically selected in a preset programme. The considerations underlying this selection are discussed. In addition to the accuracy and resolution required, the measuring time required and the effect of temperature variations are among the factors that have to be taken into account.

An 8 mm reflex klystron of simple design

621.385.623.5

Radar in the 8 mm band is now widely employed for applications such as navigation in narrow waterways. In the system developed for this purpose and described a few years ago in this journal [1] a reflex klystron is used as the local oscillator in the receiver. With a view to applications for reflex klystrons in this wavelength region with somewhat different requirements — e.g. in various kinds of measuring equipment — an experimental 8 mm tube has been developed, which has much the same characteristics as the tube mentioned above [1] but offers more possibilities for tuning and is much simpler in design.

The tube is shown diagrammatically in *fig. 1*. The upper and lower halves of the tube are almost identical in construction: most of the components are identical or derived from the same basic form. The number of components is also relatively small. Another feature of the design is the use of two large springs, *1* and *2*, which keep the gun and the reflector pressed against the central block containing the resonant cavity. This makes assembly of the tube fairly simple.

The resonant cavity (*fig. 2*) is made up from the section C_1, which is symmetrical about the axis of the tube, the tapered section C_2 and a relatively long, straight waveguide $C_3 + C_4$, which is located partly inside and partly outside the vacuum of the tube, and is terminated by a highly reflecting element, e.g. a coupling iris or a non-contacting piston. The klystron delivers power through this element to the external circuit. The tuning system is in the outside part of the resonant cavity. This arrangement contributed greatly towards simplification of the design; it was possible as in the form chosen the cavity could be several wavelengths long.

As in every reflex klystron, most of the beam electrons ultimately strike the nozzle-shaped piece of the cavity wall, and almost all of the beam power is dissipated here. In the new tube this part is enclosed on all sides by thick copper walls, so that the heat generated is easily removed. As a result, variations in beam current cause relatively insignificant temperature variations. The springs mentioned above ensure that variations in the thermal expansion of the glass parts do not alter the relative alignment of gun, reflector and cavity. The frequency of the tube is therefore extremely stable: it varies by only 0.45 Mc/s for a 1% change in beam current, and by 0.14 Mc/s for a 1% change in heater power. At the same time, this method of fixing the gun and reflector sections makes the tube exceptionally insensitive to mechanical vibrations and shock.

[1] See J. M. G. Seppen and J. Verstraten, An 8 mm high-resolution radar installation, Philips tech. Rev. **21**, 92-103, 1959/60. The construction of the reflex klystron used in this installation is broadly identical with that of the reflex klystrons for wavelengths of 4 and 2.5 mm, described in the article by B. B. van Iperen in Philips tech. Rev. **21**, 221-228, 1959/60.

trol knob can be used for indicating whether the element in question is to be analysed or passed over. Once this programme has been set for a particular sample, all that has to be done is to introduce the sample and push the starting button. The evacuation and the analysis programme are then carried out automatically. The results of the analysis are printed out on a strip of paper.

If the instrument is regularly called upon to analyse samples of different types for which it has not normally been preset, but which involve the analysis of no other elements apart from the 24 to which the braking arms have been set, a programme for each type can be set up on an interchangeable board with plug-leads and sockets. This dispenses with the necessity of setting a fairly large number of controls.

Summary. After a brief review of the principles underlying the operation of an X-ray spectrometer, an instrument is described which automatically analyses a number of samples for their content of 15 different elements. The number of samples per measuring cycle is four in absolute measurements, or three in ratio measurements. Seven different analysing crystals are available for use in the measurements, and there are two collimators, and two detectors, one of which is a flow counter and the other a scintillation counter. These components are automatically selected in a preset programme. The considerations underlying this selection are discussed. In addition to the accuracy and resolution required, the measuring time required and the effect of temperature variations are among the factors that have to be taken into account.

An 8 mm reflex klystron of simple design

621.385.623.5

Radar in the 8 mm band is now widely employed for applications such as navigation in narrow waterways. In the system developed for this purpose and described a few years ago in this journal [1] a reflex klystron is used as the local oscillator in the receiver. With a view to applications for reflex klystrons in this wavelength region with somewhat different requirements — e.g. in various kinds of measuring equipment — an experimental 8 mm tube has been developed, which has much the same characteristics as the tube mentioned above [1] but offers more possibilities for tuning and is much simpler in design.

The tube is shown diagrammatically in *fig. 1.* The upper and lower halves of the tube are almost identical in construction: most of the components are identical or derived from the same basic form. The number of components is also relatively small. Another feature of the design is the use of two large springs, *1* and *2*, which keep the gun and the reflector pressed against the central block containing the resonant cavity. This makes assembly of the tube fairly simple.

The resonant cavity (*fig. 2*) is made up from the section C_1, which is symmetrical about the axis of the tube, the tapered section C_2 and a relatively long, straight waveguide $C_3 + C_4$, which is located partly inside and partly outside the vacuum of the tube, and is terminated by a highly reflecting element, e.g. a coupling iris or a non-contacting piston. The klystron delivers power through this element to the external circuit. The tuning system is in the outside part of the resonant cavity. This arrangement contributed greatly towards simplification of the design; it was possible as in the form chosen the cavity could be several wavelengths long.

As in every reflex klystron, most of the beam electrons ultimately strike the nozzle-shaped piece of the cavity wall, and almost all of the beam power is dissipated here. In the new tube this part is enclosed on all sides by thick copper walls, so that the heat generated is easily removed. As a result, variations in beam current cause relatively insignificant temperature variations. The springs mentioned above ensure that variations in the thermal expansion of the glass parts do not alter the relative alignment of gun, reflector and cavity. The frequency of the tube is therefore extremely stable: it varies by only 0.45 Mc/s for a 1% change in beam current, and by 0.14 Mc/s for a 1% change in heater power. At the same time, this method of fixing the gun and reflector sections makes the tube exceptionally insensitive to mechanical vibrations and shock.

[1] See J. M. G. Seppen and J. Verstraten, An 8 mm high-resolution radar installation, Philips tech. Rev. 21, 92-103, 1959/60. The construction of the reflex klystron used in this installation is broadly identical with that of the reflex klystrons for wavelengths of 4 and 2.5 mm, described in the article by B. B. van Iperen in Philips tech. Rev. 21, 221-228, 1959/60.

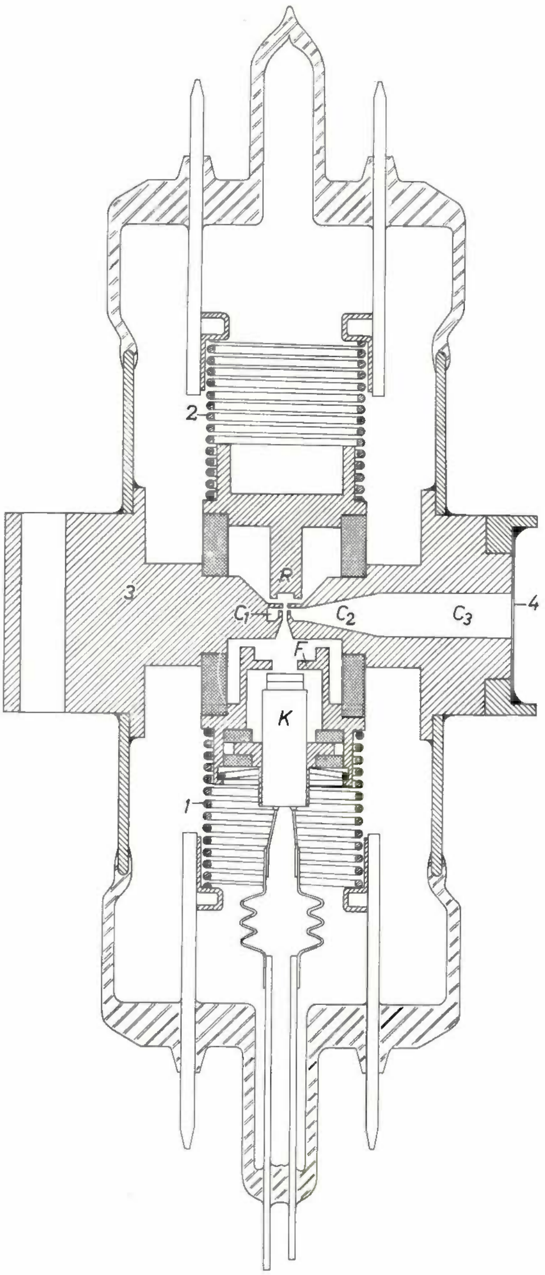


Fig. 1. Simplified cross-section of the experimental 8 mm reflex klystron. The gun section (with gun *K* and focusing electrode *F*) and the section with reflector *R* are pressed by springs *1* and *2* against the central copper block *3*. The springs also form electrical connections. The correct relative alignment of the various components is ensured by ceramic spacer rings (shaded). C_1 central section of the resonant cavity, C_2 tapered section, C_3 waveguide section. *4* vacuum-seal window.

The tube can be tuned in several ways. Four of these are mentioned here and illustrated in *fig. 3*. a) The length of the waveguide section of the resonant cavity can be varied by means of a piston. b) A vane of dielectric material can be inserted to an adjustable depth through a slot in the broad wall of the guide. With these two methods a tuning range of up to 3%, or approximately 1000 Mc/s can be achieved; this is less than with the old tube, but sufficient for many

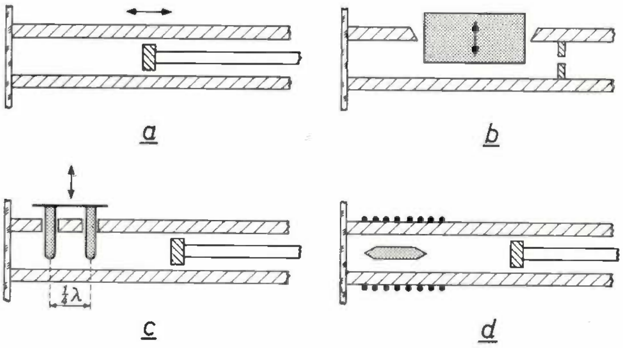


Fig. 3. Four methods of tuning the klystron.
a) Mechanical, with an adjustable piston.
b) Mechanical, with a dielectric vane projecting through a slot in the broad wall of the waveguide.
c) Mechanical, with sapphire pins through holes in the broad wall of the waveguide.
d) Electrical, by placing in the waveguide a ferrite rod which is magnetized by an external coil.
Methods (a) and (b) allow tuning over about 1000 Mc/s, method (d) over about 500 Mc/s. Methods (c) and (d) are suitable for frequency modulation.

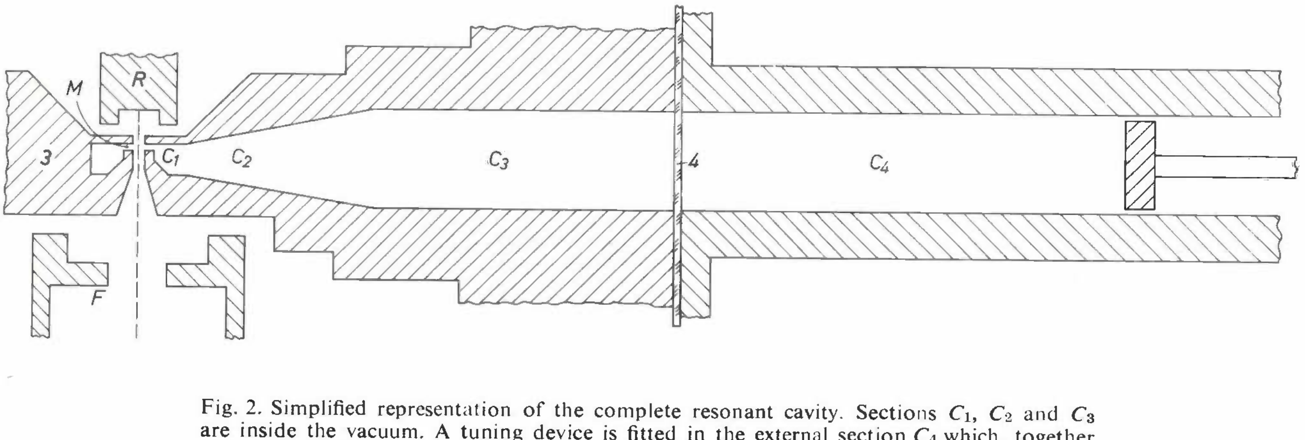


Fig. 2. Simplified representation of the complete resonant cavity. Sections C_1, C_2 and C_3 are inside the vacuum. A tuning device is fitted in the external section C_4 which, together with C_3, constitutes the waveguide section. *M* interaction gap.

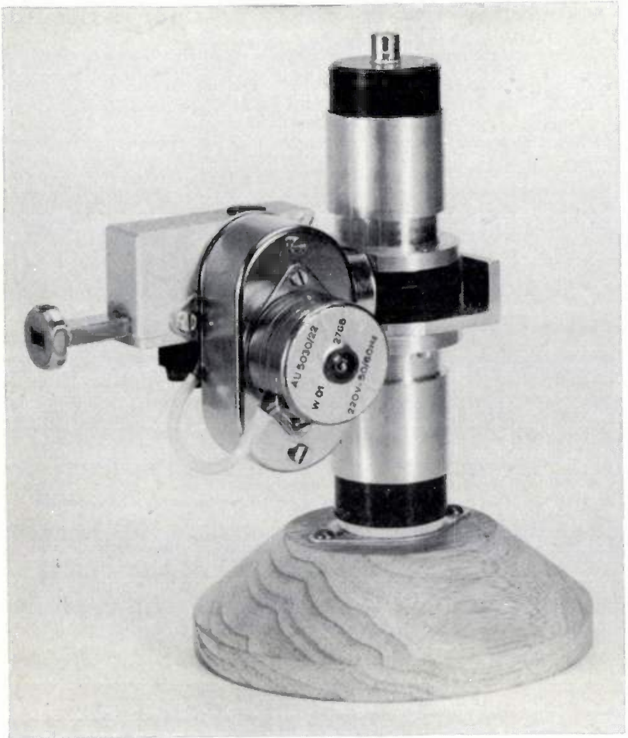


Fig. 4. The reflex klystron complete with external section of the resonant cavity, which is fitted with a tuning system as in fig. 3*b* and with a small motor and gearbox for operating the tuning system by remote control. Operating values: beam current 15 mA, anode voltage 2500 V, heater power 5.8 W. Output power about 100 m W, tuning range about 1000 Mc/s.

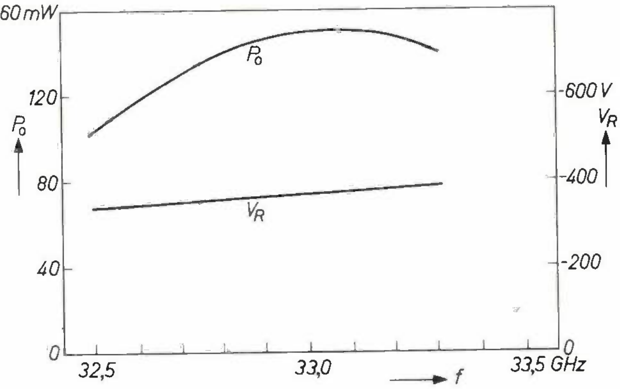


Fig. 5. Variation of the output power P_0 with frequency f for tuning by the method in fig. 3*b*, from measurements on a random sample. V_R corresponding reflector voltage.

applications. For tuning by method (c) two sapphire pins are inserted into the waveguide through holes in the broad wall. This method (like method d), is particularly suited for *rapid periodic* frequency variation (frequency modulation): this can be obtained by mounting the pins, which are extremely light, on a speaker diaphragm [2]. The upper limit of the frequency sweep can then be set by method (a) or (b). Finally, in method (d) the frequency is varied by varying the magnetization of a ferrite rod mounted in the waveguide; this is done with the aid of a solenoid wound around the guide. This is therefore an electrical method of tuning. The frequency can be varied by this method in a range of 1 to 1½% (about 500 Mc/s). Other electrically-tuned oscillators, backward wave oscillators in particular, permit a greater frequency sweep but are much more complicated. If the frequency sweep required is not so great — for many dynamic measurements 150 to 200 Mc/s is quite sufficient — the tube described here, with modulation method (d), is very attractive.

Fig. 4 shows a photograph of a complete tube fitted with the tuning system of method (b). The vane is displaced by an eccentric disc, driven by a small motor. This version of the new tube is particularly useful in radar equipment; the tube is then directly beneath the aerial and therefore has to be operated by remote control. *Fig. 5* gives a curve of the output power P_0 against frequency for this version, together with the corresponding reflector voltage, which is set for maximum P_o at every frequency.

G. H. Plantinga
J. W. Rommerts
Th. J. Westerhof

[2] A similar construction (including the principle of a resonant cavity partly outside the vacuum) has previously been used in this laboratory for a multi-reflex klystron: see B. B. van Iperen and J. L. van Lidth de Jeude, Philips tech. Rev. **24**, 184-195, 1962/63.

Ir. G. H. Plantinga, now with Philips Nijmegen Semiconductor Works, was formerly with Philips Research Laboratories; J. W. Rommerts was with Philips Research Laboratories until his untimely death in June 1966; Th. J. Westerhof is with Philips Research Laboratories.

Volume 27, 1966, No. 11 pages 285-312 Published 19th Januari 1967.

PHILIPS TECHNICAL REVIEW

VOLUME 27, 1966, No. 12

Masers for a radio astronomy interferometer

F. W. Smith, P. L. Booth and E. L. Hentley

621.375.9:522.617.3

The maser, with its very low equivalent noise temperature, has found a number of applications, particularly in radio astronomy. An interesting feature of the maser system described here is the superconducting magnet; the current in this magnet is adjusted by a superconducting dynamo following the principle described in this journal some time ago.

Introduction

In applications where a receiving aerial is directed toward cold space, so that the background noise is low, full use may be made of the low-noise properties of solid-state maser amplifiers. One such application is the detection and measurement of the discrete sources of radio noise which are present throughout the universe. In recent years the study of these sources has led to the discovery of a previously unknown type of stellar body, the "quasi-stellar object" or "quasar", and the investigation of these bodies is proving of great importance to the astronomer. Some of the stellar sources of radio noise have been positively identified with visible objects such as distant galaxies. Such identification requires precise measurement of the position and the angular diameter of the radio source.

One method which is particularly applicable to this purpose is the use of a radio interferometer. The noise from the source is received in two aerials and the correlation is measured between the noise outputs of the two aerials [1]. In principle correlation will be appreciable only if the difference in the pathlengths from the source to the two aerials is around one wavelength or smaller. A high precision is obtained by spacing the aerials many wavelengths apart: correlation will then be measured only if the source is in a very small angle around the normal to the base line. (The system may be aimed in a direction different from the normal by introducing the appropriate signal delay in one of the branches.) The diameter of the source is obtained by measuring the correlation as a function of the distance between the aerials [2].

To be able to measure weak sources a low noise temperature of the receiver is required, and it is desirable to use a maser, which can have an equivalent noise temperature of only a few °K, as the first stage. The sensitivity of the system may be expressed in terms of a minimum detectable source temperature ΔT given by:

$$\Delta T = T_{\rm s}/\sqrt{2B\tau}, \quad \ldots\ldots \quad (1)$$

where $T_{\rm s}$ is the receiver system noise temperature (mainly determined by the first stage), B the bandwidth of the receiver and τ the post-detector integration time. At given values of $T_{\rm s}$ and B, long integration times are often required to obtain the desired sensitivity.

To measure accurately the correlation which may exist between the two signals received, any distortion of the signals during amplification must be avoided; or, more precisely, the two signals should not be distorted differently. It is therefore a requirement that the maser gain and phase characteristics match, and, in view of the possibility of long integration times, that they are highly stable.

The radio interferometer at Defford, near Malvern, England, has two 25 m diameter parabolic aerials

F. W. Smith, B.Sc., P. L. Booth, B.Sc., and E. L. Hentley, A.M.I.E.R.E., are with Mullard Research Laboratories, Redhill, Surrey, England.

[1] R. R. E. Journal, No. 50, Oct. 1963.

[2] This principle and its application to "optical" stars have been treated in: R. Hanbury Brown and A. Browne, The stellar interferometer at Narrabri, Australia, Philips tech. Rev. 27, 141-159, 1966 (No. 6).

(*fig. 1*). One of these can be moved along a substantially East-West railway track, the other along a track at 67° to the first. This arrangement allows the length and orientation of the base line to be varied.

In this article a short description is given of the masers that have been built for this system. These are travelling wave masers, operating at 3.025 Gc/s (approximately 10 cm). In many respects they resemble the travelling wave maser developed in this laboratory for use in satellite communications, which has been described extensively in a previous article [3] in this journal. Two possibilities for improved performance which were discussed in the earlier article have been realized in the present masers. The first is the use of *persistent current superconducting electromagnets* which ensure a very stable magnetic field at the maser crystal, and thus a high stability of phase and gain. The second is the use in the helium dewar vessel of waveguide for the input lead instead of a coaxial line: this reduces the microwave losses of the input and thereby the noise temperature.

As the interferometer system requires a considerable distance between the aerials to obtain a high angular resolution, it is desirable to have remote tuning of the centre frequency of operation of the masers, to enable their gain-frequency characteristics to be made identical. This is achieved in the present system by the use in each maser of a *superconducting dynamo* controlling the magnetic field of the superconducting magnet, which in turn determines the centre frequency.

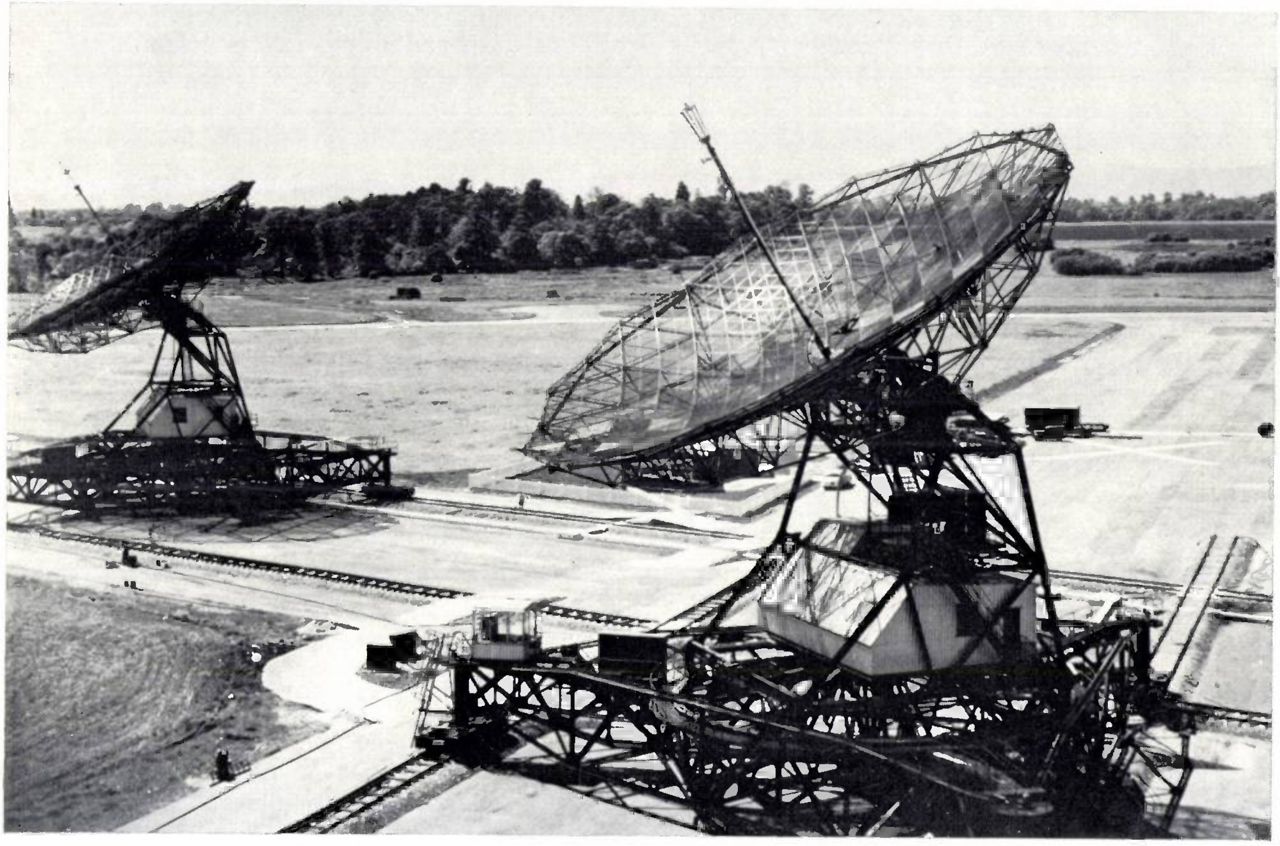


Fig. 1. The radio interferometer at Defford, near Malvern, England. Noise from a radio source is received at each of the two parabolic aerials (diameter 25 m). The noise signal is amplified in each channel and the receiver noise can be strongly reduced by using a maser as the first amplifier stage. Finally the correlation between the signals is measured. To improve the noise performance and to facilitate the fitting of the masers the aerials are being modified to a Cassegrain system. The aerials can be moved along the railway tracks in two different directions to vary the base line in length and orientation.

