FIFTY YEARS OF THE GAS-FILLED LAMP

by J. C. LOKKER *).

621.326.72

The invention of the gas-filled lamp, now half a century ago, was one of the more important advances — perhaps the most important — in the evolution of the incandescent lamp. This 50th anniversary gives us occasion to trace once again the development of the gas-filled lamp. The share which Philips had in this development is also recalled in the article below.

Mr. Lokker, who wrote this article at our request and whose photograph appears here, took an active and leading part in the development of the incandescent lamp at Philips from its earliest days. He was the first graduate engineer to be appointed by Mr. G. L. F. Philips, joining the company in 1908. In later years he managed the department now referred to as the "Lighting Division", until his retirement in 1945.

In the year 1879, Thomas Alva Edison solved the problem of how to produce light with electricity in a reasonably practical form. The invention was demonstrated with great success at the Paris World Exhibition in 1881. This was the beginning of the carbon-filament lamp. Several firms and engineers set to work to make similar lamps and installations, but since electricity networks were few and far between, the development and spread of electric light made slow headway at first.

G. L. F. Philips, born in 1858 and who graduated at Delft in 1883 as a mechanical engineer, was so interested in the principles of electricity, and especially in the carbon-filament lamp, that after a few years of gaining practical experience, and after experimenting in a primitive laboratory in his parents' home at Zaltbommel, he began in 1891 to manufacture carbon-filament lamps in a former buckskin factory at Eindhoven [1]). At that time other firms were already producing these lamps in large quantities, and in the early years G.L.F. Philips had to contend with numerous difficulties. If his brother A. F. Philips had not come to his assistance in 1895 to organize the selling side of the business, he might well have had to close down production. Gradually the business began to prosper. Since there were hardly any electric power stations in the Netherlands in those days, the two brothers turned their attention to the German market, to such good effect that the Düsseldorf Gewerbeausstellung (Industrial Exhibition) in 1902 was lit entirely by Philips lamps.

The bitter competitive struggle fought with other manufacturers led in 1903 to the setting-up in Berlin of the "Verkaufsstelle Vereinigte Glühlampenfabriken" (Associated Lamp Manufacturers' Selling Agency). While this brought some peace on the commercial side, there was no easing-up of pressure on the production side, caused by the demand for new and better lamps. Although the production of carbon filaments was substantially improved by changing from zinc-chloride cellulose to collodion acetate as the basic material, the carbon filaments nevertheless consumed too much power for a given light output, and the lamps tended after some time to turn black. Efforts made to find other materials for the filament led for example to the Nernst lamp, which used a slender rod of thorium-cerium oxide (1897), the osmium lamp (1900), the tantalum lamp of Siemens (1904) and, in 1906, the tungsten-filament lamp (*fig. 1*).

Although the melting point of tungsten is not so high as that of carbon, its rate of evaporation at high temperature is much lower. This made tungsten better suited as a material for lamp filaments. The melting point of tungsten metal was too high, however, for it to be melted in any material known at the time (graphite was ruled out for chemical reasons), and therefore a special method had to be devised for obtaining tungsten in the form of wire. For this purpose a very fine powder of tungsten was mixed with an organic binder to form a paste which was "squirted" through fine holes in diamond dies. After pre-heating in an inert gas to remove the organic binder, followed by heating to a very high temperature (the preparation process), filament wire with a bright metallic surface was obtained. The filaments were put on special mounts and sealed in glass bulbs. The lamp so produced, which came out in about 1906, was called the "squirted"-tungsten-filament lamp.

This lamp was a very considerable improvement

*) Formerly with Philips, now in retirement.

[1]) See N. A. Halbertsma, The birth of a lamp factory in 1891, Philips tech. Rev. **23**, 222-236, 1961/62.

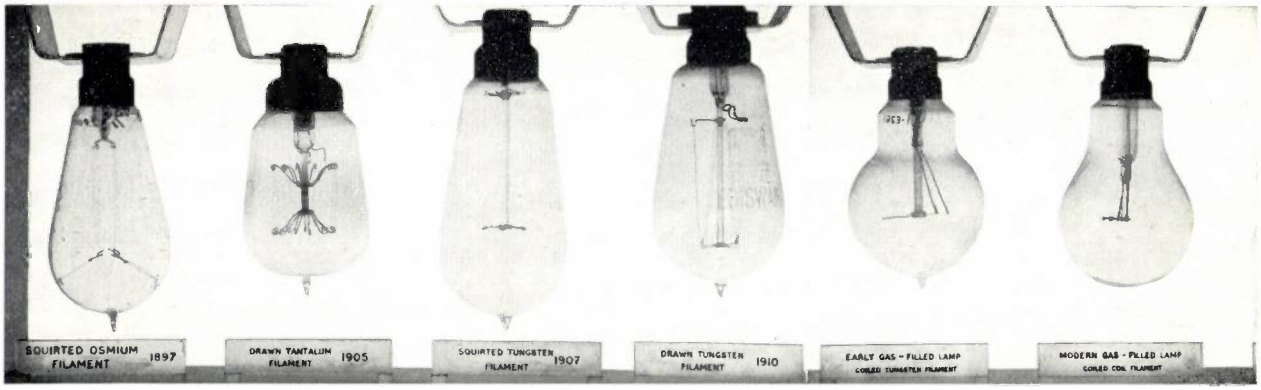


Photo Science Museum, London

Fig. 1. Six metal-filament lamps from the years 1897 to about 1937.

on the carbon-filament lamp, which therefore gradually disappeared from the market, although it continued to be used for years in places where the lamps were subject to severe vibrations. The reduction of the power consumed, from 3.5 W per candle in the carbon-filament lamp to about 1 W in the tungsten lamp, quickly proved decisive both as regards the potential uses of the electric lamp and the spread of power stations. The method of manufacturing the new lamps, however, more resembled laboratory work than factory production, and the fact that they were not well able to withstand shocks and transportation proved to be a drawback. Every possible endeavour was therefore made to find a method of making stronger tungsten filaments, viz, *by drawing*.

The first to succeed was the American Coolidge in 1908. In his process the tungsten powder was pressed into thin bars, which were pre-heated to make them conductive and to give sufficient coherence for handling, subsequently further heated in an inert gas to just below their melting point, and then machine-hammered white-hot (swaged) into thin rods (*fig. 2*). These operations made the material

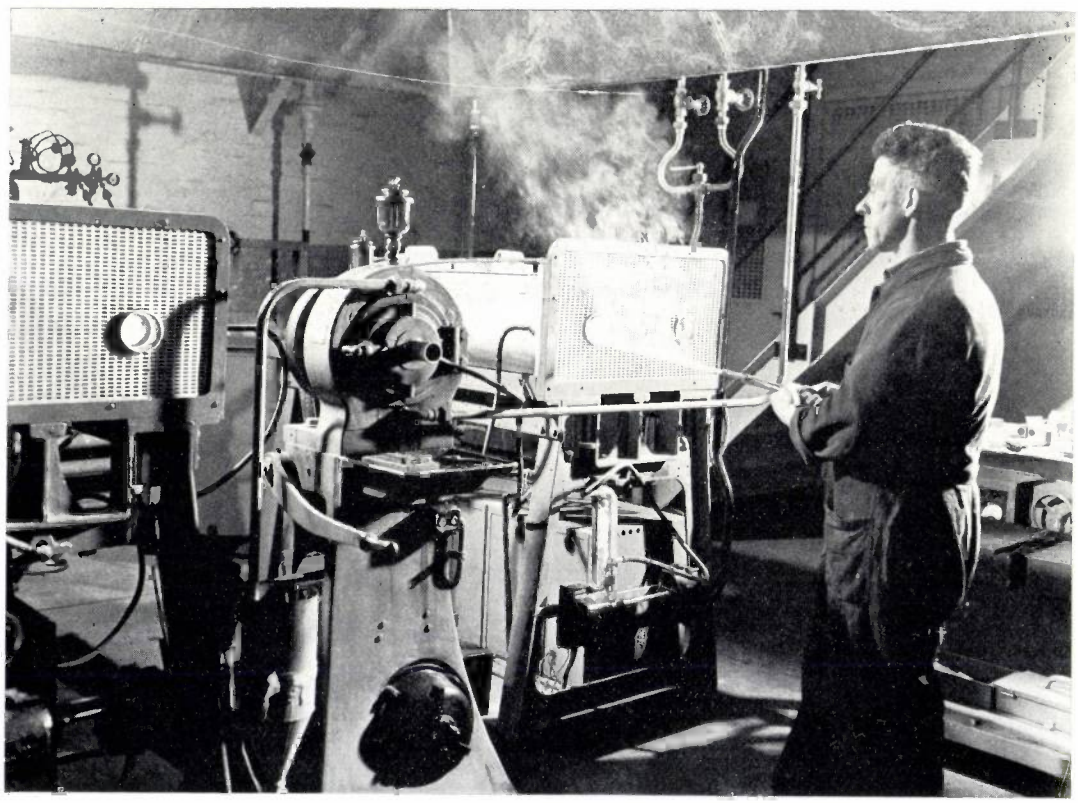
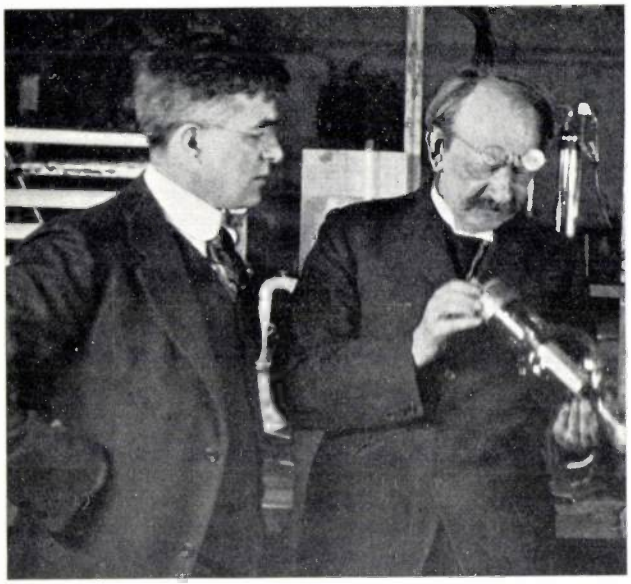


Fig. 2. Ductile tungsten wire is made by machine-hammering (swaging) sintered tungsten bars to increase their density after which they are drawn to the required thickness through hard-metal or diamond dies. The photograph shows a swaging machine, with the tungsten bar being introduced manually after having first been raised to a very high temperature in the adjoining furnace; after some passes through this machine, the bar is passed through an automatic swaging machine.



Volume XXXIV. *June, 1912.* *No. 6*

THE

PHYSICAL REVIEW.

CONVECTION AND CONDUCTION OF HEAT IN GASES.

By Irving Langmuir.

Part I. Historical.

THE loss of heat by convection from a heated body has apparently always been looked upon as a phenomenon essentially so complicated that a true knowledge of its laws seemed nearly impossible. A. Oberbeck[1] gives the general differential equations for this problem but finds it impossible to solve them for actual cases. L. Lorenz[2] for the case of vertically placed plane surfaces is able to obtain some approximate

Fig. 3. Irving Langmuir (left), in conversation with Sir. J. J. Thomson, the discoverer of the electron. (The photo, taken in 1923, is by courtesy of General Electric Research Laboratories Schenectady.)

Right, the opening lines of the first of Langmuir's publications that led to the development of the gas-filled lamp.

so ductile that at high temperature it could be drawn into wire of any required thickness, down to the very finest.

The Philips factories, too, very soon adopted this process, and their first drawn-tungsten-filament lamp was made on 5th December 1911. Owing to this great effort the new lamp was brought out by Philips almost simultaneously with those of competitors. In July 1912 the production of squirted-filament lamps was stopped altogether, and from then on only lamps with drawn filaments were put on the market.

The new technique using drawn tungsten wire had hardly been introduced in the factory when a new discovery was announced — the incandescent lamp with a gas-filled bulb. That was at the beginning of 1913, now half a century ago. As the new lamp consumed about half a watt per candle, it soon became fairly generally known as the "half-watt lamp". The credit for the invention of this lamp was due to the distinguished physicist Irving Langmuir, who was working in the laboratories of the General Electric Company in Schenectady (U. S. A.); see *fig. 3*.

At first the invention related only to large lamps, of 600 to 3000 candles. The gas filling was not yet suitable for lamps of lower power, which claimed the lion's share of the production of the lamp factories. Intensive research throughout the world, however, enabled the major categories of these lamps to benefit from the same principle within a few years.

The steps that led to this invention, and the research which fundamentally widened its potentialities, may still be regarded as a classical example of applied physics. The fact that this year marks the 50th anniversary of the gas-filled lamp has prompted us to review this interesting work once again. We shall first consider the invention itself, and then trace the subsequent development.

Langmuir's invention

A critical study of experiments carried out by Nernst concerning the formation of nitric oxide on an incandescent wire in air [2]), induced Langmuir to investigate the loss of heat by convection from a wire heated to incandescence in a gas.

The attraction of burning the filament in an inert gas (i.e. one which would not react with the white-hot tungsten) instead of in a vacuum as done previously, was that the surrounding gas considerably slows down the evaporation of the tungsten which is responsible for the blackening of the bulb. This made it possible, while maintaining the same useful life, to heat the filament to a very much higher temperature, the higher the filament temperature the better being the conversion of the electrical energy into light. With a gas filling, however, heat is lost by conduction via the gas. If nothing is done about this, it will completely offset the gain of better energy conversion. It was precisely the object of Langmuir's investigation to reduce the heat losses

[2]) W. Nernst, Chemisches Gleichgewicht und Temperaturgefälle, Festschrift L. Boltzmann, published by Barth, Leipzig 1904, pp. 904-915.

caused by the gas filling. His results appeared in a number of now famous publications [3]) (see fig. 3). We shall here very briefly summarize the results of his experiments and theoretical work [4]).

The viscosity of a gas increases with increasing temperature. In the immediate neighbourhood of an incandescent body the viscosity is so high even that a gas no longer flows. For quantitative purposes it is convenient to assume that every incandescent body in a gas atmosphere is surrounded by a *stationary layer of gas*, whose outer surface has roughly the temperature T_1 of the ambient atmosphere, and inner surface the temperature T_2 of the incandescent body.

On the basis of this hypothesis, in which the incandescent filament loses its heat, apart from radiation, solely by conduction through the stationary layer of gas, Langmuir found that the heat loss W per unit length of filament (in W/min) can be defined by the formula:

$$W = \frac{2\pi}{\ln \frac{b}{a}} (\varphi_2 - \varphi_1), \quad \ldots\ldots (1)$$

where

$$\varphi_i = 4.19 \int_0^{T_i} k \, dT. \quad \ldots\ldots (2)$$

Here a is the diameter of the filament, b the diameter of the cylindrical, stationary gas layer and k the coefficient of thermal conductivity of the gas (in cal/m degree s). For calculating the factor

$$\frac{2\pi}{\ln \frac{b}{a}}, \quad \ldots\ldots (3)$$

occurring in (1) and usually denoted by s, which contains the unknown diameter b of the stationary gas layer, Langmuir gave the formula:

$$\frac{a}{B} = \frac{s}{\pi} e^{\frac{-2\pi}{s}}, \quad \ldots\ldots (4)$$

where B is the thickness of the stationary layer of gas produced on an incandescent *flat plate* in the same gas. *Fig. 4* shows a plot of s versus a/B.

The value of B was of fundamental importance for the later practical application of the results. Langmuir showed that B is proportional to the viscosity of the surrounding gas, inversely proportional to the density of the gas, inversely proportional to the 0.75 power of the gas pressure, and finally roughly proportional to the absolute temperature of the gas. At values of B and of the filament diameter a likely

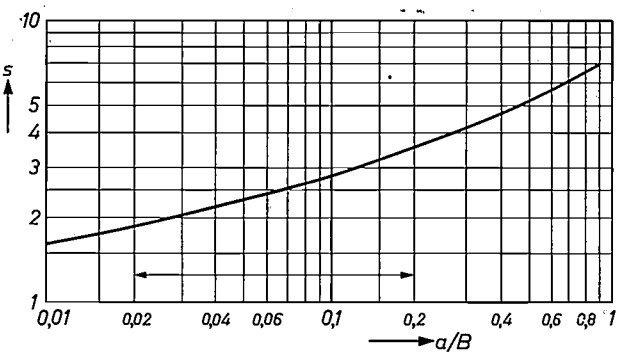


Fig. 4. Graphic representation of equation (4). In practice a/B varies only within the limits indicated along the abscissa.

to be encountered in practice, a/B is found to lie between 0.02 and 0.2. (We shall presently see why the term "filament" is now used and not wire.) In this range of values the curve in fig. 4 can be represented with reasonable accuracy by the equation:

$$\frac{2\pi}{\ln \frac{b}{a}} = s = C \left(\frac{a}{B}\right)^{0.3}, \quad \ldots\ldots (5)$$

C being a constant. The usefulness of this formula was later confirmed by numerous experiments.

For a filament of length l and diameter a (both in mm) the total heat loss W_g (in watts) is now:

$$W_g = C \, l \left(\frac{a}{B}\right)^{0.3} (\varphi_2 - \varphi_1). \quad \ldots (6)$$

It can be seen from this that the heat losses to the gas are primarily determined by the *length* of the filament, while its diameter is of subordinate influence. For an ambient temperature T_1 of 300 °K and a filament temperature T_2 between 2500 and 3300 °K, we can express $\varphi_2 - \varphi_1$ as a function of T_2 for the gases nitrogen, argon and krypton by:

$$\varphi_2 - \varphi_1 = \alpha \left(\frac{T_2}{2700}\right)^{\beta}, \quad \ldots\ldots (7)$$

in which the constants α and β have the values stated in *Table I*. *Fig. 5* gives a graphic representation of equation (7).

From the foregoing it may be concluded that it would be advantageous to have a *short, thick* filament, and from equations (6) and (7) one would then be able to calculate the heat losses and from this the luminous efficiency. But if one wishes to make an

[3]) I. Langmuir, Convection and conduction of heat in gases, Phys. Rev. **34**, 401-422, 1912; also Proc. Amer. Inst. Electr. Engrs. **31**, 1011-1022, 1912. Idem, Convection and radiation of heat, Trans. Amer. Electrochem. Soc. **23**, 299-332, 1913.

[4]) Many years later, experiments were done to determine the influence of deviations from certain simplifying assumptions introduced by Langmuir; see I. Brody and F. Körösy, J. appl. Phys. **10**, 584, 1939. Further: W. Elenbaas, Physica **4**, 761, 1937 and **6**, 380, 1939.

Table I.

Gas	a in W/min	β	atomic weight
Nitrogen	0.170 *)	1.65	14
Argon	0.122	1.65	40
Krypton	0.073	1.65	83

*) This was calculated without taking into account the dissociation of the nitrogen molecules N_2. Experiments show that the value, accounting for dissociation, is about 0.22.

incandescent lamp for a given voltage and power, the length and diameter of the filament are already fixed. This brings us to the crux of the whole problem. After filling in (3), we can write equation (4) in the form [5]):

$$\frac{b}{a}\ln\frac{b}{a} = 2\left(\frac{a}{B}\right)^{-1}. \quad . \quad . \quad . \quad . \quad . \quad (8)$$

Since a/B, as mentioned, lies between 0.02 and 0.2, we see that b/a, the ratio of the diameter of the stationary gas cylinder to the diameter of the filament, is for all practical purposes much larger than 1 (see *fig. 6*). If we wind a long thin wire into a short helix, so that the pitch of the successive terms is about 1.5 times the wire diameter, then the stationary gas layers of the separate windings completely *overlap* each other. When coiled in this way the long thin wire thus behaves, as far as heat transfer to the gas is concerned, like a short cylindrical incandescent body with a diameter equal to the outside diameter of the coil. A simple calculation, using formula (6), shows that this makes it easily possible to reduce the heat loss to the gas to 15%, which largely overcomes the drawbacks of the gas filling — at least for lamps of high wattage. For smaller lamps, as will appear in the following, the balance was at first still unfavourable.

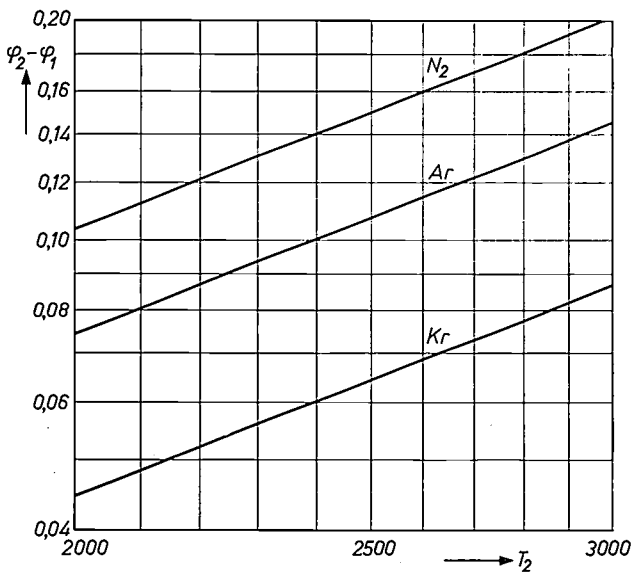


Fig. 5. Plot of the quantity $\varphi_2-\varphi_1$ in Langmuir's heat-loss equation (eq. 1) versus the filament temperature T_2 in °K, representing the theoretical values for different gases (in the case of N_2 no acccount is taken of dissociation).

[5]) Langmuir really gives the relation:

$$b\ln\frac{a}{b} = 2B,$$

which is not, however, as clear as (8).

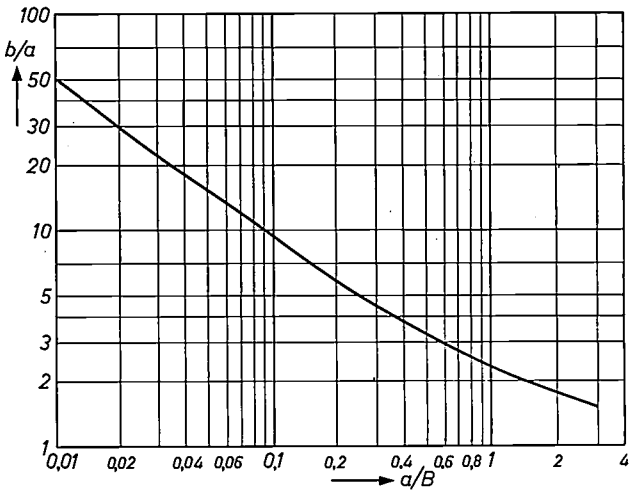


Fig. 6. Ratio of the diameter b of the stationary gas layer to the diameter a of the (cylindrical) incandescent filament, as a function of a/B, calculated using Langmuir's equation (8).

Realization of the invention

To put Langmuir's invention into practice it was necessary to coil tungsten wire into a close helix. This called for ductile wire, so that Coolidge's process for drawing tungsten wire had arrived just in time.

Production brought other problems, however, the first among which was the blackening of the bulb — the very effect the gas filling was meant to overcome.

Years of experience had already been gained in combating the blackening of incandescent lamps. In the manufacture of carbon-filament lamps it had been discovered that the bulbs, and the glass vacuum system used to evacuate the lamps, should be thoroughly dried. This was done by heating the bulbs under the pump hoods to several hundred degrees centigrade and removing the liberated water vapour with phosphorus pentoxide. The residual water vapour and the remaining air in the bulb were removed after seal-off: for this purpose a small quantity of red phosphorus was placed in the exhaust stem and heated to evaporation while the carbon filament was burning.

When the change-over was made from the carbon-filament lamp to the tungsten-filament lamp, bulb blackening sometimes occurred to a very serious

extent. It was first thought that the black deposit consisted of carbon originating from the organic binder used when making the squirted tungsten filaments; later it was found that the deposit was not carbon but tungsten. Since the blackening varied from lamp to lamp, it was probable that it was not only attributable to the normal evaporation of tungsten but also to a substance present in some lamps and not in others. It was assumed that water vapour was again the culprit, apparently somehow being able to accelerate the transport of tungsten from the filament to the wall. When the temperature under the pump hoods was appreciably increased, so as to release the vapour still present on the bulb wall, and when this vapour was then exhausted, the irregular blackening was in fact no longer found.

With the gas-filled lamp the situation as regards blackening is in itself much more favourable than with the vacuum lamp: the hot gas near the filament — outside the stationary layer — rises and carries with it the evaporated tungsten, which is therefore mainly deposited on the bulb wall directly above the filament. In lamps mounted in a hanging position the deposit thus forms in the neck, where it is least troublesome. On the other hand, all glass parts of the lamp, particularly on the inside are heated by the hot gas to much higher temperatures than in the vacuum lamp. This apparently was the reason why in the early attempts at production residual water vapour was still being released during life, resulting once again in excessive blackening. It was now no longer sufficient simply to raise the tempera-

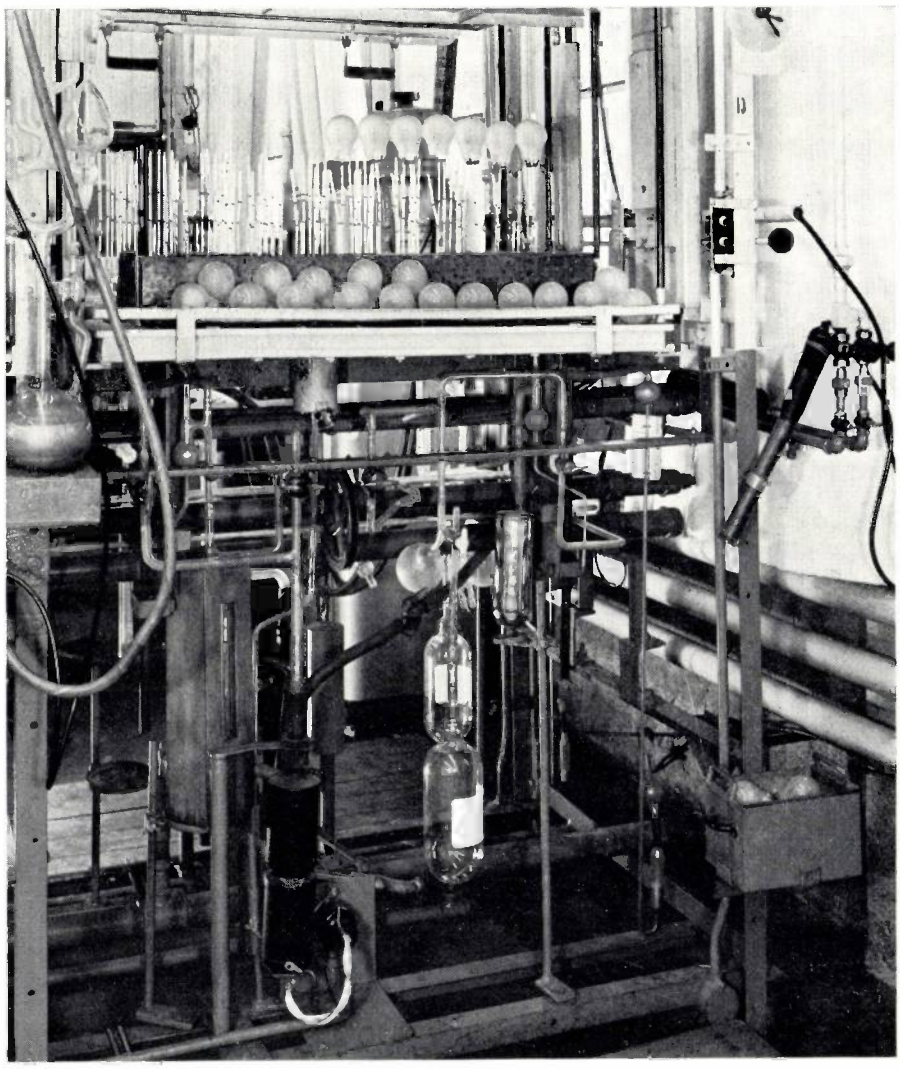


Fig. 7. When evacuating the lamp bulbs on the pumps the last traces of water vapour are removed by a liquid-air trap in the pump line (see the Dewar flask roughly in the middle of the photograph).

ture under the pump hoods during evacuation. A better means of drying had to be found. Only when liquid air was adopted as the drying agent was it possible to reap the benefits of the gas filling, which are due to the reduction of the rate of evaporation of the tungsten (*fig. 7*).

Blackening is still one of the fundamental problems in the construction of incandescent lamps — particularly in connection with bulb dimensions, as we shall presently see. This is demonstrated by the recent development of iodine incandescent lamps, where the problem is tackled again [6]).

[6]) See e.g. J. W. van Tijen, Philips tech. Rev. **23**, 237, 1961/62.

On 19th November 1913 the Philips factories announced that they were now able to supply gas-filled lamps of 600 to 3000 candle-power (*fig. 8*).

NAAMLOOZE VENNOOTSCHAP

PHILIPS' METAAL-GLOEILAMPENFABRIEK.

TELEGRAM-ADRES: „META EINDHOVEN"
A B C CODE 4e EN 5e UITGAVE.
LIEBER'S CODE.
MASTER CODE.
TEL. INTERC. No. 92.
In uw antwoord te verwijzen naar:

Gezien 22 NOV 1913
Beantwoord

Eindhoven, 19 November 1913. (Holland.)

PRIJZEN

der

PHILIPS „HALFWATT" LAMPEN.

50- 70 Volt	300 Watt	600 N.K.	Fl. 8.75 p. stuk	passend in Armaturen P1 & P2
50- 70 / 95-135	500	1000 N.K.	10.50	
50- 70 Volt / 95-135 / 200-250	1000 Watt / 1500	2000 N.K. / 3000 N.K.	Fl.15.75 p. stuk / 21.--	passend in Armaturen PH1 & PH2

ARMATUREN.

P1	zonder reflector bestemd voor	600-1000 N.K.	Fl. 11.50 p. stuk.
P2	met	600-1000	12.50
PH1	zonder	2000-3000	14.50
PH2	met	2000-3000	15.50

De armaturen zijn zwart geëmailleerd met helderen of melkglasballon.

Levering van lampen en armaturen FRANCO-HUIS incl. verpakking.

ONDERDEELEN.

heldere of melkglasballon	No:L1 voor armaturen	P1 & P2	230 m/m door-	Fl. 1.30 p.st.	
	No:L2	PH1 & PH2	300 snede	1.75	
reflector	No:S1	P1 & P2	450	1.10	
	No:S2	PH1 & PH2	550	2.--	

Onderdeelen af fabriek exclusief verpakking.

Transportbreuk wordt uitsluitend vergoed bij franco terugzending der lampen onmiddellijk na ontvangst.

Fig. 8. Price quotation of N.V. Philips' Metaal-gloeilampenfabriek, dated 19th November 1913, with the first offer of gas-filled lamps. As can be seen, the lamps were rather expensive: approx. Fl. 10 (≈ £ 1 today) per lamp. The lamps for 2000 and 3000 candles could be made for voltages up to 250 V, lamps for 600 candles only for the lower voltages of 50 to 70 V.

This was hardly a month after Langmuir and J. A. Orange had presented a paper in New York to the Americal Institute of Electrical Engineers, in which they reported on the realization of the gas-filled lamp [7]).

The lamps of the candle-powers mentioned were made for the lighting of streets and large enclosed

[7]) I. Langmuir and J. A. Orange, Tungsten lamps of high efficiency, Proc. Amer. Inst. Electr. Engrs. **32**, 1893-1926, 1913.

spaces; they were competing here with the carbon-arc lamp, which was in use for these purposes. Arc lamps burnt very irregularly and called for a great deal of maintenance, so that it is not surprising that from then on they were gradually superseded by the gas-filled lamp.

The fact that the gas filling was at first used only in high-power lamps may be understood as follows. For a higher power at a given voltage, a thicker and longer filament is needed. Such a filament can be wound on an appreciably thicker mandrel than the thin wire for a smaller lamp without the coil becoming too limp. Consequently the length of the coil in both cases can be roughly the same, only the thickness of the filament being greater. Since, however, the heat transfer to the gas increases, as appears from formula (6), by only the 0.3 power of the filament thickness, this loss is of much less importance in large lamps than in small ones.

For a given power at a *lower voltage* a thicker and shorter filament is needed. For this reason Philips introduced in 1914 lamps of 100 candles for a voltage of 14 V. Here the filament could again be coiled and a gas filling used with advantage. A gas-filled lamp of low power was thus obtained, but a transformer was necessary in order to use the lamp with the normal mains of 220 or 110 V. In 1914 part of the Kerkstraat in Amsterdam was lit by 21 lamps of this kind, each lamp being provided with a 220/14 V step-down transformer (see *fig. 9*). Although the Municipal Electricity Corporation was well satisfied with this installation (see the extract from a report reproduced in *fig. 10*), it was not a satisfactory solution for private users.

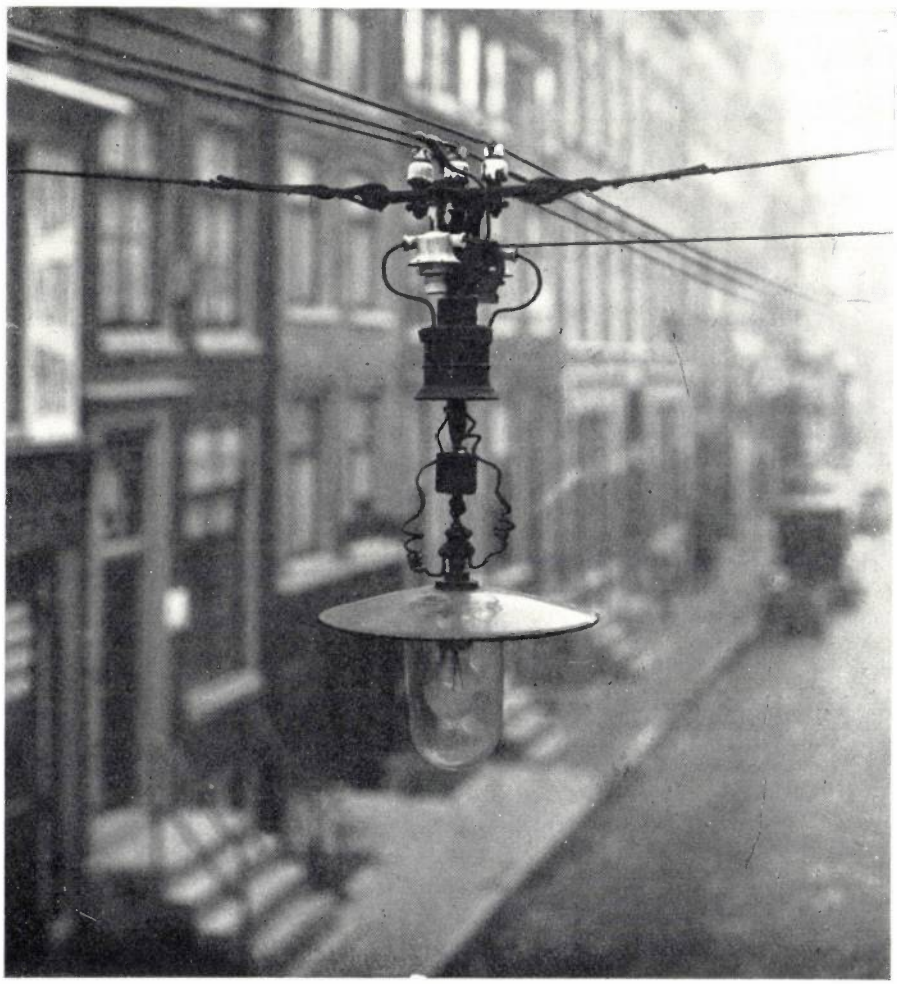
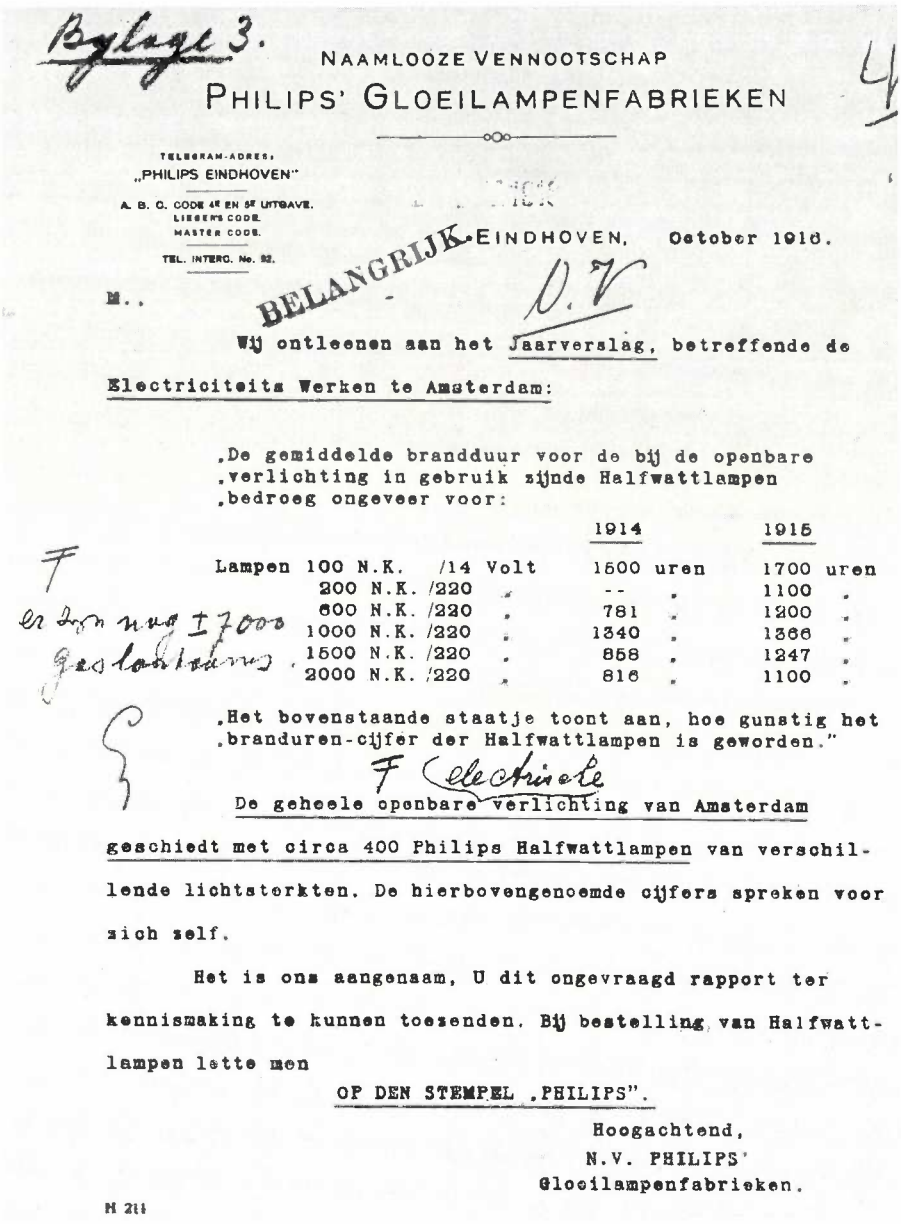


Fig. 9. Fitting containing a 100-candle gas-filled lamp, part of the lighting installation in an area of the Kerkstraat in Amsterdam in 1914. Each lamp required its own step-down transformer for 220/14 V. (Photograph by courtesy of the editor of "De Koppeling"; see also that journal, vol. **8**, 149, 1953.)

Again in 1914, Philips announced a 200-candle lamp for 220 V and a 100-candle lamp for 110 V, and as early as 1st November of the same year the price of these lamps was drastically lowered (*fig. 11*).

This rapid improvement was partly due to the introduction of argon for the gas filling instead of the nitrogen originally employed. It was evident that gases of greater molecular weight would be more suitable because of their lower coefficient of difference between monatomic argon gas and diatomic nitrogen gas does not seem very considerable (respective molecular weights 40 and 28), but an undesirable effect of nitrogen is that it dissociates at high temperature, so that in fact nitrogen compares



Bylage 3.

NAAMLOOZE VENNOOTSCHAP
PHILIPS' GLOEILAMPENFABRIEKEN

4

TELEGRAM-ADRES:
„PHILIPS EINDHOVEN"
A. B. C. CODE 4e EN 5e UITGAVE.
LIEBER'S CODE.
MASTER CODE.
TEL. INTERC. No. 92.

BELANGRIJK. EINDHOVEN, October 1916.

N.V.

M..

Wij ontleenen aan het Jaarverslag, betreffende de Electriciteits Werken te Amsterdam:

„De gemiddelde brandduur voor de bij de openbare „verlichting in gebruik zijnde Halfwattlampen „bedroeg ongeveer voor:

		1914	1915
Lampen	100 N.K. /14 Volt	1500 uren	1700 uren
	200 N.K. /220 „	-- „	1100 „
	600 N.K. /220 „	781 „	1200 „
	1000 N.K. /220 „	1340 „	1366 „
	1500 N.K. /220 „	858 „	1247 „
	2000 N.K. /220 „	816 „	1100 „

F
er zijn nog ± 7000 gaslampen.

„Het bovenstaande staatje toont aan, hoe gunstig het „branduren-cijfer der Halfwattlampen is geworden."

F Electriciteit

De geheele openbare verlichting van Amsterdam geschiedt met circa 400 Philips Halfwattlampen van verschillende lichtsterkten. De hierbovengenoemde cijfers spreken voor zich zelf.

Het is ons aangenaam, U dit ongevraagd rapport ter kennismaking te kunnen toezenden. Bij bestelling van Halfwattlampen lette men

OP DEN STEMPEL „PHILIPS".

Hoogachtend,
N.V. PHILIPS'
Gloeilampenfabrieken.

H 211

Fig. 10. Copy of an extract from an annual report for 1916 relating to the Amsterdam electricity works. The extract shows that in 1915 Philips were already supplying 200-candle (N.K.) gas-filled lamps for 220 V, which lasted for 1100 hours. The report also states that the complete (electric) lighting of Amsterdam was done with about 400 Philips half-watt lamps. The handwritten part adds that about 7000 gas lamps were also used.

thermal conductivity k (see equation 2). Subsequent investigations, including work done in the Philips Research Laboratories, which had meanwhile been established [8]), showed that a gas of greater molecular weight has a further advantage in that it reduces even more the rate of evaporation of tungsten. The even more unfavourably with argon than was predicted in the theory (see the values of α in Table I). The use of argon therefore considerably improved the heat balance.

As air contains a relatively large percentage of argon (about 1%), it was possible to use this inert gas on a large scale. The argon was supplied to Philips by a German firm, "Gesellschaft für Lindes

[8]) E. Oosterhuis, Chem. Weekbl. **14**, 595, 1917. See also W. Geiss, Philips tech. Rev. **6**, 334, 1941.

EINDHOVEN, 1 November 1914.

Belangrijke Prijsverlaging.

Philips' „1/2 Watt" Lampen van 100 Kaarsen

voor Winkel-, Etalage-, Restaurant- en Huisverlichting.

M

Door de zeer groote vraag naar PHILIPS' 1/2 WATT LAMPEN in kleine kaarssterkten, zijn wij tot massa-fabricage kunnen overgaan en wenschen wij de daaraan verbonden voordeelen in den vorm eener belangrijke prijsverlaging den verbruikers ten goede te doen komen.

De prijs der 100 N.K. lampen is gebracht van f 4,20 op **f. 2,70 per stuk.**

Onderstaande stroomberekening bewijst Uw voordeel bij het gebruik van deze lampen.

100 KAARSEN $\frac{125 \text{ VOLTS}}{110 \text{ VOLTS}}$ f 2,70 per stuk.

De stroomkosten van 2 metaaldraadlampen 110 of 125 Volts, 50 Kaarsen bedragen bij een tarief van 20 ct. per K.W.U. en een gemiddelden brandtijd van 1000 uren per seizoen en per lamp:

1000 × 2 × 50 × 1 1 . . .	= 110 K.W.U.
ad Fl. 0,20	= Fl. 22,—
Lampenverwisseling na gemiddeld 1000 uren: 2 × Fl 0.55 . .	= „ 1,10
Totale verlichtingskosten . .	Fl. 23,10

De stroomkosten van EENE „1/2 Watt" lamp, 110 of 125 Volts, 100 Kaarsen bedragen bij een tarief van 20 ct. per K.W.U. en een gemiddelden brandtijd van 1000 uren per seizoen en per lamp:

1000 × 100 × 0,6 . . .	= 60 K.W.U.
ad Fl. 0,20	= Fl. 12,—
Lampenverwisseling na gemiddeld 800 uren: 3/4 × Fl. 2.70	= „ 3.37
Totale verlichtingskosten	Fl. 15,37

EENE TOTALE BESPARING DUS VAN Fl. 7,73.

Spaart dus stroom en geld en vervangt Uwe metaaldraadlampen door:

PHILIPS' „1/2 WATT" LAMPEN van 100 KAARSEN.

Levering geschiedt uitsluitend door bemiddeling van H.H. Installateurs en Wederverkoopers.

Hoogachtend

N.V PHILIPS'

Metaal-Gloeilampenfabriek

Fig. 11. Announcement of 100-candle gas-filled lamps at substantially reduced prices in November 1914: price of a 100-candle lamp dropped to Fl. 2.70.

Eismaschinen" of Höllriegelskreuth (near Munich), the gas being a byproduct in the preparation of oxygen and nitrogen by the fractional distillation of liquid air. The outbreak of the first world war, however, quickly put an end to these supplies. Fortunately, a member of Philips' staff at that time had acquired considerable experience in the liquefying of inert gases, and within a remarkably short time the Philips factories were consequently able to build and operate their own fractional distillation equipment, thus ensuring sufficient argon for their requirements.

An important and at first sight perhaps unexpected consequence of the gas filling was that lamps could be made much *smaller* for a given power (*fig. 12*). This is bound up with the fact, already mentioned, that the evaporating tungsten is now carried

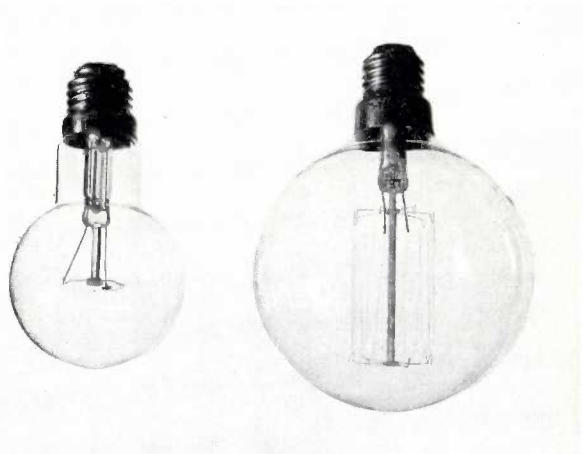


Fig. 12. The compact form of the filament and the favourable distribution of the gradually forming black tungsten deposits made it possible to produce gas-filled lamps (left) with a much smaller bulb than vacuum lamps of the same power.

upwards by the rising gas, and settles almost entirely in the upper parts of the bulb. In the vacuum lamp the tungsten is deposited all over the bulb wall, and therefore this wall must be given a large surface area to ensure that the tungsten deposit remains sufficiently transparent. In this case, in fact, the useful life is governed by the increasing blackening; on the other hand the useful life of the gas-filled lamp is limited by the occurence of thin or weak spots in the filament after a certain amount of tungsten has evaporated, causing the tungsten wire to break. Consequently, the choice of bulb diameter of a gas-filled lamp does not depend on blackening and useful life, but can be made as small as the temperature of the bulb wall permits (the bulb of course becoming hotter as its diameter decreases).