The travelling wave maser

For an extensive description of the maser principle and of considerations for the design of a travelling wave maser the reader is referred to the article [3] mentioned

[3] J. C. Walling and F. W. Smith, Solid state masers and their use in satellite communication systems, Philips tech. Rev. **25**, 289-310, 1963/64.

above. We will confine ourselves to a short description of the present masers.

Let us recapitulate in a few words the main elements of the travelling wave maser. The microwave signal to be amplified travels from input to output along the *maser crystal*, a paramagnetic crystal, which delivers power to the wave. The crystal is mounted in a *slow wave structure* to intensify the interaction of the crystal with the wave. The maser crystal is activated by a microwave *pump signal* usually at an appreciably higher frequency than the signal to be amplified. The crystal is placed in a *magnetic field* tuning it to the frequencies of signal and pump. Finally a *low temperature* is essential to ensure appreciable gain and low noise: the maser is usually operated in a bath of liquid helium.

In a travelling wave maser, the "electronic gain" G (the gain produced by the maser material not counting structure losses) is given (in dB) by:

$$G = 27.3 \frac{fLF}{v_g} \frac{1}{Q_m}, \quad . \; . \; . \; . \quad (2)$$

f being the frequency, L the length of the slow wave structure, v_g the group velocity, F the fraction of the total magnetic energy stored in the maser material, and Q_m the magnetic quality factor of the maser material.

The slow wave structure of the present masers is of the comb type. The arrangement and the dimensions are shown in *fig. 2*. The length of the structure is 12 cm, the group velocity is approximately 1/80 times the free space velocity of light. One side of the structure is loaded with maser material. The other side contains yttrium iron garnet (YIG) discs held in place by a dielectric material (with a dielectric constant of 9, which approximately matches that of the maser material). By this arrangement the device is made non-reciprocal: under operating conditions forward waves interact substantially only with the maser material and are amplified whereas backward waves interact only with the YIG discs and are attenuated (cf. ref. [3]). The comb structure is attached directly at one end to the waveguide transmitting the pump power.

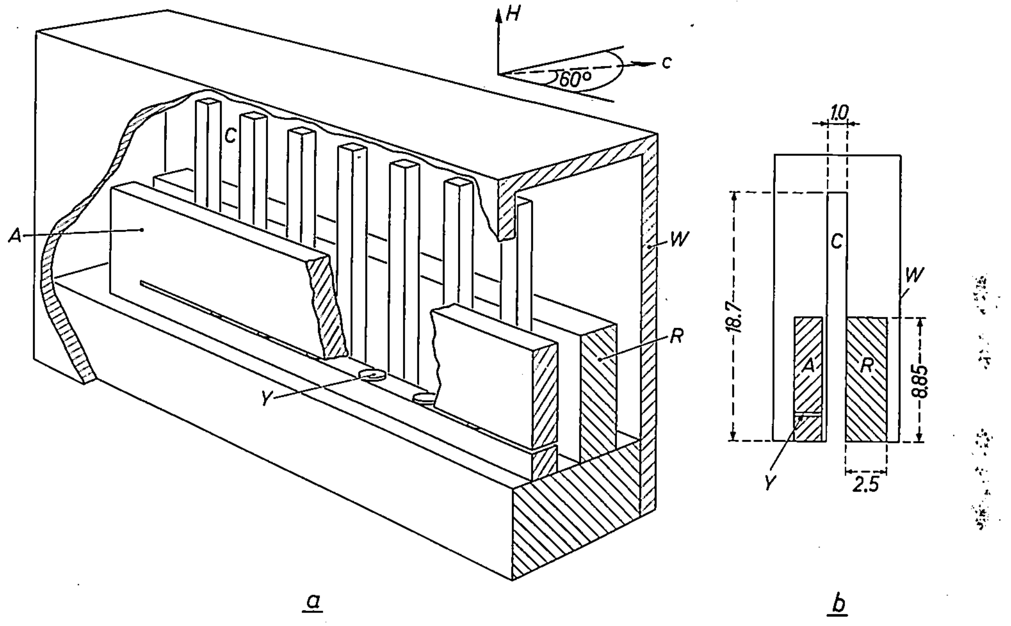


Fig. 2. Cut-away view and cross-section of the 3.025 Gc/s travelling wave maser. The dimensions are in mm. Electromagnetic waves are propagated slowly (group velocity 1/80 times the free space velocity of light) along the comb structure *C*; forward waves are amplified by the maser crystal *R* (ruby), backward waves are attenuated by the YIG discs *Y* held in place by the dielectric slab *A*. The maser crystal is tuned by a magnetic field *H* and activated by a pump signal (26.4 Gc/s) transmitted by the waveguide *W*. The orientation of the optic axis *c* is indicated.

The maser material is a ruby single crystal (0.04% chromium by weight), about 12 cm long, with the optic axis at 60° to the longitudinal direction (i.e. the direction of growth). In the structure the ruby is mounted with the optic axis perpendicular to the pins of the comb; these are parallel to the static magnetic field.

Fig. 3 shows the paramagnetic energy levels of the ruby in a magnetic field perpendicular to the optic axis; the pumping scheme is indicated.

Both at the input and at the output the comb structure is matched to a rigid low loss (helical membrane) coaxial line. Final adjustment of the passband of the slow wave structure is made by means of slight alterations to the dimensions of the dielectric loading.

The intrinsic loss of the structure is 12 dB; when

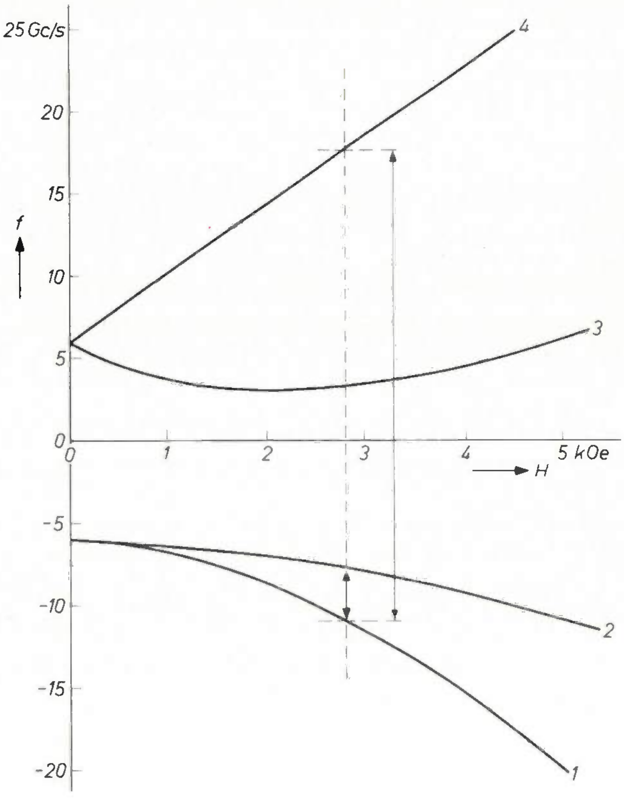


Fig. 3. Paramagnetic energy levels of the maser crystal (ruby) shown against H when H is perpendicular to the optic axis. The pump transition *1-4* (26.4 Gc/s) and the signal transition *1-2* (3.025 Gc/s) in a field of 2800 Oe are indicated.

operated at 1.5 °K in a uniform magnetic field the electronic gain is 48 dB, giving a net gain of 36 dB. Under these conditions the 3 dB bandwidth is 14 Mc/s. The bandwidth may be increased at the expense of gain by using a non-uniform (staggered) static field, so that different parts of the crystal experience different magnetic fields and thus have different centre resonant frequencies. A greater bandwidth thus obtained is advantageous (cf. eq. 1) if the lower gain is still sufficient to render the noise contributions from later stages negligible. The overall noise temperature is not affected as long as the pump levels remain saturated in all sections of the crystal and the maser noise temperature is itself not increased. The maser noise temperature is not affected provided all sections of the crystal contribute significant net gain over the increased bandwidth [3].

The superconducting magnet and the superconducting dynamo

The magnetic field which tunes the maser crystal to the signal and pump frequencies is provided by a superconducting magnet, built closely around the slow wave structure, the magnet and the slow wave structure being immersed in the helium bath together as one unit (*fig. 4*).

The magnet [4] consists of a yoke of mild steel with two superconducting coils; a cross section is shown in *fig. 5*. The gap volume is 3.5×3.0×22 cm. A slab of lead-bismuth alloy (50Pb50Bi) is placed on either side of the pole pieces and the coils to act as a superconducting diamagnetic screen and reduce the leakage from the gap [5]. Fig. 5*a* and *b* show the effect of these shields. The uniformity as measured over the ruby crystal volume is better than 1 part in 10^3 (cf. *fig. 6*).

The coils are made of copper-covered niobium-zirconium wire (75Nb25Zr) 0.25

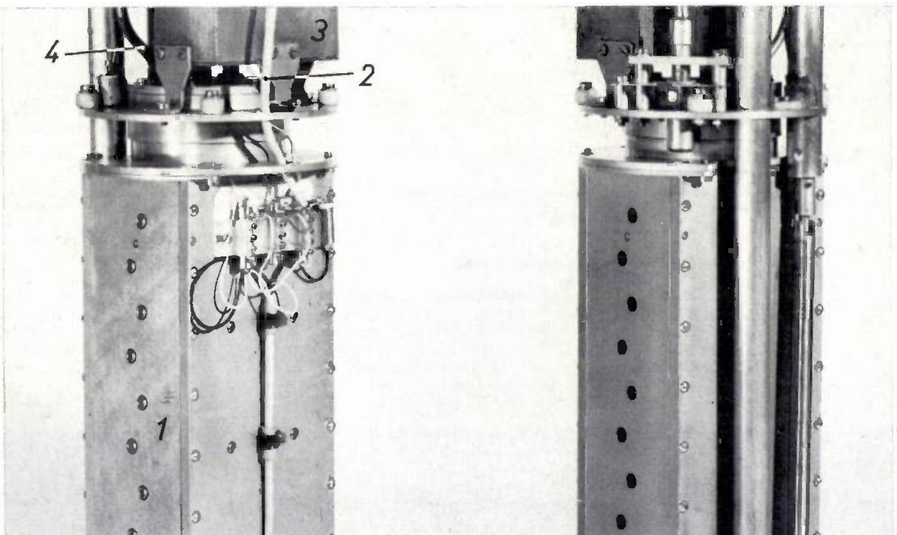


Fig. 4. The unit containing the slow wave structure and the magnet. The magnet *1* can be rotated over a few degrees by the gears which can be seen at the top of the maser on the right. *2* lead from external magnet current supply. *3* input waveguide. *4* output coaxial line.

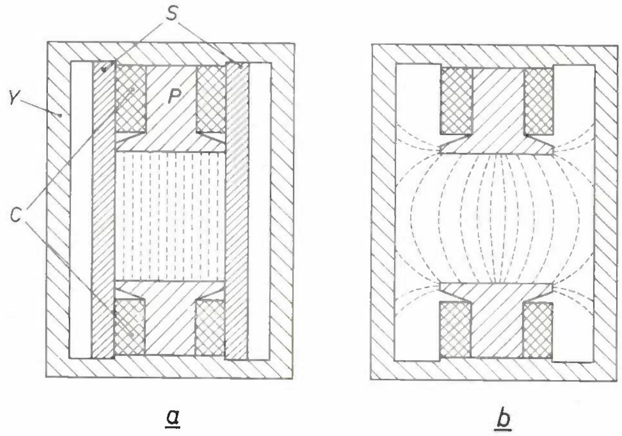


Fig. 5. *a*) Cross-section of the superconducting magnet. *Y* yoke and *P* pole pieces of mild steel. *C* superconducting coils of niobium-zirconium wire. *S* superconducting shields of lead-bismuth. The lines of magnetic field are shown dashed.
b) To show the effect of the shields *S*, the field pattern that would be obtained without the shields is shown.

mm in diameter, and nylon-insulated. Each coil contains 600 turns. They are wound on coil formers of a titanium alloy with the same thermal expansion coefficient as the wire.

The coils make up a superconducting closed circuit together with a superconducting dynamo which is used to obtain a fine variation of the magnet current. This dynamo is described below. In principle it could be used to build up the whole magnetic field; this, however, is not a practical procedure in this case as it would take some 9 hours. The arrangement by which the current in the superconducting magnet is established initially is shown in *fig. 7*. An external power supply is connected to the coils, in parallel with the superconducting dynamo. The superconduction in the dynamo branch is disturbed by a *thermal switch*, and the current delivered by the power supply is taken completely by the magnet coils. After the current has been set to the correct value the thermal switch is de-energized; superconduction is restored in the dynamo branch and a persistent current flows in the superconducting circuit. The external supply can then be disconnected.

The thermal switch consists of a coil of niobium-zirconium wire interwound with a heater coil of constantan wire; a current of 50 mA in the latter will cause normal conduction in the niobium-zirconium coil.

The superconducting dynamo, originally developed at Philips Research Laboratories, Eindhoven [6], is

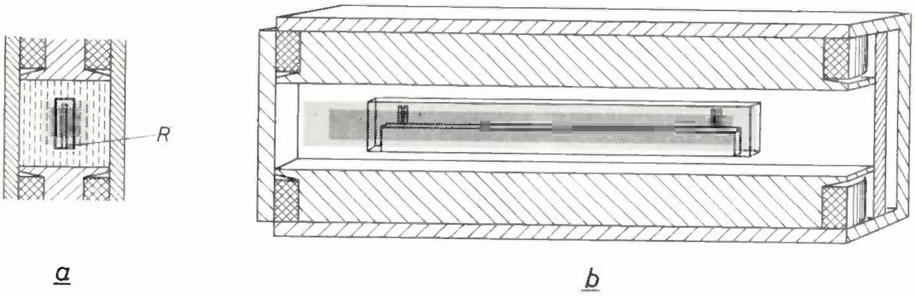


Fig. 6. The distribution of the magnetic field in the magnet gap. In the dark grey area the field homogeneity is better than 1 : 10^4, in the light grey area better than 1 : 10^3. On the left a cross-section of the magnet (as in fig. 5*a*) is shown. *R* is the ruby crystal.

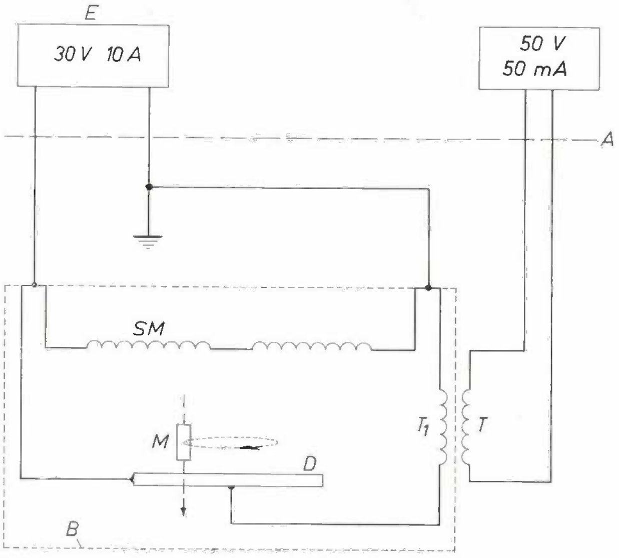


Fig. 7. Magnet current supply diagram. The persistent current in the superconducting circuit (within the contour *B*) is established by: a) activating the "thermal switch" TT_1 (i.e. disturbing the superconduction in the coil T_1 by a small current through *T* heating *T* and T_1); b) connecting the external supply *E* and setting the current; c) de-energizing the thermal switch TT_1. As soon as T_1 is superconducting again, a persistent current flows in *B* and *E* can be disconnected. *SM* superconducting coils. *M*, *D* superconducting dynamo. The parts drawn above line *A* are outside the cryostat.

[4] E. L. Hentley, Mullard Research Laboratories Report, Sept. 1964.

[5] P. P. Cioffi, J. appl. Phys. **33**, 875, 1962.

[6] J. Volger and P. S. Admiraal, Physics Letters **2**, 257-259, 1962. See also J. Volger, Philips tech. Rev. **25**, 16-19, 1963/64, and J. van Suchtelen, J. Volger and D. van Houwelingen, Cryogenics **5**, 256-266, 1965 (No. 5).

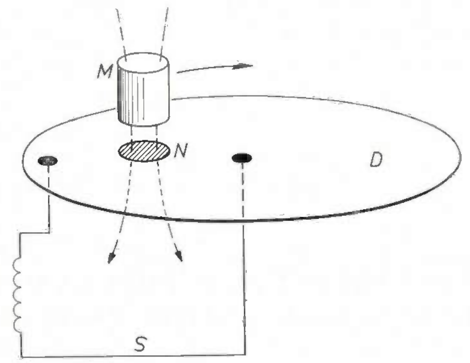


Fig. 8. Diagram of the superconducting dynamo. The permanent magnet M induces a normal, non-superconducting region N in the superconducting disc D and some flux from M penetrates through N. When M rotates around the centre of D so that N passes repeatedly through the terminals of the superconducting circuit S, flux is pumped into S and the persistent current is changed.

shown diagrammatically in *fig. 8*. The dynamo (or "flux pump" as it is often called) consists of a superconducting disc — connected to the superconducting circuit at its centre and at its circumference — and small permanent magnets rotating close to the disc. These magnets induce normal, non-superconducting, regions in the disc and part of their flux penetrates through these regions. Each time a normal region passes the "terminals" of the circuit, flux is pumped into the circuit, causing a change in the persistent current. The current may be increased or decreased by rotating the magnets either clockwise or anticlockwise with respect to the disc. The flux pump is mounted at the base of the maser magnet (*fig. 9*). The disc is lead, 25 μm thick, mounted on a Terylene support disc. Two "Ticonal" permanent magnets provide a field of 1000 Oe at the disc; they are mounted in an aluminium block which can be rotated at 500 r.p.m. by a motor on the cryostat top plate. This corresponds to a field variation of 5 Oe/min which provides adequate control over the maser centre frequency.

The maser crystal resonant frequencies are dependent not only on the magnitude of the magnetic field but also on its direction with respect to the crystal axes. To enable the field to be orientated accurately during operation of the maser the magnet is mounted so that it may be rotated a few degrees around the longitudinal direction of the crystal; it can be aligned by a gear drive adjusted from the cryostat head.

In a superconducting circuit such as the above the making of superconducting joints requires great care. To join two niobium-zirconium wires, they are cleaned carefully, twisted together tightly, wrapped in a piece of 0.08 mm copper sheet and mounted in a stainless steel clamp. Non-superconducting connections (the leads of the external supply) can be soft-soldered to the copper sheet. The connection between the lead disc of the flux pump and the niobium-zirconium wire of the coils is made using niobium sheet with a layer of niobium-tin as an intermediate conductor. The layer is obtained by heating the sheet in tin surrounded by an inert atmosphere. The niobium-zirconium wires are spot-welded to the niobium sheet; a lead-bismuth wire is soldered both to the niobium-tin layer and to the lead disc.

The cryostat

The double dewar vessel of stainless steel is shown in *fig. 10*, and the maser assembly, as it is mounted in the dewar vessel, in *fig. 11*. Details are indicated in the

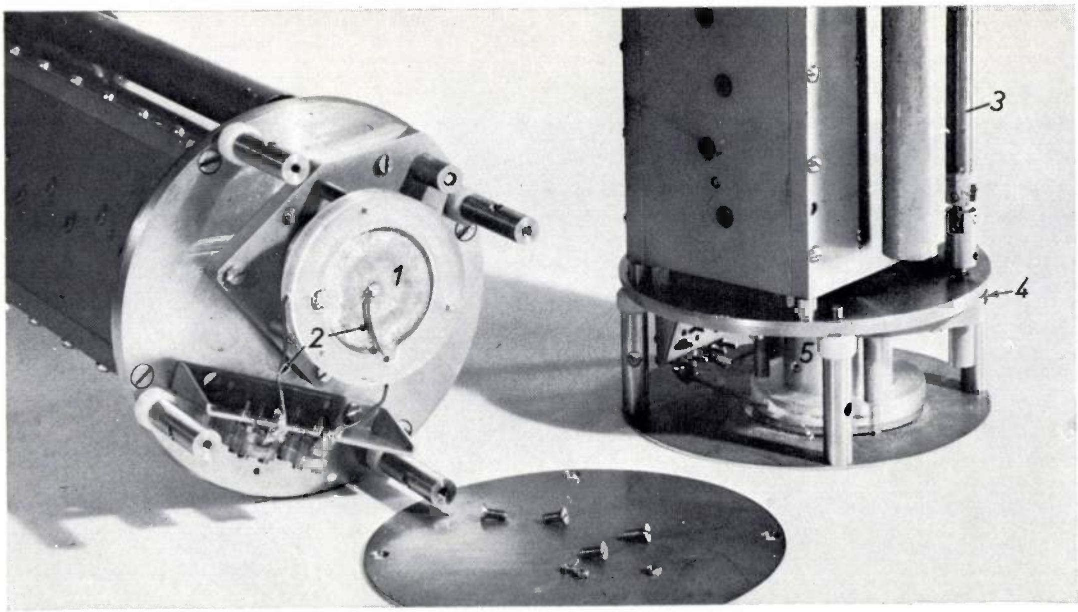


Fig. 9. The superconducting dynamo. *1* lead disc. *2* lead-bismuth wires connecting the dynamo to the magnet coils. *3* shaft, *4* gear for rotating the "Ticonal" magnets in aluminium block *5*.

captions, and the figures do not require much further comment. A few points, however, may be noted.

At a certain height a thick copper block is attached to the neck of the helium vessel. This block is kept at 77 °K by a length of copper sheet reaching into the liquid nitrogen. At this level the maser assembly contains a copper radiation shield, with spring fingers for thermal contact to the inner dewar wall, keeping all parts at this level roughly at 77 °K.

As mentioned in the introduction, the signal input

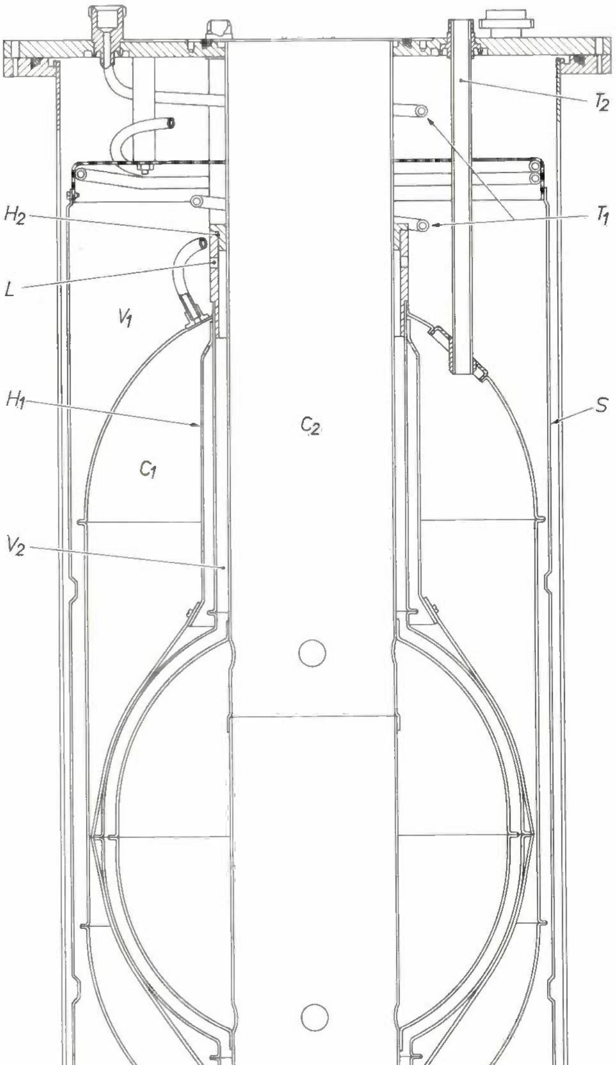


Fig. 10. Sectional drawing of the cryostat. Internal diameter 11.5 cm, maximum diameter 28 cm, length 95 cm. C_1 liquid nitrogen container. C_2 liquid helium container. V_1, V_2 vacuum spaces linked by the holes L in the copper block H_2. The "tie-point" H_2 is kept at 77 °K by the copper sheet H_1 reaching into the liquid nitrogen. S radiation shield. The top of S, connected to the nitrogen exhaust tubes T_1, is cooled by nitrogen vapour. T_2 nitrogen filling tube. The helium capacity is 17 litres, the nitrogen capacity 21 litres.

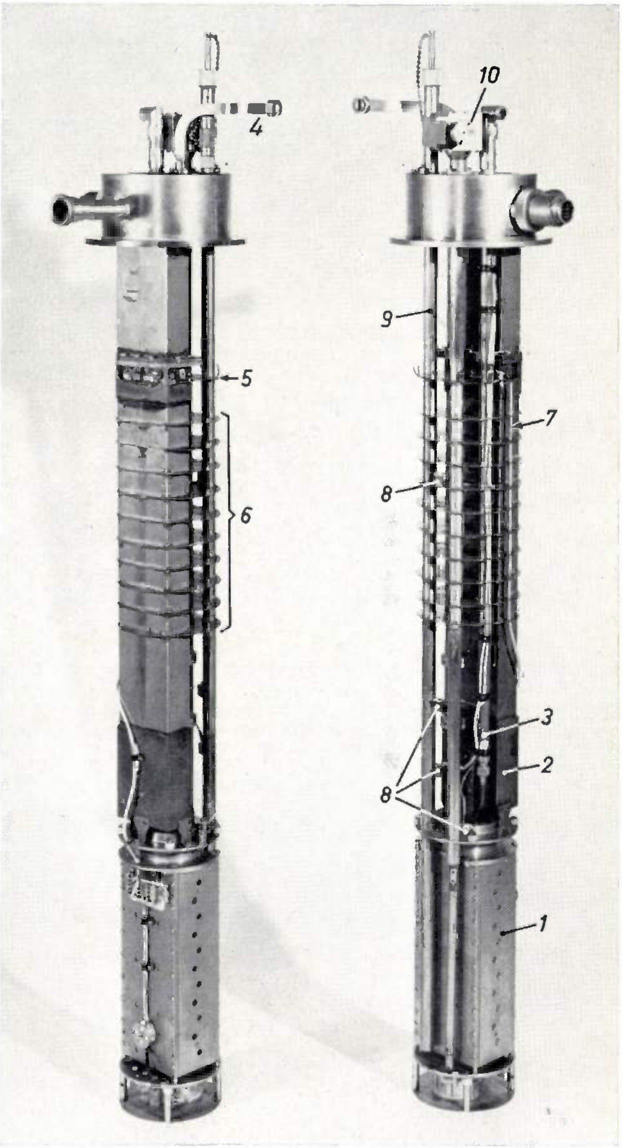


Fig. 11. The maser assembly. *1* maser magnet surrounding the slow wave structure. *2* input waveguide. *3* output coaxial line. *4* pump power waveguide. A horizontal radiation shield is kept at 77 °K by spring fingers *5* contacting the 77 °K tie-point of the cryostat (H_2 in fig. 10). The copper baffles *6* obstruct the helium vapour, leaving only a meandering path through irregular holes, thus providing a good heat exchange between the vapour and the assembly of leads. The lead *7* from the external magnet current supply which has to carry 6.8 A is of copper; it is kept at 77 °K at the radiation shield *5*. Platinum wire resistance thermometers are used as helium level indicators (*8*) [7]. *9* guide tube for filling with liquid helium. *10* gear box connecting to the motor for driving the superconducting dynamo.

[7] E. L. Hentley, Mullard Research Laboratories Technical Note, Jan. 1966.

lead is a waveguide, to reduce the input loss, and thereby keep the noise temperature at a low value. It is made of stainless steel to prevent an unacceptable heat influx to the liquid helium. The section from the 77 °K tie-point level upwards is copper plated on the inside. A waveguide-coaxial transition (standing-wave ratio 1.1) connects the waveguide to the input coaxial line of the maser. The output lead to the top of the cryostat is a thin-walled stainless steel coaxial line. The pump power is fed to the structure through a thin-walled copper-nickel waveguide. Both waveguides and the coaxial line are vacuum sealed at the top of the cryostat.

The cryostat is mounted in a cradle fixed to the aerial. When charged with liquid nitrogen and helium it can be operated at angles up to ±45° from the vertical position. The extreme positions correspond to the aerial being directed to the zenith and to the horizon.

The top of the cryostat is shown in *fig. 12* and the two complete masers in *fig. 13*.

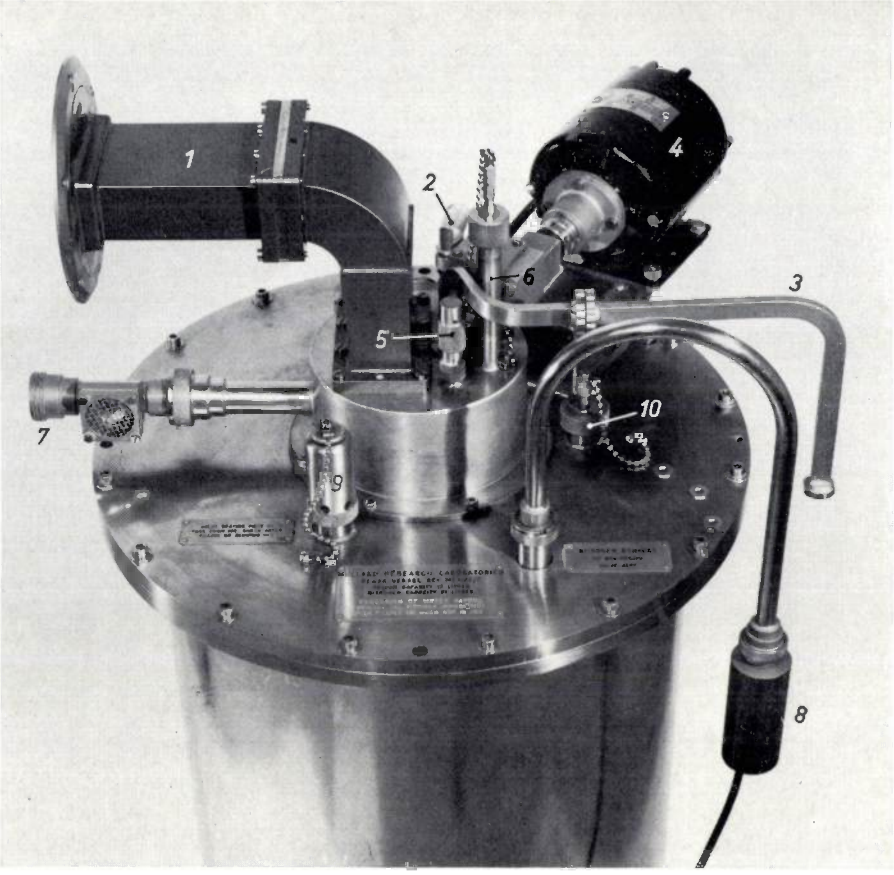


Fig. 12. The top of the cryostat. *1* input waveguide. *2* coaxial output. *3* pump waveguide. *4* motor for driving the superconducting dynamo. *5* magnetic field orientation adjustment. *6* helium filling tube. *7* helium exhaust. *8* electrically heated "bunsen valve" preventing water from entering the nitrogen dewar. *9* nitrogen safety valve. *10* nitrogen filling tube.

Performance

A few data concerning the masers are shown in the following table:

signal centre frequency	3.025 Gc/s (9.92 cm)
pump frequency	26.4 Gc/s (1.15 cm)
pump power	120 mW
magnetic field	2800 Oe
magnetic field current	6.8 A
operating temperature	1.5 °K

The following performance figures for the two masers have been obtained, using a staggered magnetic field acting on the maser crystal:

	Maser A	Maser B
Net gain	26.2 dB	28.8 dB
3 dB bandwidth	24.5 Mc/s	23.0 Mc/s
Noise temperature	4.5 ± 3.0 °K	6.0 ± 3.0 °K

The noise temperature has been obtained by a well known method [8]. Matched loads at room temperature and at 77 °K are connected to the input, and the difference in noise output is recorded.

The gain stability of the maser is affected mainly by variations in the temperature of the helium bath and by variations in the power and the frequency of the microwave pump. Gain changes of the maser will also affect the phase. Another important cause of change of phase is the variation of the effective dielectric constant in the input waveguide and the output coaxial line as the helium level drops in these leads.

The phase stability of a single maser was measured in a phase bridge. In a period of 19 hours a steady drift in phase of 19° was observed, 11° of this drift could be accounted for by the change in the helium level. Some of the residual phase drift appears to be due to a slight reduction of the helium bath temperature as the amount of the liquid in the dewar decreases. This reduction in temperature contributes to the phase change through the resulting increase in electronic gain.

The effect of variations in the pump is slight: a 5 Mc/s change in pump frequency introduced a phase

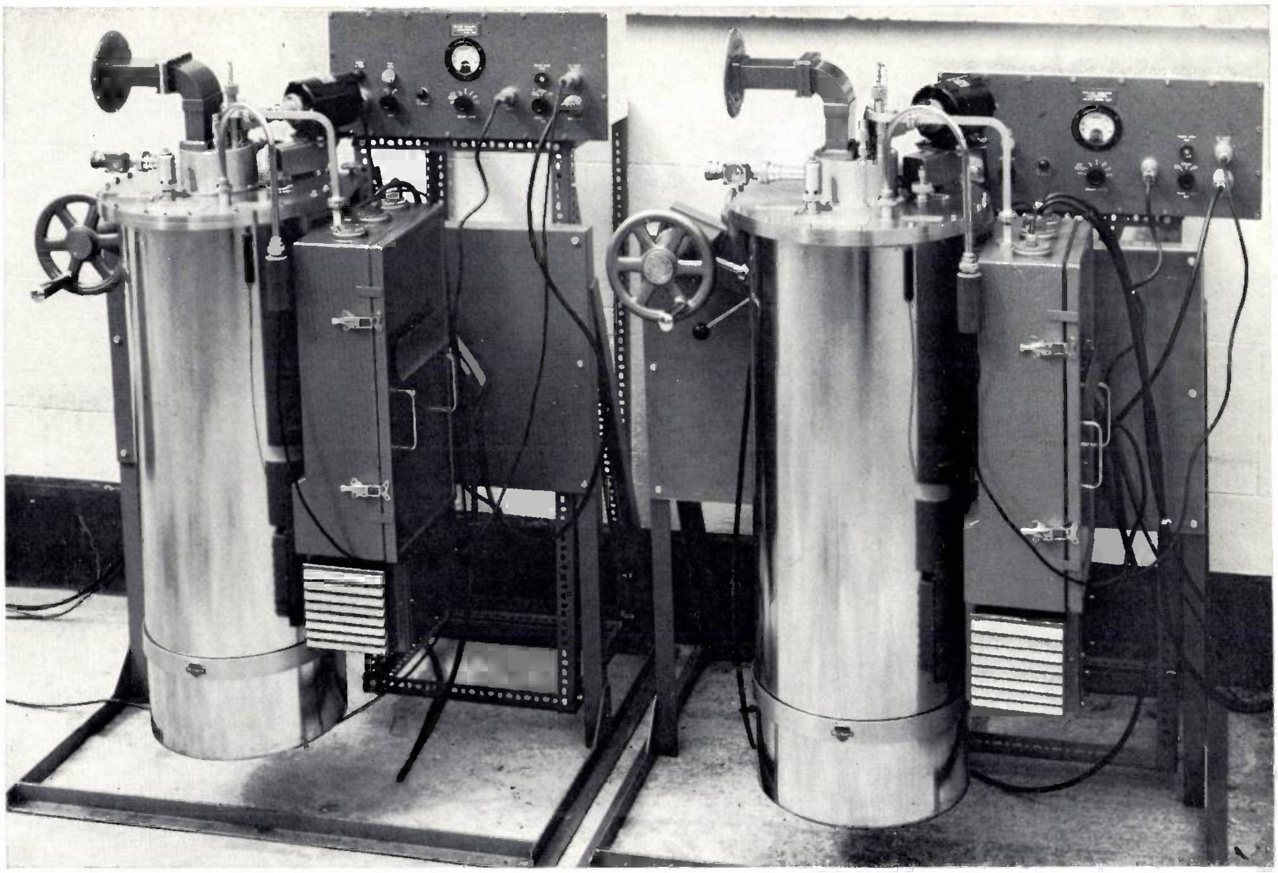


Fig. 13. The two complete masers.

change of 1°, and a 1 dB reduction in pump power resulted in a phase shift of 1.2°.

A second experiment was carried out with one maser in each arm of the phase bridge. The two cryostats were filled with liquid helium to the same level and a common vacuum pump was used to reduce the pressure above the liquid. The *differential* phase drift between the two masers, which is the parameter of importance in the interferometer application, amounted to only 5° over a period of 18 hours.

In continuous operation one helium filling of the masers will last about 44 hours. A significant fraction of the evaporation of the helium is due to the microwave pump power. The duration of a helium filling can therefore be increased by switching the pump power into a matched load when no measurements are being taken. In the radio interferometer application this is a useful procedure as there are long periods in which the maser is not being used.

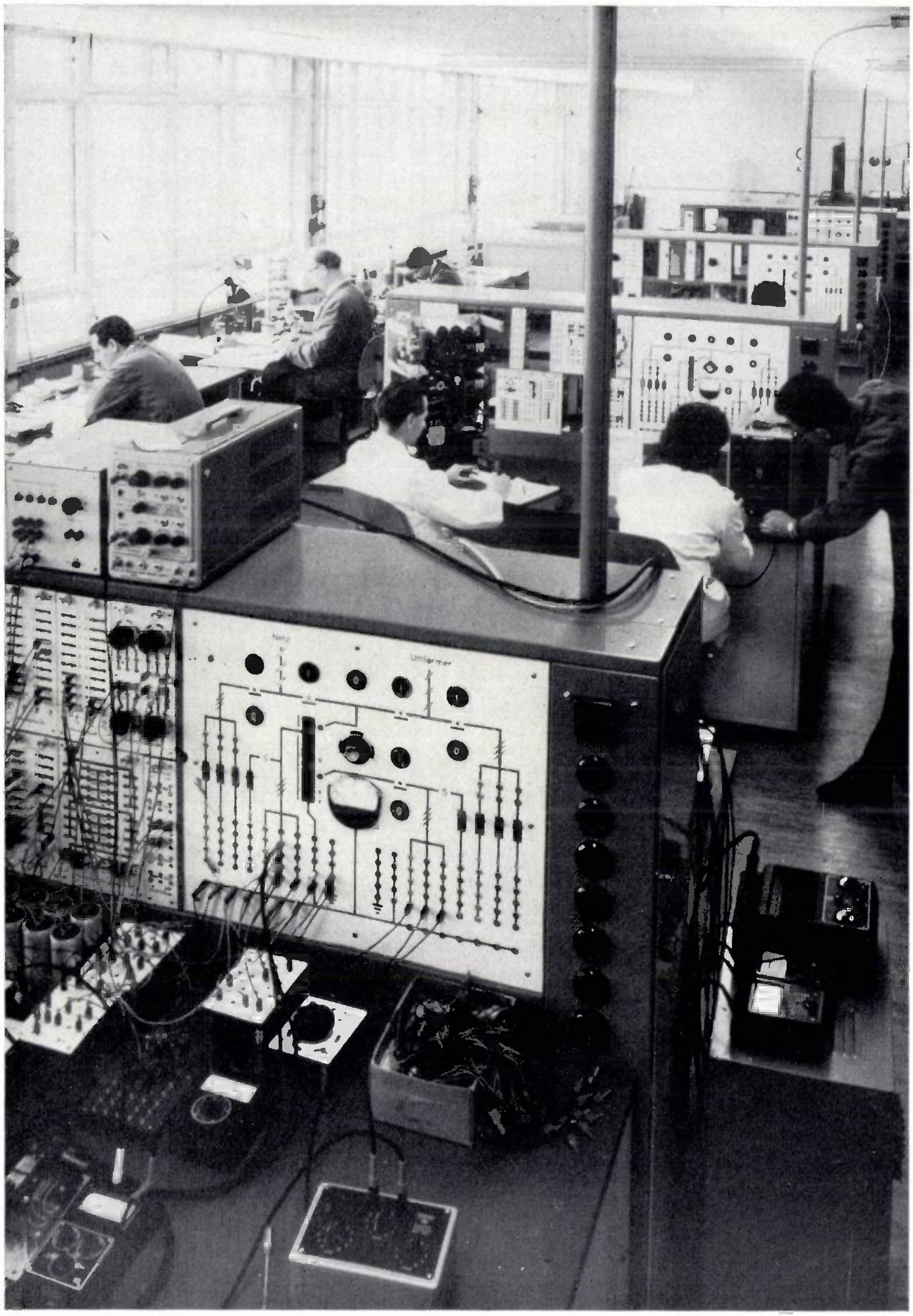
A further increase of 30 % in operational time can be achieved by completely filling the cryostat with liquid helium under reduced pressure. For this purpose special sealing valves have been constructed which make it possible to connect the cryostat to the storage dewar whilst maintaining a pressure of a few torr in both vessels.

The work described in this article was carried out on behalf of the British Ministry of Defence, Navy Department.

Summary. A description of two travelling wave masers developed by Mullard Research Laboratories for the radio astronomy interferometer at Defford, England. This interferometer, which is used for determination of the position and angular diameter of sources of radio noise, does so by measuring the correlation between the signals received by two large parabolic aerials. In order to obtain a very high receiver sensitivity an extremely low noise temperature is required, and a maser is therefore used for the first stage of each receiver. The masers must be very stable and have closely similar characteristics to enable the correlation between the two received signals to be measured accurately. High stability was obtained by making use of a superconducting magnet to give the magnetic field which tunes the ruby maser crystal to the signal and pump frequencies. The magnet and maser are arranged in a single unit in a helium cryostat. The magnetic field of each of the masers is adjusted by a superconducting dynamo, and this enables the centre frequencies of the two masers to be individually set to the same value by remote control. The masers operate at 3.025 Gc/s at an operating temperature of 1.5 °K, in a field of 2800 Oe. With a single maser, a phase shift of 19° was measured in a 19 hour run; the differential phase shift measured between the two masers was only 5° in an 18 hour run.

[8] A. L. McWhorter, J. W. Meyer and P. D. Strum, Noise temperature measurement on a solid state maser, Phys. Rev. **108**, 1642-1644, 1957.

The laboratory of an X-ray equipment factory



The photograph shows part of the electrotechnical laboratory of the C. H. F. Müller GmbH X-ray equipment factory in Hamburg. Work is carried out here on the development of low-voltage circuits for X-ray generators; the equipment meets the widely diverse requirements of modern X-ray technology.

Each row of laboratory benches is provided with variable stabilized voltage sources, which deliver a voltage at a frequency of 50 c/s or 60 c/s, as required, so that the appropriate voltage is also available for investigating export equipment. Each bench has a separate rack for holding various X-ray generator components, which are available as units of standard dimensions. Equipment such as voltage sources, stabilizers, time switches and counters can thus be set out tidily above the table. The units used are principally types in current production, or units derived from these types. When this photograph was taken a new fast-starting drive was being developed for a tube with a rapidly rotating anode ("Super-Rotalix").

The flexibility achieved with this lay-out is indispensable in order to keep pace with present-day progress in X-ray technology. Typical of this progress is the increasing employment of electronic components where, only a few years ago, almost exclusive use was still being made of conventional power components.

Semiconductor detectors for ionizing radiation

W. K. Hofker

539.1.074.55

Following the discovery of semiconductor devices that could perform the circuit functions of thermionic valves and photoelectric cells — the transistors and photoresistors now so widely used — there has been considerable success in the last few years in the development of semiconductor devices which can be used for the detection of ionizing radiation. These detectors have some exceptionally valuable features, and it can already be said that their introduction has very considerably advanced the technique of radiation measurement. By way of introduction to future articles, in which the applications of semiconductor detectors will be discussed, the article below deals with the operation and distinctive features of these new devices.

Principles and characteristic features of semiconductor detectors

In the last six years or so, great changes have been taking place in the instrumentation used for detection and energy-measurement of charged particles, and a similar situation has recently arisen in γ-ray spectrometry. The conventional instruments for energy measurement, such as ionization chambers, proportional counters and, to a lesser extent, scintillation counters, are being superseded by detectors of an entirely different type, i.e. semiconductor counters. These solid-state detectors (*fig. 1*) combine small size and ready interpretation of the output signals with exceptionally high resolution of the energy of the detected particles or quanta. This high resolution is particularly useful in many branches of research in nuclear physics. The use of semiconductor detectors has for example made possible experimental analysis of certain disintegration processes whose details could previously only be assumed on theoretical grounds. Really accurate measurement of particle or quantum energy is also often important in certain investigations directed more towards practical purposes, such as process measurements using radioactive indicators, analyses by means of chemical activation, and reactor investigations.

In principle a semiconductor detector can be regarded as an ionization chamber in which the sensitive volume is a solid instead of a gas: it consists basically of a semiconductor wafer with an electrode on each side. An incident particle or quantum produces a certain ionization charge in the semiconductor, just as it would in a gas, and this gives a pulse of current in an external circuit. This in turn causes a voltage pulse to appear across a resistance included in the circuit (*fig. 2*). The great attraction of using a solid as the detection medium is its very much higher absorption, which en-

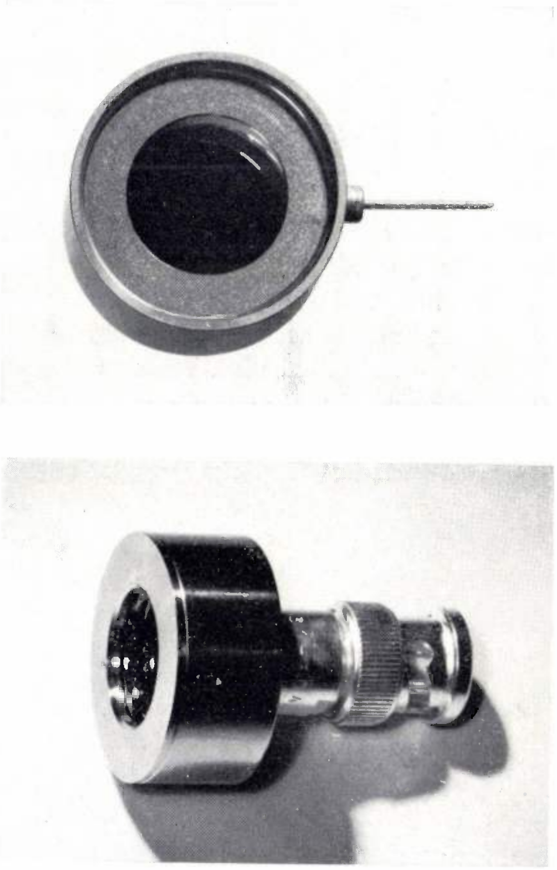


Fig. 1. Two types of experimental semiconductor detectors. As their outward form is mainly determined by their mounts (outside diameter about 3 cm), the visible differences are slight. The window has a surface area of 2 to 3 cm^2. The lower photograph shows an α-particle counter.

Ir. W. K. Hofker is with Philips Research Laboratories, Amsterdam division.

ables even fast β-particles to be stopped in a relatively thin layer, making it possible to measure their energy with an instrument of small dimensions. An incidental but extremely important advantage is that a semiconductor counter can be very much faster than a gas-filled ionization chamber. Of course, the solid used as the detection medium has to be a material in which the ionization-charge carriers — electrons and holes — can move freely under the influence of an electric field and in which they are not lost prematurely due to the presence of impurities or other crystal imperfections.