When much later, with a view to the further improvement of luminous efficiency, the even heavier inert gas krypton was considered as the filling gas instead of argon, the high price of krypton made the bulb volume itself an important consideration. On this subject reference may be made to the article by Geiss quoted above [8]). Similarly, the change brought about in the *radiation properties* of the filament as a result of coiling will not be dealt with here. This too was the subject of extensive and much more recent investigations, including joint research carried out by the Osram and Philips factories [9]).

Further developments, notably of the filament

Although the drawn tungsten filament possessed very good mechanical properties and could readily be coiled, an unforeseen difficulty in the early days of the gas-filled lamps was the tendency of the tungsten filament to *sag*. At the very high temperature to which the coil is heated in operation the metal became so soft that the windings of the coil gradually opened out and sagged under their own weight, sometimes even resulting in an almost straight wire. Owing to the lengthening of the filaments which accompanied this sagging, the heat losses increased and the light output therefore dropped sharply. If the advantages of the coiled filament are to be maintained during the whole life of the lamp, the spirals must keep their original shape as far as possible. This is indeed still a problem calling for constant attention, particularly in lamps of smaller wattage [10]).

This problem led to intensive research all over the world to find a filament that would not sag — especially after it was found that tungsten containing certain impurities showed far less tendency to sag than completely pure tungsten.

When pure tungsten is used as the starting material in the manufacture of the filament wire it is

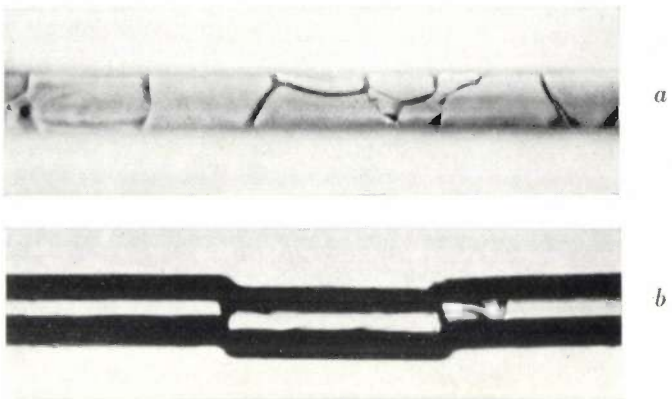


Fig. 13. *a*) Structure of a wire of pure tungsten after recrystallization.
b) "Offsetting" in a tungsten filament, occurring when a lamp having a filament with the structure shown in (*a*) had burnt for some time.

seen after recrystallization, which occurs at a high temperature, to give the wire a structure as shown in *fig. 13a*. After the lamp has burnt for some time, vibrations and the force of gravity cause sliding along the boundaries of the tungsten crystals, producing the effect called "offsetting" (fig. 13*b*). This effect promotes the sagging of the coil. Since it also promotes local irregularities in the filament temperature, it has a disastrous influence on the life of the lamp.

In the early development of the gas-filled lamp the Philips factories prepared tungsten by the "Battersea process". In this process tungstic acid was heated in closed, refractory crucibles at high temperature, about 1200 °C, removing the H_2O and producing a very compact WO_3, which was then reduced to tungsten in the normal way. The WO_3 absorbed impurities from the crucibles, mainly K_2O, SiO_2 and Al_2O_3. The presence of these impurities evidently produced a different recrystallization texture from that in pure tungsten wire. *Fig. 14* illustrates this texture.

The quantity of impurities taken up varied considerably, however. By mixing portions of tungsten from different crucibles it was possible to arrive at a certain favourable quantity. It was fortunate for our work that this insight into the behaviour of tungsten was already available in our laboratories at the very time we were developing the gas-filled

[9]) G. Holst, E. Lax, E. Oosterhuis and M. Pirani, Leuchtdichte und Gesamtstrahlungsdichte von Wolframwendeln, Z. techn. Physik **9**, 186-194, 1928.
[10]) See e.g. E. W. van Heuven, Shock testing of incandescent lamps, Philips tech. Rev. **24**, 199-205, 1962/63 (No. 7).

lamp. The Battersea wire could be used with good results in the gas-filled lamp, and the sagging of the coils was thus kept within reasonable bounds.

Even better results were obtained by applying an American process [11]) in which, before the reduction, a controlled quantity of Na-K silicate was added to the tungstic acid. In this way tungsten wire was

Fig. 14. Recrystallization texture of a tungsten wire containing impurities resulting from the Battersea process.

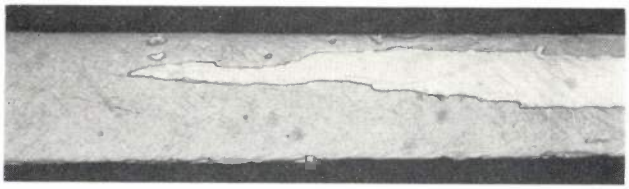


Fig. 15. Recrystallization texture of "doped tungsten", i.e. tungsten containing controlled amounts of additives. The elongated, wedge-shaped crystals give the wire considerable strength and prevent sagging when the wire is coiled.

obtained which, after recrystallization, consisted of long overlapping crystals; see *fig. 15*. This process was soon adopted in the Philips factories. With the new wire all sagging difficulties were overcome, *before* the reduction process — but none of them are entirely satisfactory [12]).

The argon gas offered, as described, the advantage of removing less heat from the filament than nitrogen. A disadvantage, however, was that its breakdown potential was considerably lower than that of nitrogen. Consequently, if the lead-in wires came too close to each other, arcing occurred between them, prematurely ending the life of the lamp. Fortunately it was found that this effect could be suppressed if a certain percentage of nitrogen was added to the argon (about 10%), provided the argon pressure was not too low. This breakdown problem was primarily encountered in the European countries, where the mains voltage is generally higher than e.g. in the U.S.A.

As time went on, more and more types of lamps with argon filling appeared, and they were produced in very large quantities. Tungsten-lamp manufacture in the 'twenties was therefore characterized by increasing mechanization of the production process.

In the 'thirties a further important improvement was made to the gas-filled lamp. In Europe the standard tungsten lamps for voltages above 200 V have a long filament wire, and even when coiled the filament is still relatively long. With the idea of reducing the cooling area, it was decided to coil the filament doubly (*fig. 16*), thus producing the "coiled-coil" lamp. This resulted in a quite appreciable

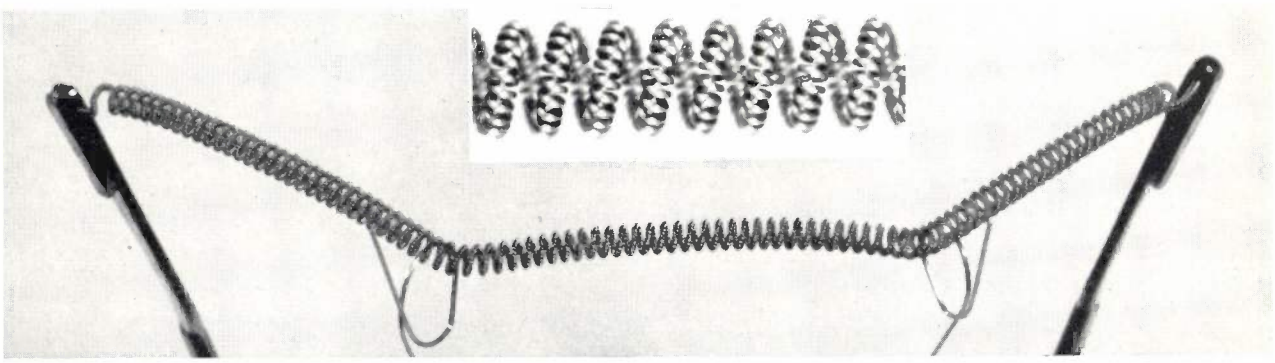


Fig. 16. Coiled-coil filament for a 100 W lamp. The inset shows a piece of the coiled-coil at higher magnification.

making it possible later also to meet the requirements of the coiled-coil lamp (see below).

Many theories have been put forward to explain the action of the additives ("dope") — N.B. added

improvement of luminous efficiency particularly at lower wattages [13]). The coiled-coil lamp for low watt-

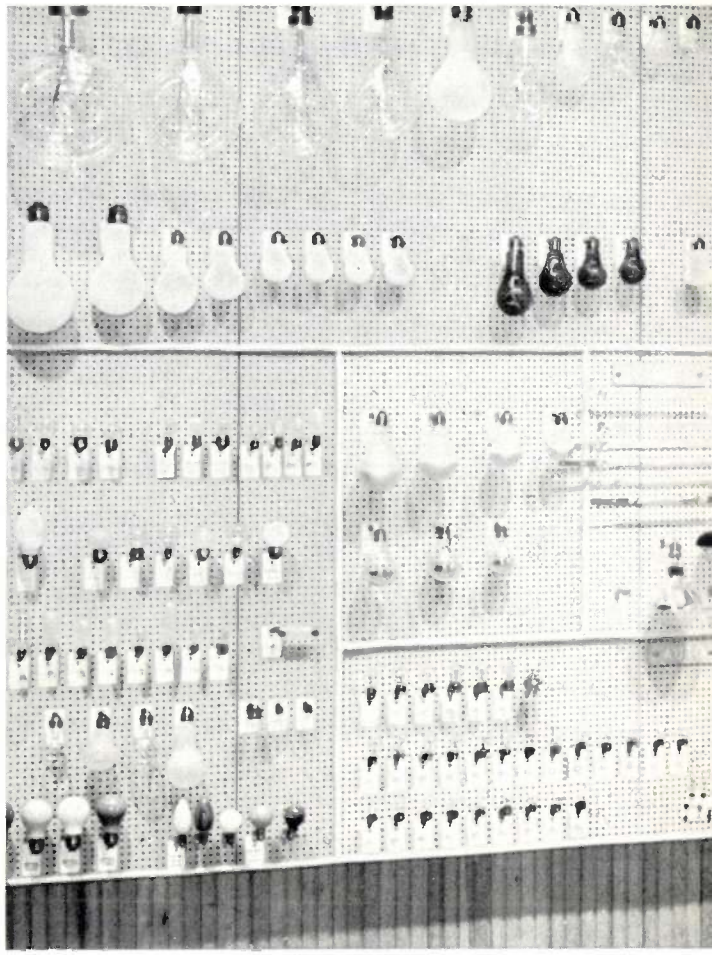
[11]) U.S.A. Patent 1410499, filed Feb. 1917, granted March 1922 in the name of A. Pacz.

[12]) See e.g. J. L. Meijering and G. D. Rieck, The function of additives in tungsten for filaments, Philips tech. Rev. **19**, 109-117, 1957/58.
H. L. Spier, Influence of chemical additions on the reduction of tungsten oxides, thesis Technische Hogeschool Eindhoven, 1961.

ages is in fact a specific European contribution to the development of the incandescent lamp.

The coiled-coil lamp makes even more stringent demands on the gas filling and on the non-sagging properties of the filament than the single-coil lamp. A coiled-coil can only be made with the best non-sagging wire. It is produced by first winding the tungsten wire on a molybdenum wire mandrel of the appropriate thickness, and then winding the coil thus obtained, together with its mandrel, around a second thick molybdenum mandrel. After heating the filament to incandescence for some time to "set" the wire, the two mandrels are removed by dissolving them in a suitable acid. A good solution was found in the Philips factories for the technological problems which this involved.

Finally, to illustrate the achievements to date, *fig. 17* shows the luminous efficiencies (in lumens per watt) of three types of lamps manufactured today: vacuum lamps, single-coil gas-filled lamps, and coiled-coil gas-filled lamps. All three curves relate to lamps for 225 V having a useful life of 1000 hours. It can be seen that the gas filling, in conjunction with a coiled filament, is now used with advantage for lamps of 40 W and higher. Also clearly to be seen is the gain obtained by coiled-coil filaments [13]), particularly at the lower wattages. The name "half-watt" for the two latter categories of lamps is not properly relevant for lamps given on this graph. Broadly speaking a luminous flux of 10 lumens is equivalent to a luminous intensity of one candle. Thus it is only at luminous efficiencies of about 20 lm/W, encountered when extending the graph to 1000 W, that one has half-watt lamps.



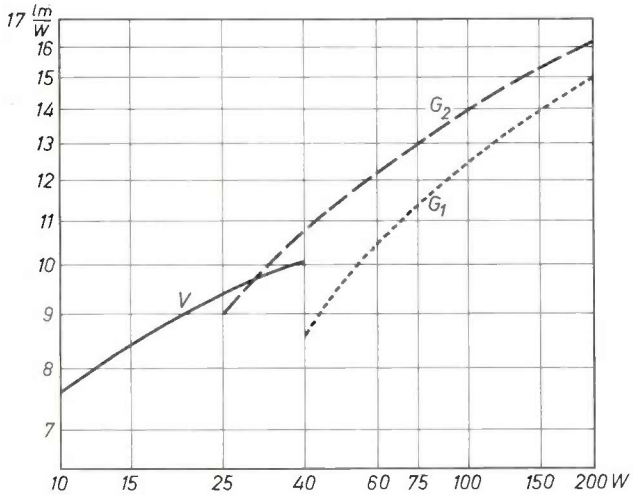


Fig. 17. Luminous efficiency of present-day vacuum lamps (V), of gas-filled single-coil lamps (G_1) and gas-filled coiled-coil lamps (G_2), as a function of wattage.

The combination of a gas filling with a coiled filament proved to be desirable also for tungsten lamps other than those used for general lighting, and in fact nearly all types of vacuum lamps were replaced by gas-filled types. Indeed many kinds of lamp only became possible because coiling allowed the construction of a sufficiently compact filament. One of the most striking examples is the tungsten lamp for film or slide projection. The very compact filament provided the necessary high average luminance, while the possibility of making the bulb very small was essential to the effective design of the optical system and for limiting the size of the whole projector. The same factors were decisive in the development of special lamps for car headlights and of many other similar types. The invention of the gas-filled lamp has therefore contributed in no small measure to the extraordinary diversity of incandescent lamps now available; at the present time, for example, Philips produce some tens of thousands of

[13]) W. Geiss, On the development of coiled-coil lamps, Philips tech. Rev. **1**, 97-101, 1936.

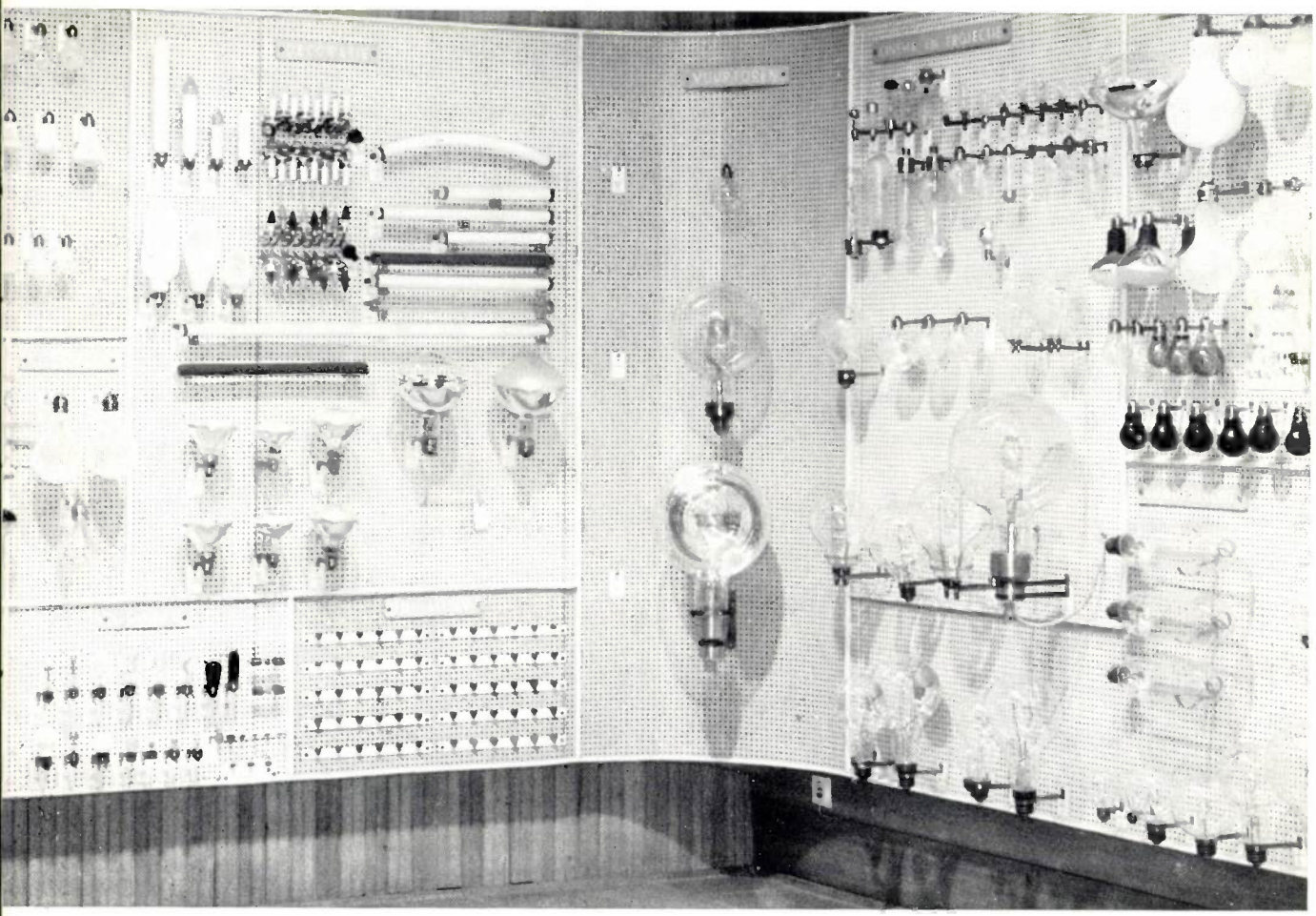


Fig. 18. A small selection from the tens of thousands of types of incandescent lamps nowadays produced by Philips. Most types come into the many categories of "special lamps" (projection lamps, car bulbs, window-display lamps, airfield, sports-field and lighthouse lamps, studio lamps, infrared-heating lamps, signal lamps, bicycle bulbs, etc. etc.); but a large number of types also come into the category of standard lamps for domestic and street lighting, made with numerous variations of wattage, voltage, kind of bulb, etc.

types. To conclude this review, our last figure (*fig. 18*) shows a small selection from this enormous variety.

Summary. The first gas-filled incandescent lamps appeared at the beginning of 1913, now half a century ago. The invention of the gas-filled lamp, which followed from the work of Irving Langmuir, is recalled in this article. After a brief history of the development of the incandescent electric lamp, a short account is given of the theory underlying Langmuir's invention. The problems involved in the manufacture of gas-filled lamps are discussed (named half-watt because the power per candle was reduced from about 1 W to about ½ W — at least for lamps of more than 2000 candles), in particular the blackening of the bulb and the sagging of the coiled tungsten filament. A concise review is given of subsequent developments, which led to the use of inert gases for the gas filling, and to the invention of the coiled-coil filament, which allows the use of a gas filling even in lamps of relatively low wattage.

A DYNAMO FOR GENERATING A PERSISTENT CURRENT IN A SUPERCONDUCTING CIRCUIT

by J. VOLGER *).

621.313.291:537.312.62

In various fields of research a considerable demand has arisen in recent years for stationary magnetic fields of exceptional strength. In solid-state research, for example, it would be useful to be able to study the Hall effect and various resonances in fields of 5 to 10 Wb/m² (50 000 to 100 000 gauss). Nuclear physicists require strong fields for aligning atomic nuclei by the "brute force" method. And finally, investigations in the field of plasma physics, which it is hoped will one day lead to controlled nuclear fusion and to the building of thermonuclear power stations, are confronted with the problem of confining extremely hot, highly ionized gases (plasmas) in a space where they are not in contact with a material wall. It is hoped to achieve this with what are called "magnetic bottles"; in view of the elevated temperature of the gas (of the order of 10^7 °C) the magnetic fields required for this purpose must be exceptionally strong, perhaps greater than 10 W/m². In some of the cases mentioned the problem of generating such a strong field is increased because it is required in a large volume, e.g. of several cubic decimetres.

In order to generate such fields using normally conducting coils — coils of course without an iron core — enormous power is needed and hence an enormous cooling capacity, for the power is almost completely converted into Joule heat [1]). The use of *superconducting* coils, to reduce the power required and ease the cooling problem, was not possible until recently because the superconducting state in the materials then known was destroyed by even a fairly weak magnetic field; a coil made of such material therefore reverts spontaneously to normal conductivity if the current through it exceeds a certain value.

The discovery of "hard" superconductors has changed this situation. These materials can be exposed to a very strong field, and in the form of wires they can carry extremely high currents [2]).

A difficulty of working with a superconducting coil is the problem of generating the current, the reason being that the coil has, of course, to be contained in a cryostat. Normally the current source will be outside the cryostat and the current has to be supplied through cables that are not superconducting. This raises cryogenic problems. If relatively thin supply cables are used, too much heat is generated in them; if, to avoid this, thick cables are used, the result is an impermissible leakage in the thermal insulation of the cryostat. It is obvious, therefore, to look for a means of generating the current *inside* the cryostat. If one used for this purpose a current source which is itself superconducting, a "*persistent current*" can then flow in the circuit. The current source then does no more than set this current in motion, after which it can be switched off.

In the Philips Research Laboratories, Eindhoven, an experimental version of such a current source was recently successfully put into operation. It is in fact a kind of dynamo which works on a new principle [3]). Referring to *fig. 1*, we shall briefly describe the essentials of its construction, after which we shall explain the operation.

A major part of the dynamo is the thin lead disc D, which forms part of the superconducting circuit in which the persistent current is generated. The rest of this circuit is formed by the wire W (for the time being we can ignore the coil L in this wire), which is fixed to the rim of the disc at a and to the centre at b. The current is generated by turning the shaft S, thereby causing the bar magnet M — one pole of which is immediately under the disc — to describe a circular path. For the dynamo to function properly, the pole should be close enough to D to enable the magnetic field of M to destroy the superconducting state of the disc in the immediate vicinity of the pole. Further, part of the flux of M must pass through this zone of normal conductivity. This part above M is represented schematically in the figure as a hole (H). As M revolves, the "hole" must revolve with it. The current that now flows in W is in the first instance proportional to the number of revolutions that the shaft S has made. When the shaft is stopped, the current retains the value it has reached at that instant. *Fig. 2* gives a rough sketch of the actual construction.

*) Philips Research Laboratories, Eindhoven.

1) The highest power which can at present be supplied to a normally conducting magnet coil is in the region of 10 megawatts. This can produce, for example, a field of 20 Wb/m² in a coil having an inside diameter of no more than 4 cm.

2) A survey will be found e.g. in R. H. Kropschot and V. Arp, Superconducting magnets, Cryogenics 2, 1-15, 1961.

3) J. Volger and P. S. Admiraal, A dynamo for generating a persistent current in a superconducting circuit, Physics Letters (Amsterdam) 2, 257-259, 1962 (No. 5).

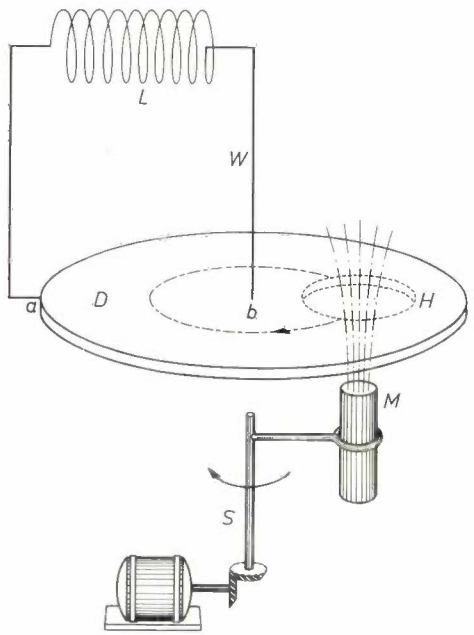


Fig. 1. Schematic diagram of the dynamo for generating a persistent current in a superconducting circuit. *D* lead disc. *W* circuit of "hard" superconducting material, connected to *D* at the rim (*a*) and in the centre (*b*). The shaft *S* mounted below the centre of *D* carries on an arm the bar magnet *M*. The upper pole of *M* is so close to *D* that the field is strong enough to destroy the superconductivity in the zone *H*. Part of the flux of *M* passes through the zone *H*. The rotation of *S* causes *H*, together with the flux, to describe a circular path. Consequently a current flows, as explained in the text, through the circuit formed by *D* and *W*. *L* coil forming part of *W*.

To explain the operation of this device, we can best take as a starting point the property that it is impossible to cause any change in the magnetic flux enclosed by a superconducting ring. If one tries to do this, for example by bringing a magnet near to it, a current starts to flow in the ring that has the effect of exactly cancelling the change of flux produced by the change of position of the magnet. (It is tacitly assumed here that the field of the magnet is not so strong as to interrupt the superconducting circuit.)

It is, however, possible to alter the spatial field distribution inside a superconducting ring. Suppose we have a complicated ring circuit as sketched in *fig. 3a*, and that it has been sufficiently cooled in the presence of the magnet *M* to become superconductive. Owing to the eccentric position of *M*, the flux enclosed by the ring passes mainly through the part *x*; for simplicity we suppose that the entire flux passes through *x*. If we now constrict the hole

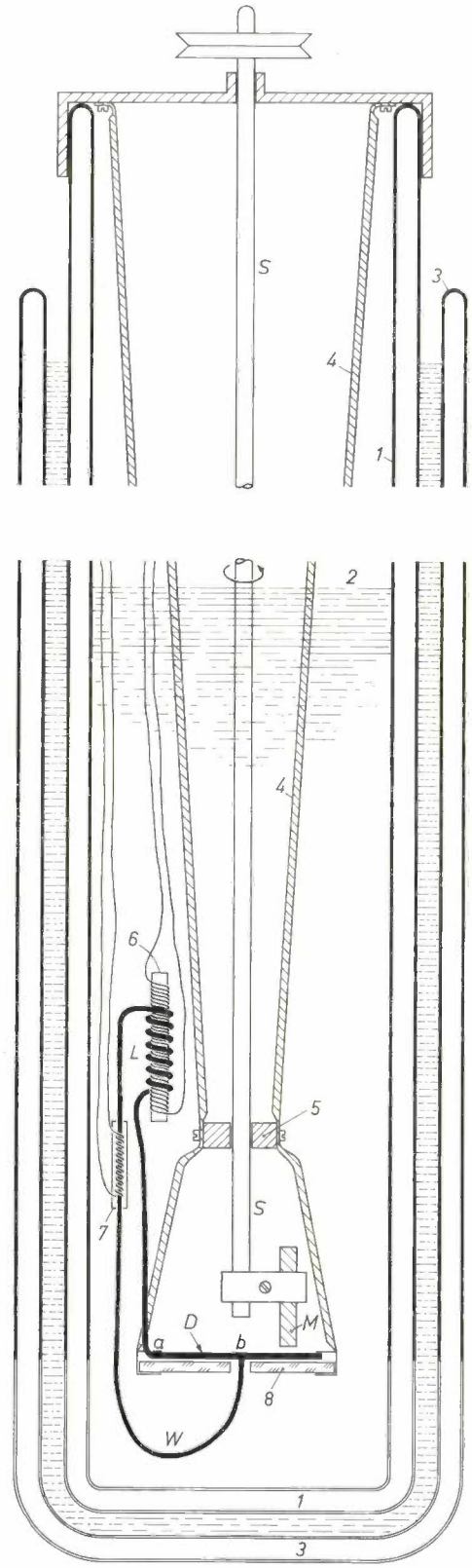


Fig. 2. Sketch showing the construction of the new dynamo and its positioning in a cryostat. The letters have the same meaning as in fig. 1. Other symbols are: *1* Dewar vessel of inner cryostat. *2* surface of the liquid helium. *3* Dewar vessel of outer cryostat; the space between *1* and *3* is filled with liquid nitrogen. *4* tubular rods screwed to the cap, carrying the bearing *5* and the disc *D*.

For measurement of the current in *W*, this circuit is coupled to a ballistic galvanometer via a transformer *6* (of which *L* is the primary). If the current in *W* is interrupted, the galvanometer gives a deflection proportional to that current. The current is interrupted by means of a heating element *7*, with which the superconductivity of *W* can be removed locally. *8* cerium-glass plate (explained later).

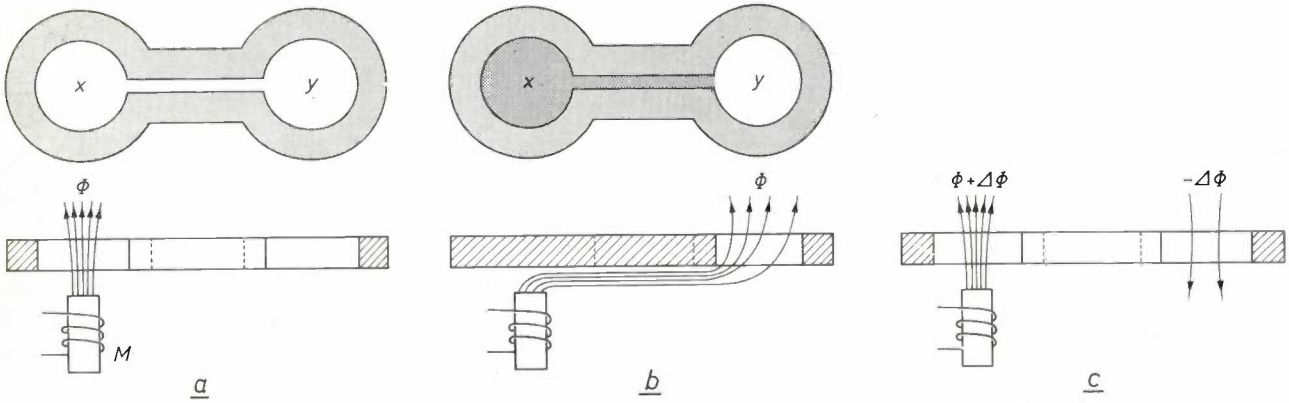


Fig. 3. Illustrating the properties of a superconducting ring circuit of complicated shape, in one part of which a flux is enclosed.

in x, for example by gradually closing it with a superconducting stop, the flux then shifts to part y of the ring (fig. 3*b*). If on the other hand we increase the flux at x by an amount $\Delta\Phi$, e.g. by more strongly energizing the magnet M, the current induced in the ring is such that y encloses a flux $-\Delta\Phi$; the total flux through the ring remains equal to Φ (fig. 3*a*). It should be noted here that in a *soft* superconductor, such as lead, these currents flow in a surface layer. The interior of such a superconductor is always free from currents and magnetic fields.

A further extension of the property of magnetic-flux conservation is that it holds not only for rings but equally for the holes in e.g. a triply or quadruply connected body. We have found in fact that the flux through each hole remains constant when such a body is subjected to a change of shape, even when the holes become unrecognizably deformed in the process.

The operation of the dynamo can now easily be explained by applying the latter property to the complicated superconducting body sketched in *fig. 4a*. This body, in which the circuit of fig. 1 is easily recognized, contains two "holes". The flux contained in both holes is invariant. According to the theory of superconductivity this must be formulated thus: magnetic fluxes are invariant which are contained within the contours *1* and *2* running within the (currentless) interior of the superconductor. Where exactly these contours are drawn is immaterial. For contour *1* the enclosed flux has a finite value, which we again call Φ, and for contour *2* the value is taken to be zero.

If we now displace the hole contained by *1* in the same way as in the dynamo with the zone H of normal conductivity (fig. 1), then contour *2*, in order to pass through superconducting material carrying no current, must alter in shape (fig. 4*b*). Owing to the required invariance, the moving hole continues while in motion to enclose the flux Φ. When the hole has completed one revolution, we have the situation drawn in fig. 4*c*. This differs from the starting situation in that the flux Φ is now also contained within contour *2*. Since, however, the total flux inside *2* must remain zero, this contour will now have to contain elsewhere a flux $-\Phi$, which calls for the flow of a current in the circuit formed by W and D. After two revolutions the latter flux is -2Φ, and the current must now be correspondingly larger, and so on. If a coil is included in the circuit W (fig. 1), a field is then generated in it [4]).

[4]) In this article we have not mentioned variants of the method described. In one variant the plate in which the flux is rotated is smaller than the disc D discussed here, in such a way that during its revolution the flux alternately leaves and re-enters the plate. For the dynamo to work well it is only necessary that the magnetic flux during its passage through the plate should not cause any complete interruption of the superconducting path between the connection points of the coil.

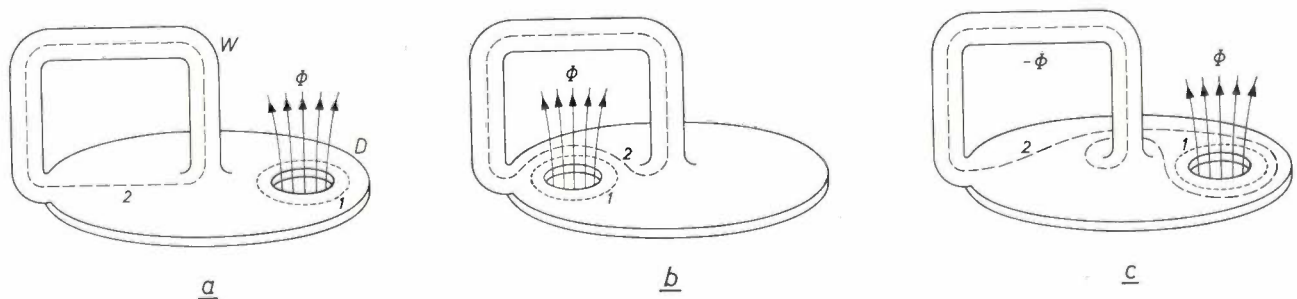


Fig. 4. Illustrating the operation of the new dynamo.

A practical difficulty is encountered if, with a dynamo of the type in fig. 1, it is desired to feed a coil of a large magnet, the difficulty being the high self-inductance of such a coil. In general, when a constant voltage E is applied to a superconducting circuit having a self-inductance L, a current i will flow which increases with time in accordance with the equation $i = Et/L$. The time which, for a given L, elapses before i reaches the required value is thus inversely proportional to E. In order to be used for energizing coils with which strong fields are to be generated, a dynamo working on the new principle must therefore be capable not only of delivering the required current, but should also have not too small an e.m.f. In a set-up as shown in fig. 1, this e.m.f. is proportional to the speed of revolution of the shaft S and to Φ. With our experimental set-up it is not yet possible to achieve a voltage high enough to produce a current of 10 A in a coil of, say, 1 H within an acceptable time. The reason is that the proportionality between the e.m.f. and the speed of revolution is not in practice unlimited: if the shaft speed exceeds a certain value, the e.m.f. gradually moves towards a maximum value. In circuits having a small L, on the other hand, very strong currents (more than 100 A) have been generated in a short time.

The reason why the e.m.f. shows a maximum is that the zone of normal conductivity does not follow the movement of the magnet at unlimited speed. This was observed with the aid of a plate of cerium glass with a reflecting surface on one side, which was mounted immediately under the disc D in the set-up in fig. 2. At low temperatures cerium glass shows a marked Faraday effect (rotation of the plane of polarization of light transmitted through a magnetized medium). When polarized light is directed onto the plate and the reflected light is passed through an analyser, the part immediately under the zone of normal conductivity — which is therefore exposed to the magnetic field of M — is seen to be darker (or lighter) than the rest. When the shaft is rotated fairly fast, the dark patch is seen to acquire a "tail". Upon very fast rotation, the tail fills the complete circumference of a circle and the zone of normal conductivity becomes ring-shaped. With a thin disc this effect is less strong than with a thick one.

The maximum current that can be produced in a coil by this new method of current generation is determined either by the current-carrying capacity of the coil (the "hardness", see above), or by that of the dynamo itself, or by the quality of the junctions. One should therefore try to bring these three roughly into correspondence with one another.

Summary. The extremely strong magnetic fields (of the order of 10 Wb/m^2) needed in various fields of research can now in principle be generated, without requiring enormous power, by making coils from a "hard" superconductor. The current generation in such a coil should preferably take place inside the cryostat. A description is given of the principle of a superconducting dynamo which can generate a persistent current in a superconducting circuit, and an experimental version of such a dynamo is discussed. This consists of a lead disc to which the remaining part of the circuit is connected in the centre and at a point on the periphery. One of the poles of a bar magnet is situated eccentrically under the disc, producing in the disc a small zone of normal conductivity. Part of the flux of the magnet passes through this zone. If the latter is rotated around the centre of the disc, a current is produced in the circuit. The e.m.f. (which governs the speed at which the current can grow in a circuit possessing finite inductance) is in principle proportional to the speed of rotation.

LUMINESCENT *P-N* JUNCTIONS IN GALLIUM PHOSPHIDE

535.376

In *P-N* luminescence, as in all forms of electroluminescence, light is produced by the direct conversion of electrical energy. The effect has been briefly described earlier in this journal [1]). Polycrystalline gallium phosphide was then used for the experiments and the *P-N* junctions needed for the luminescence occurred naturally at the grain boundaries. Thus virtually no control could be exerted over the nature and situation of the *P-N* junctions. It is clear that under such conditions the results obtained were not readily reproducible.

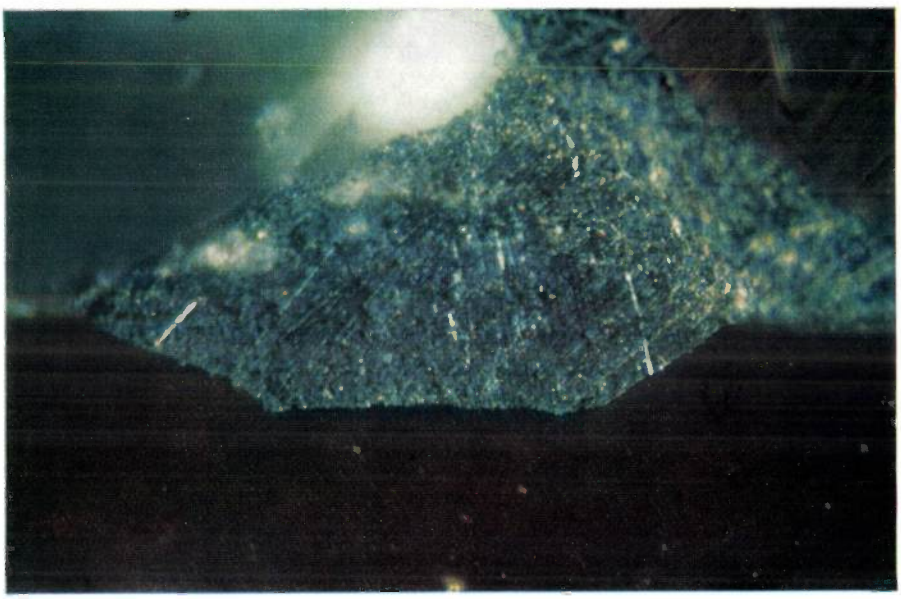


Fig. 1. Cross-section of a *P-N* junction in GaP with Au contact, photographed in polarized light on Kodachrome film (artificial-light reversal film), exposure 30 seconds. The *P-N* junction is roughly 0.5 mm long.

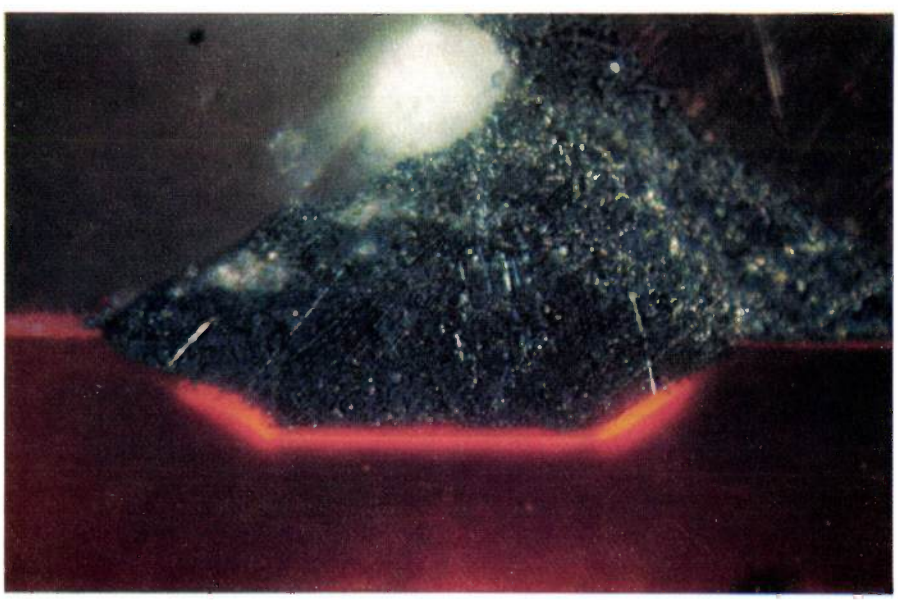


Fig. 2. The same *P-N* junction with forward current flowing (10 mA). Exposure 30 seconds.

In subsequent experiments an attempt has been made, using an alloying process, to replace the random *P-N* junctions by ones whose properties and situation are precisely established by the method of fabrication. This indeed proves possible if two

[1]) H. G. Grimmeiss and H. Koelmans, Philips tech. Rev. **22**, 360, 1960/61.

contacts of e.g. tin and gold (+ 4% zinc) are alloyed into a GaP wafer by a short heat treatment at about 600 °C. During heating the GaP goes into solution in the alloying metal. Most of the GaP thereby dissociates and the phosphorus produced disappears. Upon cooling the wafer, the GaP still dissolved crystallizes out, but because this quantity is very small only small recrystallization zones appear. These are relatively strongly doped with Sn, which is a donor, and with Zn, which is an acceptor. If the GaP wafer is *P*-type (hole conduction) the Sn contact forms a *P-N* junction and the Au contact an ohmic contact. If the GaP is *N*-type (electron conduction) the Sn and Zn then exchange functions, the *P-N* junction now being formed at the Au contact and the ohmic contact at the Sn.

At first we used single-crystal GaP for these experiments. Later, however, we found that similar results could be achieved using polycrystalline wafers obtained from a Ga melt by a special method. Although these wafers are polycrystalline, all the crystallites grow in the [111] direction with a misorientation of less than 0.25 °. For the future characteristics of the wafer as a diode, it is important that the alloying should start from the $(\bar{1}\bar{1}\bar{1})$ crystal face (which is occupied by *P* atoms and can easily be distinguished from the opposite (111) face by etching the GaP wafer in HCl [2])). Only in this case can *P-N* junctions be obtained which mainly run parallel with the surface [3]). Further, these junctions show good electrical properties (reverse current generally less than 10^{-10} A, and sometimes even less than 10^{-13} A, at voltages up to 8 V).

Fig. 1 shows a colour photo of the cross-section of such a *P-N* junction. At the bottom of the photo can be seen the orange *P*-type GaP, above it the Sn contact with the solder, and top right a part of the copper wire conductor. The small recrystallization zone between the Sn and the GaP can be seen as a dark strip; to make the zone more easily visible the photo was taken in polarized light.

To obtain strongly luminescent *P-N* junctions *P*-type GaP was used which, due to doping with certain impurities (activators), emits light mainly of 7000 Å wavelength [4]). In the *N* region of these diodes the donor atoms are considerably in excess of the acceptor atoms in the *P* region, so that when a forward bias is applied the current through the *P-N* junction is caused mainly by electron injection in the *P* region, and it is here that recombination takes place, resulting in the emission of light. *Fig. 2* shows a photo of the same *P-N* junction as in fig. 1, but now with forward current flowing. The forward bias was about 2 V and the current 10 mA. It can be seen that the luminescent zone is indeed in the *P* region.

As mentioned earlier, light-sources of this kind are eminently suited for use in opto-electronic circuit devices [5]). In conjunction, for example, with a photoconductor of cadmium selenide, it is possible to make power amplifiers, light relays, choppers, frequency multipliers and, by suitable combination, flip-flop circuits. The switching time of a light relay designed on this principle is about 1 ms. It is governed solely by the speed of response of the photoconductor, as the light-sources can work with light-pulses of less than 10^{-7} s; the switching time of the diodes in themselves, which depends almost entirely on their *RC* constant, is less than 5×10^{-8} s.

W. GLÄSSER *),
H. G. GRIMMEISS *),
H. SCHOLZ *).

[2]) See Philips tech. Rev. **24**, 61, 1962/63 (No. 2).
[3]) M. T. Minamoto, J. appl. Phys. **33**, 1826, 1962 (No. 5).
[4]) H. G. Grimmeiss and H. Koelmans, Phys. Rev. **123**, 1939, 1961.
[5]) See G. Diemer and J. G. van Santen, Philips Res. Repts **15**, 368, 1960.
*) Philips Zentrallaboratorium GmbH, Aachen Laboratory.

A LOW-FREQUENCY OSCILLATOR WITH VERY LOW DISTORTION UNDER NON-LINEAR LOADING

by G. KLEIN *) and J. J. ZAALBERG van ZELST *).

621.373.421

*This article is the second of a series on electronic circuits for special measuring instruments **). It deals with an oscillator (with fixed frequency) whose output voltage is required to meet exacting demands with regard to amplitude constancy and freedom from distortion under non-linear conditions that cause severe distortion of the current through the load.*

Valve oscillators under non-linear loading

The output voltage of valve oscillators is in most cases reasonably sinusoidal provided the load consists of a linear element. Non-linear loading, on the other hand, can cause severe distortion of the output voltage, and also make it very difficult to meet the condition for oscillation continuously. For certain purposes, however, a voltage is required which, even under severe non-linear loading, should contain only a very small fraction of higher harmonics. This was the case, for example, in a set-up used in this laboratory for measurements on magnetic amplifiers in which ferromagnetic cores undergo varying DC magnetization. For this purpose an oscillator was needed the current from which could show distortion up to 20% of the maximum current taken (r.m.s. value of the total higher harmonics 20% of the r.m.s. value of the fundamental component under full load) while the voltage distortion was not to exceed the very low value of 0.01%. Under varying load, and also under constant loading for longer periods, the amplitude of the output voltage was allowed to vary by no more than 0.1%. The specification of this oscillator was therefore as follows:

Output voltage	U_o	= 50 V approx.
Maximum power	P	= 2 W
Distortion in output voltage at 8 mA distortion current (20% of the nominal current under full load)	d	= max. 0.01%
Variation in output voltage	$\Delta U_o/U_o$	= max. 0.1 %
Frequency	f	= 80 c/s

To this specification should be added that no current forms were to occur with a peak value higher than 120 mA. This limitation rules out those cases in which the current would contain no more than 8 mA in higher harmonics, but where the voltage distortion requirement would not be met. A case in point would be a current consisting, for example, of a sinusoidal component plus a more or less pulse-shaped component. Keeping to a given distortion percentage one might, by taking the pulse narrow enough, make its amplitude arbitrarily high; in which case it would become increasingly difficult to keep the influence of the pulse on the voltage below a predetermined limit. The distortion requirement can therefore better be formulated by specifying that the *internal resistance* of the oscillator should be particularly low for multiples of the fundamental frequency, namely smaller than $10^{-4} \times 50$ V : 8 mA = approx. 0.6 ohm. The specification of a maximum voltage variation of 0.1% under fluctuating load also boils down to a low internal resistance, now however at the fundamental frequency.