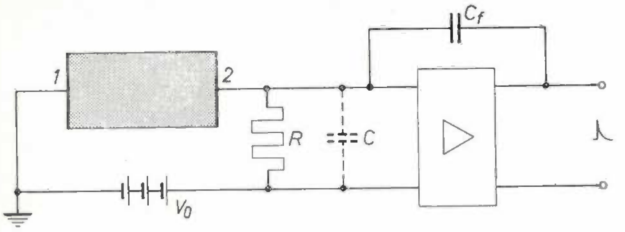


Fig. 2. Diagram and circuit of a semiconductor detector. The detection medium (shaded) is between two electrodes, *1* and *2*, to which a voltage V_0 is applied. An incident radiation quantum or charged particle frees pairs of charge carriers in the detection medium. Under the effect of the electric field the charge carriers move to the electrodes, giving rise to a voltage across R. The effect of variations in V_0 is virtually eliminated by means of capacitive feedback in the amplifier (via capacitor C_f).

The first indication that a solid can act as an ionization medium was found as long ago as 1913 by Röntgen and Joffé [1]. They established that an insulating crystal becomes slightly conductive when brought into the vicinity of a radioactive source. With the electrical instruments available at that time, however, it was not possible to measure the particles individually.

The first solid-state detector capable of counting particles individually, and which could be used for measuring their energy, was not reported until 1945, by Van Heerden [2]. The ionization medium used was a silver chloride crystal which was held at the temperature of liquid air. The resolution, however, was rather poor.

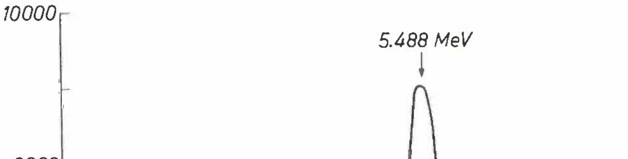
The great advances made in this respect in recent years have mainly come about because intensive semiconductor research has made it possible to produce single crystals of silicon and germanium which combine extreme purity with a very high degree of crystal perfection. With such starting materials, and using techniques partly derived from diode and transistor manufacture and partly new, solid-state detectors can now be made that meet exacting requirements.

The chief characteristic features

The pulse spectra obtained with a semiconductor counter are very easily interpreted, because the height of the pulses is proportional to the energy of the detected particle, while moreover the proportionality factor is the same for all kinds of particles — except very heavy ones like fission products.

The pulse rise-time, i.e. the time in which the ionization charge is produced plus the time it takes to travel to the electrodes, can be very short: in certain circumstances as short as 10^{-8} s, and never longer than 10^{-5} s at the very most. As this time mainly determines the number of incident particles which the counter can handle per second, semiconductor counters show a very good performance in this respect. The short rise time is also important in the study of coincident phenomena; the shorter the rise time the more accurate is the time discrimination. This will reduce the number of situations in which two events taking place in rapid succession are mistakenly regarded as coincident.

The energy resolution, as in all ionization chambers, is determined by 1) the electrical noise from the detector and amplifier, and 2) the statistical fluctuations in the ionization process. Both effects are relatively weaker in those counters in which particles of a particular kind and energy bring about stronger ionization, i.e. in



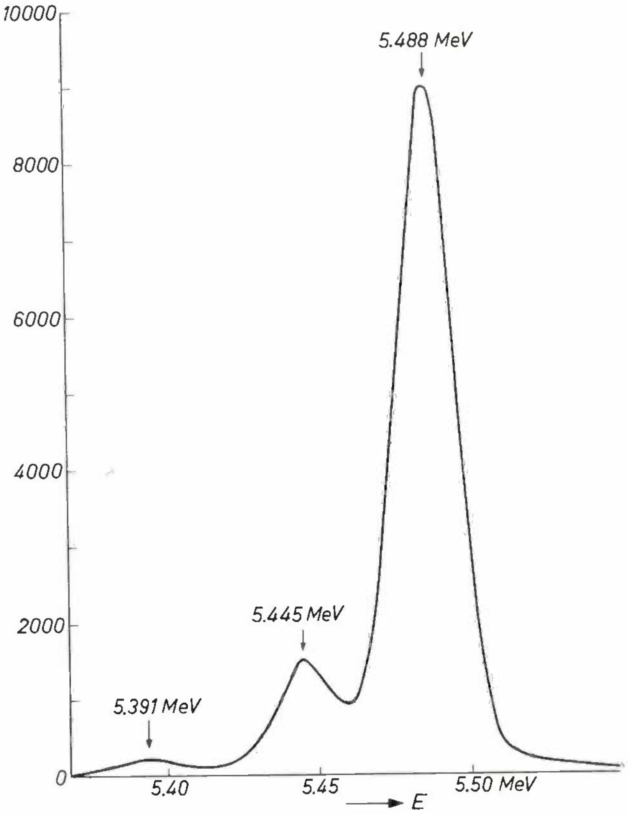


Fig. 3*a*. The α-spectrum of ^{241}Am recorded with a silicon detector for α-particles. The particle energy E is shown on the horizontal axis, and the number of pulses per energy interval of 0.25 keV is shown on the vertical axis. The width of the strongest line is about 20 keV at half height. Although the energy separation between the lines is about 45 keV (less than 1 % of the particle energy) it is easy to discriminate between lines.

which the average energy required for producing a pair of charge carriers is lower. Semiconductor counters owe their exceptionally high resolution principally to the very small average ionizing energy of silicon and germanium (3.6 and 2.9 eV) respectively, compared with about 30 eV for argon). For the detection of X-rays and γ-rays the resolution of the detector is in fact so good that even the best low-noise amplifiers that can now be built are barely good enough to do full justice to it. Some semiconductor counters do, however, have to be cooled, e.g. to liquid nitrogen temperature (77 °K), to limit the noise contribution from the bias current. *Fig. 3* shows α- and γ-spectra recorded with semiconductor counters, and for comparison a γ-spectrum of the same radioactive substance, recorded with a scintillation counter.

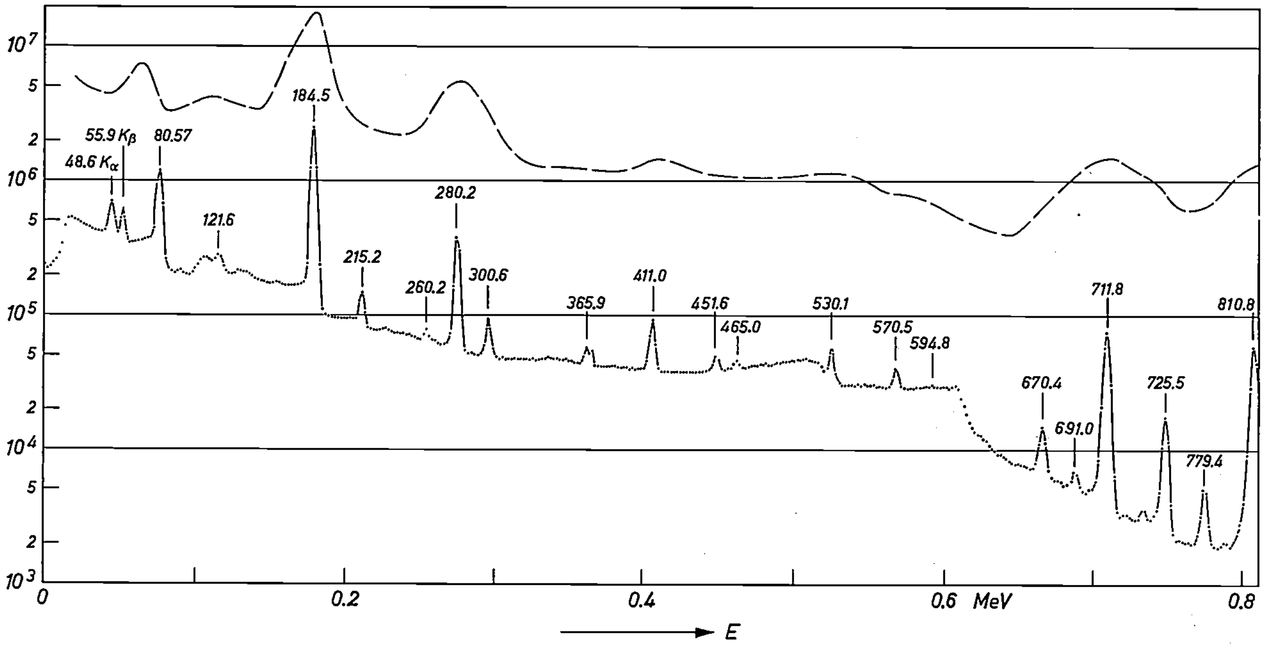


Fig. 3*b*. Part of the γ-spectrum of ^{166m}Ho, recorded with a germanium counter. Each point represents the number of pulses in an energy interval of 2 keV. A very large number of separate lines can be distinguished; the quantum energy for most of them is indicated separately. The dashed line indicates the spectrum found with a scintillation counter. (For clarity this is shown at about 100 times the measured value.)

The design of a semiconductor detector can readily be adapted to the nature and energy of the particles to be detected and to the purpose for which it is to be used. There are detectors for α-particles, for hard γ-radiation, for X-rays, and so on.

Another feature is that the single electrodes on each side of the semiconductor wafer can be replaced by a series of separate strips. If the strips on one side are aligned perpendicular to those on the other, the detector is divided into a large number of subdetectors, so that the place where a particle enters can be determined accurately [3] ("checker-board counter", *fig. 4*). This approach is of interest in nuclear physics and also in various branches of nuclear engineering.

A feature useful in some types of research is the fact that semiconductor detectors can be made so thin that the particles are able to pass through them with very little loss of energy. The detector is then used, of course, not to measure the energy of the particle but the energy loss per unit path-length. This quantity is of interest in establishing the identity of an unknown particle.

The last important advantage to be mentioned is the very small effect which variations in the supply voltage have on the output signal when a suitable circuit is employed; the scintillation counter and the proportional counter, on the other hand, require highly stabilized power supplies.

In the following sections of this article we shall deal first with the processes taking place in the detection medium. We shall then briefly consider the principal effects determining the resolution and discuss the merits of various counter configurations. Finally, the characteristic features of semiconductor counters will be compared with those of proportional counters and scintillation counters.

[1] W. C. Röntgen and A. Joffé, Ann. Physik **41**, 449, 1913 and **64**, 1, 1921.

[2] P. J. van Heerden, The crystal counter, thesis, Utrecht, 1945. See also P. J. van Heerden, Physica **16**, 505, 1950 and P. J. van Heerden and J. M. W. Milatz, Physica **16**, 517, 1950.

[3] This type of counter was developed in co-operation with a team of physicists from the Instituut voor Kernfysisch Onderzoek (Institute for Nuclear Physics Research), Amsterdam.

The ionization process in a semiconductor

When a high-energy charged particle or a quantum of radiation moves through a medium it dissipates energy in it. In the process, atoms of the medium may become ionized. In this section we shall examine more closely just how this takes place, particularly in semiconductors, and show that in a given substance the average energy w required for a single ionization is independent, in the first instance, of the nature of the radiation and of the energy of the particles (quanta).

A charged *heavy* particle (proton, α-particle), loses most of its energy in collisions with electrons, and a small part in collisions with atomic nuclei. The energy transferred to an electron upon a collision is relatively low owing to the differences of mass involved — for example a 6 MeV α-particle can transfer at the most only 3 keV to a stationary electron — but even so it is usually more than sufficient to free the electron from the atom. This is known as primary ionization. A much greater part of the ultimate ionization charge is due, however, to secondary ionization. This is largely brought about by the electrons freed by primary ionization. Some electrons are also freed by the Auger effect: when an electron freed from a deeper shell by the primary radiation is replaced by an electron from a peripheral shell, the released energy is not always emitted in the form of a quantum of radiation but is sometimes transferred to another electron, which then

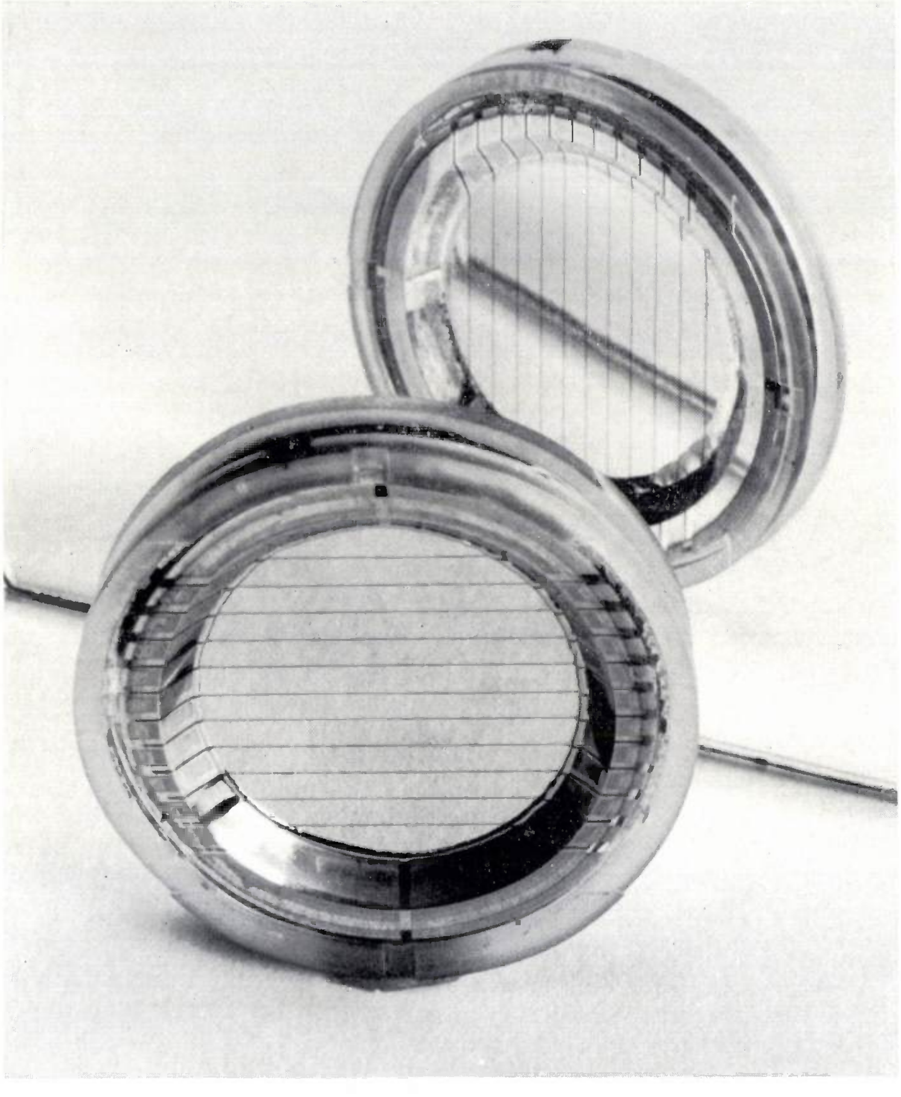


Fig. 4. Semiconductor detector for determining directional distribution ("checker-board counter"). This detector has a series of electrodes on opposite sides in the form of strips. The direction of the electrodes on one side is perpendicular to that of the electrodes on the other side (see mirror image; the line crossing this is the lower edge of the mirror, twice reflected).

leaves the atom. If we assume that the primary and secondary ionizations take place one after the other, we can represent the situation arising after each of the two phases in a semiconductor by an energy band diagram like that in *fig. 5*.

Ionization by γ-particles or accelerated electrons takes place basically in the same way. Owing to the equality of the masses involved, however, a single collision can result in a much greater part of the energy (up to 100%) being transferred to a stationary electron. A charged particle travelling at very high speed also loses some of its energy through the emission of electromagnetic radiation. The intensity of the radiation is proportional to the square of the atomic number of the medium. For electrons in germanium this effect only becomes of significance when the particle energy is greater than 10 MeV. For heavy particles it can be entirely disregarded, since it only becomes noticeable in the GeV region.

Because of the relatively high density of solids the range of particle radiation in them is very much smaller than in gases. For example, the range of 1 MeV β-particles in silicon is 2 mm, compared with 3.75 m in air; the range of 10 MeV α-particles — the highest energy particles emitted by the nuclides known at present — is only about 70 μm in silicon. A thin wafer of the detection medium can therefore be used; for α-detectors this need only be very thin.

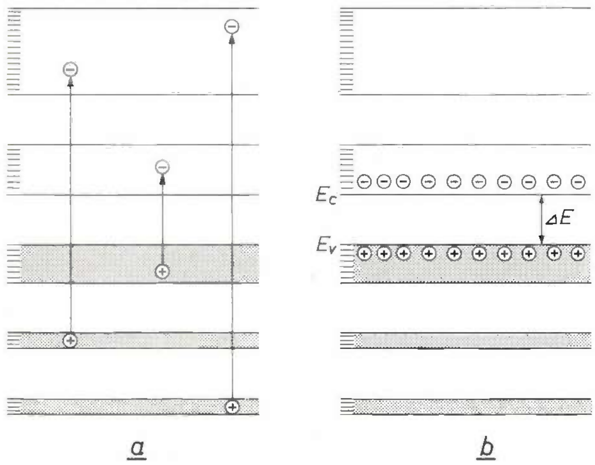


Fig. 5. *a*) Immediately after the incidence of a quantum or particle, there are freed charge carriers in the conduction and valence bands (*C* and *V*) as well as electrons in higher bands and holes in lower bands.
b) A short time later, all electrons are in the lowest levels of the conduction band (lower limit E_c) and the holes are in the highest levels of the valence band (upper limit E_v). In addition large numbers of secondary electrons and holes have been freed. The difference $E_c - E_v$ is the energy gap ΔE.

The ionization due to γ- or X-radiation differs from that due to charged particles in that the primary ionization is brought about by three different effects: the photoelectric effect, the Compton effect and pair formation .In the photoelectric effect the quantum energy E is imparted entirely to the freed electron. In pair formation an electron and a positron (positively charged electron) are created, which together receive the energy E — 1.02 MeV. In the Compton effect a variable part of the quantum energy is transferred to the electron. The probability of the occurrence of these processes is a function of E and of the atomic number Z of the ionization medium. The probability of the photoelectric effect, for instance, is proportional to $Z^5E^{-3.5}$. It thus increases sharply with increasing Z and decreases sharply as E is increased. The probability of pair formation is proportional to Z^2, and increases slightly with increasing E, initially in proportion to E — 1.02 MeV, but less at greater E. Because both effects increase with increasing Z, the relatively heavy germanium ($Z = 32$) is preferable to silicon ($Z = 14$) for a γ-detector. Gamma radiation penetrates deeper into the substance than particle radiation of the same energy. For energy measurements on γ-quanta of about 1 MeV the semiconductor layer must be several mm thick.

Relation between ionization charge and energy

In a fairly wide range of energies the ionization charge, within the accuracy of measurement, is proportional to the energy of the particles (quanta) and independent of the nature of the radiation. This independence is understandable when we consider that ionization is entirely due to electrons for γ-quanta and β-particles and very largely due to electrons for heavy particles.

The average energy w required for a single ionization is, as we have seen, much lower in solid-state detectors than in gases, but it is still considerably greater than the energy gap ΔE. For germanium $w = 2.9$ eV, against $\Delta E = 0.65$ eV; in silicon $w = 3.6$ eV, against $\Delta E = 1.1$ eV. There are two reasons for this inequality: 1) part of the energy of the freed electrons is used for lattice vibrations and 2) electrons whose energy has decreased to less than ΔE can no longer cause ionization, so that this residual energy plays no part in the ionization process. In any case, then, w will be greater than ΔE.

For heavy particles there is in theory some deviation from linearity in the relation between E and the ionization charge, because these particles can no longer cause ionization when E has dropped to a value which is still well above ΔE. For α-particles in silicon this threshold value is about 1 keV. Although this is a large multiple of ΔE, it is nevertheless such a small fraction of the normal initial value of E that the deviation is not detectable in the experimental results. If the ionization medium is a gas, this threshold value may be very much higher, and a marked deviation is indeed found for heavy particles of not unduly high energy [4].

From ionization charge to pulse

We shall now examine what happens when the applied field causes the ionization charge to move towards the electrodes. This determines the form of the pulse observed in the external circuit in the detection of a particle. We shall assume to begin with that none of the ionization charge is lost on the way, that the detection medium has a large energy gap (so that, in the natural state, there are virtually no free charge carriers in the medium), and that no charge carriers can penetrate into the medium from the electrodes.

In every ionization event *two* charge carriers of opposite sign are produced. When an electric field is applied (due to a voltage V_0) the freed positive and negative charge carriers (each with a total charge Q_i) move away from one another and a charge is induced in each of the two electrodes *1* and *2* (fig. 2). If the two kinds of charge carrier have become separated by a distance equivalent to a potential drop of ΔV, then the induced charge [5] has a magnitude of $Q_i \Delta V/V_0$. The voltage across the resistor R (fig. 2) at that moment is $(Q_i/C)(\Delta V/V_0)$, where C is the capacitance of the electrode *2* with respect to earth. Now if the time constant RC is large compared with the time in which the ionization charge is collected at the electrodes, then the final value of the voltage across R — i.e. the height of the voltage pulse — is equal to Q_i/C, hence proportional to Q_i and independent of V_0. (If the charge collection time is *not* short with respect to RC, then the proportionality with Q_i remains but the final value is lower than Q_i/C and no longer independent of V_0.)

The above picture of a semiconductor detector is, of course, over-simplified: some charge carriers from the electrodes will always penetrate into the medium, and even without this the medium in the natural state contains a certain number of free charge carriers.

In counters for particle and gamma spectrometry, which should therefore have a very high energy resolution, these effects must be suppressed as far as possible. In counters, on the other hand, which are only required to measure the average intensity of a stream of particles (quanta), such as counters for dosimetry, there may be advantages in a ready flow of charge carriers from the electrodes (injection contact), since the sensitivity can be increased by this [6]. Counters with injection contacts are called conduction counters, and the others are known as barrier-layer counters. In the following we shall be concerned purely with barrier-layer types, unless otherwise stated.

A complication may arise from the effect in which clouds of positive and negative charge carriers, as soon as they are separated by the electric field, are in principle neutralized by charge carriers of opposite sign which come from the surroundings and from the electrodes. This neutralization varies exponentially with time. The time constant $\tau_{\rm rel}$, the dielectric relaxation time, is equal to ε/σ, where ε is the dielectric constant and σ the conductivity of the material. The above treatment of the charge collection applies, strictly speaking, only for $\tau_{\rm rel} = \infty$. In detectors in which the natural material contains hardly any charge carriers, and where none can come from the contacts, the neutralization is virtually negligible.

Life of charge carriers and pulse height

In the foregoing we have assumed that all charge carriers freed by the incident radiation do in fact reach the electrodes. In reality a number of them will not, but will be trapped on the way and may even recombine. Obviously, there will be less probability of their recombining if the travelling time of the charge carrier — maximum value equal to the carrier transit time — is short compared with its average life. For a group of charge carriers that has to cover the whole distance d between the electrodes it has been calculated that a fraction $[1 - \{1 - \exp(-d/\mu\tau F)\}\mu\tau F/d]$ are lost. In this expression F is the electrical field strength, assuming the field to be uniform, and μ the mobility of the charge carriers. At a given d and F the product $\mu\tau$ therefore determines the magnitude of the loss. It should be remembered here that $\mu\tau$ does not in general have the same value for holes and electrons. The greatest signal loss is found when the particle to be detected is incident at a place such that the charge carriers with the lowest $\mu\tau$ value have to travel through the whole crystal.

Since the signal loss depends on the place of incidence, pulses produced by particles of identical E do not all have the same height. For a good energy resolution, therefore, the signal loss must be kept relatively small. It follows from the formula just given that we must make $\mu\tau F$ large with respect to d, i.e. d/F small with respect to $\mu\tau$. (If this condition is satisfied, the expression for the relative loss approximates very closely to $d/2\mu\tau F$.)

If we now reduce the condition above to $d/\mu F \ll \tau$ we see at once the relation to the qualitative approach at the beginning: since the velocity of a charge carrier is μF, $d/\mu F$ is the carrier transit time, and therefore, taking a value for μ corresponding to that of the slowest charge carriers, $d/\mu F$ is the maximum value $t_{\rm c\,max}$ of the charge collection time. The condition for a relatively small signal loss is simply $t_{\rm c\,max} \ll \tau$.