Interference signals induced in the output voltage should be kept to the same low percentage as the higher harmonics.

In the following we shall discuss an oscillator which meets the conditions mentioned.

Division into an oscillating and a power-output section

The combination of 1) positive feedback for the fundamental frequency such that the oscillation condition is constantly fulfilled, 2) strong negative feedback for the higher harmonics to keep the internal resistance at the low value required, and 3) measures for keeping the amplitude of the voltage constant, entails well-nigh insuperable difficulties in the design of a non-linearly loaded oscillator which is required to deliver energy.

A much more attractive principle is *to let the oscillation and the delivery of energy be carried out by two separate sections.* The first section, which remains

*) Philips Research Laboratories, Eindhoven.

**) For the first article see Philips tech. Rev. 24, 275-284, 1962/63 (No. 9).

unloaded, should generate a constant alternating voltage which has the required frequency and a distortion well below the specified limit; we shall call this voltage the *reference voltage*. The second section should closely "follow" the reference voltage. It contains a control system in which a certain fraction of the output voltage is compared with the reference voltage; the difference is amplified and drives the output stage which, even under non-linear loading, must be capable of delivering the required output voltage and power. Here, then, the principle of the stabilized power supply is applied, with this difference that the reference and output voltages are not direct but alternating voltages.

The first section will be called the *reference oscillator*, and the second the *output stage*.

The reference oscillator

When designing an oscillator one can generally choose between the *LC* and the *RC* type. Since the frequency in the present case is low (80 c/s), a coil with a particularly high inductance would be needed for an *LC* oscillator. Since the reference voltage must be just as free from interference voltages of outside origin as from higher harmonics, the screening of such a coil against stray alternating magnetic fields could be difficult. *RC* oscillators however do not require heavy screening, which was one of the main reasons for choosing this type.

The voltage delivered by conventional *RC* oscillators is unsatisfactory as a reference voltage, both in regard to distortion and constancy. The distortion is unsatisfactory because with the conventional *RC* oscillator the selectivity is obtained by using passive *RC* networks, which do little to reduce the distortion introduced by non-linear elements, such as valves. Given a signal around 10 V a distortion of about 1% is therefore normal. To keep the voltage *constant*, an amplitude-limiter is used — e.g. a thermistor (resistor with negative temperature coefficient, incandescent lamp) — which suppresses the gain as the amplitude increases. Over long periods a constancy better than about 1% is difficult to achieve in this way. This is primarily due to the influence of the ambient temperature, the operation of thermistors being based on an energy balance.

In the solution adopted a further subdivision is made: *the reference oscillator consists of an amplitude-limiting part and a part which ensures that the oscillation condition is fulfilled.* As will presently be shown, the amplitude-limiting part delivers a voltage which, although its amplitude is very constant, is at the same time severely distorted. Consequently, steps had to be taken at the same time *to free the reference voltage from higher harmonics.*

The limiter used to keep the amplitude satisfactorily constant is a balanced stage (double triode T_1-T_2, *fig. 1a*) with a high cathode resistance R_k (e.g. 0.1 MΩ) in the common cathode lead (known as a "long-tailed pair"). T_1 and T_2 are biased in such a way that half the total cathode current I_k flows through each of them. A slight difference V_d (a few volts) between the voltages on the grids is sufficient to cause the total current I_k to flow to one or the other anode; this is illustrated in fig. 1*b*, where the two anode currents are plotted as a function of the difference $V_{g1} - V_{g2}$ between the grid voltages. If we superimpose on the grid of T_1 an alternating voltage v_i (fig. 1*c*) with an amplitude several times that of V_d, then T_1 and T_2 will pass the current I_k alternately. On the anode of T_2 which has a resistance R_a in series with it, a voltage will then appear with a more or less square waveform. The amplitude of the square wave voltage is $I_k R_a$, and is thus independent of the amplitude of v_i. By stabilizing the supply voltages and using

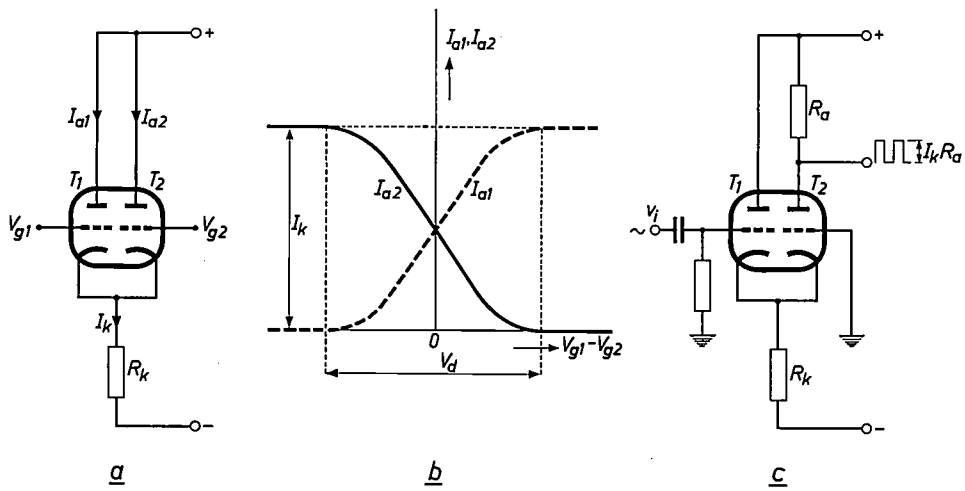


Fig. 1. *a*) Balanced stage with high resistance — here an ordinary resistance R_k — in the common cathode lead ("long-tailed pair").
b) The anode currents I_{a1} and I_{a2} of the triodes T_1 and T_2 in (*a*) are plotted versus the difference $V_{g1} - V_{g2}$ of the grid voltages. The difference V_d needed in order for the whole cathode current $I_k = I_{a1} + I_{a2}$ to flow from one anode to the other is only a few volts.
c) When an alternating voltage v_i with an amplitude several times that of V_d is applied to the grid of T_1, a square-wave voltage with amplitude $\frac{1}{2}I_kR_a$ appears at the anode of T_2, which has a resistance R_a in series with it.

metallic resistors (wire-wound or metal-film) for R_k and R_a, we can make I_k and R_a sufficiently independent of variables such as the mains voltage and the ambient temperature. Once a constant square-wave voltage has been obtained in this way, its fundamental component — which we shall use further here — also has a constant amplitude.

The shape of the sides of the approximately square-wave voltage has very little influence on the amplitude of the first harmonic. This is easily seen if one considers the way in which the first harmonic of a periodic odd function $f(\omega t)$ is determined by Fourier analysis; the function is multiplied by $\sin \omega t$ and the product is integrated with respect to ωt between the limits 0 and 2π. If $f(\omega t)$ is a more or less square-wave function, with zero transitions at $\omega t = 0$, π and 2π, then the sides coincide with small values of $\sin \omega t$. Thus they contribute very little to the amplitude of the first harmonic.

Referring to the block diagram of the reference oscillator in *fig. 2a*, we shall now explain how the oscillation is generated and how the reference voltage is kept free from higher harmonics. On the far left in fig. 2*a* can be seen the balanced stage shown in fig. 1*c*. We assume that there is a sinusoidal alternating voltage of 10 V ($10/\sqrt{2}$ V amplitude) at the input p of this stage. At the point q there then appears a square-wave voltage of constant amplitude. Point q is connected via a resistor R_1 to the input of an amplifier A_1, the output r of which has

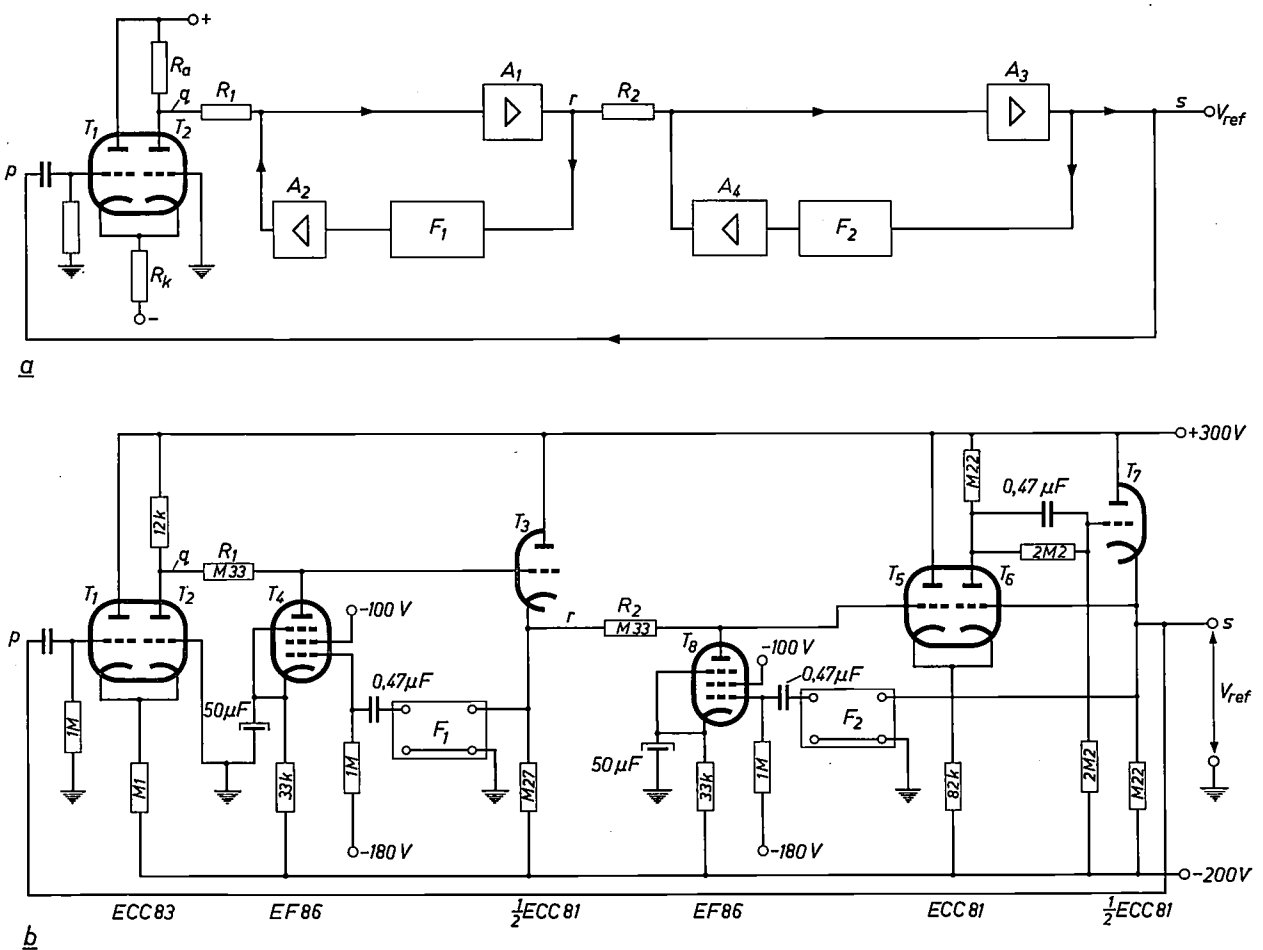


Fig. 2. *a*) Block diagram of the reference oscillator. T_1-T_2, R_k and R_a correspond to fig. 1*c*. At point q there is a constant square-wave voltage (frequency 80 c/s), the fundamental component of which passes through the amplifiers (cathode followers) A_1 and A_3. For the higher harmonics in the square-wave voltage, strong negative feedback is applied, A_1 and A_3 each containing in the feedback path a filter (F_1, F_2) which passes the higher harmonics but blocks the fundamental component. A_2 and A_4 amplify the transmitted higher harmonics. The voltage at the output s shows less than 0.01% distortion. As s is connected to the input p, the circuit constitutes an oscillator.
b) The complete circuit diagram. The various components are shown as far as possible underneath the corresponding blocks in (*a*). Notation of resistances: 82k means 82 kΩ, M1 means 0.1 MΩ, 2M2 means 2.2 MΩ, etc.

negative feedback through a filter F_1 and an amplifier A_2. The filter — to which we shall presently return — passes the higher harmonics almost without attenuation but does *not* pass the fundamental component. The higher harmonics therefore undergo strong negative feedback, while the filter blocks the feedback path for the fundamental component. The result is that the higher harmonics are considerably attenuated in proportion to the fundamental. The voltage at point r thus approximates fairly closely to a sine wave; its distortion is about 0.3% — a great deal less than the roughly 50% distortion present in the square-wave voltage at point q.

The distortion at r, however, is still much greater than the permissible maximum of 0.01 % for the reference voltage. For this reason the above procedure is repeated: point r is connected via a resistor R_2 to an amplifier A_3, which has again negative feedback for the higher harmonics through a filter F_2 (identical with F_1) and an amplifier A_4.

Since the output s of A_3 is connected to the input p, the whole circuit forms an oscillator. The condition for oscillation is now fulfilled: the loop gain is automatically equal to 1 and the frequency is that at which the total phase shift (of amplifier and filters together) is zero. The phase shift in the amplifiers is negligibly small, and that in the filters (assumed to be ideal) is zero at the frequency at which the transfer is zero (the zero frequency). The whole system therefore oscillates with the zero frequency.

Fig. 2*b* shows the complete circuit of the reference oscillator. An important feature is that in the "main line", between the points q and s, very little happens that can endanger the constancy of the amplitude of the sine-wave voltage at s. In the main line there are two amplifiers (A_1 and A_3) and two voltage dividers (R_1-T_4, and R_2-T_8). As can be seen from fig. 2*b*, A_1 and A_3 are cathode followers (T_3 and T_6 respectively); their gain is therefore a little less than unity ($1-A^{-1}$, where $A \gg 1$) and very constant. The same applies to the voltage divisions; the division ratios are similarly governed by an expression of the form $1 - A^{-1}$ with $A \gg 1$. The amplifiers A_2 and A_4 need to have a high gain for suppressing the higher harmonics. They are situated, however, in the negative feedback paths, which are blocked to the fundamental component by the filters F_1 and F_2; therefore their gain need not be particularly constant. It would have been much more difficult to obtain a constant output voltage using a perhaps more obvious circuit containing a filter in the main line which passed the fundamental and blocked the higher harmonics. Again, to avoid induced interfering voltages in coils having a high inductance, such a filter would have to be an *RC* and not an *LC* type. The low selectivity of such a filter would have to be improved by considerably amplifying the fundamental component in the main line. It would then have been difficult to get the required amplitude constancy.

In fig. 2*b* it is seen that the amplifier A_1 consists of a single cathode follower (triode T_3) but that A_3 has a more complicated circuit. The reason lies in the magnitude of the distortion introduced by the valves as a result of the curvature of their characteristic. For A_1, where the signal — as stated — still shows a distortion of about 0.3%, the distortion introduced by a single cathode follower is relatively insignificant. For A_3 the distortion in the output voltage is required to be better than 0.01%. This makes it necessary to take careful account of the distortion which A_3 itself produces. This distortion is smaller the higher is the impedance in the cathode lead; a limit is set to this impedance, however, by the input impedance of the filter F_2, which is rather crucial if the low value resistances in this filter are to be metallic resistors. Taking for these the largest values available we find that the distortion (mainly second harmonic) introduced by a single cathode follower at an output voltage of 10 V is such that the distortion in the output voltage comes very close to the specified limit of 0.01%. A single cathode follower might therefore have been sufficient. It was decided, however, that a wider margin was desirable for the reference voltage. This was obtained by using a more elaborate circuit for A_3, which produces much less second harmonic distortion than the single cathode follower.

As shown in fig. 2*b*, A_3 consists of a single cathode follower T_7 preceded by a balanced stage T_5-T_6 having a common cathode resistance. The second harmonic in the output voltage of this combination is only 2.5×10^{-3} % of the fundamental component; the other higher harmonics are even weaker.

Distortion introduced by the cathode followers

The distortion introduced by the curvature of a valve characteristic can be calculated by a method described elsewhere [1]). Using this method the distortion introduced by the cathode followers employed in this circuit will now be calculated.

The equation of the valve characteristic (anode current i_a as a function of the "total driving voltage" v) can be written as a power series:

$$-i_a = av + \beta v^2 + \gamma v^3 + \ldots, \quad \ldots \quad (1)$$

[1]) J. Rodrigues de Miranda and J. J. Zaalberg van Zelst, New developments in output-transformerless amplifiers, J. Audio Engng. Soc. **6**, 244-250, 1958.

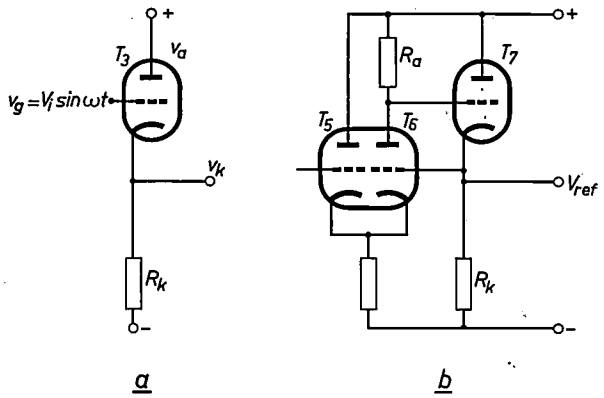


Fig. 3. *a*) Single cathode follower (like T_3 in fig. 2*b*): triode T_3 with cathode resistance R_k.
b) complex cathode follower (like T_5-T_6-T_7 in fig. 2*b*). The inherent distortion here is much lower than in the single type.

where

$$v = v_g + \frac{v_a}{\mu} - \left(1 + \frac{1}{\mu}\right) v_k . \quad . . . . . \quad (2)$$

The voltages v_a, v_g and v_k are indicated in *fig. 3a*, and μ is the amplification factor of the valve.

We can express (1) in the form:

$$v = a\, i_a + b\, i_a^2 + c\, i_a^3 + \ldots, \quad . . . . . \quad (3)$$

where

$$a = \frac{1}{\alpha}, \; b = \frac{\beta}{\alpha^3}, \; c = \frac{2\beta^2}{\alpha^5} - \frac{\gamma}{\alpha^4}. \quad . . . . \quad (4)$$

For the single cathode follower (fig. 3*a*), $v_a = 0$ and $v_k = i_a R_k$. Assuming $(1 + \mu^{-1})R_k = r$, we find from (2) and (3) for the single cathode follower:

$$v_g = (a + r) i_a + b\, i_a^2 + c\, i_a^3 + \ldots . \quad . . . . \quad (5)$$

If we expand i_a in a power series to v_g:

$$i_a = k_1 v_g + k_2 v_g^2 + k_3 v_g^3 + \ldots, \quad . . . . \quad (6)$$

the relation between (6) and (5) is the same as that between (1) and (3). By analogy with (4) we can therefore write:

$$a + r = \frac{1}{k_1}, \; b = -\frac{k_2}{k_1^3}, \quad c = \frac{2k_2^2}{k_1^5} - \frac{k_3}{k_1^4}.$$

Combination with (4) gives:

$$k_1 = \frac{\alpha}{1 + \alpha r}, \; k_2 = \frac{\beta}{(1 + \alpha r)^3}, \; k_3 = \frac{-2\beta^2 r}{(1 + \alpha r)^5} + \frac{\gamma}{(1 + \alpha r)^4}, \text{ etc.}$$

If v_g is a purely sinusoidal voltage, $v_g = V_0 \cos \omega t$, this current i_a is given by the following series:

$$\begin{aligned} i_a = {} & (k_1 V_0 + \tfrac{3}{4} k_3 V_0^3 + \ldots) \cos \omega t + \\ & + (\tfrac{1}{2} k_2 V_0^2 + \ldots) \cos 2\omega t + \\ & + (\tfrac{1}{4} k_3 V_0^3 + \ldots) \cos 3\omega t + \ldots \end{aligned}$$

From this we find the ratio d_2 of the second harmonic to the fundamental wave of the current i_a (and hence of the output voltage $v_k = i_a R_k$):

$$d_2 = \frac{2k_2 V_0}{4k_1 + 3k_3 V_0^2} = \left[\frac{2\alpha}{\beta V_0}(1 + \alpha r)^2 - \frac{3\beta r\, V_0}{(1 + \alpha r)^2} + \frac{3\gamma V_0}{\beta(1 + \alpha r)}\right]^{-1}. \quad . . . \quad (7)$$

To determine which of the three terms between square brackets is the most important, we shall fill in some practical values. For the valve ECC 81 at 1 mA and 100 V anode voltage we have: α = approx. 1.5 mA/V, β = approx. 0.80 mA/V^2 and γ = approx. 1.6 mA/V^3. If r is 50 kΩ and $V_0 = 14$ V, we find for the three terms respectively 1550, − 0.3 and 1.1, so that in this case the first term is by far the most dominant. To a good approximation we can therefore simplify (7) to:

$$d_2 = \frac{\beta V_0}{2\alpha(1 + \alpha r)^2}. \quad . . . . . \quad (8)$$

In this case, then, $d_2 = 1550^{-1} = 0.065\%$. The negative feedback further reduces this distortion by a factor which is difficult to calculate but turns out to be of the order of 10. The result is a distortion which remains just below the specified limit. As mentioned above, a certain margin was thought desirable.

From (8) we see that the distortion decreases if r, and thus the cathode resistance, R_k, is increased. R_k consists of the actual cathode resistance in parallel with the input impedance of the filter. The latter sets an upper limit to R_k.

The cathode follower preceded by a balanced stage, as used for A_3, is represented in fig. 3*b* (omitting the elements for biasing T_6). Calculation shows that, with good dimensioning and minor simplifying assumptions, the inherent distortion given by (8) is reduced by a factor

$$\left(\frac{1}{\mu'} + \frac{2}{SR_a}\right)^{-1} \gg 1. \quad . . . . . \quad (9)$$

In this expression μ' is the amplification factor and S the transconductance of the triode T_6 in fig. 3*b*. With the adjustment used here, this valve (½ ECC 81) gives $\mu' = 60$ and $S = 2$ mA/V. With $R_a = 0.22$ MΩ, the factor given by (9), with which the inherent distortion is reduced, is roughly 50.

The filters

For the same reasons that prompted us not to use an *LC* oscillator, we also avoided using coils in the filters (F_1 and F_2 in fig. 2) which therefore consists entirely of resistors and capacitors.

As we have seen, the filters block the fundamental component but pass the higher harmonics. A filter network that possesses this characteristic is the twin-T type, composed of resistors and capacitors (*fig. 4a*). It is so called because it can be regarded

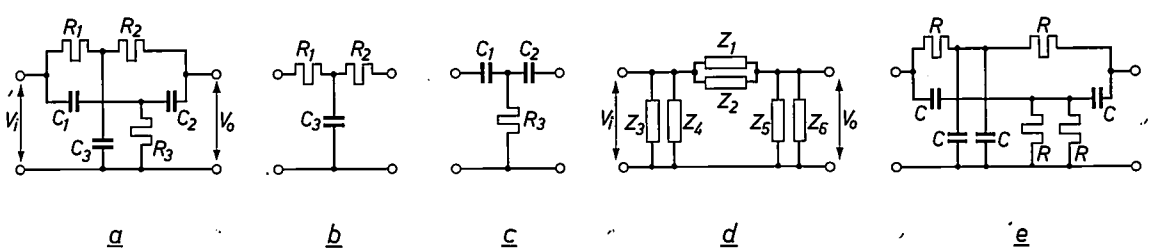


Fig. 4. *a*) Twin-T filter, consisting of resistors and capacitors. *b*) and *c*) The two T sections from which (*a*) is formed. *d*) The two T sections of (*a*) are transformed into *Π* sections. *e*) In the case of symmetry the network (*a*) can be formed from four identical resistors R and four identical capacitors C.

as two T sections "one on top of the other" (fig. 4*b* and *c*). At one frequency (the zero frequency f_0) the filter passes no signal if the resistance and capacitance values satisfy the following condition:

$$\frac{R_1R_2}{R_1+R_2}C_3 = R_3(C_1+C_2). \quad . \; . \quad (10)$$

In that case

$$f_0 = \frac{1}{2\pi\sqrt{(R_1+R_2)R_3C_1C_2}}.$$

To derive the condition (10) we transform the two T sections (or star networks) of fig. 4*b* and *c* into π sections (fig. 4*d*). The six impedances $Z_1 \ldots Z_6$ are easily expressed in terms of R_1, R_2, R_3, C_1, C_2, C_3 and the angular frequency ω. The ratio $1/A_F$ of the input to the output voltage of the filter we read from fig. 4*d*:

$$\frac{1}{A_F} = \frac{V_i}{V_o} = 1 + \frac{Z_1Z_2(Z_5+Z_6)}{(Z_1+Z_2)Z_5Z_6}.$$

From this formula it is seen that the only case in which the filter blocks the signal completely ($A_F = 0$, $V_i/V_o = \infty$) occurs when $Z_1 + Z_2$ is zero. Expressing Z_1 and Z_2 in terms of the resistances and of the capacitances, and putting both the real and imaginary parts of $Z_1 + Z_2$ equal to zero we obtain the following equations:

$$\omega^2 = \frac{1}{(R_1+R_2)R_3C_1C_2} \quad . \; . \; . \; . \; . \; . \quad (11)$$

and

$$\omega^2 = \frac{C_1+C_2}{R_1R_2C_1C_2C_3}. \quad . \; . \; . \; . \; . \; . \; . \; . \quad (12)$$

At the frequency f_0 the filter passes no signal if ω in the left-hand members of (11) and (12) is equal to $2\pi f_0$. The right-hand members are then likewise identical, which leads to equation (10).

For filters that exactly fulfil the condition (10) it can be deduced that the transmission-ratio A_F as a function of frequency is given by:

$$A_F = \frac{V_o}{V_i} = \frac{jQ\beta}{1+jQ\beta}, \quad . \; . \; . \; . \quad (13)$$

where β is the relative detuning:

$$\beta = \frac{f}{f_0} - \frac{f_0}{f},$$

and the figure of merit Q is:

$$Q = \frac{\sqrt{(R_1+R_2)R_3C_1C_2}}{R_1C_3 + (R_1+R_2)C_2}.$$

If the filter is symmetrical ($R_1 = R_2$, say $= R$, and $C_1 = C_2$, say $= C$; the symmetry also implies: $R_3 = \frac{1}{2}R$ and $C_3 = 2C$), we find $Q = \frac{1}{4}$, so that this kind of filter cannot be expected to have a very great selectivity. With an asymmetric filter a somewhat higher Q can be obtained, theoretically a maximum of $\frac{1}{2}$. The difference is so small as to be unimportant compared with the practical advantages of a symmetrical filter ,whose input and output impedances are more favourable and which can be built with four identical resistors R, and four identical capacitors C (fig. 4*e*).

Small deviations in the values of C and R cause deviations of the same order of magnitude in the oscillator frequency and voltage amplitude. For this reason, only mica capacitors and metallic resistors are used in the filters.

Instead of the behaviour of the filter itself we shall now consider the behaviour of the filter together with the amplifiers with which it works, i.e. the combination A_1-F_1-A_2 in fig. 2*a*. This *active filter*, shown separately in *fig. 5*, has an input voltage V_q and an output voltage V_r (cf. points q and r in fig. 2*a*). The transmission A_{act} of this active filter can be found with the aid of (13):

$$A_{act} = \frac{V_r}{V_q} = B\frac{1+jQ\beta}{1+j(A_2+1)Q\beta}. \quad . \quad (14)$$

Here A_2 is the gain of the pentode amplifier A_2, and B a factor, not otherwise relevant to our considerations, which is independent of frequency and differs little from unity. It follows from (14) that

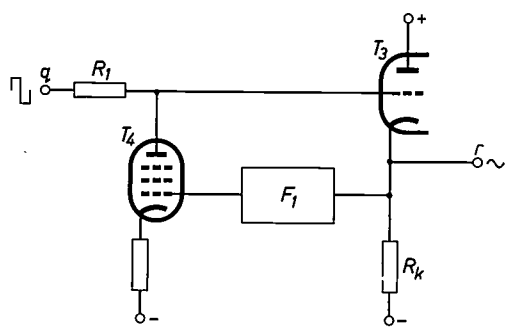


Fig. 5. The "passive" filter F_1 and the amplifiers A_1 and A_2 (fig. 2*a*) together form an active filter, with input terminal q and output terminal r. The fundamental component of the square-wave voltage at q is passed. The higher harmonics are considerably attenuated owing to the high gain of A_2.

for the fundamental component ($\beta = 0$) the transmission A_{act} is equal to B (so that the fundamental is passed virtually unattenuated) and that the higher harmonics are better suppressed the higher is the gain A_2. Given $A_2 \gg 1$ and $Q = \frac{1}{4}$, the absolute value of A_{act}/B at the frequencies f_0, $2f_0$, etc. is found from (14):

$f = f_0$	$2f_0$	$3f_0$	$4f_0$ . . .	∞,
$\beta = 0$	$1\frac{1}{2}$	$2\frac{2}{3}$	$3\frac{3}{4}$ . . .	∞,
$\left\lvert\frac{A_{act}}{B}\right\rvert = 1$	$\frac{2.85}{A_2}$	$\frac{1.80}{A_2}$	$\frac{1.46}{A_2}$ . . .	$\frac{1}{A_2}$.

It can be seen that the second harmonic is

suppressed less than the third and higher harmonics. On the other hand, the second harmonic is only weakly represented in the voltage V_q (a perfect square-wave voltage contains no even harmonics at all).

Measurements of the reference voltage

Measurements of the output voltage of the reference oscillator (r.m.s. value about 10 V, frequency 80 c/s) have been made to determine the distortion and the constancy of the amplitude and frequency.

The following values were found for the contribution of the principal higher harmonics to the *distortion*:

second harmonic	d_2	$= 2.5\times10^{-5}$,
third harmonic	d_3	$= 1.5\times10^{-5}$,
remaining harmonics	d_{rest}	$< 0.4\times10^{-5}$.

To keep the amplitude from varying by more than 0.01%, the supply voltage of + 300 V had to be kept constant to within 0.5 V, the supply voltage of — 200 V to within 0.1 V, and the heater voltage (nominal 6.3 V) to within 0.3 V. These conditions can easily be met.

Although nothing was specified regarding the constancy of the frequency, this too was investigated. In ten independent measurements we counted the number of times the voltage passed through zero in three minutes, and the frequency deviations were found to be no more than 0.02% of the average frequency.

The output stage

The circuit diagram of the output stage and the associated driving stage is shown in *fig. 6*.

The output stage contains two pentodes (T_{17} and T_{18}) in a single-ended push-pull arrangement [2]). In common with the ordinary push-pull circuit, this arrangement suppresses the formation of even harmonics, but it has the additional advantage that the direct current does not pass through the load, thus making an output transformer unnecessary. The output pentodes are of the EL 86 type, specially designed for use in single-ended push-pull circuits (high anode current at a relatively low anode voltage). In the output current these valves give rise to distortion of only a few per cent, i.e. an order of magnitude smaller than the distortion caused by the non-linear load itself.

The internal resistance of the output stage is roughly 1000 ohms. As noted at the beginning of this article, the requirement as to the maximum distortion permissible in the output voltage amounts to specifying that the internal resistance should not exceed about 0.6 ohm. One of the functions of the circuit which drives the output stage is therefore to ensure that the loop gain is a few times 10^3. For this reason the driving circuit consists of two stages.

The primary function of the driving circuit is to act as a control system, i.e. to drive the output stage in such a way that a variable part kU_0 of the output voltage is kept equal to the reference voltage V_{ref} of 10 V; the difference $kU_0 - V_{ref}$ is amplified and is used to drive the output stage. The fraction k can be adjusted from 1/6 to 1/4 using a potentiometer P, so that U_0 is continuously variable from 60 to 40 V.

An important question is the *distortion in the driving circuit*. Particularly favourable in this respect are *difference amplifiers* [3]), i.e. balanced amplifiers having a high resistance in the common cathode lead.

Minimizing the distortion makes particularly severe demands on the first stage of the driving circuit. This stage has at its input an "in-phase voltage" of 10 V; the "anti-phase voltage" is the much smaller difference $kU_0 - V_{ref}$. To keep the distortion minimum in spite of this relatively high in-phase voltage, it is necessary to minimize the current variations which the in-phase voltage causes in the valves.

This is precisely what a good difference amplifier does, hence the fact that the first stage is a very carefully designed difference amplifier, possessing both a high rejection factor and a high discrimination factor.

To limit the distortion of this stage to 0.01%, the rejection factor should be of the order of 10^4. This follows from a calculation similar to that given above under the heading *Distortion introduced by the cathode followers*.

The first stage thus consists of the cascodes T_9-T_{11} and T_{10}-T_{12} in a push-pull arrangement; the common cathode lead contains the very high differential resistance of the cascode T_{13}-T_{14} [4]).

The second stage is a simple difference amplifier, consisting of a double triode T_{15}-T_{16} in a balanced arrangement with an ordinary resistance in the cathode lead.

---

[2]) See e.g. J. Rodrigues de Miranda, Philips tech. Rev. **19**, 2, 1957/58.

[3]) In the following section some terms from the technique of difference amplifiers will be used. For an explanation of the terms see G. Klein and J. J. Zaalberg van Zelst, General considerations on difference amplifiers, Philips tech. Rev. **22**, 345-351, 1960/61.

[4]) This arrangement is a combination of the circuits shown in fig. 6 and fig. 9 in the article by G. Klein and J. J. Zaalberg van Zelst, Circuits for difference amplifiers, I, Philips tech. Rev. **23**, 142-150, 1961/62.

Unequal biasing of the valves in push-pull would cause the distortion to increase considerably. The biasing is kept equal by negative direct-voltage feedback.

The total gain of the three stages is about 12 000. The loop gain is smaller by a factor of 6 to 4 (the potentiometer ratio) i.e. about 2000 to 3000. This is sufficient for reducing the internal resistance of the output stage to the required low value.

Measurements were also carried out under non-linear loading. For this purpose the circuit of *fig. 7a* was used, with different values of R. Current passes through the diode in the branch parallel with R only during that part of the period in which the instantaneous value of the alternating voltage is higher than the voltage across the capacitor in this branch. The diode current is consequently more or less pulse-shaped (fig. 7*b*) and the total

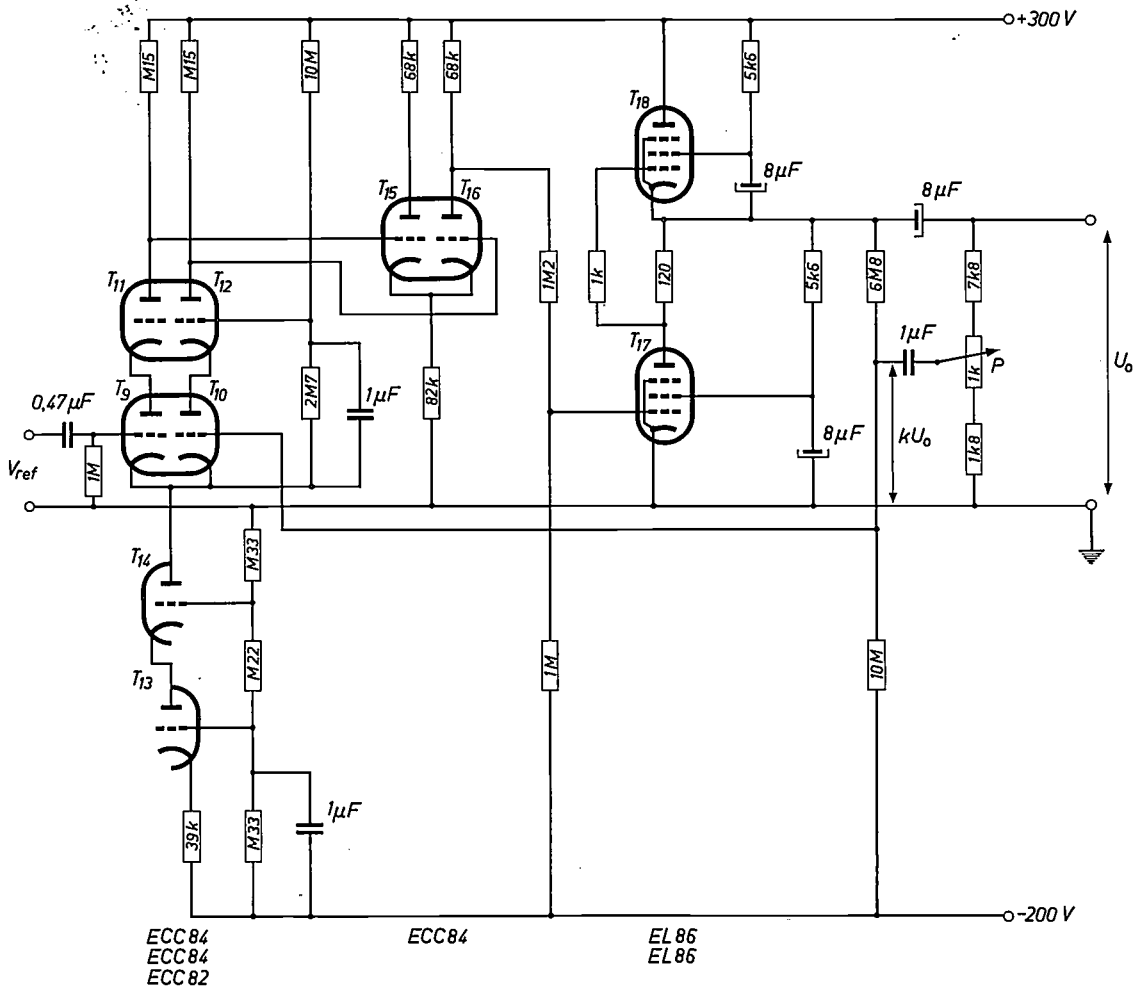


Fig. 6. Diagram of the output stage with two-stage driving circuit.
Driving circuit. First stage: difference amplifier with high rejection and high discrimination factors, consisting of cascodes T_9-T_{11} and T_{10}-T_{12} in push-pull with the cascode T_{13}-T_{14} in the common cathode lead. Second stage: simple difference amplifier, consisting of triodes T_{15}-T_{16} in push-pull, with a resistance of 82 kΩ in the cathode lead.
Output stage: single ended push-pull arrangement of pentodes T_{17}-T_{18}.
In the first stage of the driving circuit the part kU_0 of the output voltage (k being adjustable from $\frac{1}{8}$ to $\frac{1}{4}$ with potentiometer P) is compared with the reference voltage V_{ref} of 10 V. (The connection between the grids of T_{11} and T_{12} is lacking in the figure.)

Results

The following harmonic content was measured in the output voltage with a load consisting of a resistance of 1000 ohms:

second harmonic . . $2.0 \times 10^{-5} \times U_0$,
third harmonic . . $3.5 \times 10^{-5} \times U_0$,
fourth harmonic . . $3.3 \times 10^{-5} \times U_0$,
fifth harmonic . . $0.56 \times 10^{-5} \times U_0$.

current has the form of a sine wave with a superimposed peak (fig. 7*c*). The distortion in this current was measured by analysing the voltage across the series resistance of 10 ohms (fig. 7*a*).

For each of the higher harmonics the internal resistance R_i can be determined by dividing the relevant voltage component by the corresponding current component. For the second to fifth harmonics the results obtained, as the averages of measure-

ments with three values of R, varied from 0.46 to 0.66 ohm.

The lower limit of the distortion in the output voltage is determined chiefly by the distortion already present in the reference voltage (see above); the value found was 3×10^{-5}. The relative amplitude and frequency variations in the output voltage are identical with those in the reference voltage.

The oscillogram in *fig. 8* shows the peaks of the output voltage recorded in a time of roughly 1 second [5]). The amplitude of the voltage was about 70 V. The fluctuation is seen to be approximately 15 mV, or 0.02%, which is well below the specified limit of 0.1%.

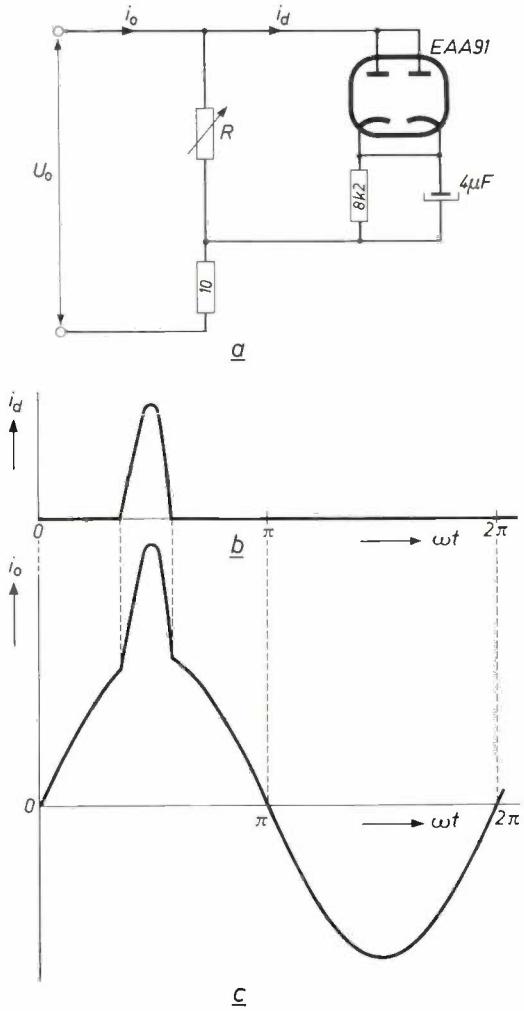


Fig. 7. *a*) Circuit used as non-linear load for distortion measurements. The load current i_o consists of a linear component, which flows through the variable resistor R, and a non-linear component i_d, which, during a small part of each period, charges up a capacitor of 4 μF through the diode EAA 91. The harmonics of i_o were found by analysing the voltage across the 10 ohms series resistance.
b) The charging current i_d. *c*) The total current i_o.

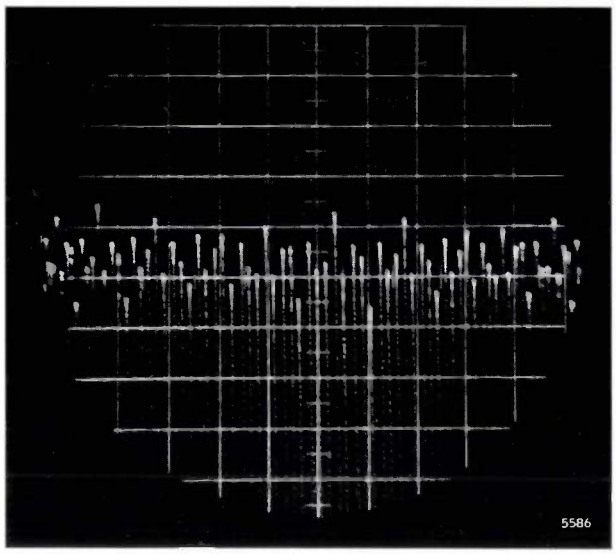


Fig. 8. Oscillogram obtained with a "voltage microscope" [5]), showing the peaks of the output voltage (amplitude approx. 70V). The height of each square corresponds to about 14 mV (the zero line of the sine wave should therefore be imagined at about 35 metres below the peaks!). The maximum fluctuations in amplitude are about 15 mV (0.02%). The width of the oscillogram corresponds roughly to one second.

[5]) G. Klein and J. J. Zaalberg van Zelst, Philips tech. Rev. **23**, 173 ff., 1961/62.

Summary. For measurements on magnetic amplifiers an oscillator was needed that could deliver 2 W at a voltage of 50 V, 80 c/s, a special requirement being that the voltage under severe non-linear loading should show no more than 0.01% distortion and fluctuate in amplitude by no more than 0.1%. This requirement means keeping the internal resistance extremely low (no higher than about 0.6 ohm).

In the solution adopted a division is made into an oscillating section and an output stage incorporating a control system. The oscillating section (the reference oscillator) delivers an almost purely sinusoidal and constant voltage of 10 V, 80 c/s, which is used as the reference voltage for the control system.

The reference oscillator begins by generating a very constant square-wave voltage of 80 c/s, the higher harmonics of which are suppressed by negative feedback via double-T section *RC* filters, which pass the higher harmonics but not the fundamental component. The result is a reference voltage with no more than 0.003% distortion and 0.02% fluctuation.

The control system of the output stage compares an adjustable fraction (1/6-1/4) of the output voltage with the reference voltage and drives the output stage (two EL 86 pentodes in a single-ended push-pull arrangement). The loop gain is so high that the internal resistance of the output stage is reduced to the required low value of about 0.6 ohm. Measures are taken to minimize the distortion in the amplifying stages of the control system.

RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF THE PHILIPS LABORATORIES AND FACTORIES [1])

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

3051: A. Smit: Investigations in the vitamin A series, IV. Some cases of abnormal decarboxylation (Rec. Trav. chim. Pays-Bas **80**, 891-904, 1961, No. 8).

3052: L. A. Æ. Sluyterman: Photo-oxidation, sensitized by proflavine, of furfuryl alcohol, N-allyl thiourea and histidine (Rec. Trav. chim. Pays-Bas **80**, 989-1002, 1961, No. 9/10).

3053: J. Bakker and H. M. van den Bogaert: Influence de la configuration sur l'absorption dans l'infrarouge des doubles liaisons dans des stéroïdes (Rec. Trav. chim. Pays-Bas **80**, 1015-1022, 1961, No. 9/10). (Influence of configuration on the absorption in the infrared of double bonds in steroids; in French.)

3054: C. J. Schoot, J. J. Ponjée and K. H. Klaassens: Acylation of aromatic ring systems with mixed anhydrides of phosphoric acid (Rec. Trav. chim. Pays-Bas **80**, 1084-1088, 1961, No. 9/10).

3055*: J. A. Kok: Electrical breakdown of insulating liquids (Philips Technical Library, 1961, XII + 132 pp.).

3056: G. P. Ittmann: Pythagorese driehoeken (Euclides **37**, 119-122, 1961). (Pythagorean triangles; in Dutch.)

3057*: P. F. van Eldik and P. Cornelius: Transformatoren, smoorspoelen, transductoren en lektransformatoren (Philips Technical Library, 1961, VIII+80 pp.). (In Dutch; 1962 published in English under the title: A. C. devices with iron cores; principles and design of transformers, chokes, transductors and leakage transformers; VIII + 86 pp.)

3058: J. B. de Boer: Die Fahrbahndecke als Lichtreflektor (Strassen- und Tiefbau **15**, Supplement Licht und Farbe im Bauwesen **4**, 27-33, 1961, No. 10). (The road surface as light reflector; in German.)

3059: G. H. Jonker: Energy levels of impurities in transition metal oxides (Proc. int. Conf. on semiconductor physics, Prague 1960, pp. 864-867, Academic Press, New York 1961).

3060: L. W. de Zoeten and O. A. de Bruin: The reactivities of the tyrosine residues in insulin with respect to iodine, I (Rec. Trav. chim. Pays-Bas **80**, 907-916, 1961, No. 9/10).

3061: L. W. de Zoeten and E. Havinga: The reactivity of the tyrosine residues in insulin with respect to iodine, II (Rec. Trav. chim. Pays-Bas **80**, 917-926, 1961, No. 9/10).

3062: L. W. de Zoeten and R. van Strik: A study of the biological activity of iodinated insulin (Rec. Trav. chim. Pays-Bas **80**, 927-931, 1961, No. 9/10).

3063: T. Kralt, W. J. Asma and H. D. Moed: Reserpine analogues, IV. Hydroxy-β-phenylethylamine derivatives (Rec. Trav. chim. Pays-Bas **80**, 932-943, 1961, No. 9/10). Sequel to **2997**.

3064: S. de Groot and J. Strating: Reserpine analogues, II (Rec. Trav. chim. Pays-Bas **80**, 944-950, 1961, No. 9/10). Part of S. de Groot's thesis, Groningen 1960.

3065: G. B. Paerels: Crystalline 3-desoxy-2-ketogluconic acid (Rec. Trav. chim. Pays-Bas **80**, 985-988, 1961, No. 9/10).

R 443: S. Duinker: Conjunctors, another new class of non-energic non-linear network elements (Philips Res. Repts **17**, 1-19, 1962, No. 1).

R 444: T. J. Viersma: Investigations into the accuracy of hydraulic servomotors (Philips Res. Repts **17**, 20-78, 1962, No. 1). Continued from **R 442**.

R 445: H. Dormont: Random effects of transit times and secondary-emission multiplications in a multiplier phototube (Philips Res. Repts **17**, 79-94, 1962, No. 1).

R 446: J. van Laar and J. J. Scheer: Influence of band bending on photoelectric emission from silicon single crystals (Philips Res. Repts **17**, 101-124, 1962, No. 2).

R 447: H.-U. Harten and D. Polder: Influence of a surface space-charge layer on the motion of recombining carriers (Philips Res. Repts **17**, 125-129, 1962, No. 2).

R 448: G. Bouwhuis: A dispersion phenomenon observable on dielectric multilayer mirrors (Philips Res. Repts **17**, 130-132, 1962, No. 2).

R 449: A. Kats: Hydrogen in alpha-quartz (Philips Res. Repts **17**, 133-195, 1962, No. 2). Thesis Delft, Nov. 1961.

R 450: A. Kats: Hydrogen in alpha-quartz (Philips Res. Repts **17**, 201-279, 1962, No. 3). Continued from **R 449**.

R 451: A. Bril and W. van Meurs-Hoekstra: Absolute efficiency measurements of infrared fluorescent zinc and cadmium sulphide activated with V-Ag and V-Cu (Philips Res. Repts **17**, 280-282, 1962, No. 3).

R 452: J. M. Stevels and J. Volger: Further experimental investigations on the dielectric losses of quartz crystals in relation to their imperfections (Philips Res. Repts **17**, 283-314, 1962, No. 3).

[1]) Beginning with this volume an abstract of each publication will no longer be given but only the title, with occasional references to other related publications.

R 453: J. Rubio: A study of some self-correcting sequential networks (Philips Res. Repts **17**, 315-328, 1962, No. 4).

R 454: A. Bril and W. G. Gelling: Conductivity induced by cathode rays in cadmium-selenide layers (Philips Res. Repts **17**, 329-336, 1962, No. 4).

R 455: F. H. Stieltjes: Relations between currents and voltages in structures containing semiconductors (Philips Res. Repts **17**, 337-343, 1962, No. 4).

R 456: M. T. Vlaardingerbroek, K. R. U. Weimer and H. J. C. A. Nunnink: On wave propagation in beam-plasma systems (Philips Res. Repts **17**, 344-362, 1962, No. 4).

R 457: J. W. Steketee and J. de Jonge: Photoconductance and spectral absorption of anthracene (Philips Res. Repts **17**, 363-381, 1962, No. 4).

R 458: C. Ducot: A comparison between thermal and quantum noise in radio reception (Philips Res. Repts **17**, 382-392, 1962, No. 4).

A 50: S. Garbe and K. Christians: Zur Gasabgabe von Gläsern (Vakuum-Technik **11**, 9-16, 1962, No. 1). (Gas desorption of glasses; in German.)

A 51: J. Schröder: Darstellung und Untersuchung der Mischkristallreihen $La(Sr)CoO_3$ und $La(Th)CoO_3$ (Z. Naturf. **17b**, 346-347, 1962, No. 5). (Description and investigation of the mixed-crystal series $La(Sr)CoO_3$ and $La(Th)CoO_3$; in German.)

A 52: R. Groth and K. Weiss: Über den Bandabstand von β- und α-AgJ (Z. Naturf. **17a**, 536-537, 1962, No. 6). (On the energy gap of β and α AgI; in German.)

A 53: P. Gerthsen and K. H. Härdtl: Halbleitereigenschaften des Lanthankobaltit (Z. Naturf. **17a**, 514-521, 1962, No. 6). (Semiconducting properties of lanthanum cobaltite; in German.)

A 54: N. Hansen: Physikalische Adsorption bei tiefen Drucken und geringen Belegungen (Vakuum-Technik **11**, 70-77, 1962, No. 3). (Physical adsorption at low pressures and coverages; in German.)

A 55: K. Weiss: Zum Leitungscharakter im α-AgJ (Z. phys. Chemie Neue Folge **32**, 256-262, 1962, No. 3/4). (Conduction in α AgI; in German.)

A 56: A. Klopfer: Gasanalysen in Vakuumsystemen (Ingenieur **74**, O 72-O 78, 1962, No. 34). (Gas analyses in vacuum systems; in German.)

A 57: H. G. Reik and H. Risken: Drift velocity and anisotropy of hot electrons in *N* germanium (Phys. Rev. **126**, 1737-1746, 1962, No. 5).

A 58: H. G. Grimmeiss and R. Memming: *P-N* photovoltaic effect in cadmium sulfide (J. appl. Phys. **33**, 2217-2222, 1962, No. 7).

A 59: M. Nacken, S. Scholz and B. Lersmacher: Beitrag zur Kinetik des Drucksinterns von Tantalkarbid (Arch. Eisenhüttenwesen **33**, 635-641, 1962, No. 9). (Contribution to the kinetics of pressure sintering of tantalum carbide; in German.)

H 13: D. Gossel: Parametrische Verstärker (Elektron. Rdsch. **15**, 91-95 and 149-152, 1961, Nos. 3 and 4). (Parametric amplifiers; in German.)

H 14: H.-U. Harten: Oberflächenleitung und Oberflächenrekombination an der Grenze Silicium-Elektrolyt (Z. Naturf. **16a**, 459-466, 1961, No. 5). (Surface conduction and surface recombination at the silicon-electrolyte boundary; in German.)

H 15: K. J. Schmidt-Tiedemann: Optische Doppelbrechung durch freie Träger in Halbleitern (Z. Naturf. **16a**, 639, 1961, No. 6). (Optical birefringence by free carriers in semiconductors; in German.)

H 16: H. Severin: Stand der Entwicklung von Ferriten und ihre Anwendung (Elektron. Rdsch. **15**, 253-258, 1961, No. 6). (Status report on ferrites and their applications; in German.)

H 17: K. J. Schmidt-Tiedemann: Symmetry properties of warm electron effects in cubic semiconductors (Phys. Rev. **123**, 1999-2000, 1961, No. 6).

H 18: K. Rohwer: Eine Wanderfeldröhre für 3 kW Dauerleistung im Gebiet der Dezimeterwellen (Nachrichtentechn. Fachber. **22**, 100-102, 1961). (Travelling-wave tube for 3 kW continuous output in the decimetre wave band; in German.)

H 19: K. J. Schmidt-Tiedemann: Stress optical constants of germanium (J. appl. Phys. **32**, 2058-2059, 1961, No. 10).

H 20: G. Schulten: Eine neue Messleitung für dielektrische Oberflächenwellen-Leitungen (Nachrichtentechn. Z. **14**, 445-448, 1961, No. 9). (A new test line for dielectric surface-wave transmission lines; in German.)

H 21: K. J. Schmidt-Tiedemann: Experimental evidence of birefringence by free carriers in semiconductors (Phys. Rev. Letters **7**, 372-374, 1961, No. 10).

H 22: H.-U. Harten: Inversionsschichten in Silicium an der Grenze zu einem Elektrolyten (Z. Naturf. **16a**, 1401, 1961, No. 12). (Inversion layers at the silicon-electrolyte boundary; in German.)

H 23: E. Neckenbürger: Zur Wellenausbreitung an einem längsmagnetisierten Ferritstab in elektrisch anisotroper Umgebung (Z. angew. Phys. **14**, 282-288, 1962, No. 5). (Wave propagation on an axially magnetized ferrite bar in an electrically anisotropic environment; in German.)

VOLUME 25, 1963/64, No. 2/3 pp. 33-80 Published 18th March 1964

Philips Technical Review

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

This number is devoted entirely to the work of the Institute for Perception Research, Eindhoven. The work of this research establishment borders on several disciplines; it uses methods and apparatus that evoke by turns the electronic laboratory and the doctor's or the psychologist's consulting room, and deals with problems that touch practically every corner of the Philips industries. These and other aspects are discussed in the introductory article by Professor Schouten, the director of the Institute. Five articles follow, which offer an insight into some fields of research on which the Institute is engaged and mention some of the results so far achieved. Finally, as on previous occasions in this journal, a brief historical survey is presented, Dr. Ten Doesschate of Utrecht having kindly contributed, at our request, an article describing how man came to measure human reaction times — a task which plays a prominent part in the work of the Institute

THE INSTITUTE FOR PERCEPTION RESEARCH

by J. F. SCHOUTEN *).

159.938

In the development of innumerable industrial products it is necessary to reckon with the properties of human hearing and vision. When these products involve operating and handling, the human capacity to perform actions also enters into account. In the realm of the Philips industries, examples are not hard to find. Radio, gramophones, hearing aids, lighting and television, are instances that first spring to mind. The same considerations apply equally, however, to measuring instruments, X-ray equipment, telephone installations and household appliances.

In developing and designing such products, then, industry has not only to consider the evolving possibilities and the limitations of technology but also the possibilities and the limitations in perception, in assimilation of perceptions and in performance of the users of the products.

Industry must also take account of the abilities of the men who are to make the products. The methods, tools and machines used must be adapted to the industrial worker so as to enable him to carry out his task as efficiently as possible. Moreover, the work must be organized in such a way that he can make the best use of his individual aptitudes.

With the increasing refinement of industrial products and methods, both these points of view are growing in importance. In fact, consideration of human aspects usually lags behind that of purely technical ones.

We are concerned here with the *subjective* phenomena that occur in humans wherever they come into contact with the outside world. On the one hand we see this in the case of the perception and the processing of the perception into e.g. a recognition, an opinion or a decision; on the other hand we see it in the action (and in the control of that action) by which man acts upon the outside world, i.e. a physical operation or the providing of information in the form of gesture, the spoken word or writing.

This cycle of phenomena, consisting of perception, processing, action and control, may be called the *informational cycle*, to distinguish it from other cycles that play a role in human life, in particular the metabolic cycle, which consists of the intake and assimilation of food and the excretion of waste products (see *fig. 1*).

*) Institute for Perception Research, Eindhoven.

With regard to the work performed in the informational cycle, we distinguish between energetic work (lifting, carrying, etc.), which can be expressed in a physical measure of energy, and perceptual work, which cannot be so expressed (listening, reading,

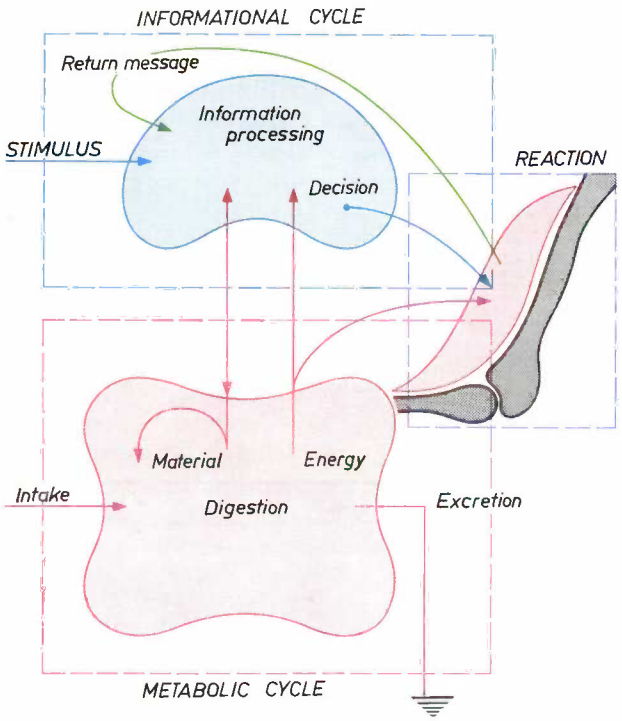


Fig. 1. Illustrating the reaction of an organism (human or animal) in connection with the informational cycle and the metabolic cycle.
This equivalent diagram is applicable to substructures of organisms, such as organs (heart, salivary gland, eye, etc.) and cells (muscle cell, nerve cell, etc.) as well as to communities of organisms and to "organizations" (family, herd, swarm; factory, club, state, etc.).

counting, calculating, sorting, mounting, checking, controlling, etc.). The increasing *mechanization* of industrial operations in the last hundred years has entailed an increasing shift from energetic to perceptual work. And considering the perceptual work in industry it can be said that *automation* in its turn is causing an increasing shift from elementary tasks to more comprehensive controlling and supervisory tasks.

The inclusion of human aspects in industrial development forms part of a separate branch of science which has been developed since 1940; in the United States it was first called "human engineering", and is now known as "human factors analysis". In Europe the term in current use is "ergonomics", the science of human work. The development referred to received impetus mainly from the advent of the new disciplines of cybernetics and information theory, founded on the now famous publications of N. Wiener [1]) and C. E. Shannon [2]). Remarkably enough, both these disciplines have their roots in problems of telecommunication engineering.

Foundation of the Institute; scope of activities

In 1956 the management of the Philips Research Laboratories drew up a plan to establish in Eindhoven a scientific laboratory that would be engaged on problems concerning the human informational cycle.

The authorities of the Eindhoven Technische Hogeschool (technological university), founded in 1956, were interested in this plan since it was part of their educational policy to lay emphasis on the interrelationships of man and technology in modern society. As a result, the Philips Research Laboratories and the Technische Hogeschool decided to form a joint laboratory for this purpose.

On 12th September 1957 the Deed of Foundation creating the "Institute for Perception Research" (in Dutch, "Instituut voor Perceptie Onderzoek" and abbreviated to I.P.O.) was effected. On 19th September 1958 a laboratory built for the Foundation was officially inaugurated. In it work members of Philips research staff together with staff and students of the Technische Hogeschool. The laboratory frequently accommodates guest workers.

The research in the laboratory covers the fields of hearing, vision, speech, human reactions and perceptual work. Many subjects are closely bound up with information theory, cybernetics or ergonomics

This diversity of subjects was adopted because it is becoming increasingly necessary to establish a relation between the results obtained in diverse branches of research. For example, the problem of optimum illumination cannot be solved by simply studying the properties of the human eye. The investigation should rather take the form of determining the perceptual load involved under different levels of illumination. Equally a study of the elementary human speech sounds would be incomplete if these sounds were not related to the behaviour of the human ear when listening to given sounds, and also to the behaviour of our brain during the production and understanding of whole words.

Broadly speaking, a study of perception should also include the associated action, and that of any action should include the associated perception: in human behaviour both are intimately interwoven.

Another aspect of great importance in dealing with

---

[1]) N. Wiener, Cybernetics or control and communication in the animal and the machine, Hermann, Paris 1948.
[2]) C. E. Shannon, A mathematical theory of communication, Bell Syst. tech. J. **27**, 379-423 and 623-656, 1948.

these subjects is to bring about cooperation between physicists, biophysicists, electronic and switching engineers on the one hand and psychologists, physiologists, phoneticians and linguists on the other. More general conclusions can be reached and results of greater practical value achieved if these studies of human behaviour are not dominated by a single line of inquiry.

Among the investigations mentioned, those concerned with perceptual work are more particularly on new ground. The practical importance of this research, however, can scarcely be stressed too much. True, in the course of the years many and various methods have been devised for measuring and evaluating factory tasks, for adapting clerical work to the human capacity to process information, and for designing tools and machines so that they can be efficiently operated by humans. These methods, however, are based mainly on empirical knowledge, whereas the aim must be, through scientific research, to gain a deeper insight into the laws that govern the optimum matching of man to his task and of the task to man.

So much for the scope of activities of the Institute for Perception Research. The Institute has been working now for about five years and a series of articles are presented in this issue which provide a survey of the research in progress. A few general observations will be given by way of introducing these articles.

Psychophysical measurements

All human behaviour is inconstant, even under the most constant conditions that can be realized in a laboratory. No human perceptions, deliberations or actions occur in a completely reproducible way.

This need not be too discouraging, for even the phenomena of inanimate nature are subject to inconstancy. In man, however, there are also conscious and unconscious factors at work, such as interest, motivation, prejudice, hope and fear. Because of these, his perceptions or reactions can differ from case to case and from one moment to another. This additional subjectivity in human behaviour seriously complicates efforts to try to establish facts about humans by means of quantitative measurements.

In an attempt to exclude these factors as far as possible, many experimenters make a point of using naive subjects, who may be expected to show the minimum of prejudice when taking part in the experiment. These subjects, however, often have an uncritical attitude, with the result that new phenomena, unknown to the experimenters and which might well be of particular importance, are in danger of being overlooked.

In our view it is therefore preferable to work with experienced observers who know what the problems are and understand them. Admittedly, their prejudice can affect the results of the measurements; where, for example, a subject has to determine the pitch of a sound, his judgement can well be influenced by what he expects to hear. This, however, is only the case within fixed and often measurable limits. It is precisely in this way that an insight is obtained into the reliability of the measurement of such subjective phenomena. After all, observation — even in the most primitive sense of simple perception — is an art that can only be learnt by practice. In the primitive case this implies no more than cultivating an automatic response; the higher plane of scientific observation calls for critical introspection, which cannot perhaps exclude the influential factors but does at least bring them to light.

The advantage of working with experienced subjects, especially in pioneering research, therefore offsets the disadvantage of their anticipations being included in the results. Later experiments with naive subjects then make it possible to check whether the phenomena perceived are also perceptible by unbiased persons.

The measurements that can be carried out in this field have, since Fechner [3]), often been referred to as *psychophysical measurements*. The term implies that phenomena of a psychological nature (such as reaction times, but also perceptions of e.g. brightness, colour, shape, loudness, pitch and timbre) are investigated by physical methods. Now, in physical measurements one can use a given instrument in a variety of ways, i.e. for determining whether a needle deflection is zero (null instrument), whether a smallest possible deflection or a change of deflection just occurs, whether the deflection is greater or smaller than in a previous case, and finally — if the scale is calibrated in units — for actually reading a deflection. Psychophysical measurements can be subdivided from the same point of view but here the person performing the test is the instrument under observation. We can thus list the following cases:

1) *Null method.* The eye, for example, is used as a null instrument and the objective relation is determined between the physical data of two coloured light-sources which produce the same impression of colour and brightness.

---

[3]) G. Th. Fechner, Elemente der Psychophysik, Leipzig 1859.

2) *Absolute threshold.* The physical threshold is determined at which a sound is just audible or a light just visible, or starts to be troublesome.
3) *Difference thresholds* (discrimination). The physical changes are determined which are needed to cause a just perceptible difference in brightness, colour, loudness, pitch, etc.
4) *Ranking.* A sequence of objective stimuli is determined according to the subjective impression they give of being larger or smaller, brighter or darker, higher or lower, etc.
5) *Gauging.* The quantitative relation is determined between the objective stimulus and the subjective impression on the basis of a subjective scale. In fact, a subjective perception can in some cases most certainly have a quantitative character. For instance, a musically trained subject can not only express differences of pitch in terms of higher or lower, but even with great precision in an interval: major third, fifth, octave, corresponding to the objective frequency ratios 4:5, 4:6, 4:8.

Finally, mention should be made here of a point to which our Institute has devoted considerable attention. Some measurements of subjective phenomena necessitate — because of the inconstancies referred to — large series of observations. While the designing and preparation of the experiments and the subsequent evaluation of the results offer the experimenter ample opportunity to exercise his ingenuity and analytical powers, the same cannot be said of the actual performance of the measurements, which is frequently dull routine. It is therefore our endeavour to make experiments — if they allow it — largely automatic. This endeavour, which is strikingly illustrated by the "DONDERS" reaction recorder, described in this issue, can be seen as an attempt to increase scientific productivity, in the sense that research workers need not spend a disproportionate amount of their time on work which scarcely makes use of their specific gifts. Viewed in this way the introduction of automatic systems fits into the context of adapting methods and tools to the worker, an aim which we mentioned at the beginning as being important to industry, but which is equally important to the work in a laboratory and indeed to the manner in which everyone organizes his own life and work.

Literature

Colin Cherry, On human communication, Chapman & Hall, London 1957.
D. E. Broadbent, Perception and communication, Pergamon Press, London 1958.
W. A. Rosenblith, Sensory communication, Wiley, New York 1961.
S. S. Stevens, Handbook of experimental psychology, Chapman & Hall, London 1951.

Summary. Short account of the origins of the Institute for Perception Research, formed about five years ago as a joint establishment of the Philips Research Laboratories and the Technische Hogeschool, Eindhoven. In this issue five articles are presented which describe some of the lines of research now being followed. The present article introduces these contributions with some general observations on the Institute's scope and the problems of psychophysical measurements.

THE PERCEPTION OF PITCH

by R. J. RITSMA *) and B. LOPES CARDOZO *).

534.321

Survey; pure tones and complex sounds

The perception of pitch, which can be remarkably accurate in persons with some amount of practice, depends on a mechanism which is as yet only partly understood. In this article we shall present a concise survey of the present state of knowledge concerning the perception of pitch and describe a number of experiments on which this knowledge is based.

To avoid misunderstanding, we begin by pointing out that the pitch of a sound is a *subjective* property and must therefore be measured by psychophysical methods [1]). Furthermore, pitch is not to be identified offhand with a sound frequency. On the contrary: it is precisely the purpose of research on pitch perception to discover the dependence of the perceived pitch on the various parameters with which a given sound can be described. Research into this subject is possible owing to the fact that pitch is a "one-dimensional" quantity: given two tones one can always ascertain whether their pitch is the same and, if not, which is higher. Thus, anyone can check the pitch he assigns to a given tone by comparing it with a reference tone, and so circumvent the difficulty that pitch, like any perception, cannot be observed directly.

From what follows it will be seen that a distinction must be made between sounds that consist of a single sinusoidal vibration (pure tones) and sounds that show a composite spectrum (complex sounds). As regards the first, the frequency can in the conventional way be taken as a measure of the pitch. This makes it possible, for example, to write a scale of musical tones as a series of numbers. In the pitch of a complex sound, however, it may be that the frequency of a pure tone that sounds just as high does not correspond to one of the frequencies contained in the spectrum, and cannot easily be derived from them. Instances are to be found in the chimes of church bells and in the human voice [2]).

We shall demonstrate below that the latter phenomenon is *not* to be explained in terms of the place theory, which was until recently generally accepted, and then indicate the probable direction in which an explanation should be sought. First, referring to *fig. 1*, we shall briefly recapitulate this theory [3]).

Fig. 1*a* gives a schematic representation of the human ear. Left, the external ear canal (auditory meatus), separated from the middle ear by the tympanic membrane, and right the inner ear. The latter consists of an oblong cavity (35 mm long) which is filled with a fluid and divided lengthwise into two by the cochlear partition, which is set in vibration by sound waves. This partition consists, among other things, of the basilar membrane in which the hair cells of Corti lie. These cells are attached by nerve fibres to the auditory nerve and detect the deflection of the relevant point of the basilar membrane. In reality the drawn section of the inner ear, the cochlea, is wound in a spiral resembling a snail-shell (which is what the term cochlea means in Latin) [4]). Now the place theory, very briefly, states firstly that every frequency is

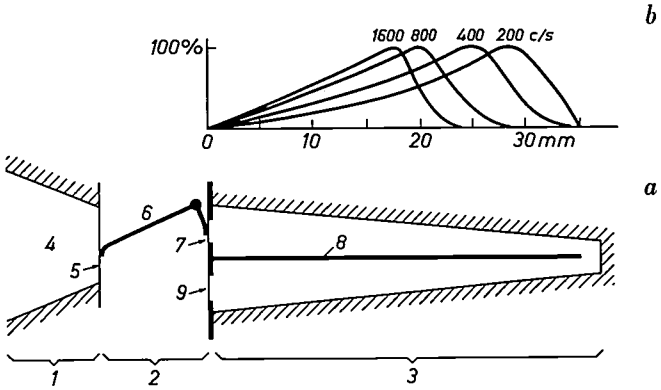


Fig. 1. *a*) Schematic representation of the human ear. From left to right the external ear (*1*), the middle ear (*2*) and the inner ear (*3*). The air vibrations pass down the ear canal *4* to the tympanic membrane *5*. The movement of the latter is transmitted via the middle ear ossicles *6* — represented as a lever with a long and a short arm — to the oval window *7* which closes the inner ear. This sets up, through the fluid in the inner ear, a pressure wave which sets in motion the cochlear partition *8*, thereby stimulating the auditory nerve terminals originating in this region. The amplitude pattern of this motion for pure tones at various frequencies is sketched in (*b*). The peak of such a curve lies closer to the end of the cochlear partition the lower the frequency. The fluid movement required to set *7* in motion is made possible by the elastic round window *9*. In reality the space *3* (the cochlea) is rolled up in a spiral.

*) Institute for Perception Research, Eindhoven.

1) See the first article in this number: J. F. Schouten, Philips tech. Rev. **25**, 33-36, 1963/64.

2) See A. Cohen, Phonetic Research, Philips tech. Rev. **25**, 43-48, 1963/64.

3) The place theory is dealt with extensively by e.g. H. Fletcher, Speech and hearing in communication, Van Nostrand, New York 1953.

4) For an anatomical description see: E. G. Wever and M. Lawrence, Physiological acoustics, Princeton University Press, Princeton 1954.

associated with a definite location on the cochlear partition, and secondly that every location is associated with a definite pitch.

For pure tones the process can in fact be summarized schematically in the way described: from the work of Von Békésy [5]) we now know that the travelling waves which, via the oval window of the cochlea, are set up in the fluid of the inner ear when a sound vibration strikes the tympanic membrane, are "damped" more rapidly the higher is the frequency; only waves of fairly low frequency "pass through" to the end of the inner ear. This is illustrated in fig. 1*b*, where the amplitude of sound waves of several frequencies is plotted (schematically) as a function of location in the cochlea. The point of the cochlea corresponding to a certain frequency can be identified with the abscissa of the peak of the relevant curve, i.e. with the place where the excitation of the cochlear partition is strongest. The inner ear can thus indeed be regarded as a frequency indicator.

Within certain limits the inner ear can even be regarded as a *frequency analyser:* to a restricted extent the ear is capable of identifying the individual pure tones contained in a complex sound. This can be demonstrated, for example, by the following experiment. If one listens to the complex sound produced by a periodic pulse, it is first heard as a
I single tone with a sharp timbre. If one of the lower harmonics is removed from the spectrum of such a tone pulse (*fig. 2*) the timbre becomes somewhat sharper. When the harmonic is now restored it is then heard, because it has been drawn to the listener's
II notice, as a separate tone.

Experiments of the following type show that, as far as complex sounds are concerned, the second postulate of the place theory — that every place in the cochlea is associated with a definite pitch — is not valid. We produce, for example, a melody (see fig. *3a*) of tones generated in the manner described above. The pitch is varied by varying the repetition frequency of the pulses. Using a filter, however, we allow only those components of the sound spectrum to pass that have frequencies between about 2000 and 3000 c/s. Nevertheless, the tones are heard to have the same pitch as the notes in fig. 3*a*, i.e. the
III pitch of pure tones of about 200 to 300 c/s.

If we displace the frequency range passed in such an experiment (fig. 3*b*) the listener hears the change of frequency range only as a change of timbre, *not* as a change of pitch. The tones still correspond in pitch to the written notes.

[5]) See G. von Békésy, Experiments in hearing, McGraw-Hill, New York 1960.

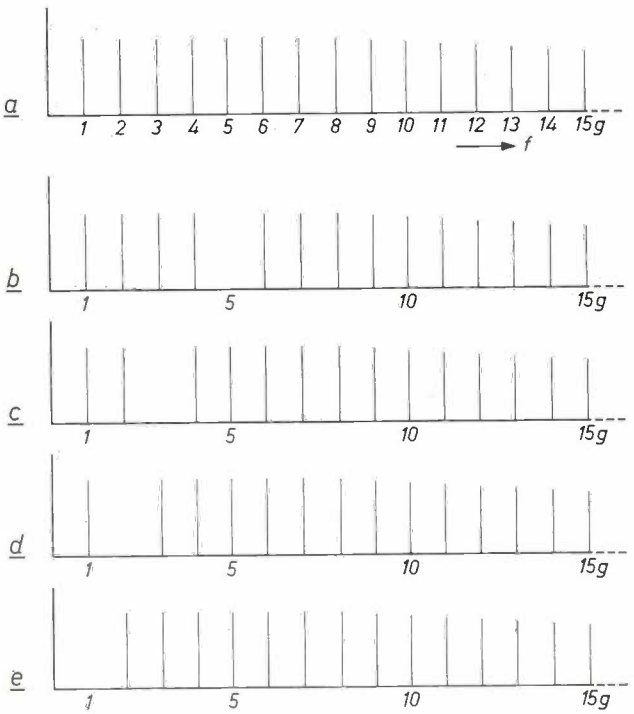


Fig. 2. *a*) Spectrum of a tone pulse, i.e. a tone produced by a periodic pulse of very short duration. The intensity of the spectral lines is plotted versus their frequency (f) expressed as a multiple of the fundamental frequency g. The spectrum consists of numerous harmonics of roughly equal strength.
b), *c*), *d*) and *e*) The same after removing, respectively, the 5th, 3rd, 2nd and 1st harmonics. In all cases the pitch corresponds to that of the fundamental; the only difference is in timbre.

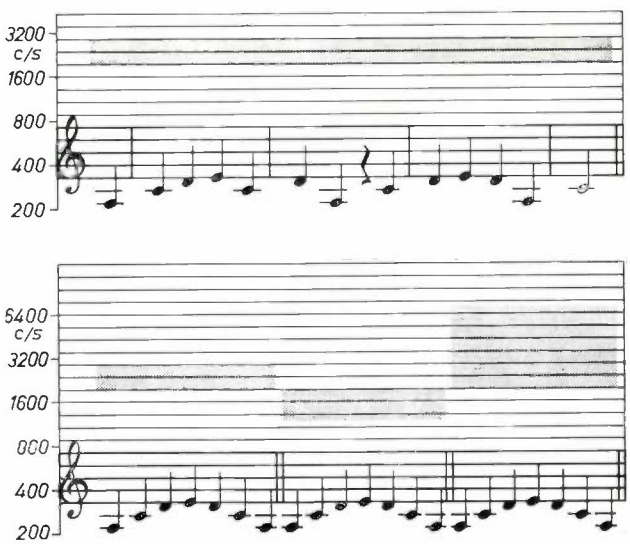


Fig. 3. *a*) When a melody is made of tone pulses (repetition frequency 200-300 c/s) of which only the harmonics between 2000 and 3000 c/s are passed, the pitch still corresponds to that of the scored notes. The range of harmonics passed is represented by the shaded strip in the lines above the staff. (The staff plus auxiliary lines can be regarded as a rectangular coordinate system, with the logarithm of the frequency corresponding to the pitch on the ordinate.)
b) When the range of passed frequencies is shifted (see the shaded strips), only the *timbre* of the tones changes, *not* the pitch.

In all cases, then, a pitch is heard which is the same as that of a pure tone having a frequency far below the lower limit of the passed frequency band, and which does not at all correspond to the excited location in the cochlea.

A complex sound which is heard as a single tone whose pitch does not correspond to one of the frequencies present has been given the name (tonal) "residue" [6]). The name arose because the pitch phenomenon described was first observed on a tone pulse (see above) from which all separately audible (lower) harmonics had been removed; the remaining group of harmonics could therefore properly be called a "residue". Later it was found that a group of only three successive harmonics, e.g. the 8th, 9th and 10th, already showed the residue effect. It may be mentioned that it is not in fact necessary to eliminate all separately audible harmonics.

Further consideration of the residue effect

For a long time it was doubted that the residue effect was an independent phenomenon. It was believed that a tone of the perceived frequency was really present in the ear; this tone, it was held, was simply a difference tone, produced by non-linear distortion in the ear. If this explanation were correct, the place theory would lose none of its validity.

For non-linear distortion the relation between the input signal $V_i(t)$ and the output signal $V_u(t)$ may in general be formulated as:

$$V_u(t) = C\,[V_i(t) + \delta_1 V_i^2(t) + \ldots] \quad \ldots\ldots \quad (1)$$

where C and δ_1 ... are constants. Given an input signal consisting of two equally strong sinusoidal components:

$$V_i(t) = A\,(\sin \omega_1 t + \sin \omega_2 t), \quad \ldots\ldots \quad (2)$$

we find from (1) for the output signal (putting $C = 1$):

$$V_u(t) = V_i(t) + \delta_1 A^2\,[1 - \tfrac{1}{2}\cos 2\omega_1 t - \tfrac{1}{2}\cos 2\omega_2 t + \cos(\omega_1 + \omega_2)t + \cos(\omega_1 - \omega_2)t] + \ldots\ldots$$

The non-linear distortion thus "contaminates" the original signal among others with a vibration of the frequency $(\omega_1 + \omega_2)$ and one of the frequency $(\omega_1 - \omega_2)$. In acoustics these are referred to respectively as the *sum tone* and the *difference tone*. The existence of these tones was propounded by musicians as far back as the middle of the 18th century (G. A. Sorge 1744; G. Tartini 1754).

The difference-tone hypothesis can be opposed on four quite distinct grounds:

1) The residue effect is found even with sounds that are so weak as to rule out non-linear distortion.
2) The tonal residue gives no beat effect with a pure tone of roughly the same pitch.
3) The tonal residue, unlike a pure tone of equal pitch, cannot be masked by noise whose frequency spectrum extends around the frequency f_p that corresponds to the perceived pitch; it can, however, be masked by noise whose spectrum contains the components of the residue [7]); see *fig. 4*.
4) An irrefutable argument against the difference-tone hypothesis is that the pitch of the tonal residue does not always correspond to the difference frequency. This can be observed, for example, in an experiment of the following type.

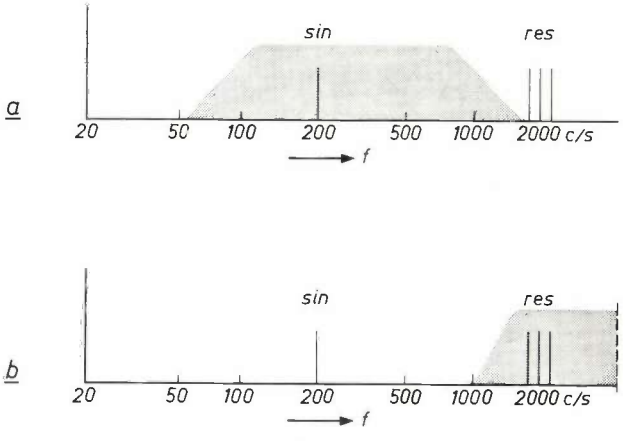


Fig. 4. Masking by noise shows that the tonal residue does not arise as a difference tone in the ear. If the noise spectrum is chosen to correspond roughly with the shaded part in (*a*), a pure tone of 200 c/s (*sin*) disappears, but not the tonal residue, which is heard to have the same pitch and originates from the complex *res* (frequencies 1800, 2000 and 2200 c/s). If the noise spectrum is as shown in (*b*), it is the residue that disappears and the pure tone remains audible.

We take a tonal residue of, say, three components with frequencies 1800, 2000 and 2200 c/s, i.e. the 9th, 10th and 11th harmonic of 200 c/s. The pitch corresponds to 200 c/s ($f_p = 200$ c/s). Next, we raise all frequencies by the same amount, e.g. 10 c/s, and repeat this a few times up to e.g. about 1850, 2050 and 2250 c/s. Although the frequency difference of the components remains constant in this operation (200 c/s), the pitch of the residue is heard to rise. When the frequencies have reached the values last mentioned, f_p has risen to nearly 205 c/s (*fig. 5*). If the frequencies are lowered in the same way, the pitch of the residue decreases.

The pitch of a complex sound of three components

The simplest complex sound on which the residue effect can be observed consists of three pure tones, each differing from the other by the same frequency interval. Apart from the loudness and the mutual intensity ratio, which are disregarded here, a complex sound of this kind can be defined with only *two* parameters — the frequency f of the centre com-

[6]) See J. F. Schouten, The perception of pitch, Philips tech. Rev. **5**, 286-294, 1940.

[7]) This was first demonstrated by J. C. R. Licklider, J. Acoust. Soc. Amer. **26**, 945, 1954.

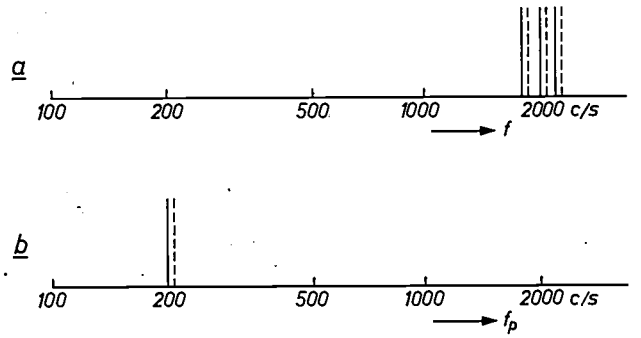


Fig. 5. When the components of a complex sound (*a*) are equidistantly shifted, the pitch of the residue (which is the frequency f_p of a pure tone of equal pitch, shown in *b*), does not remain constant but changes proportionally with the frequency f of the middle component.

ponent and the frequency difference g. Experiments of the type just described have shown that when f is varied while g is kept constant (equidistant shift) the pitch f_p varies proportionally with f [8]). This is not to say, however, that the pitch rises continuously as f is increased. Taking a complex sound consisting, as in the above experiment, for example, of the 9th, 10th and 11th harmonics ($f = 10g$), then when $f \approx 10.5g$ the pitch abruptly falls [9]). As f is further increased, f_p again rises and, at $f = 11g$, once more reaches the value g. Initially $f_p = f/10$; after the jump $f/11$, and so on. The variation of f_p with f thus has a sawtooth waveform; the pitch corresponds to the frequency difference ($f_p = g$) only when f is an integral multiple of g (*fig. 6*).

The proportionality between f_p and f can be explained in the following manner from the fine

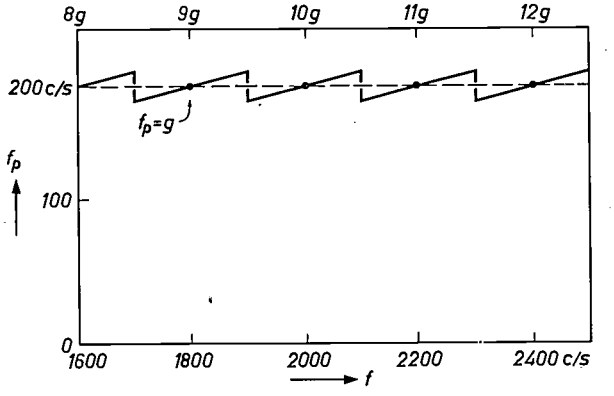


Fig. 6. When the three components of a complex sound (middle frequency f, frequency difference g) are equidistantly shifted, f_p does not rise continuously with f. When f has covered roughly half the distance to the next multiple of g, f_p makes a jump so that, when that multiple is reached, f_p is again equal to g.

structure of the acoustic signal. In our experiments this had the form shown in *fig. 7*. We have a vibration of frequency f (the solid line) the amplitude of which is modulated to a depth of 100% with the frequency g (dotted line); this was, incidentally, the way in which the signal was produced. Now the perceived pitch apparently corresponds to the periodicity of the amplitude modulation, in such a way that the ear derives this periodicity from the distance between the peaks which lie closest to the maxima of the dotted line [9]). This distance is of course proportional to the distance $1/f$. Expressed as a formula: $1/f_p = n/f$. If f rises so far that the distance $1/g$ between the said maxima is better approximated by changing to a neighbouring peak, i.e. by choosing for n a number that is larger by one, then the pitch makes a "jump". The same reasoning also explains why $fe = g$ when f is an integral multiple of g.

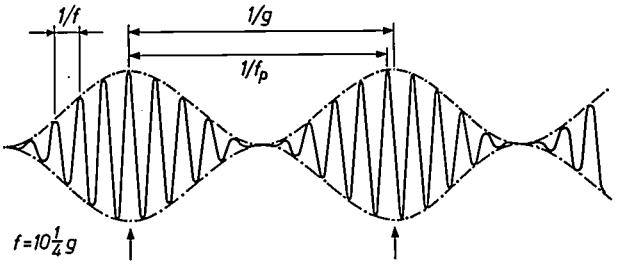


Fig. 7. Explanation of the effect that, when a complex of three components is equidistantly shifted, the pitch of the residue f_p changes proportionally with the middle frequency f. The sound signal is represented by the solid line (sinusoidal vibration of frequency f, modulated 100% by the frequency g). Although the ear derives the pitch from the periodicity of the signal, it does so only in the sense that it regards as such the distance between the peaks which, within each period $1/g$, are closest to maxima of the dotted envelope (see arrows). This distance is of course a multiple of the distance $1/f$. If f is not also a multiple of g (anharmonic complex sound) f_p will differ from g.

The foregoing leads to the conclusion that the organ which determines the pitch of complex sounds is not situated in the mechanical part of the ear, but is of a *neural* nature and must be in the auditory nerve tract or in the brain. Apparently it is not so much a spectrograph as a kind of "time measuring device" which analyses the fine structure of the signal [10]).

To conclude this section it should be pointed out that the residue effect does not occur with any arbitrary combination of frequencies [11]). *Fig. 8* shows the existence region of the tonal residue for a complex sound of three components, the

[8]) J. F. Schouten, R. J. Ritsma and B. Lopes Cardozo, Pitch of the residue, J. Acoust. Soc. Amer. **34**, 1418-1424, 1962.

[9]) See also E. de Boer, On the "residue" in hearing, thesis, Amsterdam 1956.

[10]) R. J. Ritsma, A model of human pitch-extraction based on additive correlation, Proc. 4th int. Congr. Acoust. I, paper H51, Copenhagen 1962.

[11]) R. J. Ritsma, Existence region of the tonal residue I, J. Acoust. Soc. Amer. **34**, 1224-1229, 1962.

frequency f being plotted versus the quotient n of f and g. The sloping lines are lines of constant g. We shall first consider only the solid contour $M = 100\%$. For the combinations of f and n values which fall within the area bounded by this contour a tonal residue is perceptible, but for those outside it is not. Although the form of the contour differs in details from one person to another, it is true in general that the highest f value is found at an n over ten, and the highest n value at $f = 2000$ to 3000 c/s. The lowest g — and hence roughly the lowest f_p — is about 35 c/s, the highest about 800 c/s.

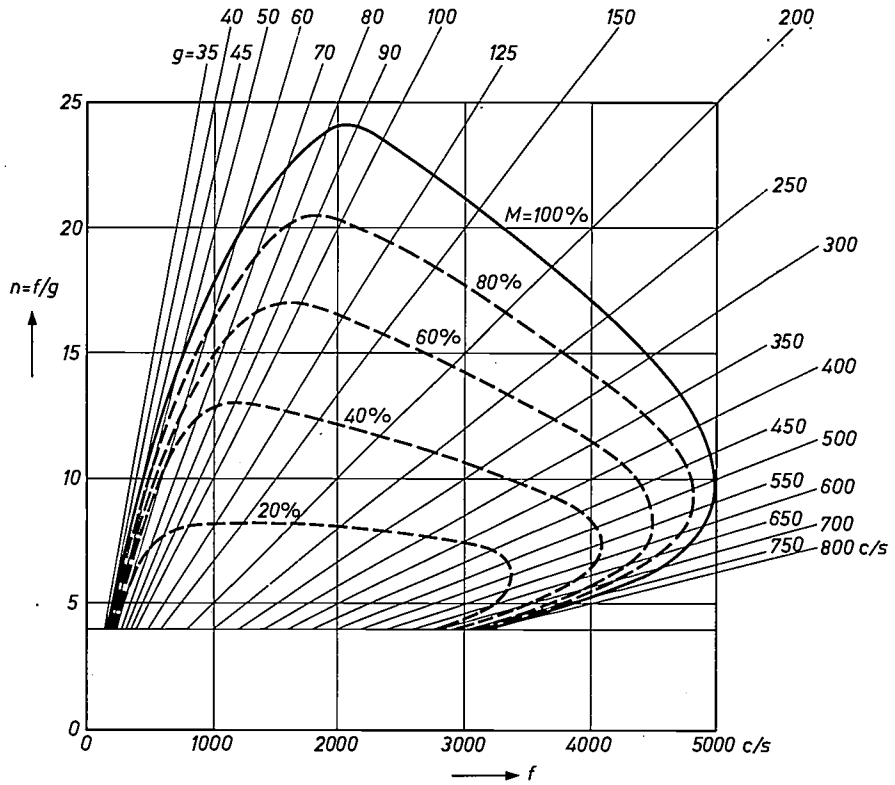


Fig. 8. Existence region of the tonal residue of a complex sound consisting of three neighbouring harmonics (middle frequency f, fundamental frequency g): the tonal residue is heard in the region bounded by the solid contour. If the two outer components are attenuated (experimentally by modulating f to a depth M less than 100%) the result is a smaller existence region.

If one experiments with a complex sound which is identical as regards frequencies but in which the components with frequencies $(f-g)$ and $(f+g)$ are weaker, the existence region is smaller. The vibration pattern of such a complex sound is in principle the same as that in fig. 7, but the modulation depth M is smaller than 100%. The dotted contours in fig. 8 give the boundaries of the existence region for the indicated values of M. The contraction of the existence region with decreasing M explains why the tonal residue is heard at a certain moment to disappear when the two outer components of a complex sound of three components are progressively reduced in strength.