Pulse rise-time

If we wish the detector to have a very high resolution in time and if we require it to be able to handle a high count-rate without the pulse height being seriously affected by the superposition of pulses, we must obviously try to obtain the shortest possible pulses. It is comparatively easy to reduce the pulse decay time: this can be done electronically, e.g. by differentiation. The rise time, however, is entirely determined by one of the processes that takes place in the detector itself: the collection of the charge. The ionization process usually takes place much faster.

We have seen that the collection time is at the most $d/\mu F$ (or $d^2/\mu V_0$). Depending on the design of the detector, values are found between 10^{-5} and 10^{-8}. On the other hand, the duration of the primary ionization process — i.e. the time that elapses between the entry of the particle and the moment at which it causes no further ionization — is only about 2.5×10^{-12} s for a 5 MeV α-particle in silicon, and the duration of the secondary ionization process, which may be similarly defined, is of the same order of magnitude.

To obtain a short rise time it is therefore necessary to impose on the maximum charge collection time $t_{\mathrm{c\,max}}$ a condition similar to that laid down for obtaining a small signal loss, the difference here being that the *absolute* value of $t_{\mathrm{c\,max}}$ must be small. This is the case if d is small and if μ and V_0 are large. The requirement for large V_0 indicates that the breakdown voltage of the counter should be as high as possible.

In the above we have tacitly assumed that the ionization charges are not dense enough to seriously affect the field in the detection medium by polarization. With heavy particles, however, which produce a very dense charge cloud, polarization may indeed occur (*fig. 6*). Inside the cloud the field strength F is then initially very small, and does not assume an appreciable value until the cloud has thinned out due to ambipolar diffusion of the charge carriers — at right angles to the direction of the field — and the loss of peripheral charge carriers.

It is evident that in such a case the collection of the ionization charge will not be so fast. For a 5 MeV α-particle, incident perpendicular to the field direction, and a field strength of 1000 V/cm we have calculated that the collection time t_{c} can be lengthened at the most by 0.3 μs. Usually, however, this time is much less, so that the effect is not particularly serious.

It should not be concluded from the above that the recombination charge loss associated with the delay is of no significance. In a dense charge cloud — a region therefore where the concentrations of both kinds of charge carrier are very high — the probability of recombination is much greater, and hence τ is much smaller than in a rarefied cloud. The recombination loss for heavy particles is therefore already greater than for light ones, so that a longer collection time is more serious here. Since the magnitude of the recombination loss will differ from case to case, the polarization effect adversely affects the resolution. This is one of the reasons why the resolution for α-particles is not as good as for β-particles of the same energy.

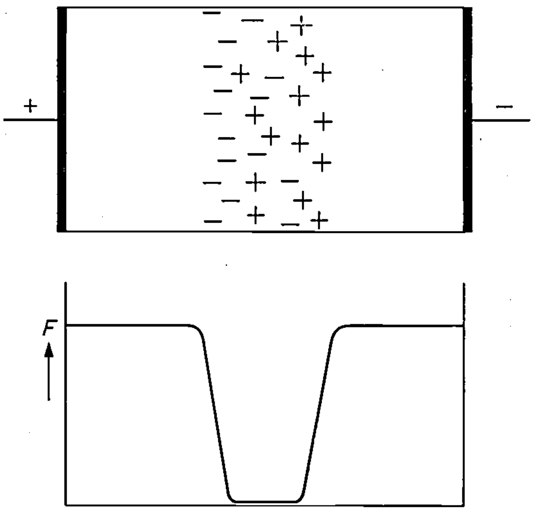


Fig. 6. If the charge-cloud is extremely dense, as it is with very heavy particles, polarization occurs as soon as the negative and positive charge carriers move apart. The field strength F is then practically zero inside the cloud; see the field variation shown below.

Energy resolution

In the preceding section we have already indicated that the resolution is limited when the collected charge, and hence also the height of the output pulse, depends to some extent on the place where the particle is incident. In this section we shall consider two other causes of fluctuations in the pulse-heights, namely phenomena which occur in the detector itself and moreover directly depend on the configuration of the detector (dimensions, nature of contacts) as well as on the nature of the detector material [7]. These are: 1) fluctuations in the ionization charge, and 2) fluctuations in the bias current through the detector, the "dark current".

The fluctuations in the ionization charge come about because only a part of the energy of an incident particle is used for ionization processes, the rest of the energy being used to set up excitations such as lattice vibrations. Since excitations are probability processes, the ionization charge produced by particles of identical energy will show statistical fluctuations. If the average ionization charge of a number of mono-energetic parti-

[4] See the article by H. W. Fulbright in Hb. Physik XLV (Springer, Berlin 1958; ed. S. Flügge), in particular fig. 5 on page 12.

[5] For the mathematical treatment, see D. H. Wilkinson, Ionization chambers and counters, Cambridge Univ. Press 1950.

[6] See L. Heijne, Physical principles of photoconductivity, Philips tech. Rev. **25**, 120-131, 1963/64, in particular page 129.

[7] For a treatment of noise in a detector-amplifier combination, see: J. A. W. van der Does de Bye, Signal-to-noise ratio of a *P-N*-junction radiation counter, Philips Res. Repts. **16**, 85-95, 1961.

cles is expressed by the number of pairs of freed charge carriers N_0, then the standard deviation δN_0 in this number is equal to $\sqrt{F^* N_0}$. The relative standard deviation is thus given by:

$$\frac{\delta N_0}{N_0} = \sqrt{\frac{F^*}{N_0}} = \sqrt{\frac{F^* w}{E}} . \quad . \quad . \quad (1)$$

In these expressions F^* is a constant, called the Fano factor [8], which depends on the material and on the type of radiation; w and E are again the average ionization energy and the energy of the incident particle, so that $N_0 = E/w$. One of the advantages mentioned in the introduction for the semiconductor detector can readily be derived from eq. (1): there is a relatively small spread in N_0 because the average ionization energy is low. Moreover, measurements have shown [9] that the Fano factor for semiconductors is also relatively small ($F = 0.1$ to 0.15 compared with 0.2 to 0.3 for gas ionization chambers). Using eq. (1) to calculate the relative standard deviation in the ionization charge of particles of 5 MeV in silicon, we find 3×10^{-4}, which corresponds to 1.5 keV. Calculating from this the half-height width of the pulse-height distribution — this width is 2.36 times as large as $\delta N_0/N_0$ and is often used as a measure of the resolution — we find 3.5 keV.

As in other semiconductor devices, in nuclear radiation detectors the *bias current* is a factor that calls for careful attention with regard to noise. The noise in the bias current has almost entirely the character of shot noise. This is obvious for the part of the bias current originating from the contacts. In the part due to thermal generation of charge carriers in the detector material, the generated charge carriers disappear very quickly from the detector volume because of the effect of the electric field, so that scarcely any recombination takes place, the transist time being short compared with the average life of the carriers. There is consequently no generation-recombination noise, but only a noise contribution from the generation process separately. This noise however also has the character of shot noise.

Another point to be noted about the current noise is that noise-smoothing due to space charge — as in thermionic valves — hardly occurs at all because the concentration of charge carriers is too low.

These considerations indicate that in calculating the output signal fluctuations which result from the bias-current noise we may start with the relationship [10]:

$$i^2 = 2eI\mathrm{d}f. \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

Here i is the effective value of the shot noise in the bias current in the frequency band $\mathrm{d}f$, e is the absolute charge of the electron and I the bias current in the detector. It is found that these fluctuations are relatively small when $\sqrt{N}$, the square root of the number of free charge carriers in the counter, is small compared with the ionization charge N_0. What this means in practice can best be seen from an example. An α- or β-particle of 5 MeV produces 1.5×10^6 pairs of charge carriers in silicon. If we now require a signal-to-noise ratio of at least 1000 : 1, then N must not be greater than 5×10^6. Now in extremely pure silicon at room temperature the concentration of charge carriers is already between 10^{11} and 10^{12} cm^{-3}. In a counter of not excessively small dimensions this means that N would be much too great if no special measures were taken to reduce it. In the following section we shall consider what measures can be taken.

Detector configuration

Semiconductor detectors can be divided into two basic groups with different electrical configurations: detectors with injection contacts — conduction counters — and detectors with blocking contacts — barrier-layer counters. This classification has already briefly been mentioned. The nature of the contacts affects nearly all the questions relating to the type and operation of the detectors, one of the most important being the manner in which the number of free charge carriers N is reduced to improve the signal-to-noise ratio. We shall therefore begin with a brief account of both kinds of detector, commenting briefly on the requirements to be met in both cases by the semiconducting material.

We shall then leave the subject of conduction counters, which have been very little used up till now, and devote the rest of this section to the different types of barrierlayer counter, and to the question of which type is most suitable for a given form of radiation.

Since conduction counters have injection contacts, the concentrations n and p of the free electrons and holes do not depend on whether or not an electric field is present: charge carriers leaving the detection medium under the effect of such a field are immediately replaced by others entering through the other contact. The values of n and p, and hence N itself, cannot therefore be affected by means of the field. In these detectors the bias current depends entirely on the nature of the detector material and not at all on the contacts.

In barrier-layer counters, on the other hand, N is certainly affected by the applied voltage, because charge carriers drawn away from the medium are only replaced to a very limited extent. Where signal-to-noise ratio is concerned a barrier-layer counter is therefore in principle preferable to a conduction counter made of the same material.

In order for the field in a counter of any kind to be reasonably uniform, the space charge should be nearly equal to zero everywhere inside it. In conduction counters this happens automatically because free charge carriers can flow unimpeded to neutralize bound charges. In barrier-layer counters, however, the space charge is zero only when the detection medium contains no bound charges, that is to say when it is an *intrinsic* semiconductor, or when the positive and negative bound charges are equal and thus neutralize each other. (In the last case, also, a semiconductor is often said to be "intrinsic", because the holes and electrons have the same concentration, so that the Fermi level also lies half-way up the energy gap.)

The materials suitable for use as the detection medium are at present few in number: for barrier-layer counters the choice is practically limited to germanium and silicon. The usefulness of these two substances is due to the following circumstances:

1) The charge carriers have a long life (0.1 to 1 ms) and at the same time a high mobility. The product $\mu\tau$ is therefore so large that, without any great loss, the ionization charge can be collected in a layer as thin as 1 cm. In other semiconductors the carrier life is generally much shorter (10^{-7} to 10^{-8} s).
2) Considerable advances have been made in the production of large single crystals of well defined and readily processed germanium and silicon.

For conduction counters, however, silicon and germanium are unfortunately not so suitable, for reasons that will presently appear. First of all, we shall consider the conduction counter in somewhat more detail, and then we shall confine our attention to the barrier-layer counters.

Conduction counters

In a conduction counter the concentrations n and p of the negative and positive charge carriers respectively are equal to those present in the medium in the absence of an applied voltage. The well-known equation.

$$n\,p \propto \exp\left(-\Delta E/kT\right). \quad . \; . \; . \; . \tag{3}$$

then applies. The product np is usually represented by $n_i{}^2$. If bound charges are present, n and p must also meet the neutrality condition:

$$n + N_a = p + N_d. \quad . \; . \; . \; . \; . \tag{4}$$

Here N_a and N_d are respectively the concentrations of the singly charged negative and positive impurity centres.

It follows from eq. (3) that the sum $n + p$ is smallest if $n = p$, i.e. if $N_a = N_d$. Quantitatively, at room temperature $n + p$ is smaller than 10^6 cm^{-3} — for a counter of 1 cm^3, i.e. a counter in which $N = n + p$, this corresponds approximately to the requirement for a signal-to-noise ratio of 1000 : 1 — only if the energy gap is larger than 1.65 eV and the difference between N_a and N_d is smaller than 10^6. Materials which satisfy this condition, and in which the charge carriers have sufficiently long life, are not yet available. For the present the only choice is to make do with materials of smaller energy gap, but this implies that cooling must be applied in order to reduce $n_i{}^2$. In the hypothetical case in which $N_a - N_d = 0$, Si would have to be cooled to —75 °C, and Ge to —140 °C.

Even the purest germanium and silicon monocrystals that can be made at present do not at this temperature satisfy the condition that $N_a - N_d$ should be less than 10^6 in a not too small volume. Further measures therefore have to be taken, such as:

1) reducing the temperature much further than would be necessary in the intrinsic case. A greater number of impurity centres will then change to the uncharged state. The extent to which it is necessary to cool the material in order to achieve a significant reduction of N_a and N_d depends, of course, on the location of the impurity levels. In silicon, for example, the boron atoms present as impurities give rise to impurity levels that are only 0.045 eV above the conduction band, and this necessitates cooling to between 10° and 20 °K.
2) reducing the difference between N_a and N_d means of appropriate doping. Use can then best be made of substances that have impurity centres lying approximately at the centre of the forbidden band, i.e. at the height where the Fermi level should be. The quantity of dope used is then less critical. A difficulty with this method is that impurity levels of this kind constitute effective recombination centres, and thus shorten the life of the carriers.

Earlier in this section it was stated that intrinsic material (type I) should be used for a barrier-layer counter to obtain uniformity of the field. It is clear from the above that the same requirement holds in principle for conduction counters, but now for the purpose of obtaining a minimum noise level.

A familiar example of the method mentioned under 2) is the doping of N-type silicon with gold, producing electron traps which are located 0.54 eV below the conduction band (ΔE = 1.1 eV). If the gold concentration has three times the value that N_d previously had, then $n \approx p$ in a wide temperature range [11]. Counters that work reasonably well have been made with a material of this kind [12].

[8] U. Fano, Phys. Rev. **72**, 26, 1947.

[9] R. L. Heath, W. W. Black and J. E. Cline, IEEE Trans. nucl. Sci. **NS-13**, No. 3, 445, 1966.

[10] See A. van der Ziel, Fluctuation phenomena in semi-conductors, Butterworths, London 1959.

[11] C. B. Collins, R. O. Carlson and C. J. Gallagher, Phys. Rev. **105**, 1168, 1957.

[12] J. D. van Putten and J. C. Vander Velde, IRE Trans. nucl. Sci. **NS-8**, No. 1, 124, 1961.

Barrier-layer counters

In barrier-layer counters the charge carrier concentration is mainly determined by the contacts and the applied voltage, and to a much lesser extent by the nature of the material. The requirement that intrinsic material should preferably be used is not in the first place due to noise considerations. The significance of the energy gap will be discussed presently.

An important question here is how to obtain contacts that have an adequate blocking action. As a rule, this can be achieved by doping. In our counters the anode is generally a semiconducting surface layer of strongly *N*-type material, and the cathode a thin layer of strongly *P*-type material. In the latter material there are very few electrons in the conduction band, and therefore the *P* contact cannot supply a large number of electrons in a short time. The same applies to the *N* contact with respect to the holes. *Fig. 7* shows the band diagram of a *P-I-N* system of this type when a voltage has been applied to it.

To get some idea of the charge-carrier concentration in the *I*-region when a voltage has been applied to the counter, one must first realize that the free charges in this region partly originate from the electrodes and partly from thermal generation. In practice sufficiently good blocking contacts can be obtained to make the contribution from the electrodes negligible, leaving only the thermal generation to be reckoned with. The electron carrier concentration resulting from this depends on the speed at which the freed charge carriers leave the detection medium under the effect of the electric field.

In our case thermal generation via an impurity level predominates — i.e. generation in two steps. If the impurity levels lie at the centre of the forbidden zone — the most unfavourable case — the number of pairs of charge carriers generated per cm^3 is $n_i/2\tau$, where τ is again the life of the carriers. The number of pairs G generated in the whole counter volume is thus $n_iAd/2\tau$, where A is the surface area. On the other hand, the average time that an electron remains in the detector is equal to $d/2\mu_nF$. For a hole this time is $d/2\mu_pF$. The number of charge carriers is now equal to G times this average time, so that, with F again put equal to V_0/d, we find for the sum of the number of holes and electrons:

$$N = \frac{d^3An_i}{4V_0\tau}\left(\frac{1}{\mu_n}+\frac{1}{\mu_p}\right). \quad . \; . \; . \; (5)$$

Applying this to a silicon counter ($n_i = 10^{10}$ cm^{-3} and $\tau = 10^{-3}$ s) with a thickness d of 5 mm, a surface area A of 1 cm^2 and an applied voltage V_0 of 500 V, we find $N = 1.7\times10^6$, which is an acceptable value. In germanium, because of the smaller energy gap, n_i is higher (10^{13} cm^{-3}) and in the conditions mentioned N is higher than desirable. To make N sufficiently low the material therefore has to be cooled.

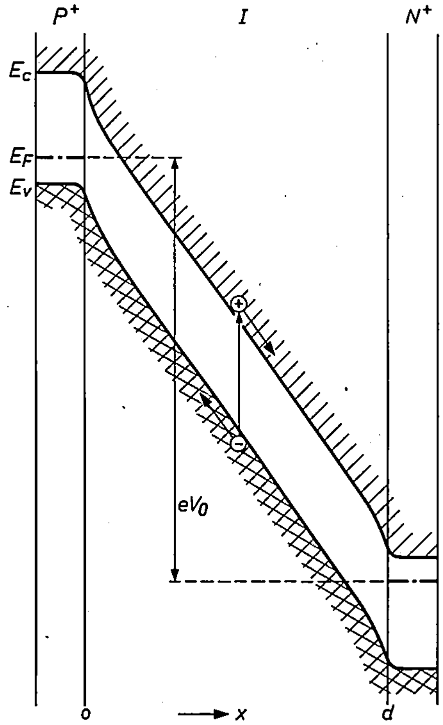


Fig. 7. Simplified energy-band diagram (potential variation in the x-direction) for a counter with *P-I-N* configuration with an external voltage V_0. (In reality eV_0 is very much greater than $E_c - E_v$.) The chain-dotted line in the contacts indicates the location of the Fermi level E_F. If the *I* region is completely free of impurities, E_c and E_v vary almost linearly with x in this region.

We shall now consider the situation when the detector material is not intrinsic but to a certain extent *P*-type or *N*-type. *Fig. 8* illustrates the case where the detection medium is weakly *P*-type. The counter configuration has now in fact degenerated to a *P-N* junction. The volume within which the field is reasonably strong and in which the ionization charge can thus be collected (the effective volume) is reduced to the barrier layer. Although charge carriers freed outside the barrier layer have a chance to diffuse towards the barrier layer and make their contribution there, this contribution is rather small, for one reason because of the relative slowness of a diffusion process.

It should not be concluded from the above that the sensitive layer of counters with a *P-N* structure is necessarily extremely thin. From the equation for the width B of a depletion layer

$$B = \sqrt{\frac{2\varepsilon(V_d+V_0)}{eN_a}}, \quad . \; . \; . \; . \; (6)$$

a depletion layer width of 0.5 mm is found for extremely pure silicon with a concentration N_a of 10^{12} cm^{-3}, at an applied voltage of 200 V. (Equation (6) is appli-

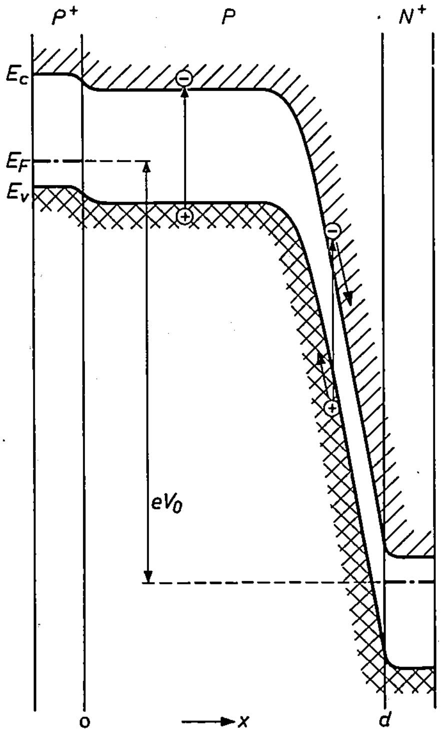


Fig. 8. As in fig. 7, but now for the case where the detection medium is not intrinsic but to a certain extent P-type. The configuration has virtually degenerated into a P-N junction, and the detection medium has a sufficiently strong field only in the layer immediately bordering on the N-contact.

cable if the concentration of impurity centres in the N-region is much greater than that in the P-region. The diffusion voltage V_d is negligible compared with 200 V.)

The last configuration to the mentioned is the one in which the depletion layer of a metal-semiconductor contact is used instead of the depletion layer of a P-N junction. The metal generally used is gold.

Method of making a P-I-N configuration

To make a P-I-N configuration [13] we start from a single-crystal wafer of fairly pure P-type germanium or silicon (resistivity about 500 Ω cm). Lithium is thermally diffused at a few hundred degrees centigrade to a depth of about 0.2 mm on one side of the material. Lithium acts in this as a donor. At the surface, where the lithium concentration is high, the material thus changes into an N-type semiconductor (*fig. 9a*).

Since the lithium ions occupy interstitial sites in the lattice, they are to some extent mobile. When a reverse voltage is applied to the resultant P-N junction, at a temperature of say 125 °C, the lithium ions in the region around the P-N junction, where the field strength is relatively high, begin to move towards the P-region. After a certain time the situation shown in fig. 9*b* is obtained: a large I region has been produced. The reason for this is briefly as follows. As long as the acceptors present are not exactly compensated by lithium, the field in the material is not homogeneous, and the lithium ion current varies from place to place in the crystal in such a way that the differences in concentration are levelled out.

This attractive method is unfortunately not applicable for the preparation of I material for conduction counters: after obtaining the P-I-N configuration the P and N layers cannot be removed. These layers are not only necessary for the production of the I layer; they are also required for its maintenance, at least when there is an electric field.

Some practical examples of barrier layer counters

Fig. 10 shows a diagrammatic representation of a semiconductor detector for α-particles. The material used is silicon. Since the range of α-particles is relatively short, a P-N junction or a blocking metal-semiconductor contact with a depletion layer width of 0.1 to 0.2 mm is sufficient in α-counters. The particles are incident through one of the electrodes, which must be very thin in order to minimize energy losses. The spec-

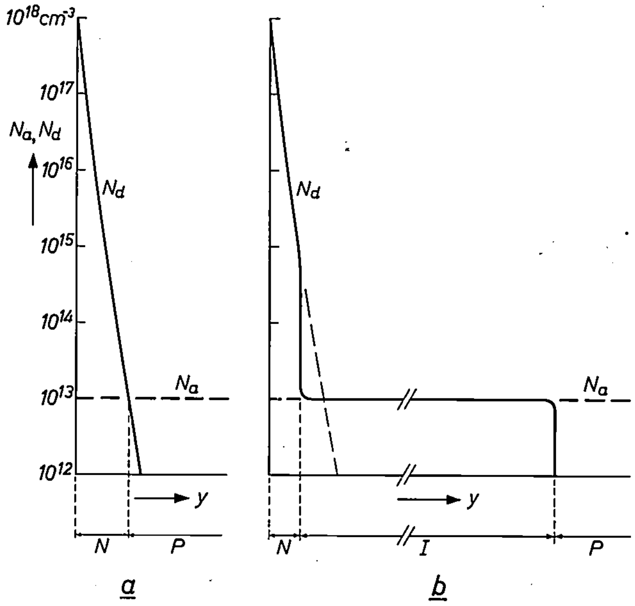


Fig. 9. Illustrating how, starting with a wafer of P-type material, a P-I-N configuration is obtained by addition of lithium, which functions as a donor in silicon and germanium.
a) Variation of the lithium concentration N_d with the distance y to the surface, after the lithium has been thermally diffused into the crystal wafer. The concentration N_a of the acceptors is independent of y. In a thin layer at the surface, N_d is greater than N_a and the material has changed from P-type to N-type. (The vertical scale is only very approximate.)
b) Variation of N_d with y after some of the lithium ions have penetrated deeper into the wafer under the effect of an electric field and at elevated temperature. In a wide region around the P-N junction in (*a*) N_d has become equal to N_a and the material is therefore intrinsic. If the lithium supply is adequate the extent of this region increases with the time the process lasts.

[13] This method is due to E. M. Pell, J. appl. Phys. **31**, 291, 1960.