The variation of f_p with f in the equidistant displacement of a complex sound of three sinusoidal (pure) vibrations need not always take the form sketched in fig. 6. A subject who *concentrates* on the change in the pitch of the residue can "postpone" the jump in f_p [8]). Some plots of this phenomenon are shown in *fig. 9*. It can be seen that the frequency regions to which the various f_p-f curves relate overlap each other to such an extent that at e.g. $f = 1800$ c/s, no fewer than four values can be assigned to f_p. *The pitch assigned by the human ear to a given sound is therefore not in all cases unambiguously determined by the physical parameters of the sound.* This is referred to as the *ambivalence* of pitch perception.

If, on the other hand, one listens without bias to such a complex sound which is equidistantly shifted from, say, $f = 10\,g$ to $f = 11\,g$, and if the perceived pitch is not regularly compared with that of a reference tone, our experience shows that one makes the jump *unconsciously*. The listener thinks he hears a continuously rising tone, and at the end of the experiment finds to his surprise that f_p is as equal to g as at the beginning.

As can be inferred from fig. 9, the change in f_p is not exactly equal to the nth part of that in f. The magnitude of the dis-

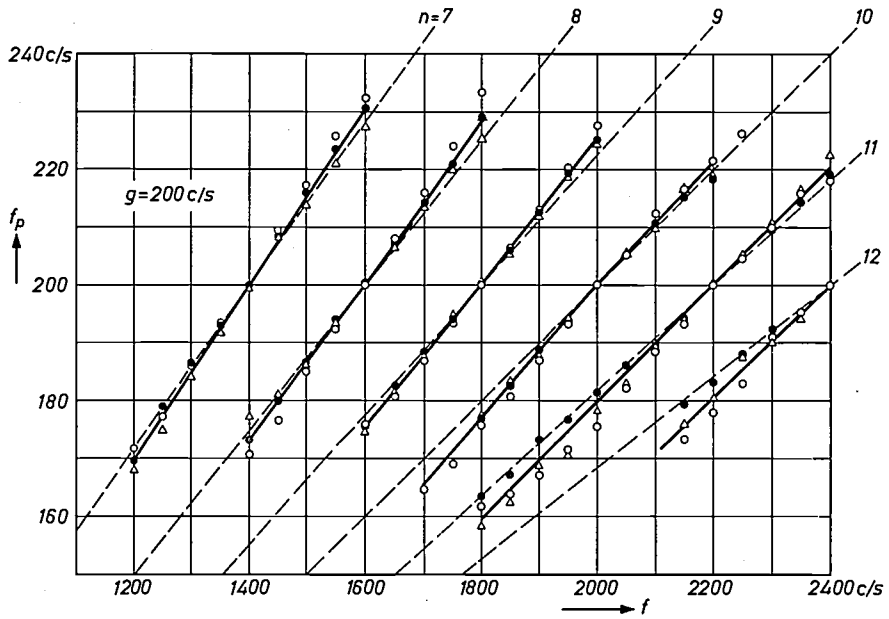


Fig. 9. The change in the pitch f_p of a complex sound of three components when the middle frequency f is varied can, if the listener concentrates on it, be followed for a short time beyond the frequency at which an unsuspecting listener hears a jump in f_p (cf. fig. 6). This means that the pitch of such a complex sound is not unambiguously established. In the case to which the graph relates, no fewer than four pitches can be assigned to the complex tone at $f = 1800$ c/s. (According to the hypothesis on the variation of f_p (see fig. 7) the change in f_p should be *exactly* equal to the nth part of the change in f, and the curves should coincide with the dotted lines. As can be seen, there is a small systematic discrepancy; no explanation of this effect has yet been found.)

crepancy appears to depend on the loudness. Research into the cause of this effect is in progress.

The pitch of sounds of short duration

So far we have been dealing with more or less sustained sounds. We shall now turn our attention to sounds of short duration.

After the foregoing considerations regarding the tonal residue of a complex sound of three components, and the proportional variation of its pitch with f when the components are equidistantly shifted, the question arises as to how many periods $1/g$ the ear needs to perceive the pitch of the residue. Experiments at this Institute have shown that the number of periods for values of f_p between 200 and 475 c/s is always *four*. At a g value of, say, 200 c/s the ear is thus apparently able to establish the pitch fairly accurately in 20 ms.

Whereas the allocation of pitch to a tonal residue is governed by the number of periods $1/g$, the decisive factor as regards *pure tones* is the duration of the sound. A relatively long "burst" or pulse is heard distinctly as a tone. If the duration is shortened, hardly anything changes at first, but at a certain
VI critical value the sound begins to change in character, gradually going over from a tone into a click.

Experiments on the perception of the pitch of short tone bursts can best be done by letting the subject hear in quick succession two equally long bursts of dissimilar frequencies. To start with, the frequency difference is very small and is gradually increased. The frequency difference is noted at which the subject only just hears the pitch of the two
VII "tones" to be no longer identical.

The results of such experiments carried out in this Institute are summarized in *fig. 10*. The sounds used were bursts of a sinusoidal vibration of 1000 c/s, the beginning and end of each of which coincided with a zero transition of the vibration [12]). The quantity Δf is the critical frequency difference just mentioned and Δt is the duration of the burst.

As can be seen, Δf is nearly constant and very small (≈ 1 c/s) provided that Δt is longer than 50 ms. When Δt is shortened still further, Δf rises,

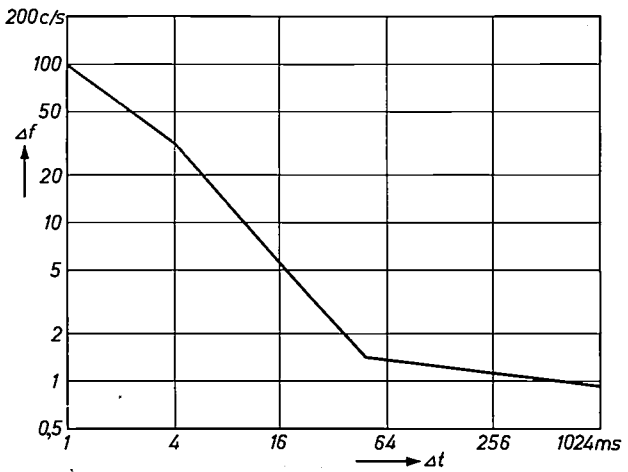


Fig. 10. The distinctness with which a short sinusoidal burst is heard as a tone depends on the duration Δt of the burst. As Δt is decreased the listener becomes more and more uncertain of the pitch. The graph shows the result of experiments in which the subject listened to pairs of tone bursts (1000 c/s) of length Δt presented in quick succession, each two bursts having a small frequency difference which was gradually increased. Plotted on the ordinate is the (threshold) frequency difference Δf, at which the subject first hears a difference in the pitch of the two tones.

[12]) B. Lopes Cardozo, Frequency discrimination of the human ear, Proc. 4th int. Congr. Acoust. I, paper H 16, Copenhagen 1962.

but not very steeply. At $\Delta t = 2$ ms, Δf is still no higher than 50 c/s, i.e. only $2\frac{1}{2}\%$ of f. The human ear is thus apparently able, when presented with sinusoidal vibrations — i.e. vibrations whose periodicity is identical with the reciprocal of the frequency — to distinguish a frequency difference of a few per cent in a few milliseconds. The plot in fig. 10 for $\Delta t < 50$ ms can be described to a good approximation by the equation $\Delta f \Delta t = \text{constant}$. The constant, which differs from one person to another, is of the order of magnitude of 0.1.

Finally, a few remarks on more irregular complex sounds than the three-component groups discussed, and on sounds which change rapidly in character after their beginning.

The properties of stationary harmonic complex sounds with more than three components can often be derived from the properties mentioned of the three-component sounds discussed. Complex sounds whose components are *not* equidistant have not yet been investigated.

As far as sounds are concerned which change rapidly after their beginning, our knowledge also shows gaps. In spite of the considerable importance of the "attack" effect — a piano note deprived of its opening is scarcely recognizable as such [13]) — this effect has not yet been extensively studied on musical instruments. On the contrary, intensive research has been carried out on the human voice. This research, however, should be classed among the phonetic fields of study.

Examples of the experiments denoted in this article by Roman numerals (I to VII) in the margin are contained on a gramophone record made by the I.P.O. [14]). With the aid of this record the reader can hear for himself the properties of pitch perception discussed in this article.

[13]) Compare, for example, the sound examples given with the article: H. Badings and J. W. de Bruyn, Electronic music, Philips tech. Rev. **19**, 191-201, 1957/58.

[14]) This gramophone record (on which the sound examples are accompanied by a commentary) can be obtained free of charge by sending in the coupon attached to the summary sheet enclosed in this number.

Summary The well-known relation between the frequency and pitch of pure tones is often not applicable to complex sounds: a group of three (or more) neighbouring harmonics (frequencies $f-g$, f and $f+g$) possesses in a wide range of f,g combinations the same pitch as the fundamental tone (residue effect). Masking experiments with noise, and the fact that the pitch changes when the three components are equidistantly shifted (i.e. with g constant) show that the tonal residue does not arise in the mechanical part of the ear but is of neural origin. The latter experiments also demonstrate that the pitch does not correspond exactly to the periodicity of the signal envelope (frequency g) but is derived from the fine structure of the signal. Where complex sounds of short duration are concerned, the pitch of the residue is already heard if the sound is four periods $1/g$ long. As regards pure tones, the *duration* of the sinusoidal pulse is the decisive factor.

PHONETIC RESEARCH

by A. COHEN *).

534.4

Phonetics is concerned with the study of speech sounds, traditionally including their production and recently also their perception. One can say that phonetics studies communication from man to man by means of speech, the language spoken being considered not as an object of study but as a datum [1]). Phonetics cannot be regarded purely as a branch of linguistics, of biology or of physics; all these sciences make their contribution to it. For if we go link by link along the whole "communication chain" we find: 1) the human vocal organ (an object of anatomical and physiological study), 2) the system by which this organ produces the speech sounds (the study of articulation), 3) the uttered sounds treated as air vibrations (acoustics), 4) the ear and its associated neural elements (anatomy and physiology). That this last link of the chain — perception — is now comprised in phonetic research is attributable to the deliberate application of the fairly recent concepts of communication theory.

In the phonetic research carried out in the I.P.O. since 1959 the question posed as central problem is: *Which physical properties of speech sounds are essential to the recognition of the linguistic content of the*

*) Institute for Perception Research, Eindhoven.

[1]) The science of phonetics is treated in various manuals and text books. See e.g. L. Kaiser, Manual of phonetics, North Holland Publ. Co., Amsterdam 1957. The science concerned with the structure of a given language (or dialect) in terms of constituents of speech that distinguish one utterance from another (called phonemes) is a branch of linguistics called *phonemics*. Since the only yardstick applied is the distinguishability of the utterances, no distinction is made in the English phonological system, for example, between the k's of *cool* and *keel*. In phonetics these are different sounds.

but not very steeply. At $\Delta t = 2$ ms, Δf is still no higher than 50 c/s, i.e. only $2\frac{1}{2}\%$ of f. The human ear is thus apparently able, when presented with sinusoidal vibrations — i.e. vibrations whose periodicity is identical with the reciprocal of the frequency — to distinguish a frequency difference of a few per cent in a few milliseconds. The plot in fig. 10 for $\Delta t < 50$ ms can be described to a good approximation by the equation $\Delta f \Delta t = \text{constant}$. The constant, which differs from one person to another, is of the order of magnitude of 0.1.

Finally, a few remarks on more irregular complex sounds than the three-component groups discussed, and on sounds which change rapidly in character after their beginning.

The properties of stationary harmonic complex sounds with more than three components can often be derived from the properties mentioned of the three-component sounds discussed. Complex sounds whose components are *not* equidistant have not yet been investigated.

As far as sounds are concerned which change rapidly after their beginning, our knowledge also shows gaps. In spite of the considerable importance of the "attack" effect — a piano note deprived of its opening is scarcely recognizable as such [13]) — this effect has not yet been extensively studied on musical instruments. On the contrary, intensive research has been carried out on the human voice. This research, however, should be classed among the phonetic fields of study.

Examples of the experiments denoted in this article by Roman numerals (I to VII) in the margin are contained on a gramophone record made by the I.P.O. [14]). With the aid of this record the reader can hear for himself the properties of pitch perception discussed in this article.

[13]) Compare, for example, the sound examples given with the article: H. Badings and J. W. de Bruyn, Electronic music, Philips tech. Rev. **19**, 191-201, 1957/58.

[14]) This gramophone record (on which the sound examples are accompanied by a commentary) can be obtained free of charge by sending in the coupon attached to the summary sheet enclosed in this number.

Summary The well-known relation between the frequency and pitch of pure tones is often not applicable to complex sounds: a group of three (or more) neighbouring harmonics (frequencies $f-g$, f and $f+g$) possesses in a wide range of f,g combinations the same pitch as the fundamental tone (residue effect). Masking experiments with noise, and the fact that the pitch changes when the three components are equidistantly shifted (i.e. with g constant) show that the tonal residue does not arise in the mechanical part of the ear but is of neural origin. The latter experiments also demonstrate that the pitch does not correspond exactly to the periodicity of the signal envelope (frequency g) but is derived from the fine structure of the signal. Where complex sounds of short duration are concerned, the pitch of the residue is already heard if the sound is four periods $1/g$ long. As regards pure tones, the *duration* of the sinusoidal pulse is the decisive factor.

PHONETIC RESEARCH

by A. COHEN *).

534.4

Phonetics is concerned with the study of speech sounds, traditionally including their production and recently also their perception. One can say that phonetics studies communication from man to man by means of speech, the language spoken being considered not as an object of study but as a datum [1]). Phonetics cannot be regarded purely as a branch of linguistics, of biology or of physics; all these sciences make their contribution to it. For if we go link by link along the whole "communication chain" we find: 1) the human vocal organ (an object of anatomical and physiological study), 2) the system by which this organ produces the speech sounds (the study of articulation), 3) the uttered sounds treated as air vibrations (acoustics), 4) the ear and its associated neural elements (anatomy and physiology). That this last link of the chain — perception — is now comprised in phonetic research is attributable to the deliberate application of the fairly recent concepts of communication theory.

In the phonetic research carried out in the I.P.O. since 1959 the question posed as central problem is: *Which physical properties of speech sounds are essential to the recognition of the linguistic content of the*

*) Institute for Perception Research, Eindhoven.

[1]) The science of phonetics is treated in various manuals and text books. See e.g. L. Kaiser, Manual of phonetics, North Holland Publ. Co., Amsterdam 1957. The science concerned with the structure of a given language (or dialect) in terms of constituents of speech that distinguish one utterance from another (called phonemes) is a branch of linguistics called *phonemics*. Since the only yardstick applied is the distinguishability of the utterances, no distinction is made in the English phonological system, for example, between the k's of *cool* and *keel*. In phonetics these are different sounds.

sounds? Like the characters of many other code systems, such as for instance the letters of the alphabet, speech sounds contain a fairly large amount of information which is not absolutely necessary for recognition. This appears from the obvious fact that one can hear not only the content of what is spoken but also draw conclusions regarding the identity of the speaker and perhaps his state of mind. It will be evident that the problem as to which physical properties are essential to the recognition of the linguistic content of speech sounds cannot be solved by means of acoustic analysis alone, but that the fourth link of the chain — the perceptual link — must play its part. In other words: the last and most important part of the measuring apparatus must be the listener himself.

Broadly speaking, there are two steps in this type of research. First, a speech sound is broken down into its components and an attempt is made to determine which components, or characteristics of it, are essential to recognition. Next, the results of this analysis are checked by using them as the basis for synthesizing speech sounds entirely by instrumental means. The manner in which all this is done will be discussed below.

The acoustic investigation of speech sounds was for a long time somewhat hampered by the inadequacies of the experimental equipment. An initial improvement came with the advent of the cathode-ray tube in the thirties, which made it possible to ascertain more exactly how the amplitude of a vibration varies with time. This was joined in the forties by the sound spectrograph, which analyses the sound from moment to moment and records the variation of the spectrum with time [2]).

[2]) A detailed description of this instrument will be found in R. K. Potter, J. A. Kopp and H. C. Green, Visible speech, Van Nostrand, New York 1947.

As regards the instrumental aspect of phonetic research, the I.P.O. takes the view that, just as in physical research, the equipment must be designed with a view to the demands imposed by the investigations and not the other way around. Although the instruments mentioned, which may now be considered conventional, are admittedly used by the Institute, they are of limited significance in carrying out the investigations outlined above; the principal instruments used have been specially designed for the purpose, and will be described in this article. Their use has made it possible, among other things, to answer the time-honoured question as to whether speech, from the phonetic viewpoint, can be subdivided or not into distinct time segments.

In the following we shall describe in turn the procedures by which speech sounds are analysed and artificial speech synthesized. We shall conclude with some observations on the practical application of the results obtained by our method of research.

Speech analysis

Before considering our method of speech analysis and the equipment employed, we shall touch briefly on the conventional instruments and the information which they can supply. As stated, with the oscilloscope one can follow the amplitude variation of speech sounds. To do so the time-base period must be made long enough for the whole word (or part of the word) to be spoken within a single period. An example can be seen in *fig. 1a*, which shows the vibration pattern produced on the screen when the word "phonetics" is pronounced. Fig. 1*b* shows the envelope of this pattern, i.e. the amplitude waveform.

Examples of recordings of the word "fine", obtained with a sound spectrograph [2]), can be seen in *fig. 2*. The instrument has two settings of resolving

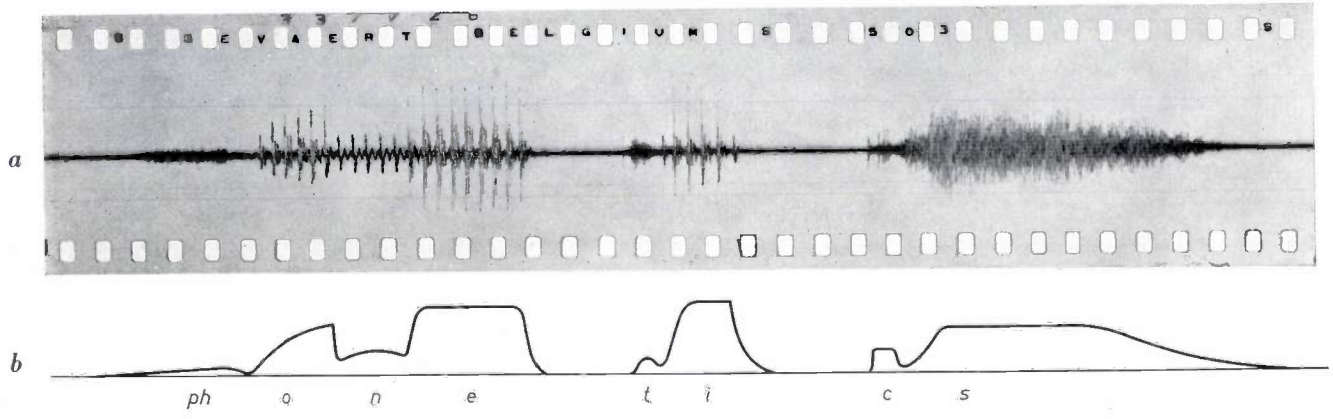


Fig. 1. *a*) Oscillogram of the word *phonetics*. Recorded on a continuously moving film while the oscilloscope time-base generator was switched off.
b) The amplitude envelope of the above pattern (giving the amplitude variation).

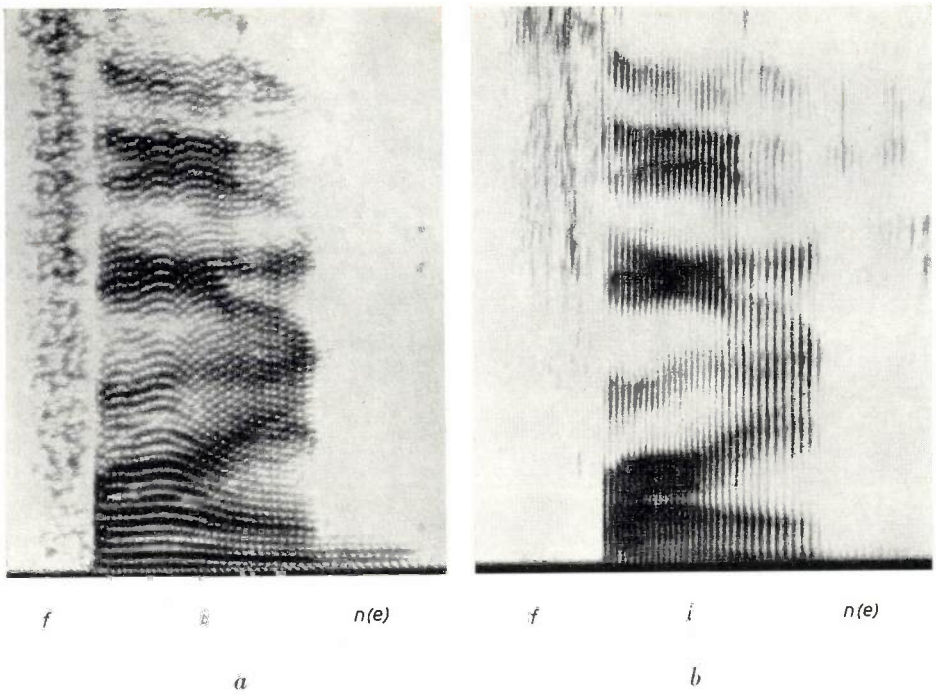


Fig. 2. Spectrograms of the word *fine*, *a*) recorded at high resolving power (approx. 50 c/s), *b*) at low resolving power (approx. 300 c/s). In (*a*) it is clear that *f* has a noise spectrum and the other sounds a line spectrum. The greater the density, the stronger is the spectral line. From (*b*) it can plainly be seen that there are frequency ranges in which the lines are strong (called *formants*), separated by ranges where they are weak.

power (50 c/s and 300 c/s). In fig. 2*a*, recorded at the higher resolving power, it can be clearly seen that the vowel has a line spectrum and the *f* a noise spectrum. The spectral lines of vowels are all harmonics of a fundamental tone, which is sometimes absent or barely visible in the spectrum. The density (blackening) of the spectral lines is greater whenever the relevant harmonic is stronger. In fig. 2*b*, which was recorded at the low resolution, the spectral lines run more or less one into the other. This clearly indicates that there are frequency regions in which the spectral lines have a relatively high intensity (called the *formants*), separated by regions where they are weak. The typical sound of a vowel is governed by the situation of these formants. This situation in its turn is governed by the position of the tongue — which largely determines the shape of the throat and mouth cavities — and is virtually independent of the fundamental tone. This can be exactly verified with another of the I.P.O.'s instruments — the "acoustic spectrum analyser" earlier described in this journal [3]).

The first of the instruments developed in the I.P.O. — referred to as a *gating circuit* — makes it possible to analyse the time structure of a word recorded on magnetic tape. *Fig. 3* shows schematically how this is done. The word is made audible by a playback head, which is switched on for only a short interval of time, while the beginning of this interval is repeatedly shifted about 10 ms in relation to the moments at which the word passes the playback head. Initially nothing is heard, then comes the first 10 ms of the word, then the first 20 ms, and so on. When the gate has almost passed the word, all that can be heard is the last 10 ms of the last sound.

There are at present three versions of the gating circuit. With the first a short loop of magnetic tape is continuously rotating. The shaft carrying the roller which transports the tape is coupled by a gear system to a second shaft whose period of rotation is longer by a time Δt (about 10 ms) than the period of rotation of the tape. A cam on this shaft actuates a contact which triggers a monostable multivibrator (univibrator). The latter switches on the playback head and governs the length of time it remains switched on. In this way, then, the gate always opens a little later with respect to the moment at which the word passes. A condition of correct operation is that the tape should not slip on its transport roller. Drawbacks of this method are the susceptibility of the tape to stretching, and the impossi-

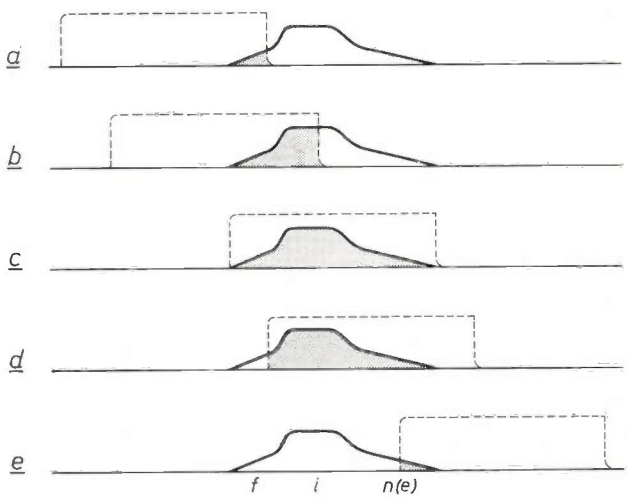


Fig. 3. Illustrating a phonetic analysis of the word *fine*, using a gating circuit. The word is represented by its envelope; the dashed line represents the response curve of the reproduction apparatus. The time interval during which the "gate" is open is shifted in steps of about 10 ms in relation to the word, so that initially a successively longer fragment of the word is heard (*a* to *c*). Thereafter the beginning is clipped (*d*) and finally only the end of the word is audible (*e*).

[3]) D. J. H. Admiraal, An acoustic spectrum analyser with electronic scanning, Philips tech. Rev. **21**, 349-356, 1959/60.

bility of changing Δt rapidly. An advantage is its great simplicity.

The second version uses an ordinary length of magnetic tape. This remains stationary, but lies along about two-thirds of the periphery of a rotating disc, in which the playback head is fitted. Again, the shaft carrying this disc has a contact which, in the same way as in the previous case, switches on the playback head. The gate can be displaced in relation to the word by moving the tape along the periphery a little. For this purpose the spool is turned through a small, constant angle by a special mechanism which is operated by hand. This method does not have the drawbacks of the previous one. Moreover, the tape containing the words to be analysed does not have to be cut, which simplifies storage. Another feature is that the gate can just as easily be run in reverse in relation to the word.

The third method is not partly mechanical but wholly electronic. The tape is again in the form of a loop, but now has two tracks which are used simultaneously. One track contains the word under analysis, the other has equidistant time marks which divide the time into units to be chosen, preferably 1 ms. After each revolution of the tape the gate opens for an adjustable number of time units.

One of the most surprising results so far found with the gating circuit is that the Dutch vowels in words such as *zijn*, *fijn*, *feit*, *zuid*, *goud*, etc., consist of two clearly distinct segments, each with its characteristic sound. The same applies to English words such as *fine*. One of the other discoveries is that when only a 25 ms fraction is passed of consonants like *f*, *z* and *s*, they sound respectively like *p*, *d* and *t* [4]).

What is surprising about the result first mentioned is that when one investigates the articulation of these sounds, scarcely any indication of distinct segmentation is found. X-ray and other studies of the movements of the mouth reveal that when the relevant words are pronounced the positions of the mouth show a gradual transition. Nor do oscillograms give any trace of segmentation (cf. fig. 3). The method of analysis described may be expected to prove useful also in investigations of the change of pitch during speaking.

Speech synthesis; the IPOVOX

To examine the perceptual value of the analysed properties of the speech sounds investigated, and in order to verify the results of the analysis, an apparatus was built — called "IPOVOX" — with which speech can be synthesized. Broadly, the apparatus consists of three sections: the first contains the sources of a number of continuous basic sounds, the second contains circuits for giving each of the sounds used the required amplitude envelope — referred to as *variable function gating circuits* (see *fig. 4*), and the third is an elaborate programme selector, which ensures that each sound fragment is made audible in the right amplitude envelope at the right moment. The continuous basic sounds consist of noise for the various explosives (*p* and *t*) and fricatives (*f* and *s*), and of tones with a harmonic spectrum for the vowels and vowel-like consonants (e.g. *l* and *m*). Both kinds of sounds are obtained respectively from one noise source and one signal generator by means of appropriate filters. The signal generator produces extremely narrow pulses. Its spectrum therefore contains a very large number of harmonics, of which certain groups can be used as required. *Fig. 5* shows schematically how the word "phonetics" is synthesized.

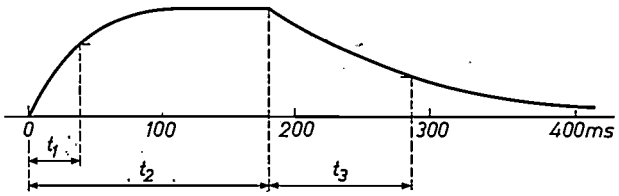


Fig. 4. In the I.P.O.'s method of speech synthesis the amplitude envelope of each sound segment is separately controlled by means of a *variable function gating circuit*. This is a special type of monostable multivibrator (univibrator) which makes it possible to adjust separately the length (t_2) of the amplitude envelope as well as the characteristic times of each of the flanks (t_1 and t_3).

The spectrograms of the spoken word and of the artificially produced version are presented in *fig. 6*. It can be seen that the vowels of the synthetic word are obtained with the aid of only two formants, whereas the natural vowels usually have four or more. This simplification does not significantly reduce the recognizability of the word. Further, the spectra of the *t* and the *s* are identical; what is essential here is the envelope of the sound heard — i.e. the setting of t_1, t_2 and t_3 (see fig. 4); of particular importance is the first segment. Finally, in the spectrogram of the spoken word the spectrum can be seen to change gradually when the sound *I* is uttered, and also to a lesser extent during ε [5]); this does not happen in the synthesized word. This simplification does not affect recognizability; the ear apparently cannot detect the difference between the two cases.

With this method of speech synthesis it proves possible to imitate live speech so closely as to bring out slight differences in pronunciation, as for instance between the Dutch *fonetiek*, the German *Phonetik* and the French *phonétique*.

[4]) A more detailed description and classification of the results will be found in A. Cohen and J. 't Hart, Segmentation of the speech continuum, Proc. 4th int. Congr. Acoust. II, Copenhagen 1962.

[5]) The phonetic symbols used are those agreed by the International Phonetic Association, London (see: The principles of the International Phonetic Association, London 1949) and will be found in the list of symbols and abbreviations given in many dictionaries.

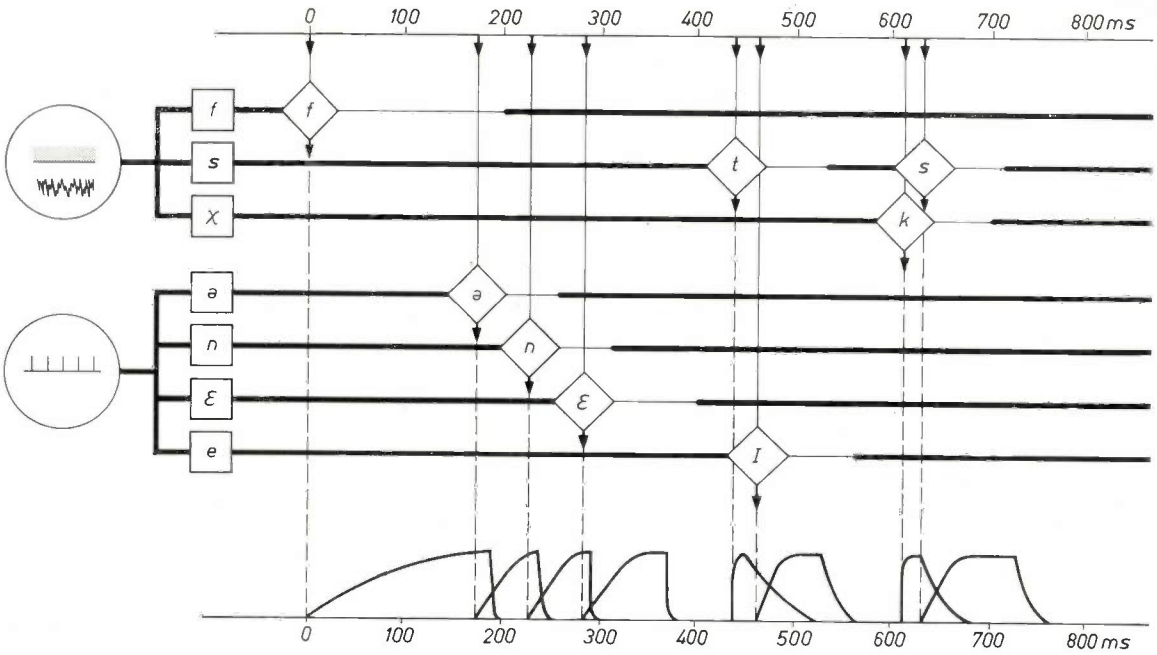


Fig. 5. Schematic illustration of how the word *phonetics* is synthesized. The consonants *f*, *s*, *t* and *k* are obtained by filtering the spectrum of a single noise source; the other sounds are produced by means of a generator of periodic very narrow pulses, of whose spectrum *two* suitably chosen regions are passed. With the aid of variable function gating circuits (cf. fig. 4) each sound is given the required amplitude envelope. A programme selector ensures that the sounds are made audible at the right moment. It will be noted that the *t* and the *s* are made with the same noise spectrum and differ only in their amplitude envelope. The same applies to *i* (written phonetically as *I*), which has the same spectrum as an *e* [5]).

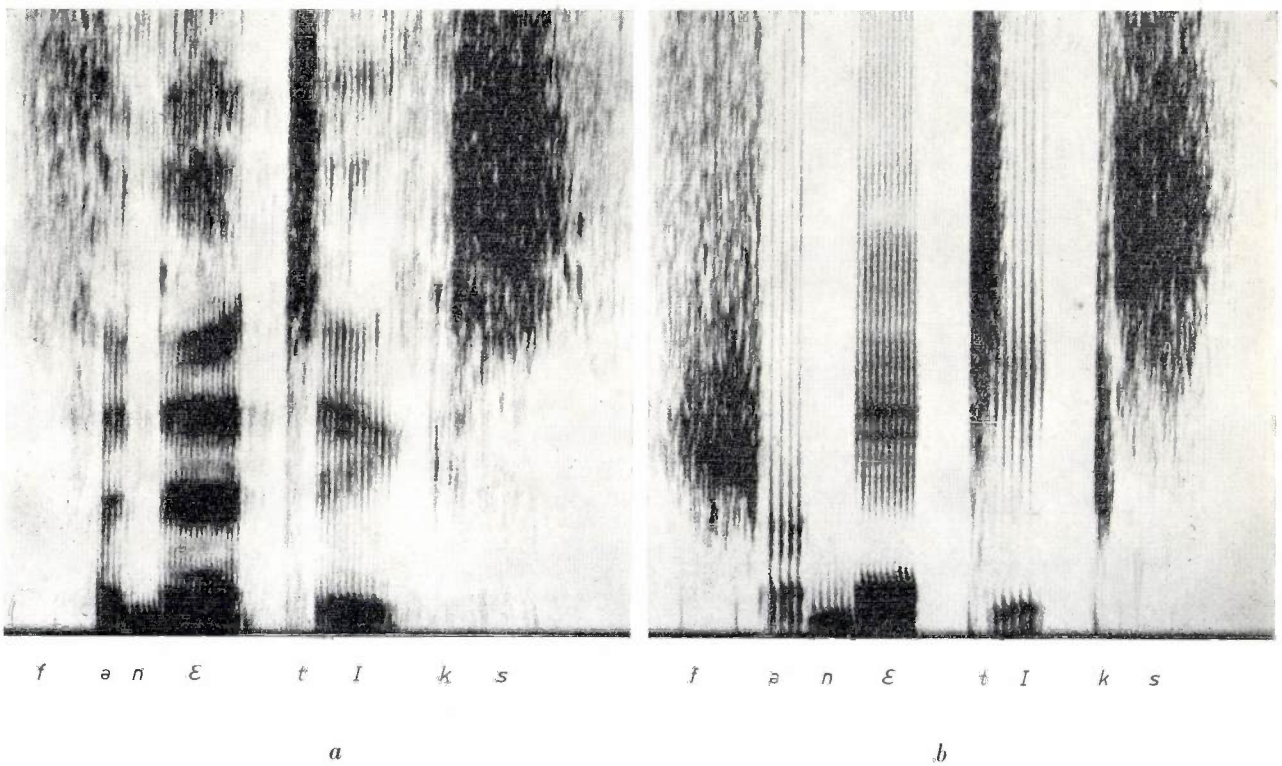


Fig. 6. Spectrograms of the word *phonetics*, *a*) spoken, *b*) synthetic. In spite of the simplification of the spectra in the synthesis (cf. fig. 5) and the absence of continuous transitions, as e.g. during the vowel sound *I*, the synthetic word is completely intelligible.

The best situation of the two formants of the synthetic vowels and vowel-like consonants was found from listening experiments with Dutch subjects [6]). The result of this investigation is summarized in *fig.* 7, in which the centre frequency of the lower formant (F_1) is plotted versus that of the second (F_2). The shaded area corresponds, in the judgement of the subjects, to sounds that cannot be produced by the human vocal organ. The ringed areas of the white field correspond to the sounds written within them. Their boundaries are of course not sharp and there is even some overlapping possible, as the boundaries were found to depend on the time during which the formant combination is made audible.

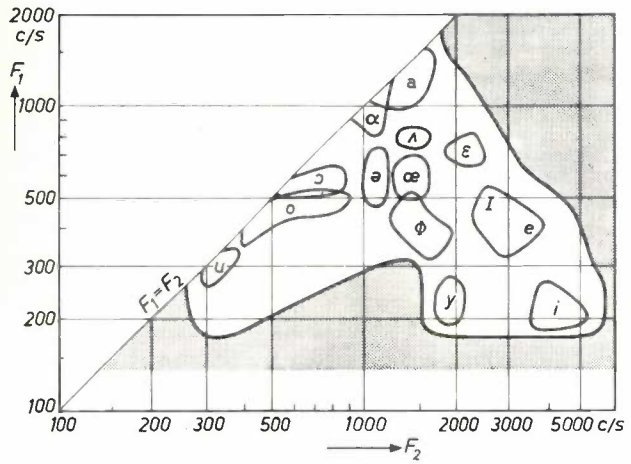


Fig. 7. In order to produce, as in the IPOVOX, vowels and vowel-like consonants with the aid of only two formants (frequency regions F_1 and F_2) it is necessary, as found by listening experiments, that in a plot of F_1 versus F_2 each of the sounds should come within the appropriate contour.

Summing up, it can be said that the results we have so far achieved irrefutably demonstrate the great perceptual significance to be attributed to the time structure of speech sound patterns. Analysis has revealed, as mentioned, a distinct segmentation of the speech continuum. It has shown that, from the viewpoint of perception, the correct amplitude variation within the segments is no less important than their spectral composition: when the amplitude envelope is well adjusted (particularly with respect to the first and last segments), the spectral composition may differ fairly considerably from the natural one without seriously impairing intelligibility. These results owe their success to the experimental methods described and to the considerable flexibility of the equipment used. Of particular value is the possibility of rapidly changing the various time intervals (variable function gating) during synthesis.

The I.P.O. has made a gramophone record of examples of the experiments discussed in the foregoing [7]). This will enable the reader to hear the various effects himself and to form some idea of the extent to which the synthetic speech approximates to human speech.

It seems possible that the results of this type of research will soon be turned to use in the field of speech tuition; this applies both to logopedics — the study and treatment of speech defects — and to the tuition of the deaf. In the long run it is not unlikely that applications will also be found in the teaching of foreign languages.

[6]) A. Cohen and J. 't Hart, Speech communication seminar, Stockholm 1962.

[7]) This is the same record that contains the demonstrations of pitch perception discussed in the preceding article. The record will be dispatched free of charge to readers who send in the coupon attached to the summary sheet contained in this number.

Summary. The instruments used in the phonetic research carried out at the I.P.O., Eindhoven, include in addition to the conventional equipment (oscilloscopes, the sound spectrograph, etc.), a specially designed "gating circuit". This consists of a magnetic recording device with auxiliary mechanisms, and it makes a word, recorded on tape, audible in successively longer sections (increase per interval about 10 ms) and finally a successively shorter part of the end of the word. This has made it possible to demonstrate, among other things, that speech sounds show a distinct segmentation, and that consonants such as *f*, *z* and *s* sound like *p*, *d* and *t* when only a 25 ms fraction is heard. The results of the analysis are checked by *synthesizing* speech with an apparatus that makes each sound segment of a word audible in the right amplitude envelope at the right moment. By means of filters the sounds are obtained from only two continuous sources — a noise generator, and a pulse generator that produces a harmonic spectrum. If the amplitude envelope is properly chosen, the spectrum of the sounds may differ fairly considerably from the natural spectrum without impairing intelligibility. The consonants *t* and *s*, for example, can be produced with one and the same noise spectrum by merely changing the amplitude envelope. The results have prospects of application in speech tuition.

WOLFGANG VON KEMPELEN'S SPEAKING MACHINE

In the 18th century scientific research in the form familiar to us — an interplay between experiment and theory — began to get under way. Associated with this was a tendency to view Man as a machine. Countless automata were built, able to perform some human function or other such as playing a musical instrument, walking, drinking, and so on.

The speaking machines described by Wolfgang Ritter von Kempelen in 1791, in a book [1]) that became famous at that time, can be included in this category. It is important to note that they were not based on mechanical and other laws that had long been known but on the results of Von Kempelen's own investigations. He was thus a real pioneer in the field of

[1]) Wolfgang von Kempelen, Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine, Vienna 1791. For a succinct description, see H. Dudley and T. H. Tarnoczy, J. Acoust. Soc. Amer. **22**, 151, 1950.

phonetic research as discussed in the preceding article in this issue, and his achievements are worth mentioning. Since 1769 Von Kempelen, who held high office at the court of Maria Theresia, had devoted his leisure to studying the mechanism of speech and to constructing the speaking machines in question. Two more or less vain attempts at such a machine were followed by the successful version shown in *fig. 1*.

Essentially, the machine consisted of a chest *A* into which air was forced with a bellows *X*, and which opened on the other side into a horn or resonator *C*. A primary sound source, the chanter reed from a bagpipe (see red dashed lines), was located in front of the neck of the resonator. Vowels were produced by suitably modifying, with the left hand, the flow of air through *C* (cf. the action of the human oral cavity). The consonant R was obtained by means of lever *r*, which lowered a wire until it just touched the vibrating reed. Separate sound sources, in the shape of pipes producing rustling sounds (*1* and *2* in fig. 1), were available for S and SH (German SCH), air from the chest being channelled to these pipes by depressing the appropriately lettered levers. Plosives P, T and K were made by shutting off *C* with the hand and quickly removing it as soon as enough pressure had built up. To obtain the required pressure without undue delay it was necessary to bypass the reed with a thin pipe (*ab* in *fig. 2*). This was still not enough to produce P, an

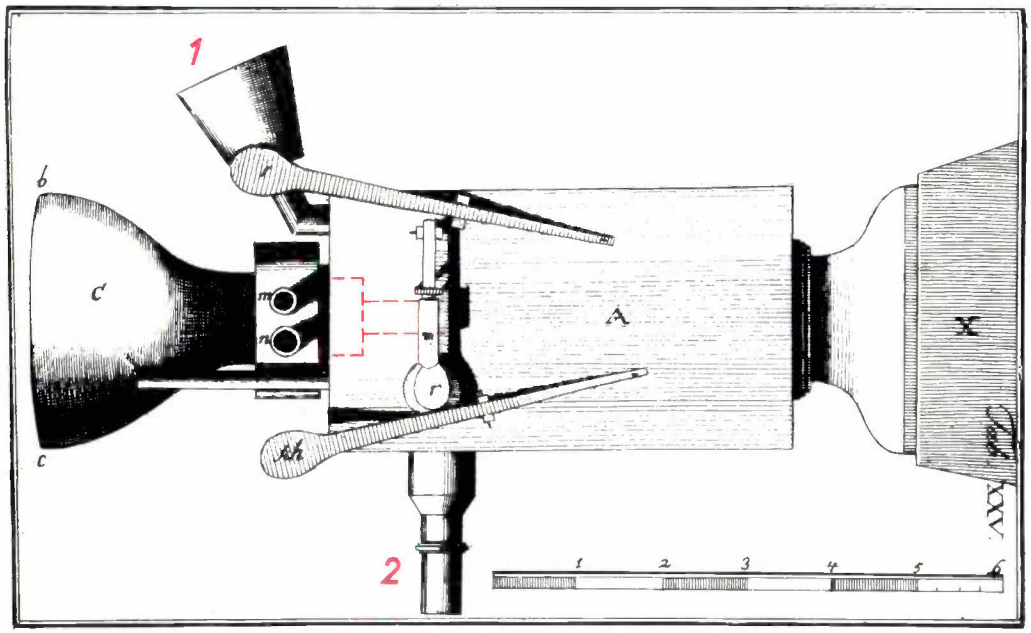


Fig. 1. Von Kempelen's speaking machine (top view), reproduced from his book [1]). The features printed in red are additions to the original drawing.

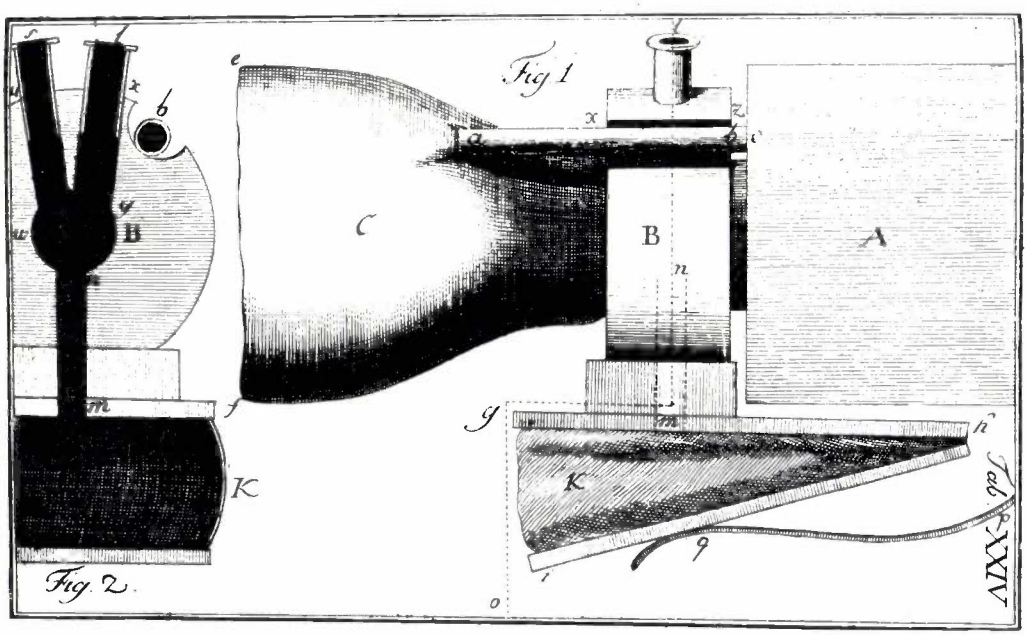


Fig. 2. The machine seen from one side, with a cross-section through the "nostrils".

auxiliary bellows *K* being necessary for that sound. The consonants N and M were likewise obtained by shutting off *C*, and at the same time opening one or both of the "nostrils" *m* and *n* (marked *s* and *l* in fig. 2). With this machine Von Kempelen was able to articulate all Latin, French and Italian words, including complicated ones such as "Constantinopolis", without any appreciable gaps between the individual sounds composing the word. The machine was unable to produce the sounds NG, CH as in "church", J as in "judge", French J as in "juge", TH as in "thin" and TH as in "then".

SOME INVESTIGATIONS OF THE VISUAL SENSORY SYSTEM

by H. BOUMA *), H. W. HOREMAN *) and J. A. J. ROUFS *).

159.931

The outside world on the one hand and the picture we form of it through our senses on the other hand may be compared with the input and output signals of an apparatus. An apparatus processes the input signal in a manner which is characteristic of that apparatus; similarly the processing of the outside world into the sensory picture of it is governed by the properties of the relevant sensory system. Thus, we learn the properties of the visual sensory system, for example, by studying the visual image in its relation to the outside world.

The light incident on the eye can not only give rise to a visual image — a process in which a conscious process is involved — but it can also cause *unconscious* bodily reactions. These include the pupillary reflex, the accommodation of the lens of the eye, certain movements of the eyeball, movements of the body to maintain balance, and so on. These *unconscious* reactions, too, can be regarded as output signals of the visual system, and can thus provide information on that system.

Studies both of unconscious reactions and of the visual image are aided to a considerable extent by knowledge obtained by other means, in particular from the study of anatomy, (electro-)physiology and psychology.

In so far as the study of the visual system is based on unconscious reactions which, as opposed to the visual image, can be observed objectively, it is possible to apply purely physical methods of measurement. In the following we shall illustrate this by describing experiments carried out in the I.P.O. on the pupillary reflex.

When studying the visual image itself with the above-mentioned aim in view, the subject must express the image in some way or another. (This is an aspect of every psychophysical test, i.e. a test designed to study mental phenomena by methods derived from physics.) As examples, two other experiments carried out in the I.P.O. will be described: the first relates to the latent time of visual observation (the perception delay), and the second to the influence of ambient lighting on the impression of brightness.

*) Institute for Perception Research, Eindhoven.

The pupillary reflex

Everyone knows from his own experience that the pupil contracts when the intensity of light increases and dilates when the intensity decreases. Less well known, perhaps, is the reaction of the pupil when one eye is closed: in that case the pupil of the open eye can be seen to grow larger, even though the intensity of light entering that eye has not changed.

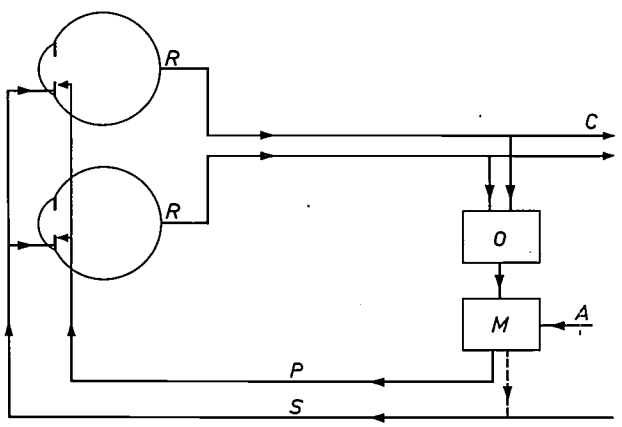


Fig. 1. Highly simplified diagram of the pupillary reflex mechanism. *R* the retinal receptors which absorb the light entering the pupil. *O* the "receptive centre" in the brain stem where the signals from both eyes are combined. *M* the motor centre, also in the brain stem, from which signals are transmitted to the pupillary muscles via parasympathetic nerve tracts *P*, and possibly also via sympathetic nerve tracts *S*.

This "control loop" can also be influenced by factors other than the incident light (e.g. the pupils dilate upon shock, contract in sleep, and contract when fixing the eyes on a nearby point, due to the "convergence reflex"). Such influences generally affect the motor centre; see arrow marked *A*. The channels which lead to conscious sensations are marked *C*.

auxiliary bellows *K* being necessary for that sound. The consonants N and M were likewise obtained by shutting off *C*, and at the same time opening one or both of the "nostrils" *m* and *n* (marked *s* and *l* in fig. 2). With this machine Von Kempelen was able to articulate all Latin, French and Italian words, including complicated ones such as "Constantinopolis", without any appreciable gaps between the individual sounds composing the word. The machine was unable to produce the sounds NG, CH as in "church", J as in "judge", French J as in "juge", TH as in "thin" and TH as in "then".

SOME INVESTIGATIONS OF THE VISUAL SENSORY SYSTEM

by H. BOUMA *), H. W. HOREMAN *) and J. A. J. ROUFS *).

159.931

The outside world on the one hand and the picture we form of it through our senses on the other hand may be compared with the input and output signals of an apparatus. An apparatus processes the input signal in a manner which is characteristic of that apparatus; similarly the processing of the outside world into the sensory picture of it is governed by the properties of the relevant sensory system. Thus, we learn the properties of the visual sensory system, for example, by studying the visual image in its relation to the outside world.

The light incident on the eye can not only give rise to a visual image — a process in which a conscious process is involved — but it can also cause *unconscious* bodily reactions. These include the pupillary reflex, the accommodation of the lens of the eye, certain movements of the eyeball, movements of the body to maintain balance, and so on. These *unconscious* reactions, too, can be regarded as output signals of the visual system, and can thus provide information on that system.

Studies both of unconscious reactions and of the visual image are aided to a considerable extent by knowledge obtained by other means, in particular from the study of anatomy, (electro-)physiology and psychology.

In so far as the study of the visual system is based on unconscious reactions which, as opposed to the visual image, can be observed objectively, it is possible to apply purely physical methods of measurement. In the following we shall illustrate this by describing experiments carried out in the I.P.O. on the pupillary reflex.

When studying the visual image itself with the above-mentioned aim in view, the subject must express the image in some way or another. (This is an aspect of every psychophysical test, i.e. a test designed to study mental phenomena by methods derived from physics.) As examples, two other experiments carried out in the I.P.O. will be described: the first relates to the latent time of visual observation (the perception delay), and the second to the influence of ambient lighting on the impression of brightness.

*) Institute for Perception Research, Eindhoven.

The pupillary reflex

Everyone knows from his own experience that the pupil contracts when the intensity of light increases and dilates when the intensity decreases. Less well known, perhaps, is the reaction of the pupil when one eye is closed: in that case the pupil of the open eye can be seen to grow larger, even though the intensity of light entering that eye has not changed.

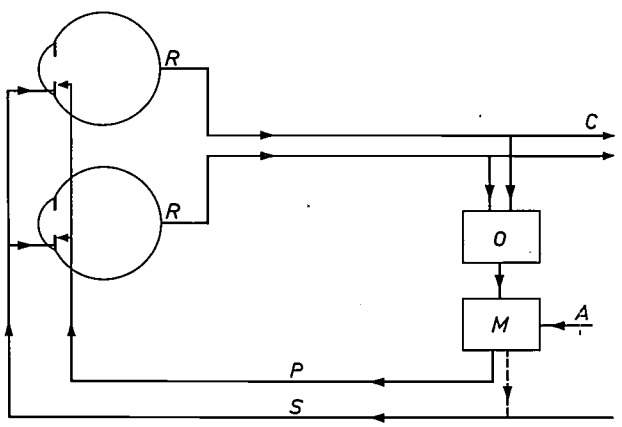


Fig. 1. Highly simplified diagram of the pupillary reflex mechanism. *R* the retinal receptors which absorb the light entering the pupil. *O* the "receptive centre" in the brain stem where the signals from both eyes are combined. *M* the motor centre, also in the brain stem, from which signals are transmitted to the pupillary muscles via parasympathetic nerve tracts *P*, and possibly also via sympathetic nerve tracts *S*.

This "control loop" can also be influenced by factors other than the incident light (e.g. the pupils dilate upon shock, contract in sleep, and contract when fixing the eyes on a nearby point, due to the "convergence reflex"). Such influences generally affect the motor centre; see arrow marked *A*. The channels which lead to conscious sensations are marked *C*.

On the basis of anatomical and (clinical) physiological investigations the mechanism of the pupillary reflex can be represented in a highly simplified diagram (*fig. 1*). The light falling on the retina is absorbed by the receptors R (the rods and cones), as a result of which a signal is transmitted to a receiving centre O in the brain stem. In this centre the signals from both eyes are somehow combined. The composite signal then passes via the motor centre M and along parasympathetic nerve tracts P (possibly also along sympathetic nerve tracts S) to the pupils. A change in the light intensity produces a change of signal which causes the pupillary muscles to react such that the initial change is partly compensated. The pupillary system may thus be regarded as a control loop. Owing to the combination of the signals from both eyes, both pupils are always equally large. The closing of one eye has the same effect as a reduction of the light intensity, so that both pupils dilate.

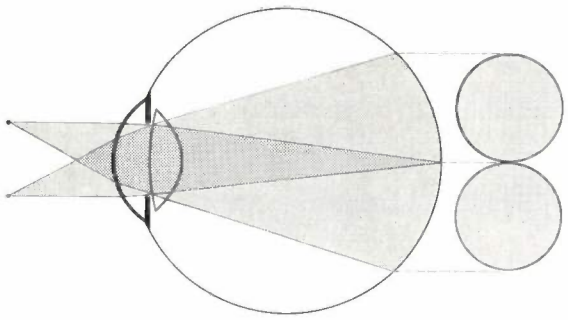


Fig. 2. Two point sources of light, immediately in front of the eye, project two beams of light on to the retina. The beams will overlap or be seen separately depending on the spacing between the light sources, the pupil diameter and the accommodation of the eye. When the eye is *not* accommodated, parallel incident rays converge on the retina. The retinal images will then just touch one another if the distance between the light sources is equal to the pupil diameter. As can be seen, the size of the retina images is immaterial.

Since the refraction of the light rays takes place mainly in the cornea and only to a slight extent in the lens, the measured diameter of the pupil is roughly 10% larger than the real diameter. The measured diameter, however, is equal to the effective diameter governing the quantity of light entering the eye.

In the I.P.O. a study has been made of the manner in which the wavelength of the light influences the pupillary reflex. For this purpose use was made of the *entoptical* method in a form developed from ideas due to Schouten [1]). We shall now briefly explain the principles of this method.

If a point source of light is placed immediately in front of a subject's eye, the light rays from that point that strike the retina cause the subject to see a circular patch of light, the edge of which is a projection of the edge of the pupil. Anyone can observe this phenomenon himself by holding close to an eye, against the light, a card in which two holes have been punctured with a needle. The subject obviously sees two projected circles and depending on the distance between the holes these circles are seen either distinct from each other or partly overlapping. If the circles just touch one another, the distance between the holes is equal to the diameter of the pupil (see *fig. 2*).

The measurement can be carried out with the aid of a pupillometer in the form of a blackened photographic negative containing a series of differently spaced pairs of dots which allow good transmission of light (*fig. 3*). One eye looks at the light whose influence on the size of the pupil is to be determined, while the negative is held in front of the other eye, and the pair of dots found at which the observed circles just touch each other. The small quantity of

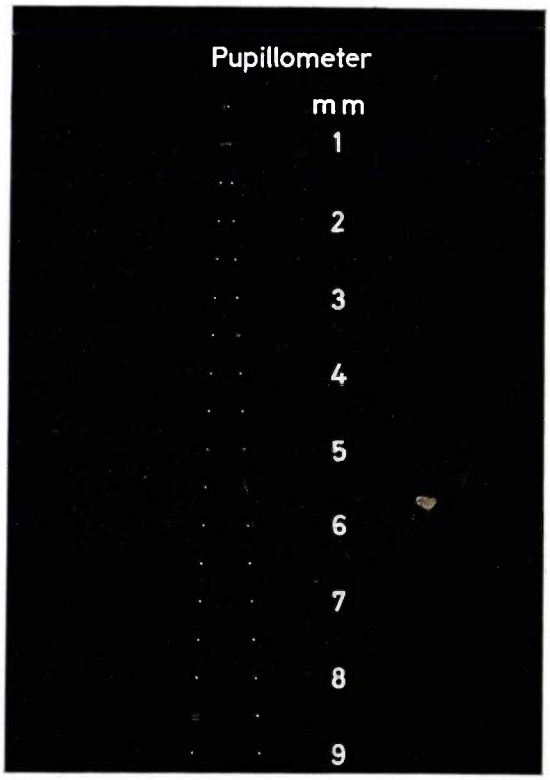


Fig. 3. Entoptical pupillometer, based on the principle sketched in fig. 2. A blackened photographic negative contains pairs of light-transmitting dots. The spacing of the dots is marked beside every other pair [2]).

The pupillometer is held against the light, with one of the pairs of dots immediately in front of one eye. This eye should be kept unaccommodated (e.g. by looking into the distance with the other eye). Two circular spots of light are then observed (see fig. 2). When the pair of dots is found at which the circles just touch each other, the pupil diameter is roughly equal to the distance between that pair of dots.

[1]) J. F. Schouten, Visuele meting van adaptatie en van de wederzijdse beïnvloeding van netvlieselementen, thesis, Utrecht 1937, p. 17 (in Dutch).

light transmitted through the dots to the eye is negligible compared with the amount incident on the other eye, and thus does not affect the measurement [2]).

Apart from measurement of the pupil diameter, entoptical projection can also be used for observing the irregularities present in every eye, such as particles on the cornea, streaks in the vitreous humour, etc. By means of two projections from different points it is even possible to localize the irregularities in depth.

If monochromatic light falls on the eye and the wavelength is varied, the above method shows that the pupil is more sensitive to one wavelength than to another. The sensitivity P of the pupil is defined as the reciprocal of the radiant flux that produces a given pupil diameter (e.g. 5 mm). This sensitivity is influenced by the wavelength because the wavelength partly governs the amount of light absorbed by the receptors. The same explanation accounts for the dependence of observed brightness on wavelength, a dependence which for photopic vision is given by the curve of relative luminous efficiency $V(\lambda)$.

The pupil diameter partly governs the illumination of the retina, so that to a certain extent the pupil controls the brightness of a given object. This is generally regarded as one of the most important functions of the pupil (the other being its influence on visual acuity). Thinking along these lines it is natural to assume that the receptors, responsible for the pupillary reflex, are the same receptors that are responsible for the observed brightness. Pupil size and brightness in that case would necessarily show the same dependence on wavelength.

Fig. 4 shows the result of an experiment in which, under identical conditions, both the relative sensitivity P of the pupil and the relative luminous efficiency V were measured as a function of wavelength [3]). It can be seen that the assumed correspondence does not exist. The pupil is most sensitive to light having a wavelength of 490 nm, whereas a wavelength of 550 nm is most effective as fas as brightness is concerned. This unforeseen disparity between the relative luminous efficiency for photopic vision and the pupil size as a function of wavelength attracted the attention of Van Liempt [4]) as early as 1937. He reported that the pupil, under identical illumination, was much wider when exposed to yellow sodium light than when exposed to bluish white mercury light, and on this finding he based a recommendation to use sodium light for portrait photography (a wider pupil being preferable for aesthetic reasons).

This wavelength dependence of the pupil sensitivity more or less corresponds to the relative luminous efficiency curve for *scotopic* vision, i.e. the curve which gives the visual sensitivity at very low light intensities. The intensities at which the pupillary reflex is measured, however, are very much greater than those at which scotopic vision operates, so that on the grounds of this correspondence alone we can draw no conclusions about a common cause.

The delay of visual perception

When a ray of light strikes the retina, we are not aware of it until a time of about 100 ms later, which the stimulus takes to reach the "perception centre". We call this time the *perception delay*. In daily life the delay in judgement which this causes is fortunately so short as to be negligible. In the observation of objects travelling at high speed, however, the situation is somewhat different: for example, we see an express train travelling towards us at 70 miles an hour about three yards farther away than it really is at that moment.

In the I.P.O. the dependence of perception delay on the intensity of the light entering the eye has been investigated [5]). The fact that dependence exists can

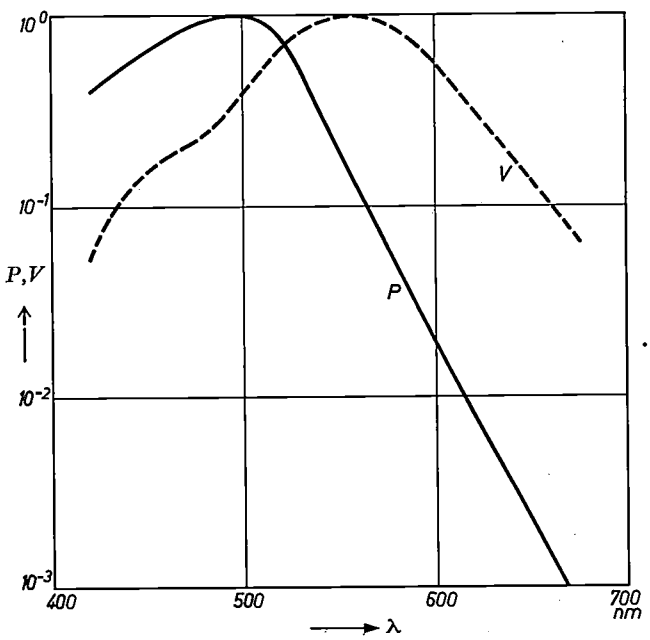


Fig. 4. The relative sensitivity P of the pupil and the relative luminous efficiency V (for photopic vision) as a function of wavelength λ. In all cases a static state was measured. The field of view for both measurements was 18°.

[2]) An example of this pupillometer will be posted free of charge to readers who send in the coupon attached to the summary sheet contained in this number.

[3]) H. Bouma, Size of the static pupil as a function of wavelength and luminosity of the light incident on the human eye, Nature **193**, 690-691, 1962 (no. 4816).

[4]) J. A. M. van Liempt, The "Philora" sodium lamp and its importance to photography, Philips tech. Rev. **2**, 24-28, 1937.

[5]) J. A. J. Roufs, Perception lag as a function of stimulus luminance, Vision Research **3**, 81-91, 1963.

easily be demonstrated with the aid of the Pulfrich effect: if we look with both eyes at a pendulum swinging in the vertical plane and hold a light-absorbing filter in front of one eye (e.g. with a transmission of 20%, equivalent to that of ordinary sun glasses), we then see that the pendulum moves out of its plane and roughly describes an elliptical cone. The direction in which this cone is traversed depends on whether the filter is held before the right or the left eye. *Fig. 5* illustrates the probable mechanism of this effect. Since the two eyes receive dissimilar

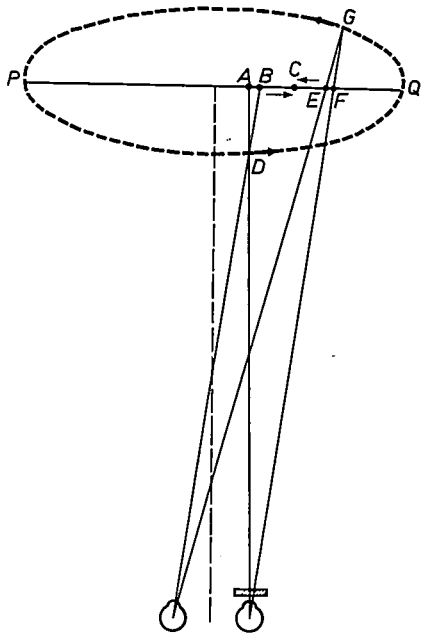


Fig. 5. The Pulfrich effect. An imaginary pendulum suspended above the drawing swings in the plane perpendicular to the paper through the horizontal line *PQ*. There is a delay in observing the pendulum — the perception delay.

In front of one eye — in the drawing the right eye — a light-absorbing filter is held. The right eye now receives less light than the left eye, and therefore the perception delay is longer for the right than for the left eye. When the pendulum is at *C* while moving to the right, the left eye sees it in the direction *B* and the right eye in the direction *A*. It therefore seems to the observer as if the pendulum is at point *D*. If the pendulum is at *C* while moving to the left, the left eye sees it in the direction *E* and the right eye in the direction *F*, and the pendulum then appears to be at G. The pendulum thus seems to describe a conc with a roughly elliptical base, as shown schematically in the figure. When the filter is held before the other eye, the apparent movement reverses direction.

quantities of light, the perception delay of each eye is also dissimilar, the weaker stimulus giving the longer perception delay. The light stimuli which are simultaneously perceived via different eyes thus correspond to different positions of the pendulum. This is stereoscopically interpreted, so that the pendulum appears to be outside its actual plane of oscillation.

Once we are aware of this Pulfrich effect, it can also be observed without the aid of a filter by partly closing one eye, thus admitting less light to it. From the apparent deflection of the pendulum due to the Pulfrich effect it is possible in theory to determine the dependence of the perception lag, defined as the difference in perception delay, on the intensity of the light. Disturbing influences, however, make this method unsuitable in practice.

The measurement might be carried out by making a light flash before a subject's eye and getting him to report when he sees the flash. The time needed for reporting must then be taken into account, however, for the reaction time of the subject contains both the perception delay and the reporting time. If we vary the intensity of the flash, the total reaction time can therefore be measured, but we do not know in how far the observed changes are attributable to the perception delay and in how far to the reporting time.

To avoid this difficulty we proceed as follows. We make an arrangement for producing two flashes of light of differing intensity in front of one eye of the subject. Apart from the light intensity we make another quantity variable, namely the moment at which the second flash is produced. The subject is now asked to adjust the moment of producing the second flash until he observes the two flashes of differing intensity *simultaneously*. The difference in the perception delay which is due to the dissimilar intensity is equal to the time interval between the two flashes.

In this way we were able to express a change in a *subjective* quantity (time of observation) as an *objectively* measurable change. We did this by asking the subject to bring two impressions into equivalence, i.e. the moments at which the flashes were perceived. Such a "null" method, as also used in physics, is widely used in psychophysical experiments [6]).

Fig. 6 shows the results of such measurements (open circles), which show that over a certain range of intensities the following relation exists between the perception delay t_1 and the light intensity I:

$$t_1 - t_0 = -T \ln I/I_0. \quad . \quad . \quad . \quad . \quad (1)$$

Here I_0 is the arbitrarily chosen intensity of the flash kept at a constant value, I is the intensity of the other flash, and t_0 a constant, being the perception delay corresponding to an intensity $I = I_0$. The factor T is a constant which is characteristic of the individual eye (4 to 8 milliseconds) and which may even differ between the left and right eye of the same person. Partly because of this fact it was concluded that the dependence of perception delay on

[6]) See the article by J. F. Schouten in this number, which gives a review of various methods used in psychophysics (pp. 35-36).

light intensity is related to processes in the retina.

We shall return for a moment to the possibility of letting a subject report his perception of the flashes. The reaction time involved can be determined as a function of light intensity by asking the subject to press a key signalling the moment he observes the flash. In fig. 6 the results obtained in this way are represented as black circles. They indicate that the reaction time t_r is related to I in the same way as defined in equation (1). We may conclude from this that the reporting time changes little if at all with the intensity of the signalled flash. This is an unforeseen result.

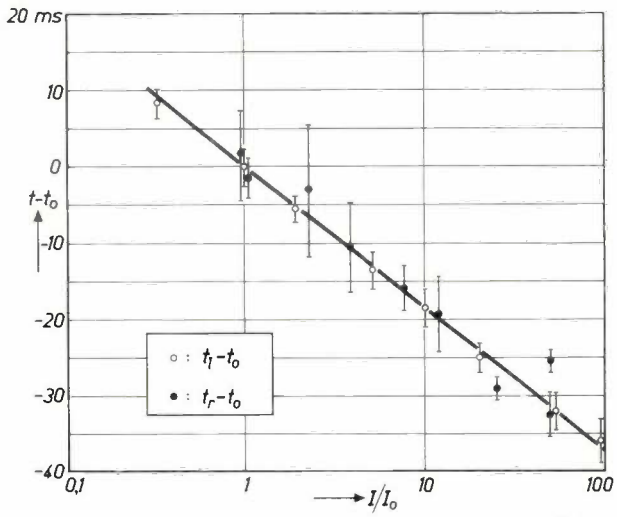


Fig. 6. In a certain range of intensities the perception delay t_1 in the observation of a flash of light decreases linearly with the logarithm of the intensity I of the flash (open circles). I_0 is the intensity of the second flash, which is kept constant, and t_0 is the perception delay for $I = I_0$.

The same relation to intensity is found for the reaction time t_r (black circles). I_0 was given the same value as in the perception delay measurements; t_0 is the reaction time pertaining to this intensity. From the equivalence of the relations found it can be deduced that the time needed for reporting a flash is not influenced by the intensity of the flash.

The figure also shows the 95% confidence interval of the results.

Finally the accuracy of the null method used for this experiment should be mentioned. The average standard deviation of the results found for a single individual adjustment is roughly 5 ms, which may be regarded as a surprisingly high accuracy especially compared with the average standard deviation of the reaction time measurements, which was about 25 ms.

Effect of ambient lighting on observed brightness

The last experiment to be described in this article is another example of the use of the null method. It forms part of an investigation undertaken in the I.P.O. into the effect of certain retinal images on other images. Effects of this kind are very common in visual observation and are often extremely complex in nature. In an extreme form the phenomenon

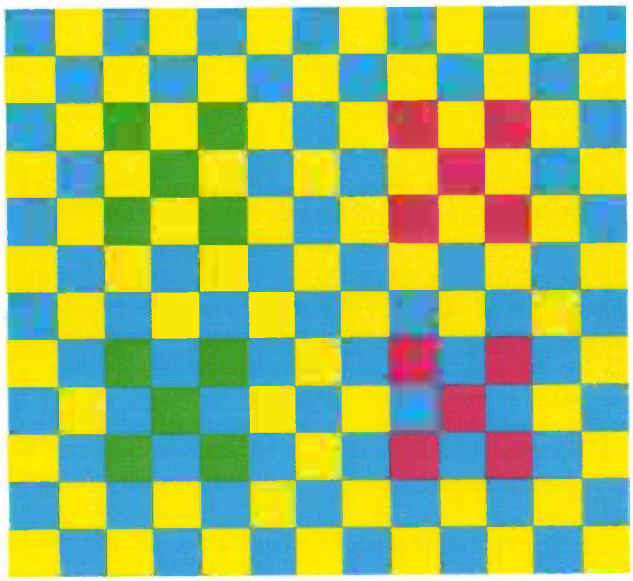


Fig. 7. The red (green) squares in the top and bottom halves of the figure are identical. The fact that they do not seem to be so is due to their different surrounds. The effect of the surround on the perception of a colour is most marked if the figure is viewed from a distance or from a small angle. (From a paper read by W. D. Wright on the Maxwell Colour Centenary Congress, London, 1961.)

is known as glare; another striking example is what one observes when looking at *fig. 7*.

The impression we receive of the upper group of red (or green) squares in the figure is not the same as that we receive from the lower group of squares of the same colour. In the lower group they seem to be darker, and moreover to differ slightly in colour. Objectively there is no difference between the two groups of squares, except that the upper squares are framed in yellow and the lower in blue. The surroundings, then, evidently have a very marked influence on colour perception and brightness. If we alter the distance from the eyes to the figure, we see how the various colour and brightness perceptions change. The effect is most pronounced if we look at the figure from a small angle (i.e. from a considerable distance or obliquely over the plane of the figure) or if we see the figure unsharply.

The experiment to be described here concerned a much simpler situation. The aim was to examine how the brightness of an object is influenced by a uniform environment.

To apply the null method, we make use of the fact that the brightness perceived by one eye is independent of that perceived by the other eye. One eye can therefore act as an "internal calibration instrument" for each measurement on the other eye.

In our experiments the subject looks by means of a special arrangement at the fields represented in *fig. 8*; with his left eye he sees the dotted calibrating field C (with a dark surround) and with his right eye

he sees the hatched test field M, as well as the ring field R, which is likewise hatched.

At given luminances L_c of the calibrating field and L_r of the ring field, the subject is now instructed to regulate the luminance L_m of the test field until the brightness of calibrating and test field are equal. In this way we can determine how the luminance of the ring influences the brightness of the test field. It is found in general that illumination of the ring results in a higher value of L_m than when the ring is not illuminated. From this observation we conclude the known fact that when the environment of a certain part of the retina is illuminated the sensitivity of that part generally decreases.

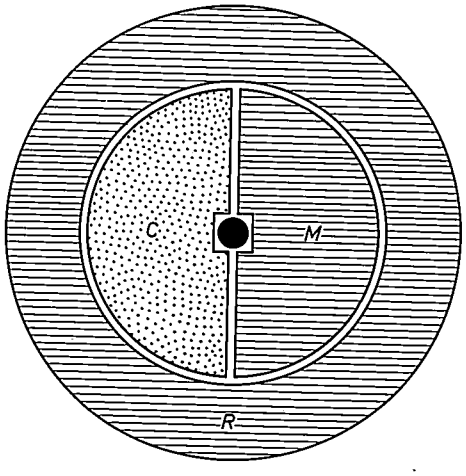


Fig. 8. The picture seen by a subject in experiments concerning the influence of the surround on the observed brightness of a given object, using the compensating method. The left eye sees only the dotted calibrating field C, and the right eye the hatched fields, i.e. the ring field R and the test field M. At a given luminance L_c of the calibrating field and L_r of the ring, the subject regulates the luminance L_m of the test field until both fields are seen equal in brightness. The central dot, seen by both eyes, is an aid to fixation.

In *fig. 9* the luminance L_m of the test field is plotted versus the luminance L_r of the ring for four luminances of the calibrating field, $L_{c1} \ldots L_{c4}$. The curves obtained are lines of constant brightness. The luminances of the calibrating field are equal to the luminances L_m at the points where the curves intersect the ordinate; here the luminance of the ring is zero, so that the brightnesses of the test and calibrating fields are equal ($L_m = L_c$). It can be seen that where the luminance of the calibrating field is high (L_{c1} and L_{c2}) there is no decrease of sensitivity as long as the luminance of the ring is lower than that of the calibrating field. It is also noticeable that the lines of constant brightness approach each other very closely as the luminance of the ring increases, implying that the stronger the surrounding illumination the smaller is the change of luminance required to produce the same change in brightness.

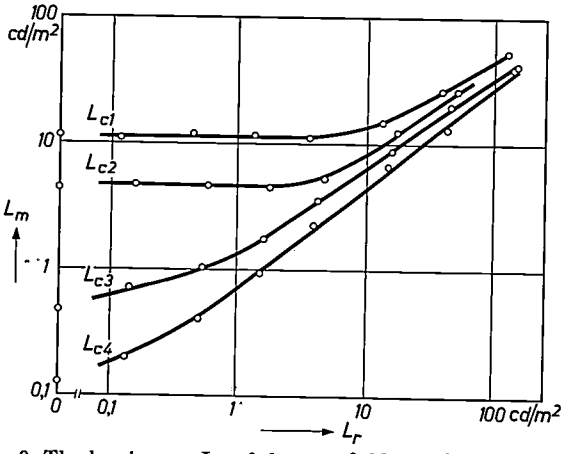


Fig. 9. The luminance L_m of the test field as a function of the luminance L_r of the ring, at four luminances L_c of the calibrating field (L_m being adjusted by the subject to produce equal brightness in left and right eyes). The upward trend of the curves indicates that the sensitivity of the part of the retina illuminated by the test field decreases with increasing illumination of the areas surrounding that part of the retina.

Similar changes of brightness have been found by other investigators [7]), although using different field configurations. The above-mentioned experiment was part of a more extensive investigation into the influence of configurational conditions on the test results [8]). It was found that the differences in configuration are of considerable importance. We shall not deal with that investigation here, however, the purpose of this article being simply to give an idea of some of the methods used for measurements concerning the visual system.

[7]) See e.g. E. G. Heinemann, Simultaneous brightness induction as a function of inducing- and test-field luminances, J. exp. Psychol. **50**, 89-96, 1955.

[8]) H. W. Horeman, Inductive brightness depression as influenced by configurational conditions, Vision Research **3**, 121-131, 1963.

Summary. The article describes three experiments on the visual system, carried out at the Institute for Perception Research. In the first the sensitivity of the pupillary reflex is determined as a function of the wavelength of the light entering the eye. The pupil diameter is measured with a simple device, an entoptical pupillometer. The results showed that, contrary to expectations, the wavelength-dependence of the pupil sensitivity does not coincide with the standard curve of relative luminous efficiency for photopic vision. In the second experiment the perception lag when a flash of light is observed is measured as a function of the intensity of the light. This is done by getting a subject to adjust the difference between the moments at which two flashes of different intensity are triggered in such a way that he sees both flashes simultaneously. The perception lag was found (within certain limits) to decrease linearly with the logarithm of the intensity. The third experiment concerned the influence of the surrounding illumination on the observed brightness of a particular object. A subject adjusted the luminance of a test field seen by only one eye, and surrounded by a ring field of a given luminance, until the brightness of the test field equalled the brightness of a calibrating field of given luminance, seen by the other eye. It was found that the stronger the surrounding illumination the greater are the changes in brightness resulting from relative changes in the luminance of an object. This experiment was part of an investigation, not discussed here, into the influence of field configuration on the test results.

IN SEARCH OF A MEASURE OF PERCEPTUAL WORK

by J. M. WESTHOFF *).

159.937:65.015.14

Right from the beginning of his existence, man has made use of tools to save himself physical work or to perform tasks which were beyond his unaided power.

This tendency is clearly visible in modern industry: an increasingly large number of operations have been taken over from man by machines. The jobs which are left are the ones which usually require very little exertion, such as sorting, checking, assembling, reading and adjusting. In all these operations, observation plays a very large role; they thus all come under the class of *perceptual work*. Outside as well as inside the field of industry the increasing importance of all kinds of perceptual work can be clearly seen. What is the work done in a laboratory but perceptual work? Another example which immediately comes to mind is modern road traffic: the road user receives a flood of information, which he must absorb and process adequately in fractions of a second.

Man is enabled to carry out mechanical work by the *metabolic cycle*, in which the chemical energy taken up from the food is transformed into mechanical energy. He is enabled to carry out perceptual work by the fact that observations, processed by the brain, lead to decisions as to what action to perform. By analogy with the above-mentioned metabolic cycle, this process may be called the *information cycle* [1]). Perceptual work, the information cycle, and the relationship between them form one of the fields of investigation covered by the Institute for Perception Research.

When studying the metabolic cycle, we are reasonably well able to calculate from a man's oxygen uptake and carbon dioxide production the amount of energy he uses and thus the amount of work he performs; these quantities can be expressed in the usual physical units, such as the calorie. It may be clearly shown that when the body does heavier work, it takes up more oxygen. This points to more intensive "combustion" of the absorbed foodstuffs, the energy source of the human organism.

The investigation of a perceptual task is also only possible when we have found a suitable measure which can be used to express the severity of the task. Now it is true that our senses are also dependent on our metabolism, but the alteration in the metabolic processes accompanying a change in the severity of a perceptual task is so slight that it cannot serve as the desired measure. The worker who wishes to investigate the action of our senses, and how the information picked up by the sense organs is processed by the human brain, must thus look for quantities other than energy or oxygen consumption in which to express his results.

Having regard to the complex structure and complicated operation of the brain and the sense organs, it is hardly conceivable that one universal measure could be found for this purpose. And in fact, various measures have been proposed of recent years, each one being reliable for a particular task or under particular circumstances, and in combination covering nearly the entire field of perceptual work. The three measures involved are:

a) the *amount of information* which must be absorbed and processed in the performance of a task,
b) the *time* needed to absorb and process the information (reaction time), or the time needed to perform the task (performance time),
c) the extent to which the performance in one task is reduced when another task is carried out simultaneously (*dual-task situation*).

In the following sections we shall describe measuring methods based on these three measures which have been developed in the I.P.O. during the past five years, and used for a number of investigations. The results of these investigations allow us to draw certain conclusions about the conditions under which a certain measure can be used, and about the comparability of the different measures.

Measuring the amount of information

The bit

If the amount of information necessary for the performance of a perceptual task is to be used as a measure of the severity of that task, we must first be able to measure the amount of information itself. One way of doing this, which was proposed by Shannon [2]), is as follows. When a message is one of a series of N possible messages (e.g. one of the

*) Institute for Perception Research, Eindhoven.

1) See the figure on page 34 of this number.

2) C. E. Shannon, A mathematical theory of communication, Bell Syst. tech. J. 27, 379-423 and 623-656, 1948.

26 letters of the alphabet), then this number N, or some suitable function of it, is a measure of the information contained in the message. In telecommunication, for which Shannon developed his theory, the function used is always the logarithm to base 2. When the series in question consists of two alternatives, the amount of information contained in the message is thus equal to $\log_2 2 = 1$; the unit defined in this way has been given the name of *bit* (a contraction of *binary digit*). Two experiments which have been carried out in the I.P.O. show that it is possible and meaningful to express in bits the amount of information absorbed and processed by man.

The first experiment [3]) was carried out in an attempt to find an answer to the question as to how much information a person can take in in a very short period of time, e.g. 0.1 s. For this purpose, use was made of the "optical square" (*fig. 1*). This consists of a square divided into 64 equal little squares, one or more of which can be illuminated simultaneously for a short time as decided by the experimenter. The illumination of a given square represents one possibility out of 64, so the amount of information which we can assign to this occurrence is given by $\log_2 64 = 6$ bits. When n little squares are illuminated simultaneously, the information content I is given by the formula:

$$I = \log_2 \frac{64!}{n!\,(64-n)!}.$$

When two squares are lit up together ($n = 2$), I is a little less than 11 bits. When the big square is divided into a smaller number of parts, the illumination of a given number of little squares corresponds to less information. This makes it possible to carry out experiments with information contents other than 6 and 11 bits.

The subject is placed in front of the vertically set square so that he can see the whole thing without moving his eyes, and is told to indicate the illuminated squares by giving their horizontal and vertical coordinates. If he mistakenly places the illuminated square one place to the right or left of its real position, or one place above or below it, we may state that 1 bit of information is lost, an error of one place along a diagonal corresponding to 2 bits, etc. If the amount of information I' taken in by the subject is plotted against the amount I presented, we get the curve of *fig. 2*. It will be seen that not more than about 9 bits of information can be taken in without error in 0.1 s. As more information is presented, a steadily increasing proportion is lost.

[3]) This investigation was carried out in the I.P.O. by W. Danziger and E. P. Köster, then attached to the Psychological Laboratory of the State University, Utrecht, Netherlands.

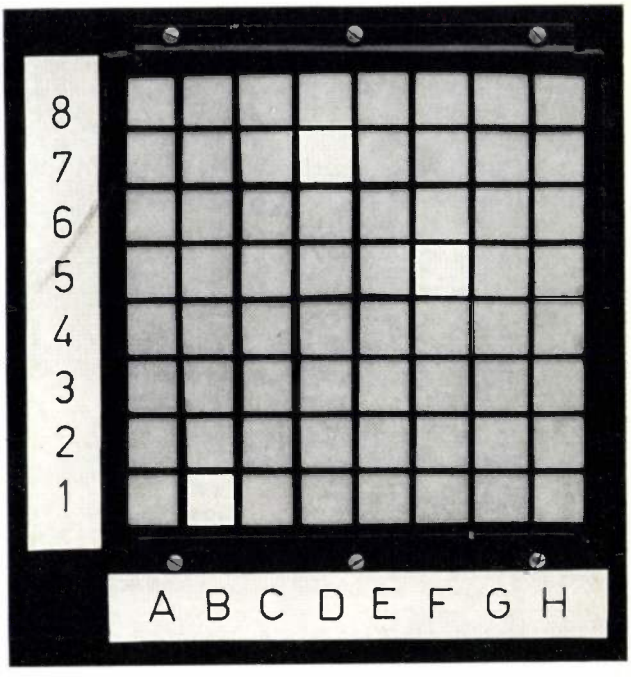


Fig. 1. The optical square. The subject must give the coordinates of the smaller squares which are illuminated for a short time (e.g. 0.1 s) by the experimenter. In this situation, the identification of one square corresponds to the intake of an amount of information equal to $\log_2 64 = 6$ bits; two and three squares correspond to 10.8 and 15.4 bits of information respectively.

The second experiment concerns the codes by which products or their components are indicated, consisting of a group of letters or numbers, or a combination of both. The I.P.O. has investigated which type of code can be remembered most easily for short periods. Leaving aside the experimental method, we shall only mention the result here: a

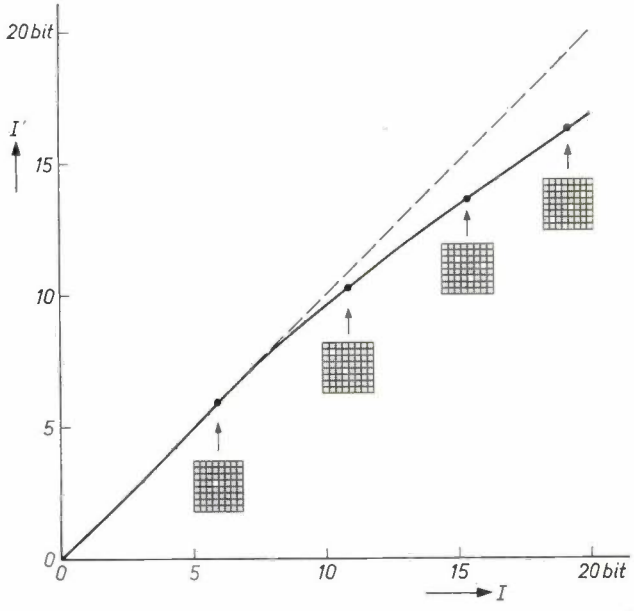


Fig. 2. The amount of information I presented in the optical square and the amount of information I' taken in by the subject are set out along the horizontal and vertical axes respectively. When more than about 9 bits are presented in 0.1 s, a steadily increasing proportion of the information is lost.

code consisting of a group of four letters is remembered as easily as one consisting of a group of five numbers. Remembering that one letter represents $\log_2 26$ bits of information, and one number $\log_2 10$, we see that both these codes contain about 17 bits. This indicates that as far as memorizability is concerned, the various codes can be compared with one another on the basis of their information content expressed in bits.

We have thus seen how use may be made of the bit in perception research. We would however like to go a step further, by stating that the severity of a perceptual task is proportional to the number of bits which must be taken in. This assumption is very reasonable in the above-mentioned experiments, so that in these cases the bit can indeed be used as a measure of the severity of the task. We shall now however consider a case which shows that things are not always so simple.

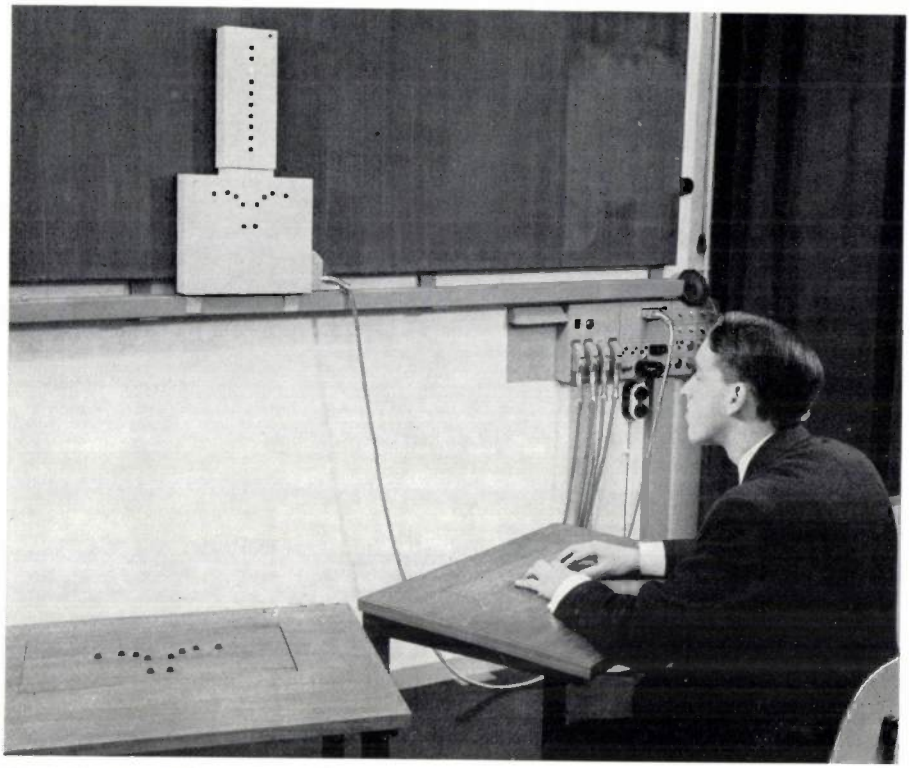


Fig. 3. Set-up which can be used to investigate the influence of the compatibility (degree of similarity between stimulus and response) on the reaction time. The subject must react as quickly as possible to the flashing of one of the lights on the stimulus panel (top left) by depressing the appropriate button of the board on which his hands rest (the reaction panel). Such a reaction panel can also be seen on the unoccupied table; the buttons are arranged so that they can very easily be operated by the fingertips. The lamps are switched on in a random sequence. If the lamp pattern is similar to that of the buttons, the average reaction time is shorter than if the two patterns are completely different, as is the case with the vertical row of lamps.

This set-up forms part of the DONDERS reaction meter described elsewhere in this number (page 71). Twenty subjects, each seated at a reaction panel, can take part in the experiment at a time. The DONDERS records the reaction time of all subjects simultaneously, as well as the button which each one depresses.

Compatibility

Many experiments have already been carried out with the aim of measuring the maximum rate at which information can continuously be taken in and processed. During the course of these investigations, it has become steadily clearer that this rate, which is round about 25 bits per second, can be influenced by all kinds of circumstances. For example, sickness or fatigue of the subject can considerably reduce the rate of intake and processing of information. Apart from these obvious influences, there are other factors which must not be neglected. We shall again demonstrate this with the aid of an experiment.

The subject sits in front of a board with ten buttons so placed that each of the ten fingertips can rest easily and relaxed on one button (*fig. 3*). He looks at a vertical panel (the stimulus panel), which contains two series of ten lamps, one with the same configuration as the buttons on the board, the other in a vertical row. He is told to press the corresponding button as soon as one of the lamps lights up, it being agreed that the top lamp in the vertical row corresponds to the little finger of the right hand. In this experiment, thus, the stimulus lamps form a pattern which in one case is similar to the reaction pattern, and in the other case is quite different. Now it may be stated that, no matter what configuration of lamps is used, the amount of information processed when reacting to the lighting up of one lamp is $\log_2 10 = 3.3$ bits. It is however found that the measured reaction time (the time which elapses between the switching on of the lamp and the pressing of the button) is on the average higher for the vertical row than for the other configuration. It thus appears that the rate at which our brain can process information is considerably slower when an unnatural or unusual relationship exists between the stimulus and the desired response. Fitts [4]), who has

[4]) P. M. Fitts and C. M. Seeger, S-R compatibility, spatial characteristics of stimulus and response codes, J. exper. Psychol. **46**, 199-210, 1953.

described this phenomenon in relationship to other experiments, introduced the term *compatibility*, by which he understands the degree of correspondence between the stimulus and the desired response.