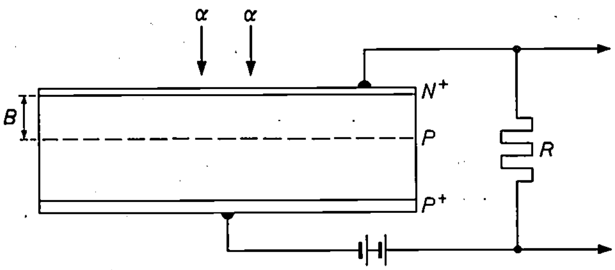


Fig. 10. Configuration of an α-detector, consisting of a *P-N* junction with a relatively wide barrier layer (width *B*).

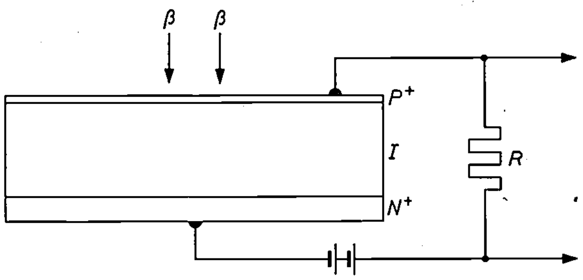


Fig. 11. In detectors for energy measurement on γ-quanta or long-range particles (high-energy β-particles) a *P-N* junction is not sufficient; such detectors should have a *P-I-N* configuration.

trum shown in fig. 3*a* was recorded with a counter of this type. A photograph of the counter can be seen in the lower photograph of fig. 1.

Because of the much greater range of high-energy β-particles, a *P-N* junction is not adequate for measuring their energy, and in this case a counter of *P-I-N* structure with an *I* layer a few millimetres thick (*fig. 11*) should be used.

Counters with *P-I-N* structure are also needed for the detection of γ-radiation. In this case, as mentioned above, germanium is to be strongly preferred because of its higher atomic number. The spectum shown in fig. 3*b* was recorded with a germanium *P-I-N* detector.

As γ-quanta can penetrate a relatively long way into light materials without any interaction taking place, γ-counters can be completely encapsulated and thus protected from atmospheric effects.

As mentioned in the introduction, extremely thin semiconductor detectors are employed in particle-identification systems. In systems of this kind separate detectors are used for measuring the energy E of a particle and the energy loss per unit path length, $\mathrm{d}E/\mathrm{d}x$. For identifying the nature of the particle, use is made of the fact that the product $E\mathrm{d}E/\mathrm{d}x$ is proportional to mZ^2 and largely independent of E (*fig. 12*). For protons, deuterons, tritons and α-particles, for example, the ratios of these products are 1 : 2 : 3 : 16. The energy loss per unit path length is measured with the thin semiconductor detector mentioned above, in which the particle loses relatively little energy (upper photograph fig. 1). In fact, of course, $\mathrm{d}E/\mathrm{d}x$ is not measured, but $\delta E/d$, where δE is the energy loss. When the particle has passed through this detector it is stopped in the other, enabling E to be measured.

The thickness of the $\mathrm{d}E/\mathrm{d}x$ detector is chosen between 25 to 250 μm depending on the nature of the experiment. A *P-N* detector with a high enough applied voltage to make the depletion layer extend to the other side of the layer forms a particularly suitable $\mathrm{d}E/\mathrm{d}x$ detector. For E measurements a silicon *P-I-N* detector is generally employed.

Apart from the electrical structure of the detector — *P-N* or *P-I-N* — the geometrical structure can also be varied in many ways. We have already mentioned the "checker-board counter", which can be used to perform very accurate directional measurements in a short time (fig. 4). With these detectors, which are divided into some 40 subdetectors per cm², a resolution of about 1° can be achieved in a directional measurement with a distance of only 8 cm between source and detector. It is also possible to design $\mathrm{d}E/\mathrm{d}x$ detectors as checker-board counters if required [14]. There

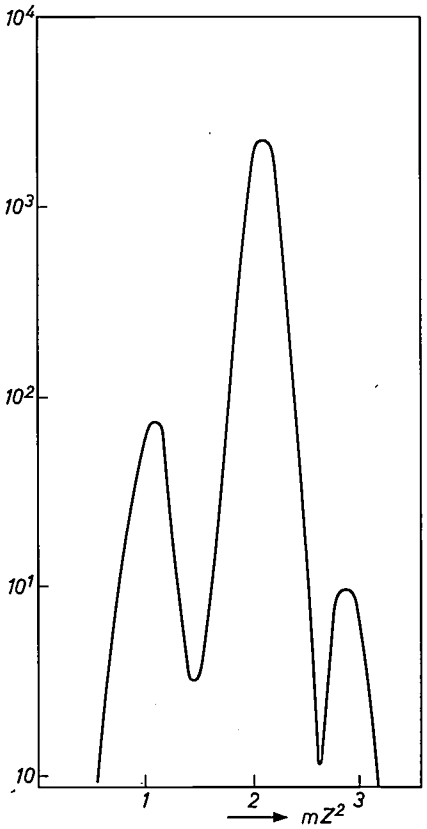


Fig. 12. Spectrum of mZ^2 values (m = nuclear mass and Z = atomic number) obtained in measurements made with a particle-identification system on a beam containing protons, deuterons and tritons; we have taken $mZ^2 = 1$ for the proton. The value of mZ^2 is calculated for each incident particle from the product of the particle energy E and the energy loss per unit path length $\mathrm{d}E/\mathrm{d}x$. These two quantities are measured with separate detectors.

are other detectors in which a beam of particles is passed through a central hole; this arrangement can be used to detect particles scattered backwards from a target situated some distance further on.

Comparison with proportional counters and scintillation counters

Linearity and energy resolution

We have seen that the pulse height in semiconductor counters is proportional to the particle energy and independent of the type of radiation, except for very heavy particles such as fission products. Proportional counters are also highly linear, but the proportionality factor for these is not entirely independent of the type of radiation. Scintillation counters are linear for γ-radiation, but not for α particles, for example, and the type of radiation here also has some effect.

The resolution of semiconductor counters in the detection of β- and γ-radiation is largely determined by the noise and is therefore to some extent independent of the energy. At an energy greater than about 0.5 MeV statistical fluctuations in the ionization charge begin to become significant and the line width increases in proportion to $\sqrt{E}$. In proportional counters and scintillation counters the resolution is primarily determined by fluctuations in the multiplication mechanism; these fluctuations are also proportional to $\sqrt{E}$, but the proportionality factor is considerably greater.

By way of illustration to the foregoing, *fig. 13* shows the resolution, for X-rays and γ-rays, of a semiconductor counter, a proportional counter and a scintillation counter as functions of $\sqrt{E}$. In this figure the line width at half height is again used as a measure of the resolution. The resolution of the semiconductor counters is seen to be better than that of the other two except for soft radiation, where the amplifier noise predominates.

It may be recalled here that the average ionization energy w for germanium and silicon is only a few electron-volts (2.9 and 3.6 eV respectively) and the Fano factor 0.1 to 0.15 as against $w \approx 30$ eV and $F^* \approx 0.2$ to 0.3 for gases. A direct comparison with the scintillation counter is of course not possible, as this does not measure the ionization charge; it is however possible to characterize the relative spread in the output signals by fictitious values for w and F^*. These are very high: 300 eV and 1 respectively.

In the detection of α-radiation with semiconductor counters the resolution is partly determined by the fact that a relatively large number of recombinations take place in the ionization track owing to the high carrier densities involved; their number fluctuates, largely because of the inhomogeneous distribution of the recombination centres. The resolution here is about 20 keV. Although for α-radiation the resolution is therefore rather worse than for β- or γ-radiation, it is still three times better than that of the detector previously used for α spectroscopy, the ionization chamber. This advantage is due to the lower ionization energy, which results in a better signal-to-noise ratio.

Pulse rise-time

The rise time of the pulses obtained with semiconductor counters depends on the thickness of the sensitive layer, on the applied voltage and on the temperature, and lies between 10^{-8} and 10^{-5} s. Scintillation counters are in general somewhat faster: depending on the nature of the crystal and on the photomultiplier tube, rise times in these counters are between 10^{-9} and 10^{-6} s. Proportional counters have a rise time of about 10^{-6} s; they are therefore generally somewhat slower than semiconductor counters. It should be remembered here that the rise time of a proportional counter is not a measure of the maximum count rate that can be handled; this is appreciably lower than would follow from the rise time because, after very pulse, it takes some time before the generated ions have disappeared again. Ionization chambers are very much slower: these have a rise time of not less than about 10^{-3} s, owing to the low mobility of positive and negative gas ions.

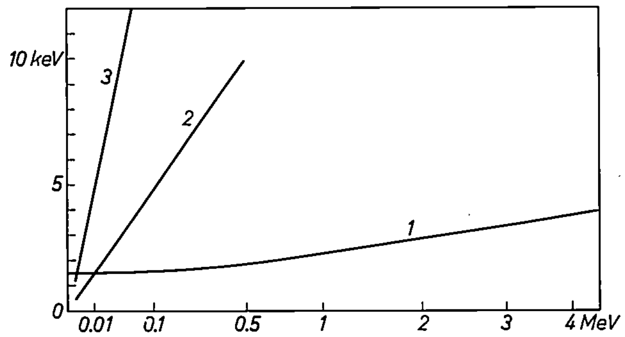


Fig. 13. Comparison of the X-ray and γ-ray resolution of semiconductor counters (curve *1*), proportional counters (curve *2*) and scintillation counters (curve *3*). The half-height width of the pulse-height distribution obtained with a good counter for monochromatic radiation is plotted against the square root of the quantum energy *E*. Except for very small values of *E*, the semiconductor counters are superior. The horizontal part of curve *1* represents amplifier noise.

Other features

In solids the range of charge particles is much shorter than in gases and the absorption of γ-rays is much greater. These effects enable semiconductor counters to be very much smaller than proportional counters and

[14] W. K. Hofker et al., IEEE Trans. nucl. Sci. NS-13, No. 3, 208, 1966.

yet still be useful for a wide range of energies. Complete semiconductor counters are even smaller than scintillation counters, which have to be used in conjunction with a light-guide and photomultiplier tube.

The small dimensions of semiconductor counters are a great advantage in various applications, such as in medical and biological experiments. On the other hand, for measurements on very large specimens, or on specimens that cannot be situated close to the detector, the small size is a disadvantage, because much of the emitted radiation is then no longer incident on the detector. This disadvantage can usually be overcome by using more than one detector, and placing them around the specimen. It also appears likely that in the future it will be possible to make larger semiconductor counters than at present.

Semiconductor detectors have the considerable practical advantage of not requiring a well-stabilized voltage source. If they are connected, as shown in fig. 2, to an amplifier with capacitive negative feedback, the height of the output pulses from the amplifier will be largely independent of the variations in the voltage V_0 across the detector. The reason for this is that variation of V_0 causes a change in the capacitance C of electrode *2* with respect to earth such that the change in the height of the pulses across R is exactly compensated by the change in the ratio C/C_f ("charge amplification").

Finally, we should mention again that semiconductor counters can be produced in a wide variety of forms. This makes it possible to adapt them more easily and more effectively than other detectors to the requirements of specific experiments.

It will be apparent from what we have said that semiconductor detectors are preferable to other types for several types of measurement. Some examples have been specifically mentioned. In nuclear physics, the advent of the semiconductor detector has brought about a considerable advance in spectroscopy and in particle identification. We have also seen that the directional distribution of radiation emitted as a result of a nuclear reaction can be determined very much faster with a semiconductor counter, with no sacrifice of accuracy.

In radiochemistry also, which mainly requires the detection of γ-rays, semiconductor counters are often to be preferred. This applies both for purely scientific research and for applied radiation chemistry such as activation analysis. In such applications the very high energy resolution of semiconductor counters often permits a simplification of procedures.

Summary. Germanium and silicon counters with *P-N* and *P-I-N* configurations are extremely useful for detecting and measuring the energy of γ-quanta and high-energy charged particles. Except with very heavy particles these semiconductor detectors are highly linear and have a particularly high resolution (about 20 keV for α-radiation, only a few keV for β- and γ-radiation). The pulse height depends only on the quantum energy (particle energy) E and not on the nature of the radiation. With extremely thin detectors it is also possible to measure $\mathrm{d}E/\mathrm{d}x$ and thus identify unknown heavy particles. The construction of the device is readily adaptable to specific requirements; to measure directional distributions, for example, a detector can be divided into numerous subdetectors (checker-board counter). The resolution is higher for larger values of the charge carrier life compared with the transit time in the counter, and for lower bias current. A *P-I-N* detector combines a large detection volume with a small bias current.

Continuous furnaces for rod material

U. H. Banga and W. Mesman

621.365.4

Furnaces in which products are subjected to heat treatment often have a temperature gradient which makes it difficult to supervise and control the process. In this article the authors show that continuous furnaces can be designed in such a way that the temperature reached by the product is very closely defined.

In many manufacturing processes a product has to to be subjected to one or more heat treatments. The degassing of thermionic valves, semiconductor diffusion processes and the soldering of components in circuits are examples from the electronics industry. In some heat treatments the furnace is charged with batches which go through a complete temperature cycle, each batch being heated to a certain temperature, perhaps kept there for a certain time, and then cooled. In another type of treatment the product is passed at a certain speed through a tubular or tunnel-shaped furnace which is kept at a constant temperature. Such *continuous* furnaces may also be used for batch treatment of products. More often, however, products travel through the hot furnace in a constant stream. During their passage through the furnace they are heated to the temperature at which they have to be processed when they emerge.

Temperature and power distributions

In many cases the temperature distribution in a continuous furnace is such that the product reaches its highest temperature somewhere inside the furnace. This applies particularly when the furnace consists of a tube with a heater winding round it, whose turns have a constant pitch. The tube then has a temperature distribution along its length as illustrated by curve p in *fig. 1a.* When a product travels through such a furnace from left to right its temperature varies with position as shown in curve q.

In a case like the one illustrated here, it is difficult to keep the heat treatment under exact control. This is because in the regulation of the temperature there is an unavoidable delay between a change in furnace temperature and the corresponding change in the final temperature of the product. Moreover, it is difficult to measure the temperature of the product *in motion.* Another difficulty is that it is by no means always permissible to let the product reach a higher temperature in the furnace than the desired final temperature.

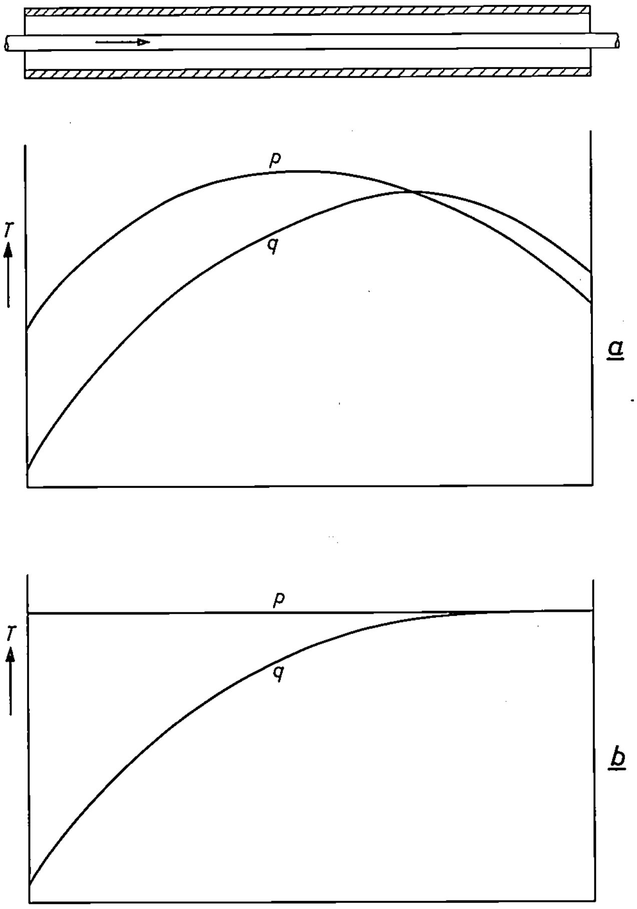


Fig. 1. Diagram showing the temperature distribution in a continuous furnace (p) and in the material passed through the furnace, (q). Curves a are applicable if a heater winding of constant pitch is wound on the furnace tube, and curves b when the turns are so arranged that the temperature inside the furnace is constant over the whole length.

The situation is much simpler if one ensures that the furnace temperature remains constant over its whole length. After entering the furnace the product then

Ir. U. H. Banga and W. Mesman are with Philips Research Laboratories, Eindhoven.

increases steadily in temperature (fig. 1*b*) and finally, if the furnace is sufficiently long, it reaches the furnace temperature. This can easily be measured and controlled: the furnace can be designed in such a way that its temperature is virtually identical with that of the heater winding, so that it is sufficient to measure and regulate the temperature of the winding. Since this temperature responds quickly to a change in the mains voltage, for example, or in the ambient temperature, the method gives a fast-acting control.

In the following we shall consider how a furnace should be designed in order to keep the temperature the same everywhere inside it, and how long the furnace should be in order to ensure that the product does in fact reach the furnace temperature. We shall consider the simplest form of furnace, namely a ceramic tube with a heater winding wound round it, enclosed by a cylindrical insulating sleeve (*fig. 2*). We shall also consider a simple shape for the product: this is taken as a cylindrical rod, and the rod is assumed to pass through the furnace at a constant speed.

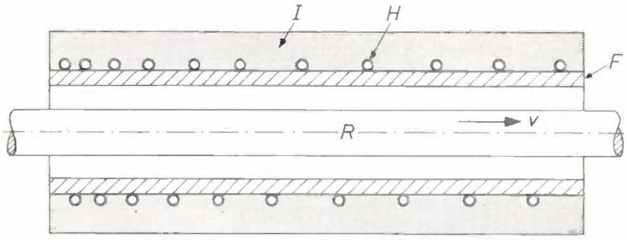


Fig. 2. Continuous furnace for rod material. *F* furnace tube. *H* heater winding. *I* insulating sleeve. *R* rod passed through the furnace at a speed *v*.

If the furnace temperature is to be the same over the whole length, the heat generated, and hence also the power supplied, should not be distributed uniformly along the length of the tube. The required distribution depends on the dimensions of the furnace, on the required temperature and also on the dimensions and material properties of the rod and the speed at which it is to travel through the furnace. The correct distribution of the power can be obtained by making the pitch of the heater winding on the furnace tube depend on position along the length of the furnace.

At the entrance, where the rod is still cold, more power is needed for heating than further along in the furnace. The proportion of the power per unit length which is effectively used, i.e. supplied to the rod, may vary as a function of position as shown in *fig. 3* by the curve $\Delta P_u/\Delta x$; at places where the rod temperature is equal to the furnace temperature "effective" power is no longer needed.

Heat losses

The heat lost through the insulating material is also not uniformly distributed over the length, even when the furnace temperature is the same everywhere inside the tube. In this case the distribution is in fact uniform over the central part of the furnace, where the heat flow is almost radial, but there are extra heat losses at the ends, and these losses increase with the diameter-to-length ratio of the furnace. The distribution of the power loss per unit length may be as shown by curve $\Delta P_l/\Delta x$ in fig. 3. The total power required per unit length is given by

$$\frac{\Delta P_t}{\Delta x} = \frac{\Delta P_u}{\Delta x} + \frac{\Delta P_l}{\Delta x}.$$

If this quantity has been calculated as a function of x for a given case, the power distribution corresponding to it can be realized by making the number of turns per unit length proportional at each point to the value of $\Delta P_t/\Delta x$ required at that point. If, for the case in question, a rod with the dimensions, material properties and speed of travel which apply in the calculation is fed through the furnace then the temperature of the furnace will be the same over the whole length.

Although it is a simple matter to calculate the heat loss occurring at places where the heat flow can be taken as radial, this is not so simple for the extra losses at the ends. A good approximation can be obtained, however, by means of an electrical analogue in the form of a resistance network [1].

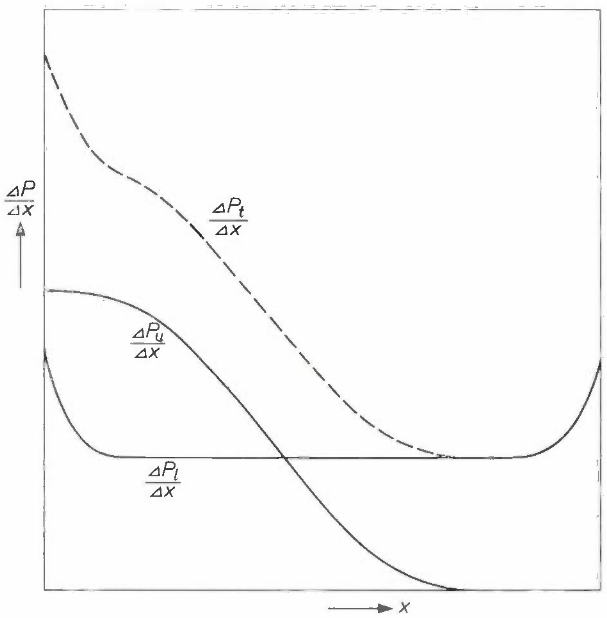


Fig. 3. Power required per unit length of furnace as a function of position x in the furnace. $\Delta P_u/\Delta x$ effective power. $\Delta P_l/\Delta x$ power to make up for the heat loss. $\Delta P_t/\Delta x$ total power.

The magnitude of the heat loss is readily controlled within very wide limits by the choice of the insulating sleeve. A thick sleeve gives a small heat loss; a disadvantage of operating a furnace in this rather economical way is, however, its high thermal capacity, which causes a considerable lag in response to temperature control. This means that when the conditions are changed (e.g. the speed of travel or thickness of the rod) fairly severe temporary variations in temperature may occur. To keep these within bounds, a relatively high heat loss therefore has to be accepted.

Calculating the temperature distribution in the rod

In order to calculate the distribution of the effective power it is necessary to know the temperature distribution in the rod. We shall therefore calculate this distribution, working on the assumption that the outside wall of the furnace tube has a uniform temperature T_f. In general, the thermal conductivity of the furnace tube is high enough to allow the inside wall temperature also to be taken as equal to T_f. We further assume that the heat transfer between the inside wall of the furnace and the rod takes place solely by radiation. Two more approximations are introduced which relate to heat transport in the rod: we assume that the thermal conductivity of the material is high enough to give a uniform temperature over the cross-section; on the other hand we neglect the thermal conductivity in the axial direction compared with the heat transport due to the movement of the rod.

The temperature T_r of the rod material will now be a function of the location x in the furnace, the speed v of the rod, the diameters D_r and D_f of rod and furnace, and the properties of the relevant materials.

For the calculation of T_r we consider a section of furnace and rod of length Δx (*fig. 4*). In a state of temperature equilibrium the heat ΔH_1 radiated from the furnace wall to the rod will be equal to the heat ΔH_2 transported by the movement of the rod. These quantities are given by the following equations:

$$\Delta H_1 = \pi D_r c_{fr} (T_f^4 - T_r^4)\, \Delta x, \quad \ldots \quad (1)$$

$$\Delta H_2 = \tfrac{1}{4}\pi D_r^2\, s\, v\, \frac{dT_r}{dx} \Delta x. \quad \ldots \quad (2)$$

Here c_{fr} is a coefficient determining the heat exchange by radiation between the furnace wall and the rod:

$$c_{fr} = \frac{c_z}{\dfrac{1}{\varepsilon_r} + \dfrac{D_r}{D_f}\left(\dfrac{1}{\varepsilon_f} - 1\right)}, \quad \ldots \quad (3)$$

where c_z is the Stefan-Boltzmann radiation constant, ($c_z = 4.96 \times 10^{-8}$ kcal/m^2 deg^{4}h), and ε_f and ε_r are the emissivities of furnace wall and rod respectively [2].