This example shows clearly that the circumstances under which the information must be taken in and processed influence the rate at which this occurs. This is connected with the fact that the information cycle contains a number of steps, of which the uptake of information via the sense organs is only one. Others are e.g. the identification of the information, the decision as to the desired reaction (e.g. depress the left index finger) and the activation of the correct muscles [5]). In some cases, where the uptake of information plays the main role, the bit can be used as a measure of the difficulty of a perceptual task. In more complicated cases, the amount of information involved is no longer the decisive factor, and we would prefer a measure which takes the whole of the information cycle into account. The experiments described below were carried out in an attempt to achieve this aim.

Time measurements

The performance time

The possibility of using the *duration* of the operation to indicate the difficulty of a perceptual task has been made use of in the I.P.O. in the investigation of the action of placing a peg in a hole [6]). The practical importance of this investigation becomes immediately obvious when one realizes how often this action is performed. Examples are, in industry: placing a screw in a hole drilled to receive it, inserting a connecting wire into a soldering tag, and placing a nut or washer on a screw. In the home: the plugging-in of an electrical appliance, putting a key in a lock, threading a needle. No one should have any difficulty in thinking of many other examples.

It has been known for quite some time that the time needed for this operation depends on the distance through which the peg (or, as it is normally called in this field, the "pin") must be moved, the diameter of the hole and the relative tolerance of the pin in the hole. Time-study engineers even have at their disposal a table, the Work-Factor assembly table, giving empirically determined standard times. We thought however that it would be interesting to check and extend these data, obtained under working conditions, with the aid of a standardized laboratory experiment in which the various parameters, in particular the tolerance, were varied over a much wider range than that given in the Work-Factor table. In this experiment, the subject is given five metal plates, each of which has a hole in the middle, of diameter 1, 2, 4, 8 and 16 mm in successive plates (*fig. 4*). A circle, the "starting circle", of radius 25 mm is marked round each hole. Each plate is provided with a set of pins, the smallest of which gives a tolerance of 50% between pin and hole, and the others being chosen so that the tolerance is successively halved. The subject is told to move the pin as quickly as possible from the starting circle into the hole. The pin and the plate form part of an

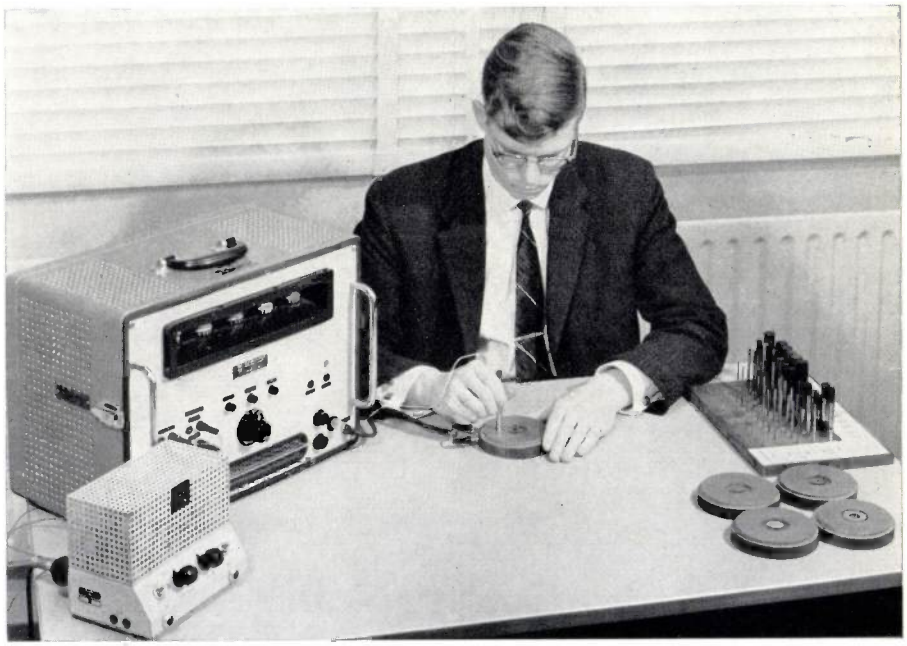


Fig. 4. Set-up for measuring the performance time for pin mounting (sticking a peg in a hole). The time taken for the operation is measured by the electronic counter on the left of the subject. This counter is set into action as soon as the pin leaves the starting circle (thus breaking an electric circuit), and counts the number of hundredths of a second which elapse before the pin reaches the bottom of the hole (when contact is restored). In order to obtain a sufficiently reliable average, the assembly operation was carried out 100 times for each pin. On the right-hand end of the table can be seen a number of plates with holes of different diameters, and pins which fit into the various holes with different tolerances.

[5]) G. ten Doesschate, Notes on the history of reaction-time measurements, Philips tech. Rev. **25**, 75-80, 1963/64.

[6]) J. F. Schouten, J. Vredenbregt and J. J. Andriessen, On the laws governing the times needed for pin-mounting as a function of diameter and tolerance, Ergonomics **3**, 275, 1960.

electric circuit, so that an electronic clock can be started as soon as the pin leaves the starting circle, and stops as soon as the bottom of the hole is reached. The time which can be read off after this operation will be called the performance time or, in this special case, the assembly time. It will be clear that, especially at low tolerances, the time taken by the same subject to place the same pin in the same hole will not always be the same. The operation was therefore carried out a hundred times per subject per pin, in order to get a reliable average. The results of this experiment are shown in *fig. 5*. Each line gives the relationship between the assembly time t_m

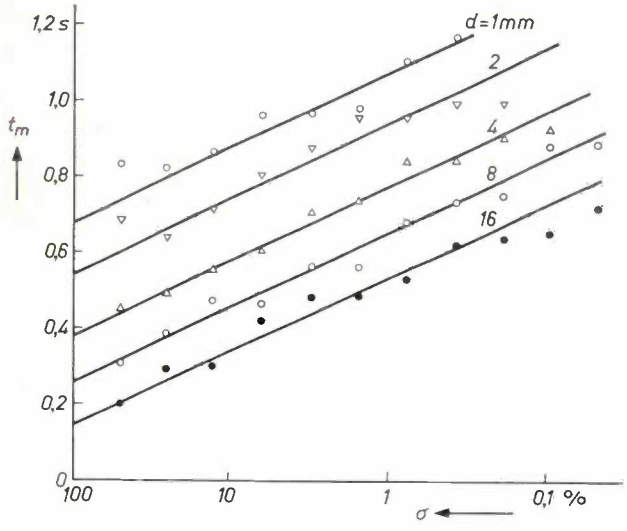


Fig. 5. Relationship between the assembly time t_m and the relative tolerance σ of the pins for various values of the diameter d of the hole. It may be seen that a logarithmic relationship exists both between t_m and σ and between t_m and d.

and the relative tolerance σ between pin and hole for a hole of a given diameter. Two conclusions may be drawn from this graph: a) There is a logarithmic relationship between the assembly time and the relative tolerance, since the lines are straight when a logarithmic scale is used on the horizontal axis. b) There is also a logarithmic relationship between the assembly time and the diameter of the hole, since the lines are equidistant while the diameters form a geometrical progression.

These experimental findings can be explained by regarding the assembly operation as a controlled process, which we may describe as follows. The hand moves the pin towards the hole in the plate. The human sense organs (the eye, the sense of touch, the sense of muscle movement) provide information about the error made during this movement (by "error" we understand here the distance from the axis of the pin to the axis of the hole). This information is processed by our brain, which then gives a correction for the direction and speed of movement. If we assume that this correction always halves the error made, then for a given hole the assembly time should increase proportionally with the negative logarithm of the relative tolerance (*fig. 6*). The same holds for different diameters, if the relative tolerance is kept constant. Supplementary experiments have shown that in this process the successive halving of the error does not begin until the pin reaches a certain, constant distance x_0 from the hole. Accordingly the moment $t = t_0$ in fig. 6 refers to the start of the second phase of the assembly. Further, it will be obvious that in fact the pin does not move along the idealized curve of fig. 6, but along a curve which oscillates about this. This is clearly seen from a stroboscopic photo of the assembly and the corresponding distance-time curve (*fig. 7* and *fig. 8*).

The results of an experiment in which the subject was told to carry out the assembly at one of three different speeds, which he had learnt before the start of the experiment, instead of as fast as possible (as in the first experiment) are also interesting. These

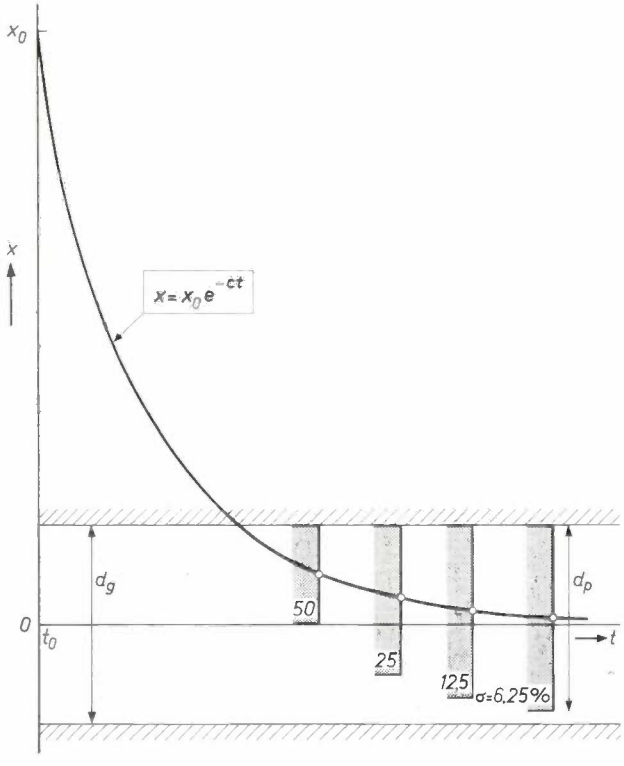


Fig. 6. In explanation of the logarithmic relationship between the assembly time t_m and the tolerance σ. If the error x (i.e. the distance between the axis of the hole and that of the pin) always decreases by the same factor in the same length of time, the pin (diameter d_p) moves to the hole (diameter d_g) along a curve like the one shown here. The graph also shows four pins with tolerances of 50, 25, 12.5 and 6.25% of the diameter of the hole, drawn at the moment when they can enter the hole. Halving the tolerance can be seen to cause a constant increase in the assembly time t_m.

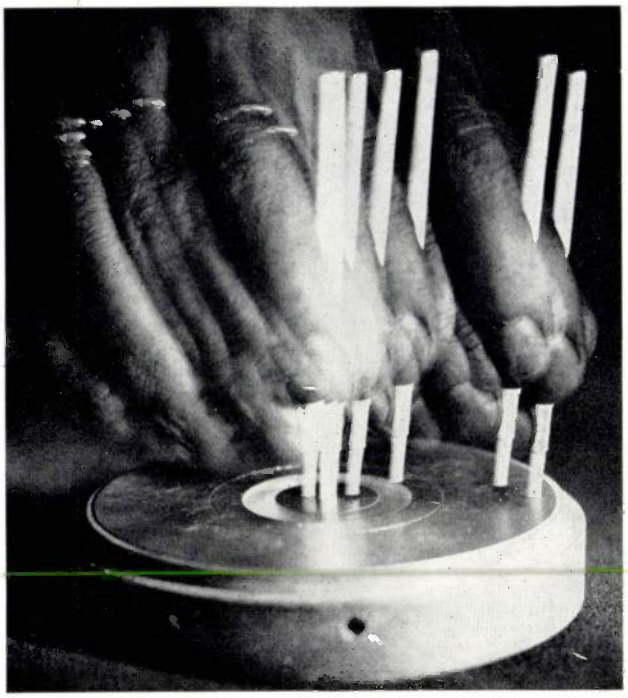


Fig. 7. Stroboscopic photo during the mounting of a pin in a hole of diameter 4 mm and a relative tolerance of 0.1%. The light was switched on at intervals of 0.1 s. It may be seen from the positions of the wedding ring that it took 0.3 s to move the pin from the top of the hole to the bottom.

results are shown in *fig. 9*, the three straight lines representing the results at low, medium and maximum speed with a constant value of the diameter of the hole. The intercepts made by these lines on the vertical axis correspond to $\sigma = 100\%$, i.e. to a pin of zero diameter. These intercepts may be taken as the times needed to move the pin from the starting circle to the hole, the slope of the lines corresponding to the increase in assembly time resulting from the greater accuracy which must be used as the tolerance is reduced. The fact that these lines are parallel means that the time needed to increase the accuracy is independent of the speed at which the operation

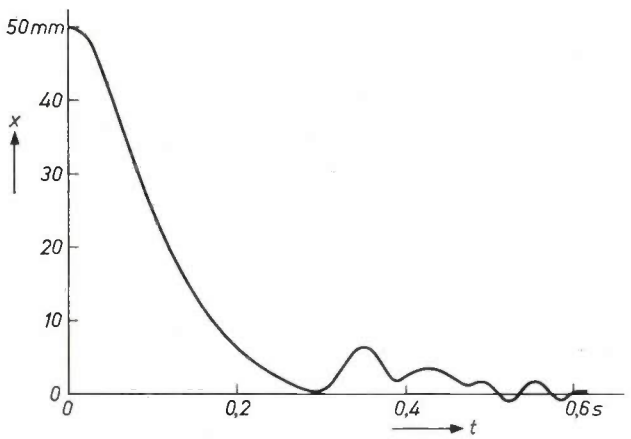


Fig. 8. The distance-time diagram for the assembly operation of fig. 7.

is carried out. In other words, the speed at which the corrections to the movement are applied is beyond the control of the human will.

The reaction time

In the previous section we have described the placing of a pin in a hole as a controlled process. This view can in fact be extended to cover all human actions: there is a mechanical system (the musculature), a measuring device (the senses) and a computer

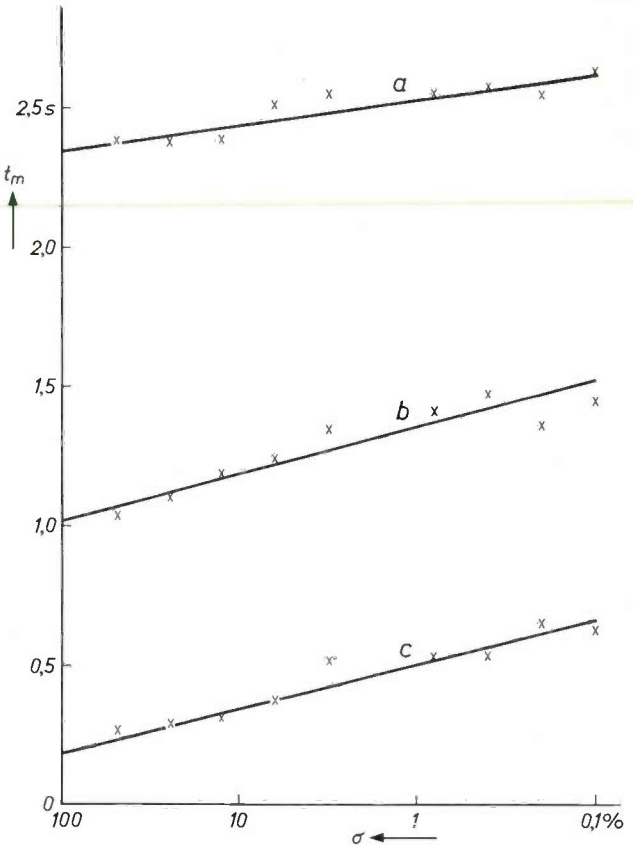


Fig. 9. Relationship between the assembly time t_m and the tolerance σ when the pin mounting is carried out *a*) slowly, *b*) at a medium rate, *c*) as fast as possible. The intercepts made by the various lines on the vertical axis represent the times needed to bring the pin from the starting circle to the hole. It follows from the fact that the lines are parallel that the actual process of "seeking" the hole takes place by a mechanism whose speed cannot be voluntarily influenced by the subject.

(the brain) which commands the mechanical system on the basis of the information delivered by the measuring device. In a case like that of pin mounting, we use mainly the mechanical system and the measuring device: the performance time is the time that these two together need to carry out the order. In other cases, the brain must first process a certain amount of information and draw appropriate conclusions, so that some time will elapse between the reception of the sense impressions and the reaction to them. This time is called the reaction time. The

relationship between this reaction time and the amount of information to be processed, in a simple case, may be seen from the following experiment, in which one hundred subjects were told to sort a pack of cards in various ways, e.g.:

a) Deal the cards one by one on the same pile.
b) Sort into black and red.
c) Sort into clubs, diamonds, hearts and spades.
d) Sort according to the thirteen different values of the cards.
e) Deal the cards out on a board, in the same arrangement as cards from another pack have been layed out on the board.

If the average time needed to deal one card is plotted against the number of piles to be made, it is found that this time increases linearly with the $\log_2$ of the number of piles (*fig. 10*). In other words, there is a linear relationship between the logarithm of the number of possibilities between which one must choose (= the number of bits) and the time needed for this choice. In this case, it is thus possible to correlate the time measurements with the measure of the difficulty of a perceptual task given in the previous chapter, i.e. the amount of information to be processed.

If we now take a closer look at the experiments described in this and the previous section, we see that the measurement of the time gives us interesting data on the load caused by the performance of a perceptual task. Moreover, we can also use the duration of the operation as a measure of the severity of the task — at least as long as we restrict ourselves to tasks with the same "work content" (e.g. pin mounting with holes of varying diameters). In fact, considerable use is already being made of this possibility in practice (the above-mentioned Work-Factor tables). But it is not possible to compare the severity of *different* tasks (e.g. an assembly and a sorting operation) by considering how long they take: it is for this reason that we talk of a performance time in the one case and a reaction time in the other. The experiments which we will now describe make use of a measure which appears to make this comparison possible.

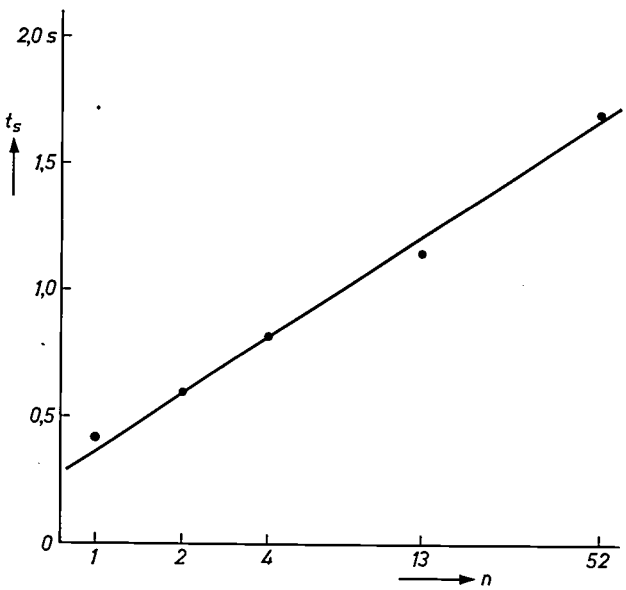


Fig. 10. Average dealing times t_s as a function of the number n of piles into which a pack of cards must be dealt. A linear relationship is found between the mean dealing time and the logarithm of the number of piles.

Dual-task situations

Every car driver is able to carry on a conversation while at the wheel of his car. He may then be said to be carrying out two perceptual tasks simultaneously. The fact that the conversation ceases whenever the state of the traffic demands the concentrated attention of the driver in order for him to carry out a rapid sequence of quick reactions, shows that man's perceptual capacity is limited and that the performance of one task influences the ability to perform another, even though quite different senses and muscles are involved. Starting from this thought, a *capacity-meter* [7]), a measuring instrument which can be used to compare different perceptual tasks as regards their difficulty, has been developed in our institute. The subject is here required to perform two tasks at the same time. One of these is the task whose difficulty is to be measured (the "trial" task), the other is always the same (the standard task). He starts off by doing the trial task on its own, and then he must do the standard task as well, in a gradually increasing tempo. The way in which the execution of the trial task is influenced by this can be used to give a measure of the difficulty of the latter.

The standard task consists in depressing a pedal under the left foot whenever a low tone (250 c/s) is heard through a pair of headphones, and one under the right foot whenever a high tone (2000 c/s) is heard (*fig. 11*). These tones are produced in a random order, at rate which can be adjusted between 10 and 120 times per minute. Before the start of the experiment, the subject practices the standard task, and the maximum rate at which he can perform it without error is measured. During the experiment, he is not allowed to make more than 1 mistake /min in the standard task. This particular task was chosen as standard because it

[7]) J. F. Schouten, J. W. H. Kalsbeek and F. F. Leopold, On the evaluation of perceptual and mental load, Ergonomics 5, 251-260, 1962.

Fig. 11. The capacity meter developed in the Institute for Perception Research. In order to measure the difficulty of a perceptual task (here the assembly of a washer and a nut on a bolt), we investigate how the performance in this task deteriorates as the standard task is simultaneously carried out at a gradually increasing tempo. The standard task consists in depressing a pedal under the right foot when a high tone is heard in the headphones, and one under the left foot when a low tone is heard, the tones being produced in accordance with a random programme (on punched tape), at an adjustable rate. On the right, next to the punched-tape reader, may be seen the counter which records the number of correct reactions, and the number of errors sorted into two groups: reactions which come too late, and wrong reactions.

leaves the hands and the eyes free, and thus allows a wide variety of trial tasks to be investigated.

Fig. 12 shows how the performance in two trial tasks was reduced when the execution of the standard task had to be increased. For both tasks, execution of the task in question alone at the maximum rate possible without error was taken as 100%. It will be seen that the assembly of a washer and nut on a bolt is much less influenced by the performance of the standard task than is simple arithmetic. By means of different experiments other investigators [8] [9] [10] came to similar conclusions.

If we assume that there is a relationship between the performance and the perceptual load in a given task, then we can take the slope of the lines in fig. 12 as a measure of the severity of the task in question: the more attention is needed for the execution of the task, the more will the performance fall off as more attention is required for carrying out the standard task.

It thus seems that this method offers good possibilities for comparing the perceptual load caused by different tasks. It could also be used on tasks which demand the use of both hands and feet (e.g.

[8]) E. Bornemann, Untersuchungen über den Grad der geistigen Beanspruchung, Arbeitsphysiologie **12**, 142-192. 1942.
E. C. Poulton, Measuring the order of difficulty of visual-motor tasks, Ergonomics **1**, 234-239, 1958.

[10]) I. O. Brown and E. C. Poulton, Measuring the spare "mental capacity" of cardrivers by a subsidiary task, Ergonomics **4**, 35-40, 1961.

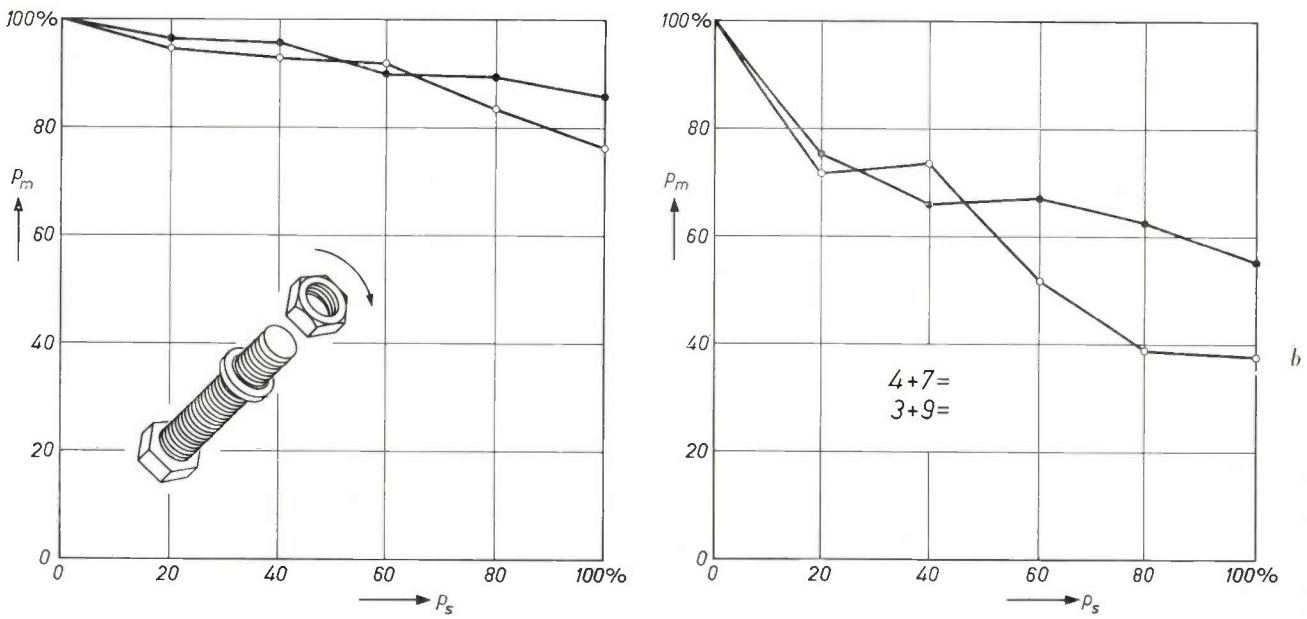


Fig. 12. Performance p_m in two perceptual tasks, *a*) assembly and *b*) simple arithmetic, as a function of the performance p_s in the standard task, in each case for two subjects. The performance is measured as the number of operations which the subject can carry out per unit time under the given conditions, the maximum performance which the test person can produce when doing only one task at a time being taken as 100%. For each experimental point, the subject is given the opportunity to find out the tempo at which he can perform the trial task, while simultaneously carrying out the standard task at the prescribed rate without making more than 1 error per minute.

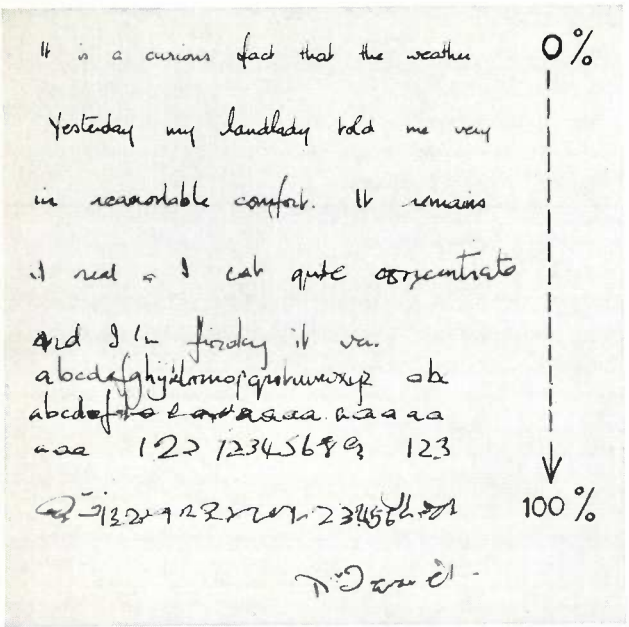


Fig. 13. Example of spontaneous writing (taken as the "trial" task) as the rate of execution of the standard task is gradually increased. Each line in the figure is taken from a whole sheet written by the subject while carrying out the standard task at a certain rate. It may be clearly seen that as the tempo of the standard task is raised the script becomes more childish, the content more concrete and banal, an increasing number of errors are made, etc.

driving a car) if another standard task were thought up. It would then be necessary to "calibrate" the different standard tasks with respect to one another.

Apart from this application in evaluating the difficulty of different perceptive tasks, the capacity-meter can also provide useful service for investigations of the way in which the mind is affected by the performance of a perceptual task [5]). For example, the subject may be told to write about whatever comes into his head. If he is then made to increase his performance in the standard task, such phenomena as increasing childishness of the handwriting, more concrete and banal content of the text, repetitions, spelling errors and inversions are observed (*fig. 13*). These phenomena are also observed in cases of mental fatigue and pathological character changes. Another possibility is to give the subject an intelligence test as trial task. The marks which he obtains are found to get worse as a higher execution of the standard task is required of him: in other words, a person who is performing a perceptual task appears to have a lower intelligence. The results of closer study of this phenomenon may well in the long run help the ergonomist to predict when a worker in a dual-task situation is likely to make errors. This can then be taken into account when distributing work or deciding on the conditions under which the work must be carried out.

Summary. The normal physical units (such as the calorie) used to express human performance in mechanical work cannot be used as measures of perceptual work. Three different attempts to provide a measure of the difficulty of a perceptual task are described. For certain kinds of tasks, the amount of information involved can be used as a measure. In other cases, the performance time (e.g. for placing a pin in a hole) or the reaction time (between stimulus and response) are suitable measures of the severity of the task. A capacity meter has been developed for comparing the severity of different tasks which do not allow the use of the above-mentioned measures. Here, the subject must carry out the "trial" task at the same time as a standard task; he is then required to increase his performance in the standard task gradually, which leads to a decrease in his performance in the trial task. The slope of the curve representing this decrease can be used as a measure of the severity of the task in question.

THE "DONDERS", AN ELECTRONIC SYSTEM FOR MEASURING HUMAN REACTIONS

by J. F. SCHOUTEN *) and J. DOMBURG **).

159.938.343

The speed of human reaction

When a subject is given a certain stimulus from a set of available stimuli and is instructed to react to this stimulus by pressing one of a set of buttons, he can only do so after a certain time: the *reaction time*. This time is shortest — somewhat less than 0.2 second — if only one stimulus is possible, e.g. a flash of light, and if there is only one button to press. The reaction time depends on the number of possible stimuli and reactions and increases by about 0.1 second per factor of 2 in this number.

It would be unfortunate for us if our perceptions and actions were to be separated by such long reaction times. The drummer would be at least 0.2 s behind the conductor's baton, or the dancers out of time with the music. In aiming at an object moving past us at a speed of only one metre per second (3.6 kilometres an hour), we should be at least 20 cm off target!

This inherent inertia of our reaction system is compensated to a large extent, however, by our abi-

*) Institute for Perception Research, Eindhoven.
**) Philips' Research Laboratories, Eindhoven.

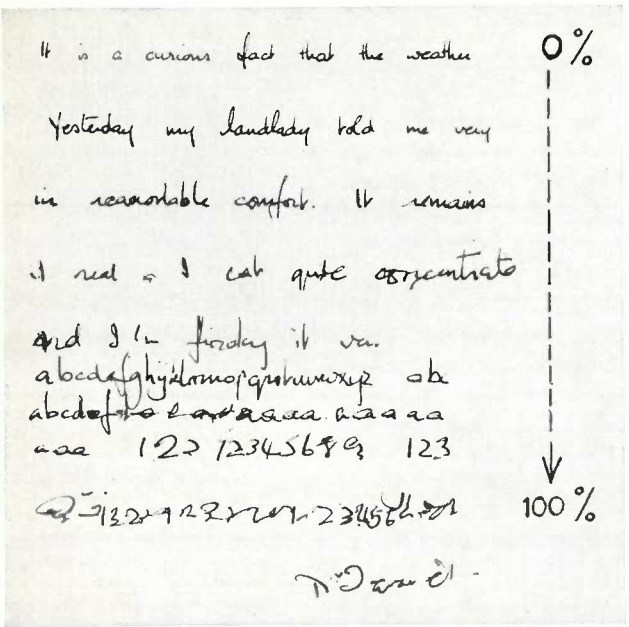


Fig. 13. Example of spontaneous writing (taken as the "trial" task) as the rate of execution of the standard task is gradually increased. Each line in the figure is taken from a whole sheet written by the subject while carrying out the standard task at a certain rate. It may be clearly seen that as the tempo of the standard task is raised the script becomes more childish, the content more concrete and banal, an increasing number of errors are made, etc.

driving a car) if another standard task were thought up. It would then be necessary to "calibrate" the different standard tasks with respect to one another.

Apart from this application in evaluating the difficulty of different perceptive tasks, the capacity-meter can also provide useful service for investigations of the way in which the mind is affected by the performance of a perceptual task [5]). For example, the subject may be told to write about whatever comes into his head. If he is then made to increase his performance in the standard task, such phenomena as increasing childishness of the handwriting, more concrete and banal content of the text, repetitions, spelling errors and inversions are observed (*fig. 13*). These phenomena are also observed in cases of mental fatigue and pathological character changes. Another possibility is to give the subject an intelligence test as trial task. The marks which he obtains are found to get worse as a higher execution of the standard task is required of him: in other words, a person who is performing a perceptual task appears to have a lower intelligence. The results of closer study of this phenomenon may well in the long run help the ergonomist to predict when a worker in a dual-task situation is likely to make errors. This can then be taken into account when distributing work or deciding on the conditions under which the work must be carried out.

Summary. The normal physical units (such as the calorie) used to express human performance in mechanical work cannot be used as measures of perceptual work. Three different attempts to provide a measure of the difficulty of a perceptual task are described. For certain kinds of tasks, the amount of information involved can be used as a measure. In other cases, the performance time (e.g. for placing a pin in a hole) or the reaction time (between stimulus and response) are suitable measures of the severity of the task. A capacity meter has been developed for comparing the severity of different tasks which do not allow the use of the above-mentioned measures. Here, the subject must carry out the "trial" task at the same time as a standard task; he is then required to increase his performance in the standard task gradually, which leads to a decrease in his performance in the trial task. The slope of the curve representing this decrease can be used as a measure of the severity of the task in question.

THE "DONDERS", AN ELECTRONIC SYSTEM FOR MEASURING HUMAN REACTIONS

by J. F. SCHOUTEN *) and J. DOMBURG **).

159.938.343

The speed of human reaction

When a subject is given a certain stimulus from a set of available stimuli and is instructed to react to this stimulus by pressing one of a set of buttons, he can only do so after a certain time: the *reaction time*. This time is shortest — somewhat less than 0.2 second — if only one stimulus is possible, e.g. a flash of light, and if there is only one button to press. The reaction time depends on the number of possible stimuli and reactions and increases by about 0.1 second per factor of 2 in this number.

It would be unfortunate for us if our perceptions and actions were to be separated by such long reaction times. The drummer would be at least 0.2 s behind the conductor's baton, or the dancers out of time with the music. In aiming at an object moving past us at a speed of only one metre per second (3.6 kilometres an hour), we should be at least 20 cm off target!

This inherent inertia of our reaction system is compensated to a large extent, however, by our abi-

*) Institute for Perception Research, Eindhoven.
**) Philips' Research Laboratories, Eindhoven.

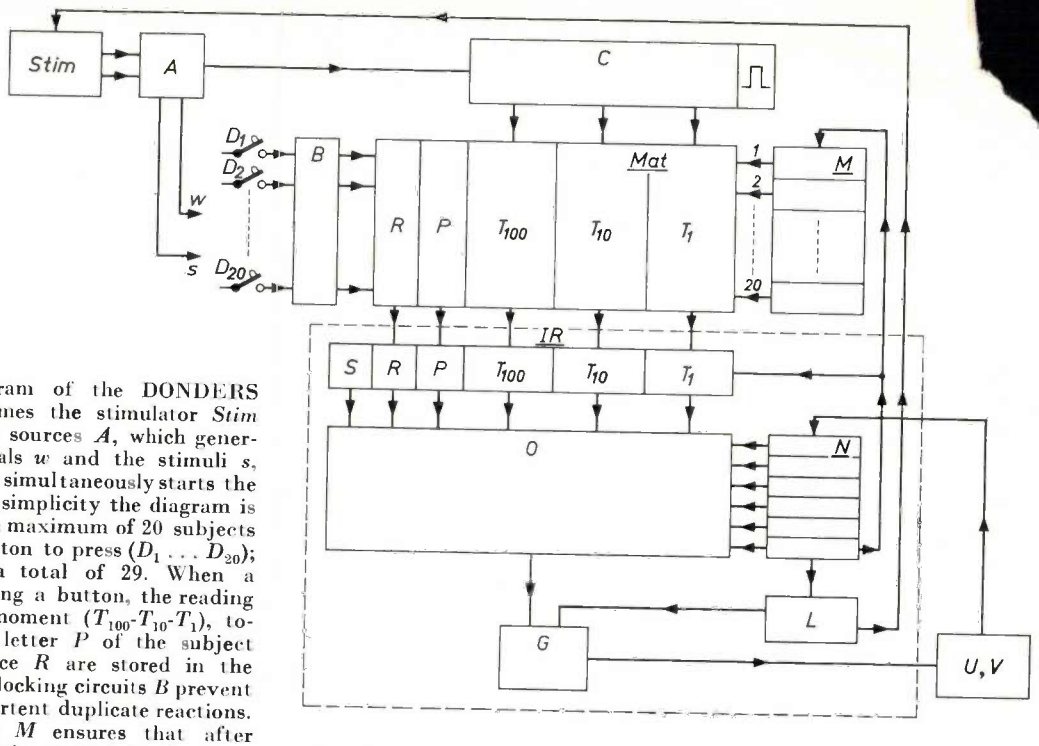


Fig. 1. Block diagram of the DONDERS (simplified). At set times the stimulator *Stim* activates the physical sources *A*, which generate the warning signals *w* and the stimuli *s*, and with each stimulus simultaneously starts the electronic clock *C*. For simplicity the diagram is drawn as if each of the maximum of 20 subjects had only one push-button to press ($D_1 \ldots D_{20}$); in reality each has a total of 29. When a subject reacts by pressing a button, the reading on the clock at that moment (T_{100}-T_{10}-T_1), together with the serial letter *P* of the subject and the reaction choice *R* are stored in the matrix memory *Mat*. Blocking circuits *B* prevent the recording of inadvertent duplicate reactions. The read-out counter *M* ensures that after each turn the information stored in the memory, with the addition of the code letter *S* (denoting the stimulus), is transferred in the correct sequence to the intermediate register *IR*. From there the information passes via the gate *G* to the print-out device (punch *U* or electric typewriter *V*). At the same time the write-out counter *N* determines which of the six quantities of one score is to be dealt with. A sequence selector *O*, consisting of 36 switches, allows the six quantities of any score to be preset in any required permutation. The punching and/or typing of each score is reported back to the write-out counter, which then moves up one step further. When a complete score has been passed, the write-out counter clears the intermediate register and signals to the read-out counter that the next score can be passed to the register. Block *L* transmits to the punch or typewriter a signal to turn to a new line and reports when the write-out is completed.

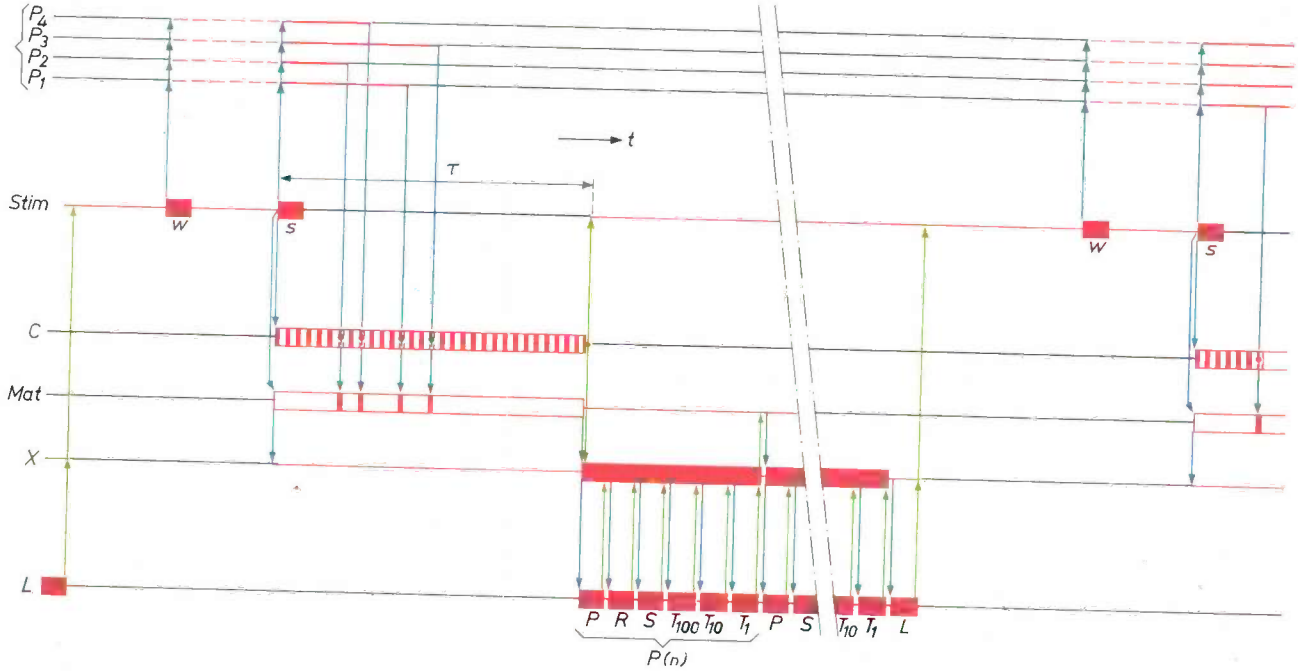


Fig. 2. Time-sequence diagram of the DONDERS. For the sake of clarity the number of subjects is limited here to four ($P_1 \ldots P_4$). Black lines denote the *resting state*, broken red lines the *alert phase*, solid red lines the *active phase*, blue arrows the *transmission of information*, and green arrows *interlock*. The time *t* runs from left to right.

A turn begins when the stimulator *Stim* (already active) gives the warning signal *w*. The subjects are thereby brought into the alert phase. A moment later the stimulus *s* initiates the active phase for the subjects, for the electronic clock *C*, for the matrix memory *Mat*, and for the write-out system *X* (includes all the blocks within the dotted square in fig. 1). The subjects each respond after an individual reaction time, and the readings on the clock at the moments of these individual reactions are stored in the memory.

As soon as the clock has reached the end of the (adjustable) thinking time τ, it signals the fact to the stimulator and the write-out system. The punching and/or typing of the scores now begins. As each quantity (*P*, *R*, etc) is punched or typed out a signal is sent back to the write-out system. The completion of a score is reported back by the write-out system to the matrix memory. At the end of a turn (at a new line *L*) a signal is sent to the stimulator, which, after an adjustable waiting time, starts the next turn.

lity to *anticipate* observations and actions. The drummer prepares his action such that the drum beat coincides with the conductor's instruction, and the dancers adapt the rhythm of their movements so that their steps are in time with the music. And just as the hunter does not aim at the moving game itself but at the point where he expects it to be when the bullet reaches the animals estimated path, so when we aim or grasp, strike or kick we do so at the place where we expect the moving object to be when our hand or foot reaches the line of its movement.

Because of this we are able, in spite of our slow reaction, to achieve a precision in time of about 0.03 seconds. In the case of the moving object mentioned this means a residual inaccuracy of only 3 cm.

Numerous factors contribute to the slowness of our reaction: the establishment of the sensory perception, the recognition of a stimulus, the decision as to the action to be performed, and the execution of the action. The Dutch physiologist Donders stated in 1868 that we might perhaps never know what human thought is, but that at least we can measure its *duration* [1]). Since then, the measurement of reaction times has remained the subject of considerable interest. Its significance has increased since communication theory formulated the concept of information quantitatively and interpreted the processing of information as a process of progressive choices.

The great difficulty in measuring reaction times is that it takes the scientific investigator a great deal of time. To obtain sufficient data it is necessary to present the subject with an extensive programme of stimuli. The reaction times must then be measured with a certain precision, recorded and finally processed. Planning and setting up a complete experiment and analysing the consolidated results may obviously be regarded as work demanding intelligence. On the other hand the execution of an experimental programme and the procedures of measuring, recording and processing are — as far as the purely repetitive part is concerned — simply dull routine.

For this reason the Institute for Perception Research drew up a plan for an electronic system that would carry out these routine operations automatically. A system of this kind was subsequently developed in the Philips Research Laboratories. The system was given the name DONDERS, in honor of the distinguished Dutch physiologist just mentioned.

[1]) G. ten Doesschate, Notes on the history of reaction-time measurements, Philips tech. Rev. **25**, 75-80, 1963/64.

Principles underlying the design of the DONDERS

Operational requirements

It is required that reaction times from 0.01 to 4.00 seconds should be measurable in steps of 0.01 second.

For the various stimuli to be administered, and the associated reactions, a number of about 30 each was envisaged. In connection with the available code the final choice was 29.

The system should be designed to allow simultaneous measurements on a reasonably large number of subjects. The maximum number was set at 20.

For each reaction the subjects should be allowed a certain maximum thinking time which should be adjustable to 1, 2 and 4 seconds.

Six quantities should be recorded per person and per "turn":

P, the serial letter of each of the 20 subjects, from *a* to *t*;

S, the nature of the presented stimulus, from 1 to 29;

R, the nature of the given reaction (reaction choice), from 1 to 29; (this number indicates which of the buttons the subject has pressed);

T, the reaction time in seconds, using three figures, of which the figures T_{100} (digits, 0-3), T_{10} (tenths, 0-9) and T_1 (hundredths, 0-9) should be recorded separately.

For reactions after the end of the thinking time, the nature of the reaction should be recorded with the sign - and the time with 000.

If a subject reacts more than once, only the first reaction should be recorded.

Although the recording takes place in a fixed sequence, it is thought desirable — with a view to simplifying the later sorting process — that the serial letter of the relevant subject should always be printed beside the record. For the same reason it is thought useful to record repeatedly the presented stimulus together with the data of each subject.

Technical requirements

In each turn the quantities P, S, R, T_{100}, T_{10} and T_1 are stored in a matrix memory consisting of magnetic cores. The six quantities together are called a "*score*".

At the end of the thinking time the scores of all subjects are transferred to punched tape for further processing. If required, this information can in addition be typed by an electrically controlled typewriter on one line, which contains a maximum of $20 \times 6 = 120$ symbols. This storage of figures consti-

tutes a directly readable record of the measurements, but is of course not suitable for automatic processing.

The *programming* of the complete experiment — e.g. of 100 trials — is performed by a *stimulator*, which is controlled by means of a programme tape. The stimulator also activates the *physical source* which produces the signals (e.g. optical or acoustical) to be used as stimuli. This means conversely that the physical sources used, and the selection of the various signals they can produce, must also be controllable by means of punched tape. An incidental advantage of this is that a given programme recorded on punched tape — e.g. of successive stimuli in a random sequence — can be used for signals of widely different kinds: now for letters, now for colours, now again for tones of different pitch, and so on.

The *automatic processing* takes place either in an electronic apparatus, called the *histometer*, the processing programme for which is likewise on punched tape, or in one of the electronic computers in Philips Computer Centre. In the initial stage of any investigation it is particularly desirable to have the data processed on the spot immediately, and for this reason the histometer is generally to be preferred.

The standard code for programming and scoring is a binary code of seven units. Each group of seven binary units forms a *heptade*.

Experimental method

By *experiment* we mean here the complete experiment on a number of subjects. An experiment consists of a number of *series*, a series of a succession of *turns*, say 100. For each subject a turn consists successively of waiting for a warning signal (in general, only before the first turn of a series), observing the stimulus and responding to it, e.g. by pressing a certain button.