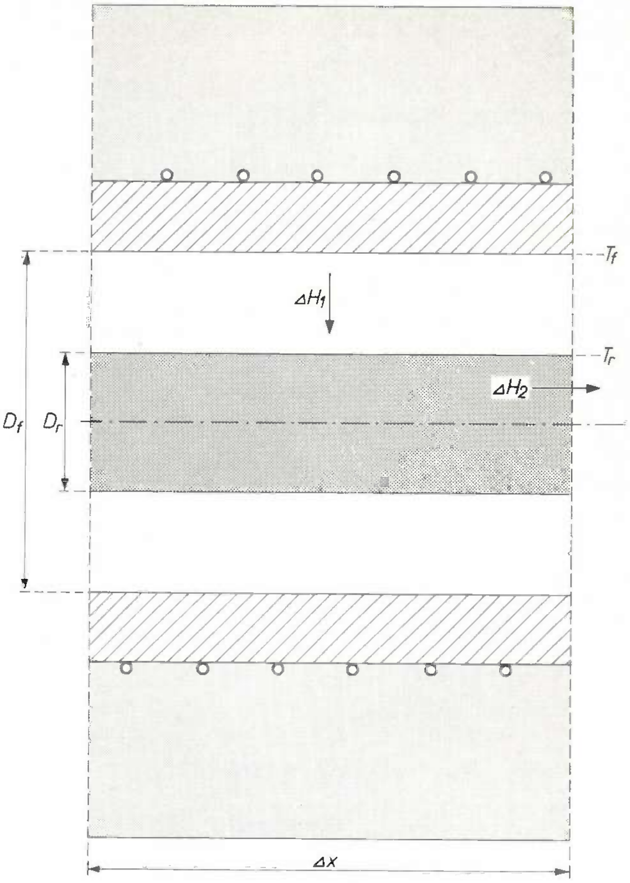


Fig. 4. Section of length Δx of rod and furnace. ΔH_1 heat flow from furnace wall to rod. ΔH_2 heat transport due to the movement of the rod.

Further, s is the specific heat capacity of the rod material (normally measured in kcal/m^{3}deg).

Equating (1) and (2) and integrating with respect to x gives:

$$\log \frac{1 + T_r/T_f}{1 - T_r/T_f} + 2 \arctan T_r/T_f = x/\xi, \quad (4)$$

where the characteristic quantity ξ, which has the dimension of length, contains all the geometric and material parameters:

$$\xi = \frac{D_r\, sv}{16\, c_{fr} T_f^3}. \quad \ldots \quad (4a)$$

This relation between the dimensionless variables T_r/T_f and x/ξ, shown graphically in *fig. 5*, may be used with any configuration to calculate the theoretical

[1] See e.g. V. Paschkis and J. Persson, Industrial electric furnaces and appliances, Interscience Publ., New York, 1960; G. Liebmann, Resistance-network analogues with unequal meshes or subdivided meshes, Brit. J. appl. Phys. **5**, 362-366, 1954; M. J. Laubitz, Design of gradientless furnaces, Can. J. Phys. **37**, 1114-1125, 1959.

[2] See H. Gröber, S. Erk and U. Grigull, Die Grundgesetze der Wärmeübertragung, Springer, Berlin, 1963, page 368 et seq.

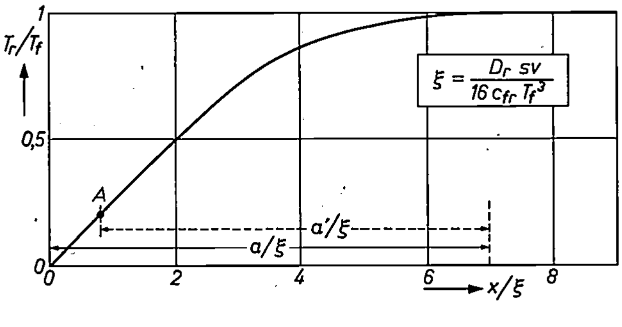


Fig. 5. Heating of rod material in a continuous furnace with uniform temperature T_f. The graph shows the ratio of the rod temperature T_r to T_f as a function of the dimensionless quantity x/ξ, in which the characteristic quantity ξ includes all the geometric and material parameters involved in the calculation. At the entrance ($x = 0$) the temperature of the material is 0 °K. It has virtually reached the furnace temperature after covering a distance a. If the rod is introduced into the furnace at room temperature T_0 the heating process begins at the point corresponding to T_0/T_f. For example, point A corresponds to $T_0/T_f = 0.2$. The material then reaches the temperature T_f after travelling a distance a'.

temperature gradient in a rod whose initial temperature is 0 °K, which enters a furnace in which the temperature is uniformly T_f. It can be seen that the rod has virtually reached the furnace temperature after travelling a distance a corresponding to

$$a/\xi = 7. \quad \dots \dots \dots \quad (5)$$

In normal conditions the rod will enter the furnace at room temperature (300 °K), and the curve should then be used from the corresponding point onwards. If, for example, the furnace temperature is 1500 °K, the heating of the rod begins at $T_r/T_f = 0.20$. According to fig. 5 this corresponds to $x/\xi \approx 0.8$ (point A). In this case the rod will therefore have virtually reached the furnace temperature after travelling a distance a' corresponding to $a'/\xi = 6.2$.

Power requirements

The useful heat flow required per unit length in order to raise the rod to the appropriate temperature during its passage through the furnace, $\Delta H_l/\Delta x$, is found directly from eq. (1) (or eq. (2)). We.now write this equation in the form:

$$\frac{\Delta H_l}{\Delta x} = \pi D_r c_{fr} T_f^4 \left\{1 - \left(\frac{T_r}{T_f}\right)^4\right\}. \quad . \; . \quad (6)$$

Converting the heat flow into electrical power (1 kcal/h = 1.16 W) we find from this the useful electric power per unit length, for which we can write:

$$\frac{\Delta P_u}{\Delta x} = \zeta \left\{1 - \left(\frac{T_r}{T_f}\right)^4\right\}, \quad . \; . \; . \; . \; . \quad (7)$$

with $\qquad \zeta = 1.16\, \pi D_r c_{fr} T_f^4. \quad . \; . \; . \; . \quad$ (7a)

With the aid of (7) and (4) — or fig. 5 — we can now again show the dimensionless quantity

$$\frac{\Delta P_u}{\Delta x} / \zeta$$

as a function of x/ξ (*fig. 6*). From this graph the required effective power per unit length can be derived for all configurations as a function of the length coordinate of the furnace.

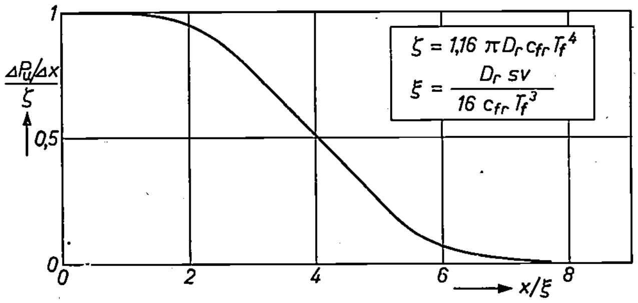


Fig. 6. Distribution of useful power per unit length, required to give the furnace a uniform temperature. The graph gives curves of dimensionless quantities in which ζ and ξ include all the geometric and material parameters.

The total effective power required could be calculated by integrating the curve in fig. 6 from the point corresponding to the temperature T_{r0} at which the rod enters the furnace. It is easy to see, however, that this power (the heat transported per second by the moving rod) is given by:

$$P_u = \frac{1.16}{4}\, \pi D_r^2\, sv\, (T_f - T_{r0}) \text{ watt} . \quad . \quad (8)$$

The *total power* required is found by adding the power loss to P_u. As already mentioned, an electrical analogue can be used for accurately calculating the latter power and its distribution over the length of the furnace. The heater coil now has to be wound on the furnace tube with a number of turns per unit length which is proportional to the total power required per unit length.

Furnace measurements

We have used the foregoing calculations in the design of a furnace for heating magnet steel in rod form (diameter $D_r = 14$ mm). The rod had to travel at a speed v of 3.6 m/h through a furnace with an inside diameter D_f of 40 mm, and had to be heated from room temperature, 300 °K, to 1200 °K. Further data are $s = 845$ kcal/m³ deg, $\varepsilon_f = 0.3$ and $\varepsilon_r = 0.38$. Using eq. (3) we calculate $c_{fr} = 1.45$ kcal/m² deg⁴h. The characteristic quantity ξ is thus 0.106 m. The heating of the rod begins at $T_r/T_f = 0.25$, which is seen from fig. 5 to correspond to $x/\xi = 1$. The rod will therefore reach the furnace temperature after travelling a distance $6\xi = 0.64$ m.

The furnace built was 1.40 metres long. *Fig. 7* shows the calculated rod temperature as a function of distance inside the furnace.

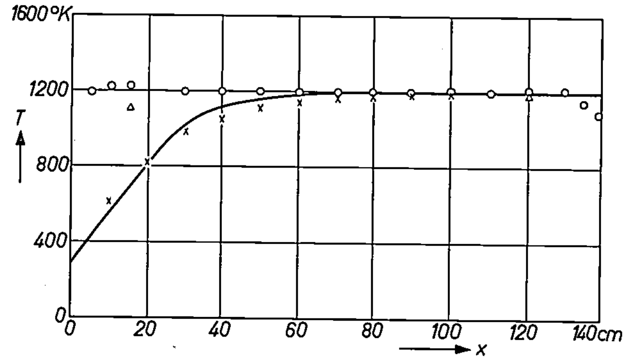


Fig. 7. Axial temperature distribution for a rod of magnet steel with a diameter of 14 mm, travelling at a speed of 3.6 m/h through a furnace in which the temperature is everywhere 1200 °K. The solid curve indicates the calculated temperature; the crosses are temperatures measured at the rod with a thermocouple. The circles indicate the temperature of the heater winding. The triangles indicate the temperature measured at two points on the inside wall of the furnace.

The useful power can be calculated with the aid of eq. (8), and is found to be 470 W. The distribution of this power over the length should follow a curve as shown in fig. 6. The curve $\Delta P_u/\Delta x$ in *fig. 8* gives this distribution on the appropriate scales.

The insulating sleeve had an outside diameter of 11 cm. The heat conductivity coefficient of the insulating material depended on temperature to a fairly considerable extent, being 0.035 kcal/m deg h at 100 °C and 0.257 kcal/mdeg h at 900 °C. For this reason, in the calculation of the heat losses the sleeve was divided into three shells, in each of which the temperature was assumed to be constant. This gave a heat loss of 840 W. The extra losses at the ends, which were determined with an electrical analogue, were 35 W, so that the total heat loss was 875 W. The distribution over the length of the furnace is shown by the curve $\Delta P_l/\Delta x$ in fig. 8.

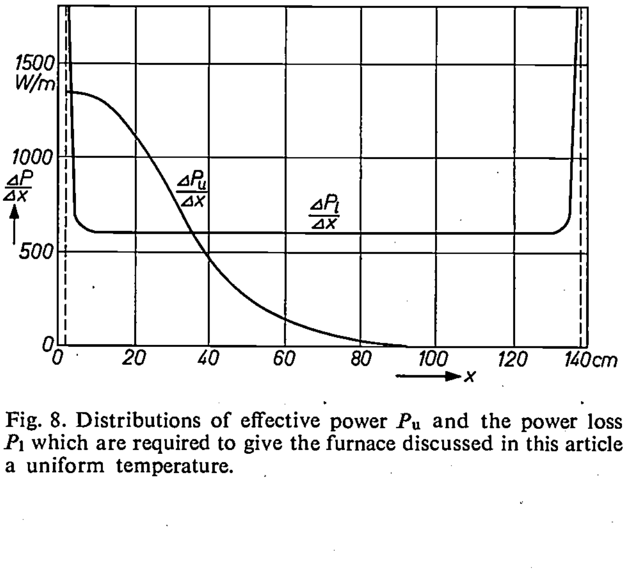


Fig. 8. Distributions of effective power P_u and the power loss P_l which are required to give the furnace discussed in this article a uniform temperature.

To find the distribution of the turns of the heater winding we must now determine the total power requirement per unit length from the sum of the ordinates of the two curves in fig. 8. This sum is plotted in *fig. 9*. The extra losses at the ends are assumed here to be constant over a small distance in order to arrive at a practicable distribution of the turns.

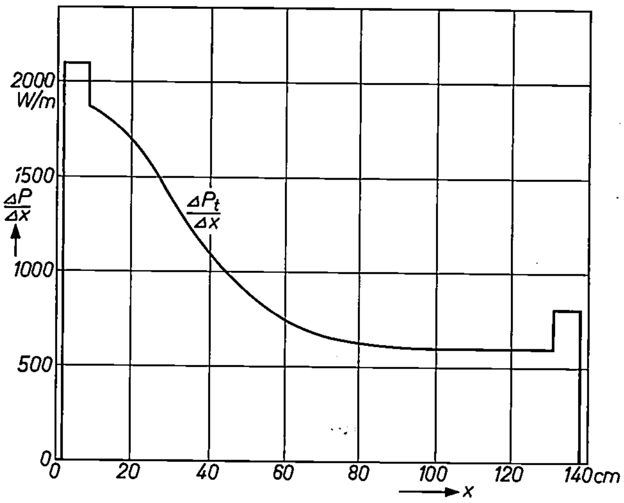


Fig. 9. Distribution of the total power required over the length of the continuous furnace described. If the heater windings are distributed in such a way that the number of turns per unit length at each point is proportional to the corresponding power required per unit length, the temperature inside the furnace is completely uniform.

In accordance with the distribution given by fig. 9, a heater winding of 136 turns was used, with a total resistance of 15 Ω. The temperature distribution in this furnace was measured by a thermocouple which travelled along with the rod. The crosses in fig. 7 indicate a number of measured points. The temperature of the winding at various points was also measured (denoted by circles). The triangles give the temperature at two places on the inside wall of the furnace. (At the entrance to the furnace the inside wall will have a somewhat lower temperature than the outside wall, owing to the heavy flow of effective heat through the furnace wall which occurs here.)

Bearing in mind that various approximations and simplications were made in the calculations, the agreement between measurement and calculation can be considered satisfactory. The actual distance which the rod has to travel before reaching the final temperature is somewhat greater than calculated: this is probably due to the difference in temperature between outside and inside wall — we took this difference to be zero — and to the fact that the emissivities ε_f and ε_r, which we assumed to be constant, in reality decrease with rising temperature.

In conclusion, it should again be pointed out that a furnace, calculated and constructed on the principle described here, shows the required temperature distribution only when the rod passed through it has the dimensions, properties and rate of travel used in the calculation. Any departure from these quantities may result in a different temperature distribution. For example, the effect of reducing the speed is to increase the temperature at the entrance end of the furnace. Calculations of these effects are rather complicated and can best be carried out by using a computer. By the same means, various refinements can be introduced into the calculation. It is possible, for example, to take into account the temperature difference between the outside and inside walls of the furnace tube, and also the conduction and convection part of the heat transfer between furnace wall and rod. The heat transport due to conduction along the axis of the rod, which we assumed to be negligible compared with the heat transport due to the forward movement, can then be taken into account as well. In many cases, however, such an elaborate treatment is not necessary and an approximate calculation gives sufficiently accurate results.

Summary. In continuous furnaces with a heater winding of constant pitch the temperature distribution may make it difficult to supervise and control the heat treatment of products passing through the furnace. This is much easier if the furnace has a uniform temperature over its whole length. This requires a winding of non-constant pitch. On the basis of a simple furnace and a product of simple form — a rod — a calculation is made of the required input power distribution over the length of the furnace. This distribution depends on the diameter of the rod, the properties of the material and the speed at which the rod travels through the furnace. An experimentally determined temperature distribution showed good agreement with the calculated distribution.

A sensitive torque meter for high frequencies

621.317.799:531.232

An instrument for dynamic torque measurements has recently been developed at Philips Research Laboratories; unlike existing mechanical meters, this instrument has a good high-frequency sensitivity. It consists basically of a shaft of magnetostrictive material (permalloy), which is clamped at one end, the torque to be measured being applied at the other end (*fig. 1*). The shaft is encircled by a coil through which an alternating current flows; this induces an alternating axial magnetic field in the shaft, and as long as there is no torque exerted on the shaft, the magnetic induction remains purely axial. When the shaft is twisted by the application of a torque, stress anisotropy occurs, and as a result of this the magnetic induction acquires a tangential component (B_t). An alternating voltage can now be measured between the two ends of the shaft,

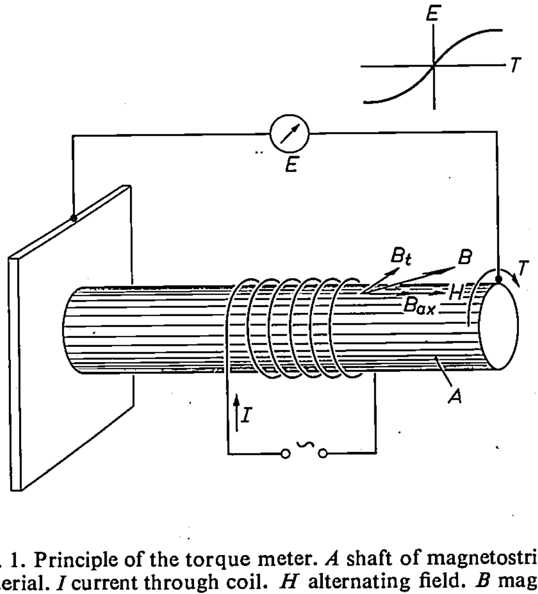


Fig. 1. Principle of the torque meter. *A* shaft of magnetostrictive material. *I* current through coil. *H* alternating field. *B* magnetic induction, B_{ax} its axial component and B_t its tangential component. *T* torque. *E* output voltage.

the rod has to travel before reaching the final temperature is somewhat greater than calculated: this is probably due to the difference in temperature between outside and inside wall — we took this difference to be zero — and to the fact that the emissivities ε_f and ε_r, which we assumed to be constant, in reality decrease with rising temperature.

In conclusion, it should again be pointed out that a furnace, calculated and constructed on the principle described here, shows the required temperature distribution only when the rod passed through it has the dimensions, properties and rate of travel used in the calculation. Any departure from these quantities may result in a different temperature distribution. For example, the effect of reducing the speed is to increase the temperature at the entrance end of the furnace. Calculations of these effects are rather complicated and can best be carried out by using a computer. By the same means, various refinements can be introduced into the calculation. It is possible, for example, to take into account the temperature difference between the outside and inside walls of the furnace tube, and also the conduction and convection part of the heat transfer between furnace wall and rod. The heat transport due to conduction along the axis of the rod, which we assumed to be negligible compared with the heat transport due to the forward movement, can then be taken into account as well. In many cases, however, such an elaborate treatment is not necessary and an approximate calculation gives sufficiently accurate results.

Summary. In continuous furnaces with a heater winding of constant pitch the temperature distribution may make it difficult to supervise and control the heat treatment of products passing through the furnace. This is much easier if the furnace has a uniform temperature over its whole length. This requires a winding of non-constant pitch. On the basis of a simple furnace and a product of simple form — a rod — a calculation is made of the required input power distribution over the length of the furnace. This distribution depends on the diameter of the rod, the properties of the material and the speed at which the rod travels through the furnace. An experimentally determined temperature distribution showed good agreement with the calculated distribution.

A sensitive torque meter for high frequencies

621.317.799:531.232

An instrument for dynamic torque measurements has recently been developed at Philips Research Laboratories; unlike existing mechanical meters, this instrument has a good high-frequency sensitivity. It consists basically of a shaft of magnetostrictive material (permalloy), which is clamped at one end, the torque to be measured being applied at the other end (*fig. 1*). The shaft is encircled by a coil through which an alternating current flows; this induces an alternating axial magnetic field in the shaft, and as long as there is no torque exerted on the shaft, the magnetic induction remains purely axial. When the shaft is twisted by the application of a torque, stress anisotropy occurs, and as a result of this the magnetic induction acquires a tangential component (B_t). An alternating voltage can now be measured between the two ends of the shaft,

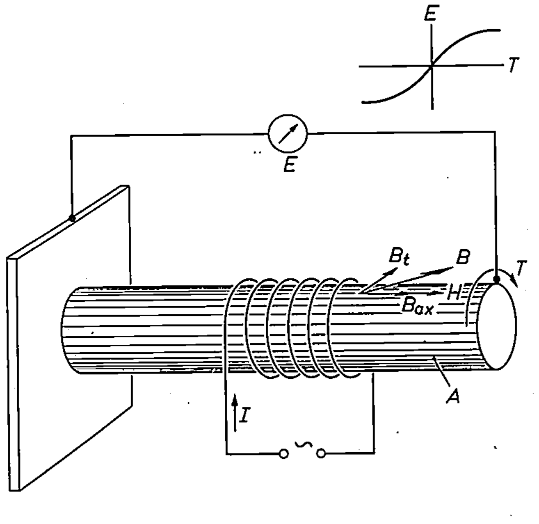


Fig. 1. Principle of the torque meter. *A* shaft of magnetostrictive material. *I* current through coil. *H* alternating field. *B* magnetic induction, B_{ax} its axial component and B_t its tangential component. *T* torque. *E* output voltage.

since the circuit thus formed now contains an alternating flux. The magnitude of the voltage is a measure of the magnitude of the torque; when the torque is reversed the voltage reverses in phase.

Compared with earlier mechanical instruments, torque meters based on this principle possess much greater stiffness for the same sensitivity, and can therefore be used up to higher frequencies with the same load. For example, a torque of less than 10^{-5} Nm (0.1 gcm) can be measured with a shaft whose stiffness is 1500 Nm/rad. Conventional torque meters of roughly equivalent sensitivity have a stiffness of only 1 to 5 Nm/rad. The resonant frequency of a loaded torque meter is proportional to $\sqrt{S/I}$, where S is the stiffness of the shaft and I is the moment of inertia of the loaded shaft. It thus follows that, for the same load, the upper limit of the frequency range of the new torque meter is some 20 to 40 times higher than that of existing meters.

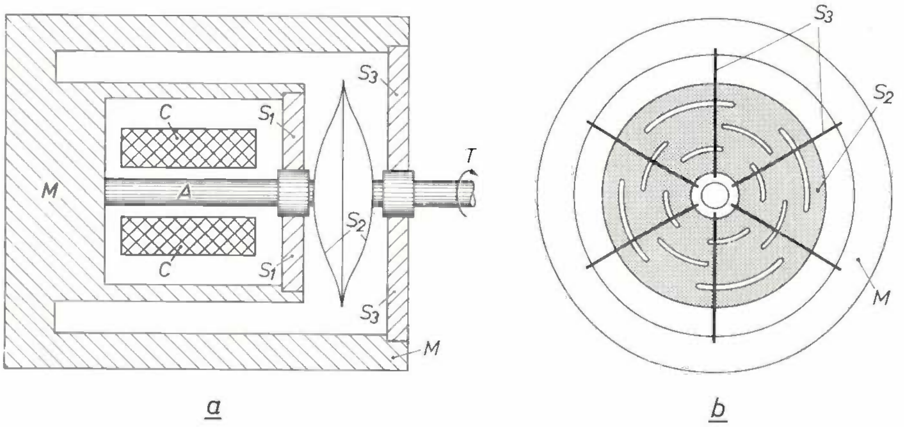


Fig. 2. Longitudinal cross-section (*a*) and plan view (*b*) of the torque meter. *A* shaft. *M* housing. *C* coil. S_1 and S_3 leaf-springs. S_2 diaphragms. *T* torque.

The use of magnetostriction is, however, not entirely free from difficulties. For instance, there are changes in the properties of the material with time; these changes are so slight, however, that there is only a very small zero drift (less than 10^{-5} Nm/h) after several hours of operation. Moreover, the direction of the magnetic induction in the shaft is not only affected by torsion but also by pressure in the axial direction and by bowing of the shaft. To eliminate these effects an arrangement of springs is employed (*fig. 2*) which transmits only the torsion. The shaft A is rigidly fixed at one end to the housing M, and its other end is mounted in a suspension system of six leaf-springs (S_1 in fig. 2*a*). The connection between the shaft and the load is formed by two circular dish-shaped diaphragms which are joined to each other at the edges (S_2), and a second set of six leaf-springs S_3. A plan view of this arrangement is given in fig. 2*b*. The sets of leaf-springs are very slack under torsion (at least in comparison with the rigidity of the shaft); under axial pressure and bowing, however, they are very stiff. Any axial or bowing forces occurring are thus largely transferred by the leaf-springs S_3 to the housing. Any residual unwanted movements made by these springs exert no forces on the shaft, since the diaphragms S_2 are very slack under pressure and bowing. Under torsion they give a firm connection between S_3 and the shaft, so that the torque is transmitted unimpeded to the shaft.

In dynamic measurements the housing of the torque meter is connected to a rotating drive mechanism, and the end of the shaft to a rotating load. The measurement now gives the torque transmitted by the shaft as a function of time. If the characteristics of the drive mechanism are known (e.g. a flywheel rotating at a constant speed) this measurement provides information about the load, for instance of the angular position at which the greatest friction occurs. If, on the other hand, the load is known (e.g. a flywheel with a known moment of inertia), conclusions about the drive mechanism can be drawn from the measurement. This technique finds application in the design of small electric motors for measuring the effect of each separate pole on the torque. For some of these measurements a frequency range up to about 1000 c/s was required; purely mechanical meters at best go no further than 50 to 100 c/s.

W. J. Schoenmakers

Ir. W. J. Schoenmakers is with Philips Research Laboratories, Eindhoven.

Mass spectrometer analysis of a solid surface

543.51

It has been known for some considerable time that an ion bombardment can be used to liberate atomic particles from the surface of a solid ("sputtering"). A fraction of these particles appears to be ionized. By a mass spectrometer study of ions liberated in this way it is possible to analyse the surface of the substance.

This principle has been used to develop both in this laboratory and elsewhere, a method of analysis [1] which offers a number of striking possibilities.

A solid specimen is placed in the source region of a mass spectrometer specially designed for the purpose (*fig. 1*). A beam of primary ions such as positive argon ions, which are obtained from an auxiliary source, and whose energy can be varied from 3 to 15 keV, is directed on to the specimen at a small angle. If for example the substance consists of the elements X and Y, then the bombardment knocks the elements out of the surface partly as atoms, partly as negative or positive ions (depending on the kind of elements). The quantity of emitted particles of each element is a measure of the concentration of that particular element in the specimen. The "secondary" ions are accelerated by an extraction voltage between the specimen and the exit slit of the source. These secondary ions are separated in the analyser section of the spectrometer according

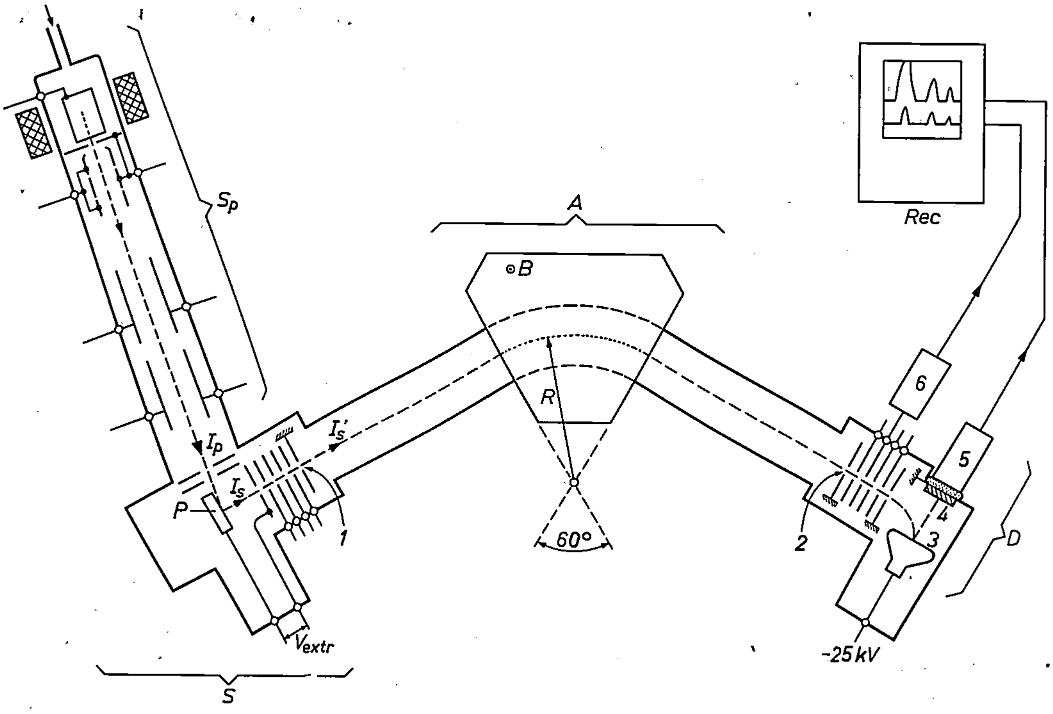


Fig. 1. Diagrammatic representation of the mass spectrometer (of the single focusing 60° type) designed for the experiments. S source section, in which the specimen P to be analysed is placed. I_p primary ion beam obtained from the auxiliary source S_p. I_s secondary ions from the specimen. V_{extr} extraction voltage. *1* entrance slit, *2* collector slit of the analyser section A. Here the accelerated ions I_s' are deflected in circular paths whose radius of curvature R is proportional to $\sqrt{(M/e)V}/B$ by a magnetic sector field. M and e are the mass number and the charge of the ion, V is the acceleration voltage, and B the magnetic induction.

V is held constant and B is varied as a function of time: the ions present with various mass numbers M then traverse the sector field consecutively with the correct radius of curvature to reach the collector slit.

The measurement of the collected ion current is simultaneously performed in our arrangement with the d.c. amplifier *6* (for currents above 10^{-14} A) and with an extremely sensitive scintillation detector D (for currents less than 10^{-13} and down to 10^{-20} A): in this the ions liberate electrons from a target plate *3*. These electrons cause scintillations in the crystal *4* which are detected with the photo-multiplier *5*.

The indications of *5* and *6* are continuously recorded so that the mass spectrum of the specimen under investigation is obtained during the variation of the magnetic field B.

to their mass-to-charge ratio. The ion currents I_X and I_Y are recorded one after the other with suitable aids (d.c. amplifier, electron multiplier) and an analysis of the specimen is thus obtained.

In order to be able to use this method for a quantitative analysis, the "sensitivity" of each element has to be known, i.e. the ratio of the secondary ion current to the primary ion current for a given concentration of the element in the substance under investigation. This sensitivity can vary from element to element and, moreover, for a single element it can depend upon the other constituents of the specimen (the matrix). For a quantitative analysis, just as in spectro-chemistry, a standard for the element in the matrix must therefore be available. Extensive studies have shown that the sensitivity for one matrix and one element is in general practically independent of the concentration of the element [2]. Our own study of 15 elements, using primary ions at 11 keV energy, has furthermore shown that the sensitivities for these elements do not differ by a factor of more than about 30.

The method has the following interesting aspects:
1) Any solid substance can be analysed: conductors, semi-conductors and insulators.
2) The method is particularly suitable for the analysis of very thin surface layers as, by suitably choosing the angle of incidence of the beam and the mass and energy of the primary ions, one can vary their average penetration depth between about 0.2 and 100 nm. Some examples of objects for investigation are evaporated films and also the adsorption films which form on any solid on exposure to the atmosphere.
3) It is possible to make a "depth analysis" of a surface: one can strip away, so to say, layer after layer of atoms from the surface with the primary beam and, in so doing, determine the concentration of the elements as a function of the depth below the surface. In doing this the intensity of the primary beam must be kept very constant and a sufficiently small penetration depth must be chosen.
4) Only a minute quantity of material is used up: a quantity of 10^{-9} grams corresponds to a secondary ion current of about 10^{-18} A for a few hours. Such a current integrated over say five seconds is sufficient to be reasonably accurately measured by means of our detection system (see fig. 1).

Points (2) and (3) are illustrated by the following experiments.

A specimen of pure aluminium was bombarded for 2½ hours with a primary ion current of 5 μA. In "pure" aluminium there is always a little sodium present. A Na^+ peak was accordingly found, whose magnitude, about 10 relative units, remained constant within about 10% during the whole time. After this the aluminium surface was contaminated with sodium by simply rubbing a finger against it. This caused the Na^+ peak to rise to a value of 300 relative units, while the magnitude of the Al peak was unchanged. Bombardment was then continued for seven hours, during which time the Al peak remained constant, but the Na^+ peak decreased gradually to 40 units, clearly because the contaminated surface layer was stripped away. This latter was corroborated by displacing the primary beam a little towards the edge of the region first bombarded: here there should have been less sodium sputtered away and in fact the Na^+ peak now increased to 110 units, to fall again to 70 units in about half an hour.

In another experiment a 100 nm SiO_2 film grown on silicon was continuously bombarded. The Na^+ peak and the $(SiOH)^+$ peak were both measured as a function of time. The resulting curves (*fig. 2*) can be considered to show the variation of the concentration with the depth below the surface. The interpretation of these results requires, however, a certain caution, as the Si^+ peak itself appears to decrease as a function of time although much less markedly than the other two.

One further remark about the detection system (fig. 1). The secondary ion current appears to remain

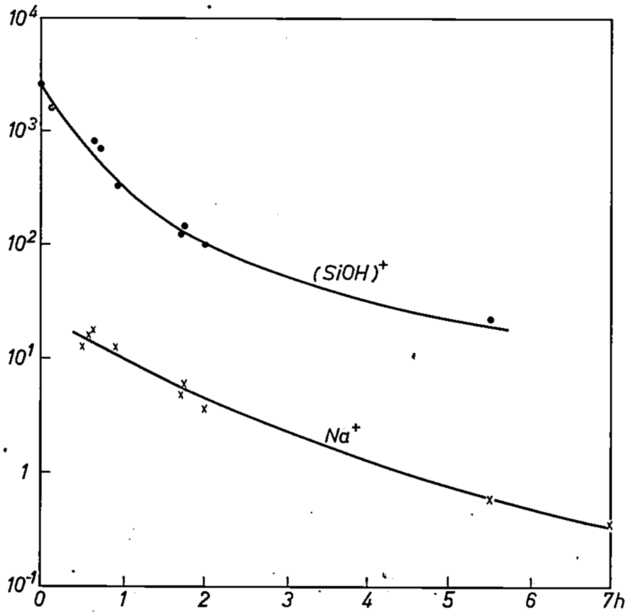


Fig. 2. Variation of the Na^+ peak and the $(SiOH)^+$ peak with time for continuous bombardment of an SiO_2 film grown on Si. The curves indicate, with certain restrictions (see text) the variation of the concentration with depth.

[1] H. J. Liebl and R. F. K. Herzog, J. appl. Phys. **34**, 2893, 1963. A. J. Smith, D. J. Marshall, L. A. Cambey and J. Michael, Vacuum **14**, 263, 1964.
H. E. Beske, Z. Naturf. **19a**, 1627, 1964.

[2] H. E. Beske, paper for the Spring Meeting, Deutsche Physikalische Gesellschaft, Mainz 1966.

constant after bombardment for about 1 hour (about 10^{-10} A for a pure substance — concentration 100% — and at a primary beam current of 5 μA). This constancy permits the time for the recording of the complete mass spectrum to run to several hours, so that each peak can be integrated over a long time; this enables us, with the aid of a scintillation detector with a photo-multiplier [3], and with the aid of counting techniques, to detect ion currents as low as 10^{-20} A. Taking the differences in sensitivity for various elements into account, we can therefore measure concentration ratios of the order of 1 : 10^9.

For reproducible measurements it is essential to prevent the scintillation detector itself from being struck by secondary ions, which would contaminate it. The arrangement of fig. 1 provides the necessary safeguard. In order to prevent the measurement of very small currents from being affected by the after-effects of earlier heavy currents, the ion current to the scintillation system is suppressed as soon as the d.c. amplifier, which measures the current continuously, registers a current greater than 10^{-13} A.

H. W. Werner

[3] H. W. Werner and H. A. M. de Grefte, paper 3rd International Vacuum Congress, Stuttgart 1965.

Dr. H. W. Werner is with Philips Research Laboratories, Eindhoven.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands	*E*
Mullard Research Laboratories, Redhill (Surrey), England	*M*
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France	*L*
Philips Zentrallaboratorium GmbH, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany	*A*
Philips Zentrallaboratorium GmbH, Hamburg laboratory, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany	*H*
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium.	*B*

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

D. Beecham: Ultrasonic scatter in metals — its properties and its application to grain size determination.
Ultrasonics **4**, 67-76, 1966 (April). *M*

G. Blasse: Compounds with α-PbO_2 structure.
Z. anorg. allgem. Chemie **345**, 222-224, 1966 (No. 3/4). *E*

G. Blasse and **A. Bril:** Luminescent properties of $NaGdO_2$:Eu.
Solid State Comm. **4**, 373-375, 1966 (No. 8). *E*

H. W. Bodmann and **G. Söllner:** Glare evaluation by luminance control.
Light and Lighting **58**, 195-199, 1965 (No. 6). *A*

G.-A. Boutry: Inauguration des nouveaux laboratoires du LEP.
Onde électr. **46**, 749-753, 1966 (No. 471). *L*

C. J. Bouwkamp: Note on diffraction by a circular aperture.
Acta phys. polon. **27**, 37-39, 1965 (No. 1). *E*

J. C. Brice and **G. D. King:** Effect of arsenic pressure on dislocation densities in melt-grown gallium arsenide.
Nature **209**, 1346, 1966 (No. 5030). *M*

J. van den Broek: Physical interpretation of a PbO-photodetector.
Solid State Comm. **4**, 295-297, 1966 (No. 6). *E*

G. Brouwer: Control of the surface potential of germanium and electromotive force against calomel with the aid of a variable *p*H electrolyte containing hydrogen peroxide.
Physics Letters **21**, 399-400, 1966 (No. 4). *E*

K. Bulthuis: Effect of local pressure on germanium *p-n* junctions.
J. appl. Phys. **37**, 2066-2068, 1966 (No. 5). *E*

K. H. J. Buschow and **J. F. Fast:** Magnetic properties of some rare-earth aluminium compounds.
Z. Phys. Chemie Neue Folge **50**, 1-10, 1966 (No. 1/2). *E*

K. H. J. Buschow and **J. F. Fast:** Magnetic and structural characteristics of some equiatomic rare-earth germanides.
Phys. Stat. sol. **16**, 467-473, 1966 (No. 2). *E*

H. P. C. Daniëls and **Th. P. J. Botden:** Ultrasonic welding in microminiaturization.
Microminiaturization in automatic control equipment and in digital computers, Proc. IFAC/IFIP Symp., Munich 1965, p. 473-484; Oldenbourg, Munich 1966. *E*

J. A. W. van der Does de Bye: Radiative recombination in *p*-type GaP doped with zinc and oxygen.
Phys. Rev. **147**, 589-599, 1966 (No. 2). *E*

C. Z. van Doorn and **G. Koch:** Abnormal green edge emission in CdS.
Solid State Comm. **4**, 345-346, 1966 (No. 7). *E*

A. van der Drift: Texture of a vapour-deposited leadmonoxide layer.
Philips Res. Repts. **21**, 289-303, 1966 (No. 4). *E*

W. F. Druyvesteyn, F. A. Staas and **A. K. Niessen:** Some experiments on the distribution of a direct transport current in sheets of type II superconductor.
Physics Letters **21**, 387-388, 1966 (No. 4). *E*

P. Eckerlin, C. Langereis, I. Maak and **A. Rabenau:** The preparation, structure and properties of $LiPN_2$.
Special Ceramics 1964 (Proc. Symp. Brit. Cer. Res. Ass.), p. 79-85; Academic Press, London 1965. *A*

P. J. Flanders, R. F. Pearson and **J. L. Page:** Magnetostriction of rare-earth iron garnets at low temperatures.
Brit. J. appl. Phys. **17**, 839-840, 1966 (No. 6). *M*

N. V. Franssen: Calculations and speculations about the hydraulic theory of hearing.
Acustica **17**, 26-33, 1966 (No. 1). *E*

J. A. Geurst: Two-dimensional theory of the unsteady motion of fully cavitating hydrofoils.
Applied Mechanics, Proc. 11th int. Congress, Munich 1964, p. 1156-1164; Springer, Berlin 1966. *E*

J. A. Geurst: Theory of space-charge-limited currents in thin semiconductor layers.
Phys. Stat. sol. **15**, 107-118, 1966 (No. 1). *E*

A. A. van der Giessen: De hydrolyse van oplossingen van Fe(III) nitraat.
Chem. Weekblad **62**, 305-309, 1966 (No. 24). *E*

J.-M. Goethals: Analysis of weight distribution in binary cyclic codes.
IEEE Trans. on information theory **IT-12**, 401-402, 1966 (No. 3). *B*

H. C. de Graaff and **H. Koelmans:** The thin-film field-effect transistor.
Ned. T. Natuurk. **32**, 80-86, 1966 (No. 3). *E*

H. G. Grimmeiss and **H. Scholz:** Optical and electrical properties of Cu-doped GaP. Part II: Photovoltaic effect.
Philips Res. Repts. **21**, 246-269, 1966 (No. 4). *A*

F. W. Harrison: Crystal data for ytterbium orthoferrite, $YbFeO_3$.
Acta crystallogr. **20**, 699-700, 1966 (No. 5). *M*

C. G. J. Jansen, A. Venema and **Th. H. Weekers:** Non-uniform emission of thermionic cathodes.
J. appl. Phys. **37**, 2234-2245, 1966 (No. 6). *E*

W. Kischio: Bildungsenthalpie von Aluminiumphosphid.
J. inorg. nucl. Chem. **27**, 750-751, 1965 (No. 3). *A*

S. R. de Kloet: Ribonucleic acid synthesis in yeast. The effect of cycloheximide on the synthesis of ribonucleic acid on Saccharomyces carlsbergensis.
Biochem. J. **99**, 566-581, 1966 (No. 3). *E*

M. van Koten-Hertogs and **J. S. C. Wessels:** Ferredoxin-stimulated photoreduction of 2,4-dinitrophenol with solubilized chlorophyll a.
Currents in Photosynthesis, Proc. 2nd W.-Europe Conf., Woudschoten-Zeist 1965, p. 207-216; Donker, Rotterdam 1966. *E*

J. R. Mansell and **J. L. Philips:** Sensitive S band travelling-wave phototube.
Electronics Letters **2**, 155-156, 1966 (No. 4). *M*

P. Marchet: Une réalisation moderne de laboratoires consacrés à la recherche appliquée.
Onde électr. **46**, 598-604, 1966 (No. 470). *L*

G. Meijer: Lumière et croissance en longueur.
Photochem. and Photobiol. **5**, 373-374, 1966 (No. 5). *E*

L. Merten: Abgeschirmte piezoelektrische Potentiale um Versetzungen in piezoelektrischen Kristallen.
Z. Naturf. **21a**, 793-798, 1966 (No. 6). *A*

M. Monneraye and **H. J. L. Trap:** Effets de l'addition de TiO_2 sur les propriétés diélectriques des verres et des produits vitrocristallins du type Cabal.
VIIe Congrès Int. du Verre, Brussels 1965, publ. No. 107/III.2. *E*

B. J. Mulder: Mean diffusion path of excitons in crystals of anthracene doped with tetracene.
Philips Res. Repts. **21**, 283-288, 1966 (No. 4). *E*

B. J. Mulder and **J. de Jonge:** Quantum efficiency of photoconduction in lead oxide.
Solid State Comm. **4**, 293-294, 1966 (No. 6). *E*

E. A. Muyderman: Bearings.
Sci. American **214**, No. 3, 60-66, 68 and 71, 1966. *E*

W. C. Nieuwpoort and **G. Blasse:** Linear crystal-field terms and the $^5D_0$-$^7F_0$ transition of the Eu^{3+} ion.
Solid State Comm. **4**, 227-229, 1966 (No. 5). *E*

W. C. Nieuwpoort, G. A. Wesselink and **E. H. A. M. van der Wee:** Thermochromic and solvochromic behaviour of cobalt (II) chloride solutions in various solvents.
Rec. Trav. chim. Pays-Bas **85**, 397-404, 1966 (No. 4). *E*

L. M. Nijland and **J. Schröder:** Generation of light by fluorine reactions in flash lamps.
Philips Res. Repts. **21**, 304-321, 1966 (No. 4). *A*

J. M. Noothoven van Goor: Charge-carrier densities and mobilities in bismuth doped with tin.
Physics Letters **21**, 603-604, 1966 (No. 6). *E*

D. J. van Ooijen and **A. S. van der Goot:** The internal friction of cold-worked niobium and tantalum containing oxygen and nitrogen.
Acta metallurgica **14**, 1008-1009, 1966 (No. 8). *E*

G. W. van Oosterhout: The identical relations between the bilinear covariants of Dirac's theory of the electron.
Physica **32**, 1090-1096, 1966 (No. 6). *E*

C. van Osenbruggen, G. Luimes, A. van Dijk and **J. G. Siekman:** Micro-spark erosion as a technique in microminiaturization.
Microminiaturization in automatic control equipment and in digital computers, Proc. IFAC/IFIP Symp., Munich 1965, p. 485-493; Oldenbourg, Munich 1966. *E*

D. A. E. Roberts and **K. Wilson:** Evaluation of high quality varactor diodes.
Radio and electronic Engr..**31**, 277-285, 1966 (No. 5). *M*

E. Roeder and **S. Scholz:** A simple hot press for laboratory investigations.
Special Ceramics 1964 (Proc. Symp. Brit Cer. Res. Ass.), p. 269-273; Academic Press, London 1965. *A*

F. C. de Ronde: The clawflange: an international standardized millimeter waveguide flange.
Microwave J. **9**, No. 5, 55-58, 1966. *E*

H. J. Schmitt and **H. Zimmer:** Fast sweep measurements of relaxation times in superconducting cavities.
IEEE Trans. on microwave theory and techniques **MTT-14**. 206-207, 1966 (No. 4). *H*

G. Söllner: Ein einfaches System zur Blendungsbewertung.
Lichttechnik **17**, 59A-66A, 1965 (No. 5). *A*

R. P. van Stapele, H. G. Beljers, P. F. Bongers and **H. Zijlstra:** Ground state of divalent Co ions in Cs$_3$CoCl$_5$ and Cs$_3$CoBr$_5$.
J. chem. Phys. **44**, 3719-3725, 1966 (No. 10). *E*

J.-P. Thiran: A class of filters with optimal response to a step of frequency.
Archiv elektr. Übertr. **20**, 388-392, 1966 (No. 7). *B*

A. G. van Vijfeijken: Resistivity and Hall angle in the mixed state of type II superconductors.
Quantum Fluids, Proc. Sussex Univ. Symp. 1965, p. 136-137; North-Holland Publ. Co., Amsterdam 1966. *M*

A. G. van Vijfeijken: On the specific heat of type II superconductors in the mixed state.
Physics Letters **21**, 140-141, 1966 (No. 2). *E*

J. Volger: Comments on pure type II superconductors.
Quantum Fluids, Proc. Sussex Univ. Symp. 1965, p. 128-135; North-Holland Publ. Co., Amsterdam 1966. *E*

K. Walther: Quantum resonances in the amplification of ultrasound in bismuth.
Phys. Rev. Letters **16**, 642-644, 1966 (No. 15). *H*

K. Walther: Ultrasonic amplification in bismuth.
Solid State Comm. **4**, 341-344, 1966 (No. 7). *H*

W. L. Wanmaker, A. Bril and **J. W. ter Vrugt:** Sensitization of Tb^{3+} luminescence by Sn^{2+} and Cu^{+} in alkaline earth phosphates.
Appl. Phys. Letters **8**, 260-261, 1966 (No. 10). *E*

W. L. Wanmaker, A. Bril, J. W. ter Vrugt and **J. Broos:** Luminescent properties of Eu-activated phosphors of the type A^{III}B^{V}O$_4$.
Philips Res. Repts, **21**, 270-282, 1966 (No. 4). *E*

G. F. Weston and **P. C. Newman:** Physics research at the Mullard Research Laboratories.
Bull. Inst. Phys. and Phys. Soc. **17**, 145-152, 1966 (May). *M*

G. Winkler: Eigenschaften und Anwendungen hexagonaler Ferrite.
Z. angew. Physik **21**, 282-286, 1966 (No. 4). *H*

W. J. Witteman: Een laser van grote intensiteit in het infrarood-gebied.
Ned. T. Natuurk. **32**, 61-72, 1966 (No. 2). *E*

W. J. Witteman and **G. van der Goot:** High-power infrared laser with adjustable coupling-out.
J. appl. Phys. **37**, 2919, 1966 (No. 7). *E*

P. Wurtz: Etude de l'émission d'un LASER au Néodyme, déclenché par effet Pockels.
Philips Res. Repts. **21**, 213-245, 1966 (No. 4). *L*

H. Zijlstra: The coercivity of permanent magnets.
Z. angew. Physik **21**, 6-13, 1966 (No. 1). *E*

H. Zijlstra and **H. B. Haanstra:** Evidence by Lorentz microscopy for magnetically active stacking faults in MnAl alloy.
J. appl. Phys. **37**, 2853-2856, 1966 (No. 7). *E*

Volume 27, 1966, No. 12 pages 313-348 Published 30 January 1967