If the warning time, the thinking time and the time needed for recording the score on the punched tape each last 2 seconds for example, the total *cycle time* per turn is 6 seconds. An experiment of 100 turns then lasts 600 seconds = 10 minutes. In this time the 20 subjects have each reacted 100 times. Since each individual reaction is characterized by the six quantities P, S, R, T_{100}, T_{10} and T_1, the punched tape at the end of the experiment contains $20 \times 100 \times 6 = 12\,000$ heptades, which take up a length of 30 metres of tape. The recording time of 2 seconds follows from the number of 120 symbols per turn and from the punch speed, which is 60 heptades per second.

It is evident that in this way a very large number of data can be collected in a short time without unduly tiring the subjects. But the experimenter needs a great deal of data if he is to be able to draw conclusions with any degree of certainty. For not only does he wish, for example, to determine the statistics of the reaction times (perhaps for the subjects individually) but also the reaction choice. No less important than the reaction times are the number and kind of errors made. Moreover, experiments of this kind involve the unavoidable — and in themselves particularly interesting — phenomena associated with the learning and tiring of the subjects, phenomena which in the long run increase or decrease their performance. In many cases this makes it desirable to analyse the 100 turns in separate groups of 10 or 25.

The rapid succession of turns — one every 6 seconds in the example just given — is in practice a considerable advantage. Long waiting times between the stimuli have an adverse effect on the subjects: in spite of all previous motivation, the attention required for the experiments tends to slacken.

Block diagram and time-sequence diagram of the DONDERS

Each turn is indicated by the *stimulator* in the DONDERS. After each turn this device resets itself to the next turn, i.e. to the next position of the programme tape in the punched-tape reader. The stimulus quantity S read from it is transmitted to the physical source and noted in the matrix memory of the DONDERS. Shortly after the warning signal has been given, the physical source is activated and at the same time an electronic clock is started, which measures in steps of 0.01 second.

The subjects perceive the stimulus, and as soon as they react, the reading on the clock (T_{100}-T_{10}-T_1) is recorded in code in the row of matrix memories allocated to the subject P. A thinking time of, say, 2 seconds has been set on the control panel beforehand, with the effect that, at the end of the thinking time, the clock stops and late reactions are not recorded. Moreover, the stopping of the clock initiates a new phase, starting the process of successively "reading out" the rows of the memory. The data read out are recorded on a punched tape. If fewer than 20 subjects are taking part in the experiment, a switch on the control panel is set to the position corresponding to their number. This avoids reading out more rows of the memory than there are subjects. When the read-out is completed, this is reported to the stimulator, which can then start a new turn.

Fig. 1 shows the block diagram and *fig. 2* the time-sequence diagram of the DONDERS. Various details of these will be dealt with presently.

Information processing by the histometer

The experimental data are recorded turn by turn, in the sequence of the subjects tested, on the first punched tape, called the *master tape*. As mentioned, they can also be typed out by an electrical typewriter; in that case they provide the first rough record, called the *master sheet* (an example of which can be seen in fig. 8).

Suppose we wish to produce a graph (histogram, see fig. 9) of the statistical distribution of the reaction times. Such a histogram can be obtained by keeping a running tally of the various times (or classes of times). If each score were recorded not on a punched tape but on a punched card, the tally in accordance with the three criteria (T_{100}, T_{10}, T_1) could be made by simply sorting the cards into stacks. The cards are then sorted first in order of the last of the three figures, that is in order of T_1, thus producing ten stacks of cards, with $T_1 = 0, \ldots, 9$ respectively. Next, the stacks in this order can be put together again to form a single stack, which is then sorted in order of T_{10}. The ten new stacks thus obtained are again combined into a single stack, which is finally sorted in order of T_{100}. This results in four stacks. If these are placed one on top of the other, the cards will be in a sequence of ascending values of time. How often each time occurs in the stack can now easily be counted.

In our case, however, the data are not available on punched cards but on punched tape. The sortability of cards could only be obtained with tape by cutting it into pieces. Another way to attain this object, however, is by making tapes of successive "generations", the original tape remaining intact. For this purpose the master tape, driven by the control tape of the histometer, is scanned several times. In the first round, only the scores with $T_1 = 0$ are copied; in the second, only the scores with $T_1 = 1$; and finally, in the tenth round, the scores with $T_1 = 9$. The first-generation tape obtained in this way contains all the scores of the master tape, sorted in ascending order of the ten values of T_1. The procedure is now repeated with the first-generation tape and produces a second-generation tape, on which the scores are likewise sorted in order of the ten values of T_{10}. Finally the sorting procedure is repeated to produce a third-generation tape in order of the four values of T_{100}.

The foregoing will be illustrated with a simple example. Suppose that the master tape (M) contains the following series of numbers:

M: 23 30 27 31 35 41 28 32 29 27 36 29 36 36 39 31.

In order to arrange these numbers in ascending order, we first make in ten rounds a first-generation tape on which, in the first round, only those numbers are taken over from M which end with an 0; in the second round, only those that end in 1 are taken, and so on until, in the tenth round, only those are taken that end in 9. On this first-generation tape (G_1) we then have:

G_1: 30 31 41 31 32 23 35 36 36 36 27 27 28 29 29 39.

Next, we make from G_1 in ten rounds a second-generation tape (G_2), on which we take over successively from G_1 only the numbers beginning with an 0, then those beginning with a 1 (both categories are missing here), next those beginning with a 2, and so on. The result is as follows:

G_2: 23 27 27 28 29 29 30 31 31 32 35 36 36 36 39 41.

In this way we have obtained the arrangement required.

It should be added that the right result would not have been attained if the numbers had been sorted first in order of the first figure and then in order of the second figure. The first-generation tape would then have contained:

G_1: 23 27 28 29 27 29 30 31 35 32 36 36 36 39 31 41,

and the second-generation tape:

G_2: 30 31 31 41 32 23 35 36 36 36 27 27 28 29 29 39.

This sequence is incorrect.

This method of sorting has been described here for the classification of reaction times. It is universally applicable, however, and is therefore just as useful for the classification of subjects, stimuli and reaction choices, as well as combinations of these criteria. The number of necessary generations from the master tape is as a rule, limited to a few. For if a consolidated result of observations is to lead to significant conclusions, it forbids — almost by definition — any subdivision into a large number of classes.

The times T_{10} and T_1 are coded as follows:

0 = 10000,	5 = 00101,
1 = 10001,	6 = 00110,
2 = 10010,	7 = 00111,
3 = 10011,	8 = 01000,
4 = 10100,	9 = 01001.

This is the conventional binary code (which readily lends itself to adding), preceded by a 1 for the values 0, . . . , 4, and by an 0 for the values 5, . . . , 9.

This code also makes it possible to sort into wider time classes than the minimum. When sorting into classes of 0.01 s, these classes obviously coincide with the measured value. The time code must then be read out in full. Sorting into classes of 0.02 s is done by omitting the last binary digit, and sorting into classes with 0.05 s by omitting all binary digits except the first. Classes of 0.1 s are sorted by completely passing over the T_1 code, etc.

For the purpose of typing out a first-generation or second-generation tape in one or another consolidated form (histogram, table) it is desirable to add texts or subscripts to them. In the coding provision is made for this in the following manner. Each com-

plete character in the punched tape consists of seven holes (heptad). Six of these carry information in the form of letters or numerals. If this information relates to the results of the experiment, a seventh hole is punched. If, on the other hand, the information is "operational", it is then characterized as such by omitting the seventh hole. (By operational information we mean both the information used to control the histometer and that subsequently used for providing the processed result with an explanatory text.)

Further particulars

The stimulator

The progress of an experiment is controlled by the stimulator, a photograph of which can be seen in *fig. 3*. For the time-division of a turn, use is made of an electronic clock (counting circuit) which, for each turn, counts from 1 to a maximum of 10 000 in steps of e.g. 0.01 s. By means of a switch some of these numbers can be preselected as "signal numbers". As each signal number is passed, the counting circuit delivers a signal (electrical pulse) which e.g. lights the warning light, triggers the stimulus, etc. The counting circuit returns to its starting position as soon as the recording equipment reports that it has completed processing the data.

The physical sources

The physical sources that supply the stimuli can have a variety of forms. One of the simplest is an array of, for example, ten lamps, which can be switched on and off by means of relays, either one by one or in certain combinations. Loudspeakers, headphones, etc. can also serve as physical sources.

In experiments concerning the recognition of letters a "letter projector" is used as physical source. This consists of a light source, a projection system and a wheel; the wheel is fitted with slides of the letters to be presented for recognition. The position of the wheel is governed by the stimulator.

The matrix memories

For recording the reaction times a register is used for each of the twenty subjects. In this register, which consists of small magnetic ring cores [2]), the reaction time is recorded on the coincidence principle. As soon as a subject presses a button, a current $+\frac{1}{2}I$ passes through a wire threaded through all cores in his register ($+I$ being the current required to reverse the magnetization of the cores). Only cores indicated by the reading on the electronic clock receive, via a second wire, an additional magnetizing current $+\frac{1}{2}I$, so that only these cores have their magnetization reversed. When this happens the reaction time is recorded. One of the matrix memories of the DONDERS is visible in *fig. 4*.

Sometimes a subject may inadvertently press the button a second time, or he may press undecidedly, so that contact is made several times. To prevent such duplicate reactions from being recorded, the following method was adopted. When a subject presses one of the 29 buttons available to him, he closes the corresponding one of 29 circuits. The 29 circuits have a common return line in which a transistor network is incorporated. When the reaction time is

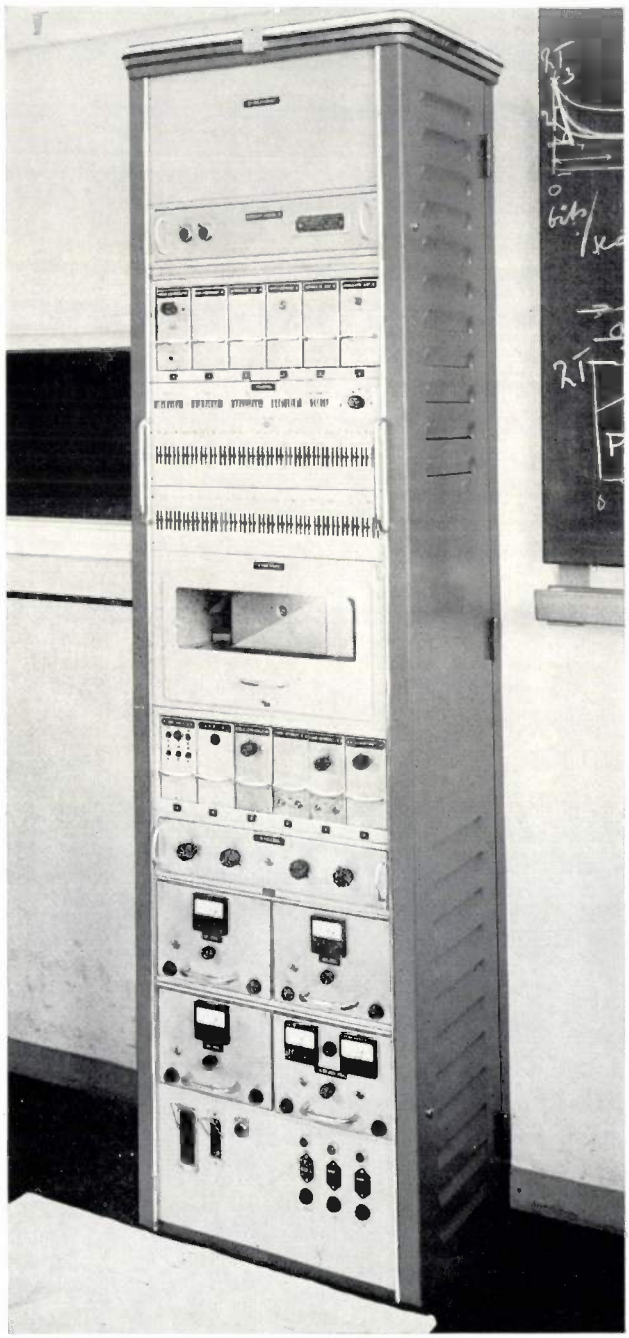


Fig. 3. The stimulator, which controls the progress of an experiment.

[2]) See e.g. Philips tech. Rev. **20**, 193, 1958.

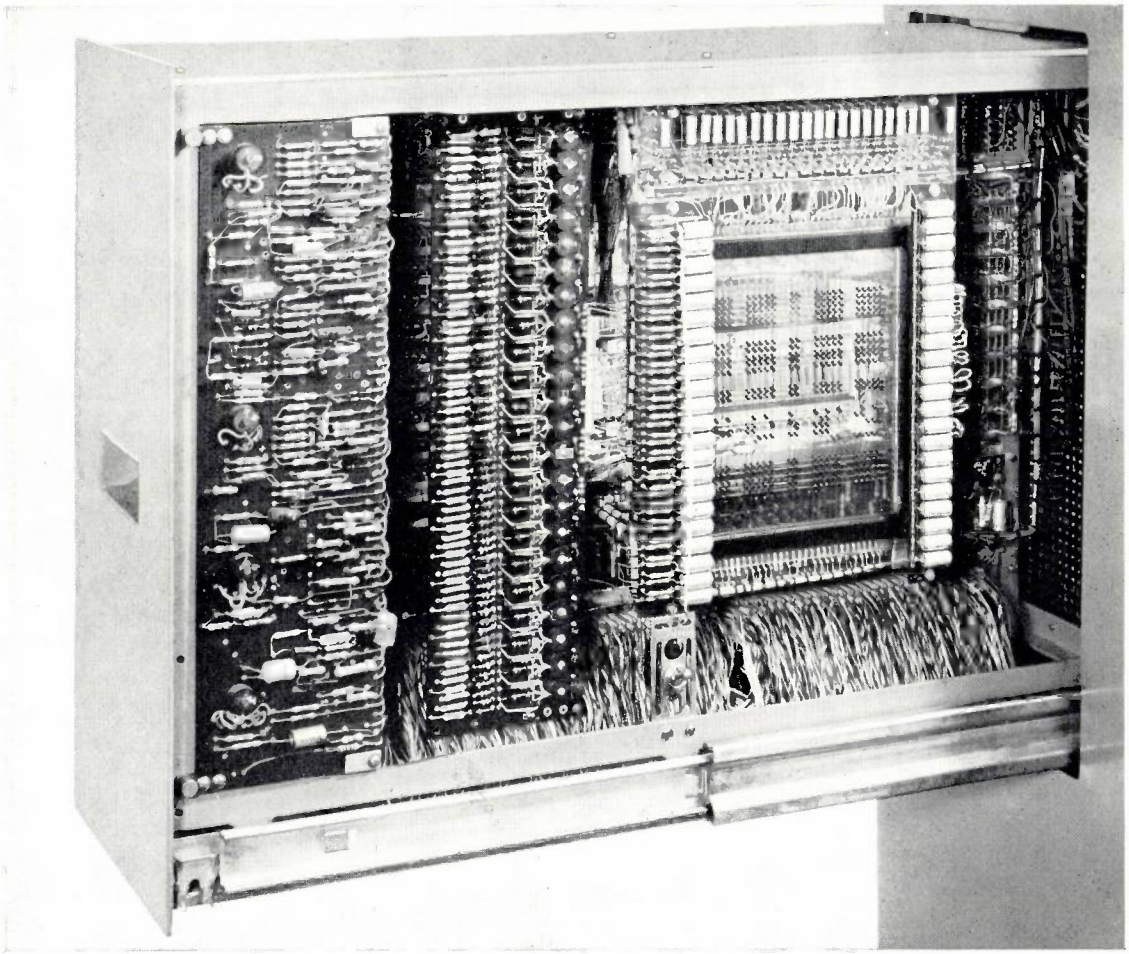


Fig. 4. The circuit wiring behind the operator's presetting panel (see fig. 7) is contained in a drawer, which is here shown open. Through the transparent panel right of centre can be seen part of the cores of the matrix memory.

to be recorded (the reading on the clock when the contact is first made), this transistor is made conductive. The first reaction of the subject, however, has the effect of blocking the transistor circuit; the return line is then broken, thus precluding the recording of any subsequent reaction. The transistor circuit is not unblocked and the path opened for recording the next reaction time, until all data relating to the turn in progress have been recorded and the next turn begins. Further details of the operation are given in *fig.* 5. The 20 blocking circuits are grouped in block B in fig. 1.

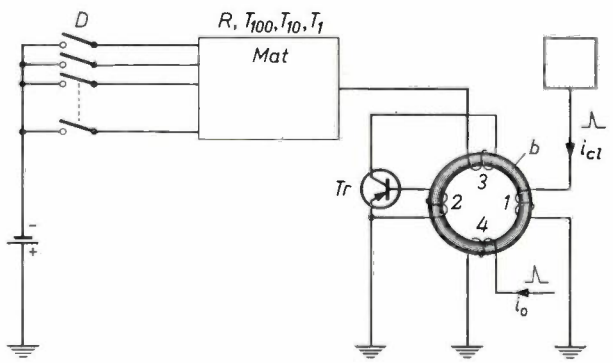


Fig. 5. Circuit which prevents the recording of inadvertent reactions and at the same time blocks the matrix memory during the movements of the electronic clock.

At the moments when the clock is in a fixed position a single pulse i_{cl} passes through the winding *1* on the blocking core *b*. In winding *2* the pulse induces a current surge which serves as the base current for the transistor *Tr*. If the subject reacts by pressing one of the 29 buttons *D* before him, a current passes via the collector through the matrix *Mat* and through winding *3* on the blocking core *b*. This current reverses the magnetization of the core, which means that the reaction choice *R* and the reaction time T_{100}-T_{10}-T_1 of the subject are stored in the matrix, and that the flux of the core, in the direction in which it is now magnetized, cannot become stronger. Consequently no base current can now be induced in the transistor and therefore no collector current can flow. The matrix, then, is unable to store a subsequent reaction until the magnetization of core *b* has been returned to its original direction by a current pulse i_0 through the deblocking winding *4*. The current pulse i_0 originates from block *L* (fig. 1) and is delivered during the read-out.

There is one such blocking circuit for each subject, i.e. a total of 20, denoted by *B* in fig. 1.

Read-out of the matrix memories

When the thinking time has elapsed, the information stored in the memory has to be "read out" in a particular sequence, which amounts to determining the direction of magnetization of each core. This direction decides whether a hole has or has not to be punched in the tape at the position allocated to the core.

The read-out is organized roughly in the following way. The subject registers are always read out in a fixed sequence. This is done by means of a *read-out counter* (block M in fig. 1), consisting of a magnetic counter with as many positions as there are subjects. For each subject the information to be read out comprises the quantities T_{100}, T_{10}, T_1 and R. In addition there are the data S and P; in each turn the quantity S is the same for all subjects, and for the same subject the quantity P is the same in all turns. The six heptades ($T_{100}, \ldots, P$) are recorded in an intermediate register (IR in fig. 1). A write-out counter (a series of six flip-flops, block N in fig. 1) ensures that the quantities constituting one score are passed successively via a gate circuit (G in fig. 1) to the print-out device (punching machine or typewriter or both). Each heptade, e.g. T_{100}, continues to be offered to the print-out device until a message from the latter has been received and processed. As far as the punching machine is concerned, this means that the punch has made one stroke and that the tape has moved up one step. As regards the typewriter, this sends back a message when one of the type bars has made a stroke and returned to rest. The receipt of this message is thus proof that the heptade has been processed.

The write-out counter now moves one position further, enabling the next heptade, e.g. T_{10}, to be presented to the print-out device. In this way all six heptades are consecutively transmitted to the punch, typewriter or to both. This being done, the whole process is repeated for the next subject.

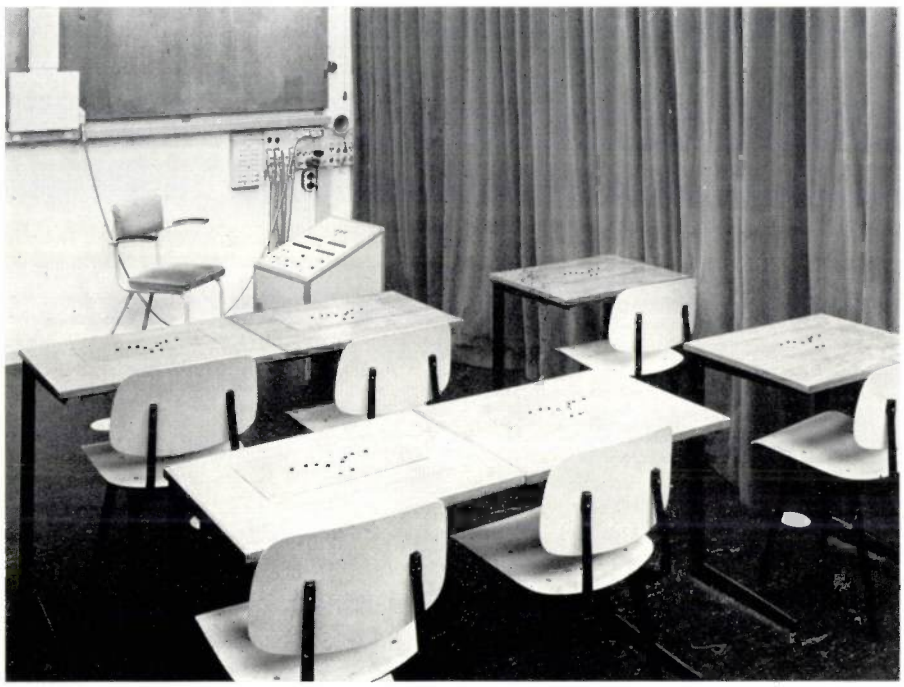


Fig. 6. View of the test room, showing six of the twenty desks at which the subjects sit. In the background can be seen the experimenter's desk. Top left, two optical stimulus sources, consisting of lamps in various configurations (more clearly seen in the photograph on page 58 of this number).

As mentioned, the sequence in which the six heptades are passed to the print-out device can be selected, permutations for these heptades being made possible by 36 selector switches (denoted by block O in fig. 1). The possibility of being able to preset any desired sequence allows better processing of the punched tape by the histometer. In addition there are six switches by which any of the six heptades can be omitted.

Experimental rooms

The subjects are seated in a test room containing 20 desks and chairs (*fig. 6*). Each desk is fitted with an interchangeable panel containing push-buttons and lamps and possibly a warning light, one or more stimulus lamps, etc. The number and arrangements of

the pushbuttons and lamps are chosen in accordance with the experiment to be performed. The panels are connected by 32-core cables to the DONDERS, which is set up in an adjacent room. One-way windows in the partition between the rooms enable the operator of the DONDERS to see that everything is in order in the test room (*fig. 7*) while remaining virtually unseen to the subjects. The windows also provide visual communication with the experimenter in the test room. In addition, they can both speak to each other on an intercom system.

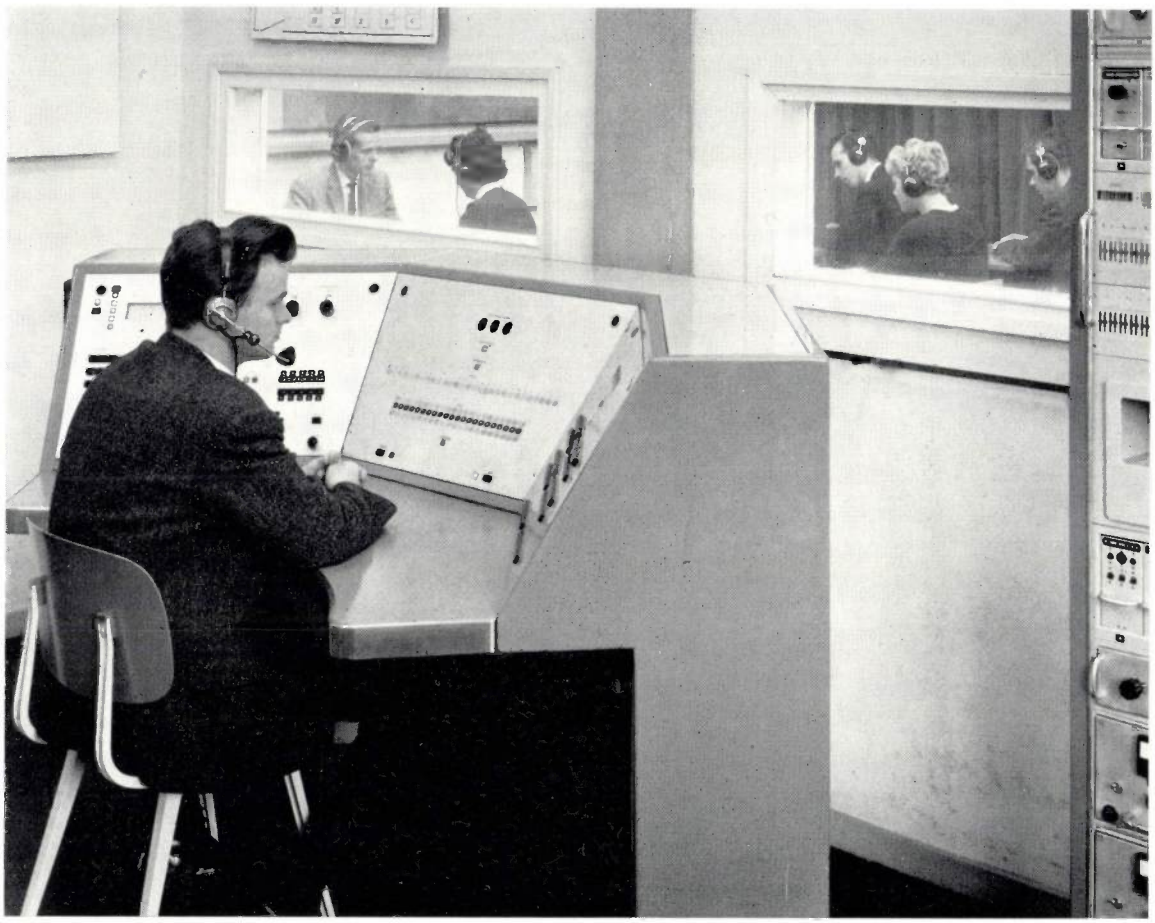


Fig. 7. The operator's console, in which the DONDERS is mounted. On his left the operator has the presetting panel, facing him is the monitoring panel, and on his right is the control panel. A duplicate of the monitoring panel is on the experimenter's desk (fig. 6). On the extreme right, a part of the stimulator. The subjects can be observed through windows in the operator's room.

Operation of the DONDERS

For each experiment various actions are required to operate the DONDERS (some before and some during the experiment) which, however, take up very little time.

Before the experiment begins, the operator has to switch on the mains voltage, check various supply voltages, set the write-out counter to the number of subjects taking part in the experiment, select the print-out device to be used (punch, typewriter or both), determine which of the quantities P, S, R, T_{100}, T_{10} and T_1 are needed and in which sequence they are to be recorded, set the thinking time and select the mode of stimulus. (The stimulus sources can if necessary be controlled manually, so that the turns can follow each other at any speed required.) The controls for these settings are contained on a *presetting panel* on the left of the console at which the operator sits (fig. 7).

In front of him the operator has the *monitoring panel*, which contains a large number of pilot lamps. From these he can see, among other things, what stage an experiment has reached at any given moment (start, warning, presentation of the stimuli, end of thinking time, print-out) and which persons have reacted. Further, a row of signal lamps indicates which stimulus has been presented. An electronic counter keeps a tally of the turns completed

and stops the experiment as soon as the preset number is reached. An indicator shows how many turns have still to come. By means of a push-button the operator can interrupt the experiment if unexpected circumstances should make this necessary.

A duplicate of the monitoring panel is contained in the experimenter's desk in the test room (fig. 6).

On his right the operator has the *control panel* (fig. 7). This contains keys with which he can control the stimuli by hand. The panel furthermore contains a complete series of push-buttons corresponding to the reaction buttons before the subjects, so that the operator can himself verify the effect of manipulating these buttons. Finally, the control panel contains push-buttons used for tracing faults.

Example of a reaction experiment

Method

We shall now, by way of example, describe the method and results of an experiment in which ten stimuli were presented to ten subjects, each of whom could give ten reactions.

The method was as follows. Each subject had a panel before him with ten push-buttons, so arranged that he could easily place his finger tips on them (fig. 6). The subjects had their eyes fixed on a board containing ten lamps, which was attached to the wall facing them. The stimuli consisted of the flashing of one lamp. The lamps were arranged in the same configuration as the push-buttons; in this way it was very easy to find out which button corresponded to which lamp.

After a practice run of 50 stimuli, four series of 100 stimuli were presented. Each stimulus lasted one second, the thinking time two seconds, and the cycle time three seconds. Each series, therefore, lasted five minutes, followed by an interval of one minute. A warning signal preceded each series, but not each turn. The complete experiment lasted half an hour, after which the subjects reported their findings and discussed them.

Results

Fig. 8 shows part of the master sheet of the first series. Each line represents the data of the ten subjects for one trial.

Fig 9 shows the histogram of the reaction times of the first series, sorted into time classes of 0.05 second. For each class the results were sorted into correct reactions (symbol 0), and wrong reactions (symbol X). This gives a histogram, from which the histometer takes its name.

A more compact though visually less clear method of recording is given by the time-stimulus diagram shown in *fig. 10*. This presents the numbers per time class, subdivided moreover according to the ten stimuli. It can plainly be seen that for the stimuli *b*, *i*, *c* and *h* (corresponding to the third and middle fingers) the distribution shifts towards longer reaction times.

The average reaction times $\bar{t}$ and the percentage errors *f* obtained from the histograms for the total experiment are shown in *fig. 11*. In the marked mutual differences, two effects are involved: the *dis-*

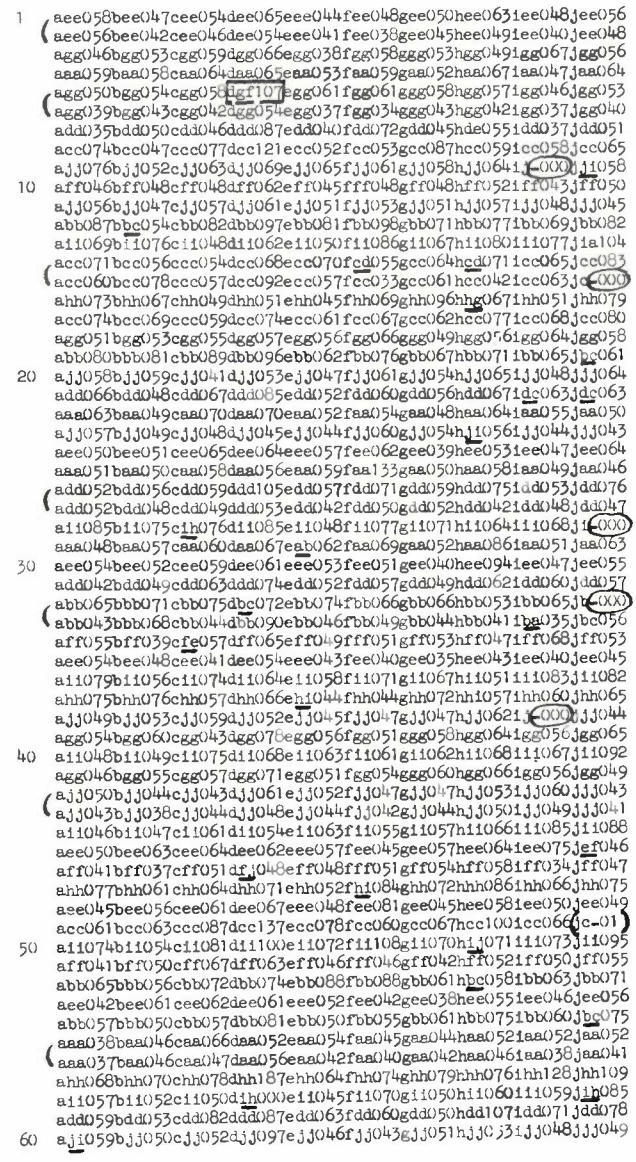


Fig. 8. Part of a master sheet for series 1. Sixty lines give the results of 60 of the 100 turns for 10 subjects *a* to *j*. For each subject six symbols (a *score*) are printed in the sequence P-S-R-T_{100}-T_{10}-T_1. Thus, the score enclosed in a rectangle, for example, means that in the fifth turn the subject *d* responded to stimulus *g* with reaction *f* in 1.07 second.

Wrong reactions are underlined, late reactions (score - 000) are encircled. The score (c-01) in the 49th turn is an error of the DONDERS. Successions of identical stimuli are marked with half a bracket on the left. Note that the reaction time is usually shorter for a repeated stimulus.

tinguishability of the stimuli (which is better for the boundary stimuli than for those in the middle) and the *adeptness of the fingers*. The good results for the little fingers and their stimuli (*a* and *j*), both in terms of reaction time and percentage of errors, are primarily attributable to the first effect, whilst those for the index fingers and their stimuli (*d* and *g*) can be ascribed mainly to the second effect.

Fig. 12 shows the transposition matrix which indicates the number of times the subjects responded to each stimulus with each reaction. The diagonal terms give the numbers of correct reactions, the extradiagonal terms the number of wrong ones. It can clearly be seen that most errors were due to reacting by pressing neighbouring push-buttons. It is noticeable that the errors of the third and middle fingers show some asymmetry: there is a marked tendency to use the middle finger for the third finger stimulus and to use the index finger for the middle-finger stimulus.

The average reaction time and the standard deviation determined for each stimulus and for each subject show values that deviate by about 25% both as regards persons and stimuli and as

Stimulus	a	b	c	d	e	f	g	h	i	j
Reactietijd in millisec.										
0100	---	---	---	---	---	---	---	---	---	---
0200	---	---	---	---	---	001	---	001	001	---
0300	025	012	016	017	049	050	020	020	011	022
0400	133	043	029	093	150	166	111	051	044	162
0500	138	086	105	147	143	115	164	090	107	149
0600	074	104	103	088	049	044	075	098	108	045
0700	017	071	071	032	009	008	016	084	066	015
0800	007	052	042	011	002	004	003	032	045	003
0900	001	020	014	001	001	---	001	011	012	001
1000	---	006	007	003	---	001	001	005	007	---
1100	---	003	004	---	---	---	---	---	001	---
1200	---	---	001	---	---	---	---	002	---	---
1300	001	001	001	---	---	---	---	---	---	---
1400	---	---	---	---	---	001	---	002	---	---
1500	---	---	001	---	---	---	---	001	---	---
1600	---	001	---	---	---	---	---	001	---	---
1700	---	---	---	---	---	---	---	---	---	---
1800	---	---	---	---	---	---	---	001	---	---
1900	---	---	---	---	---	---	---	---	---	---
2000	---	001	001	004	---	---	001	---	002	002

Fig. 10. Time-stimulus diagram for a complete reaction experiment. For each of the ten stimuli *a-j* the histometer has sorted the total number of 400 measurements on ten subjects into time classes of 100 milliseconds (indicated in the first column). The difference in the time distribution of the numbers for the various stimuli can be clearly see.

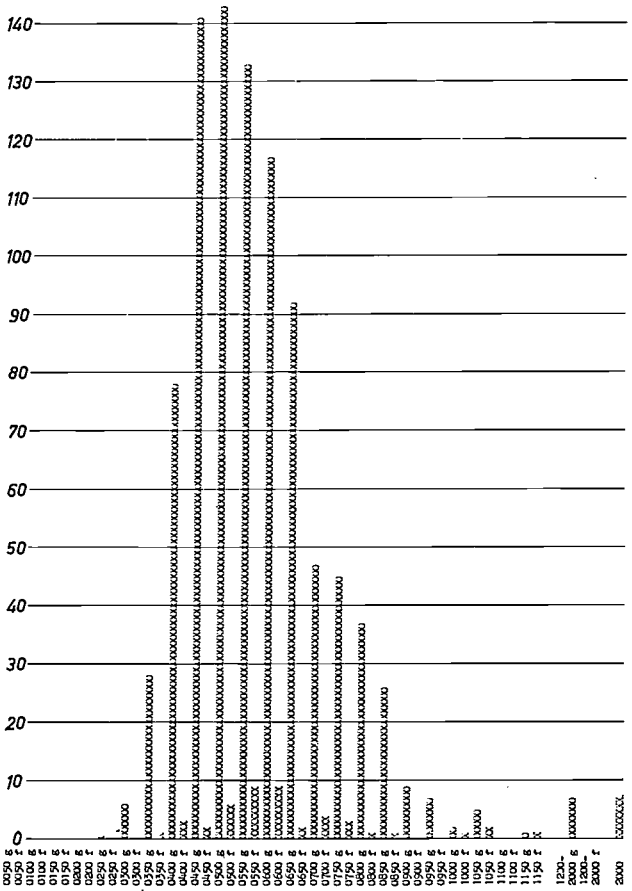


Fig. 9. Histogram of the reaction times for series 1. In time classes of 50 milliseconds the correct results (g) are first typed out with the symbol 0, then the wrong reactions (f) with the symbol X. The distribution curve is the result of 100 measurements on ten subjects.

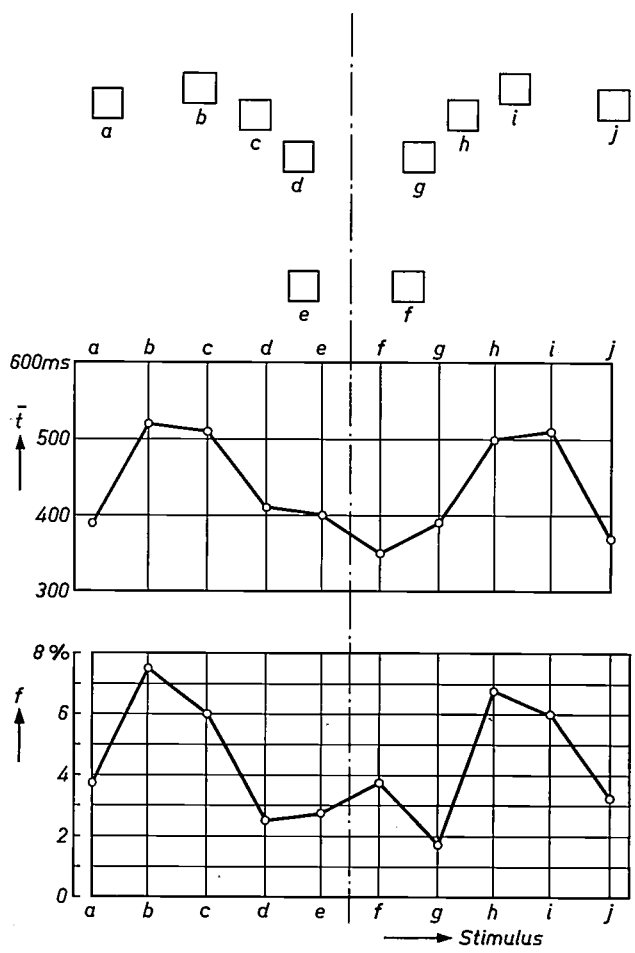


Fig. 11. The average reaction time $\bar{t}$ and the error percentage f pertaining to the stimuli *a-j* for the total experiment. Above: the configuration of the stimulus lamps (likewise of the push-buttons).

For the stimuli corresponding to the third fingers (*b* and *i*) and to the middle fingers (*c* and *h*) both the reaction time and the error percentage are distinctly greater.

Stimulus / Reactie	a	b	c	d	e	f	g	h	i	j
a	382	004	---	---	---	---	---	---	002	---
b	015	368	008	---	---	001	---	---	---	---
c	---	025	368	003	---	---	001	---	---	---
d	---	001	016	382	001	001	---	---	---	---
e	---	---	---	005	389	011	001	---	---	---
f	---	---	---	---	010	380	001	---	---	---
g	---	---	---	002	---	001	385	021	004	---
h	---	---	---	---	---	---	004	372	014	001
i	---	---	---	---	---	---	---	005	379	012
j	---	---	---	---	---	001	---	001	004	385
Totaal fout	015	030	024	010	011	015	007	027	024	013

Fig. 12. Transposition matrix for the total experiment. The histometer has now sorted the scores according to the numbers of times the subjects responded to a given stimulus with a given reaction.

The diagonal terms give the numbers of correct reactions, the extradiagonal terms the numbers of wrong reactions. Late reactions have not been counted. The bottom lines gives the sums of the number of *wrong* reactions in each column. Note the tendency to use the middle finger (*c* and *h*) instead of the third finger (*b* and *i*), and the index finger (*d* and *g*) instead of the middle finger.

regards stimuli per person. The ratio of the standard deviation to the average reaction time is found to be constant at 0.19 for eight of the ten subjects. For subjects *b* and *i*, however, the respective ratios are 0.15 and 0.17.

It is also interesting to note that when, by chance, two identical stimuli follow each other — which happened in 10 per cent of the trials — the reaction time for the repeated stimulus was about 15 per cent shorter than for the first. Evidently the subject still has the earlier reaction "at his fingertips". This can easily be seen on the master sheet (fig. 8) from the trials marked with a half bracket.

The histograms of the individual subjects show appreciable disparities, in the average reaction time as well as in the form of the distribution curve and in the number and kind of errors made. Significant differences are also found between the four series.

The results described above illustrate that statistically reliable data on the numerous aspects involved in a complete reaction experiment can be of the utmost importance for testing theories on human reactions.

The processing of the 4000 observations mentioned, which were made by ten subjects within half an hour, into four master sheets and 30 histograms would have taken many months without the aid of automatic devices. Using the histometer as a simple sorting device, all this took no more than a week, including the making of the control tapes and generations of tapes. Although this is still a long time compared with the duration of the experiment, it is nevertheless — and this is the essential point — short compared with the time spent by the scientist on interpreting the processed results.

An electronic computer could perform the computing work much faster and more flexibly, but here again the time effectively needed to prepare the programmes, which vary from one case to another, should not be underestimated.

For the scientist it remains of the utmost importance to obtain processed results of measurements as soon as possible, so that he can use the conclusions drawn from them in his further research.

In connection with the realization of the total project, special mention should be made of: G. J. J. Moonen for the electrical development, D. J. H. Admiraal and J. C. Valbracht for the electrical adaptation, B. Lopes Cardozo for the histometry, F. F. Leopold for the procedure and the ergonomic design of the desks and panels, and C. J. F. P. van den Bosch *) and W. K. Waisvisz *) for the industrial design.

*) Philips Lighting Advisory Bureau. Eindhoven.

Summary. In the Institute for Perception Research at Eindhoven an electronic system for measuring human reactions has been built in cooperation with the Philips Research Laboratories. Named after the Dutch physiologist Donders, the system can automatically perform a large amount of the routine work involved in reaction measurements.

Up to 29 different optical or acoustic stimuli can be presented to a maximum of 20 subjects. Reaction times can be automatically recorded with an accuracy of 0.01 s. The whole programme of an experiment is pre-recorded on punched tape which drives the principal control organ, called the "stimulator". The results of the measurements appear in code on a second punched tape which is suitable for further processing, either by an electronic computer or by a "histometer" — a device specially designed for the purpose. The results can also be typed on an electric typewriter.

Finally an experiment is described in which ten stimuli were presented to ten subjects who had a choice of ten reactions.

NOTES ON THE HISTORY OF REACTION-TIME MEASUREMENTS

by G. ten DOESSCHATE *). 159.938.343

A historian is someone interested in the past. It is not unreasonable to begin a historical article with this definition.

From this definition it follows that everyone who observes his environment with interest is a historian. For to observe means to open the mind to stimuli in order to acquire information. That information will always relate to a state that belongs to the *past*, even though to a very recent past, more recent than dealt with by lecturers on contemporary history. This is because the moment at which a stimulus begins to operate on a sense organ and the moment at which the effect of the stimulus is perceived, are separated by a certain very short time, called the perception lag (the latent period of perception).

The refined methods by which perception lags can nowadays be measured and correlated with other phenomena in experimental and applied psychology — methods of which some account is given in this issue — might well be apt to make us forget the labour pains that preceded the advent of reasonably clear-cut concepts and reliable quantitative data relating to perception. An important contribution to their introduction was made by a Dutchman, Franciscus Cornelis Donders (*fig. 1*). In the present short article we shall try to evaluate the significance of Donders' work in this field. To avoid misunderstanding, it should be emphasized from the outset that this outline cannot possibly do justice in any way to the great man that Donders really was [1]).

In thinking about thought the ancient philosophers also considered the process of perception, which they treated by analogy with the process of thinking. In a book on optics Claudius Ptolemy (the man who has given his name to the pre-Copernican cosmology, and who lived in Alexandria round about 150 A.D.)

1) E. C. van Leersum, Het levenswerk van F. C. Donders, Haarlem 1932, and F. P. Fischer and G. ten Doesschate, Franciscus Cornelis Donders, Assen 1958, give some impression of Donders' importance as a physiologist and opthalmologist and in other respects. According to private communication, a bibliography of Donders' scientific work, compiled by C. H. van Herwerden, covers 339 publications, those which appeared in several languages being counted as one publication. Of these, 77 deal with physiological and 135 with opthalmological subjects.

Fig. 1. Franciscus Cornelis Donders (1818-1889) at the age of 57.

*) Retired doctor, Utrecht. Until 1957 Dr. Ten Doesschate was on the staff of the National Aeronautical Medical Centre at Soesterberg.

saw it roughly in this way. As soon as an object appears before the eye it makes an impression on the visual organ, but this impression does not yet amount to perception of the object. Perception, it was thought, was based on the processing of the impression by the soul (psyche). This processing was believed to be comparable to a syllogism in logic. If, for example, a sphere appeared before the eye, the process initiated by the impression would correspond to the following chain of reasoning.

1) I receive this particular impression.
2) When I earlier received a similar impression it was due to a sphere.
3) Therefore what I see now is a sphere.

We shall pass over the question of what Ptolemy thought he had explained with this formula; what we are mainly concerned with is what Ptolemy added to it: he said that a syllogism of this kind is worked out at such a speed that one does not notice that one has in fact reasoned. For him — at least for all questions that he could pose — such a process was infinitely fast.

For many centuries this hypothesis went unchallenged, or rather the question of perception proper was provisionally more or less in abeyance because of a preoccupation with other questions. Right into the 19th century men of learning were struggling to explain the *mechanism* of sensory perception, particularly of seeing and hearing. Physical, physiological, philosophical and anatomical findings and theories appeared thick and fast, sometimes supplementing each other and sometimes contradictory. Kepler, Mersenne, Descartes and Steno on the visual side, and Coiter, Schelhammer, Mersenne, Willis, Perrault and Du Verney on the hearing side are some of the more important names in connection with this struggle [2]). It is true that the connection between perception and action — another important matter dealt with in this issue — also received attention, and this brought the scientists back near the question raised by Ptolemy, as to what perception actually is. An excellent example is offered by a famous illustration in Descartes' "Traité de l'homme" (*fig. 2*). But Descartes himself regarded the essential link in this connection, namely becoming aware and the step from this to prompting the action, as "an inscrutable secret residing in the very essence of the soul".

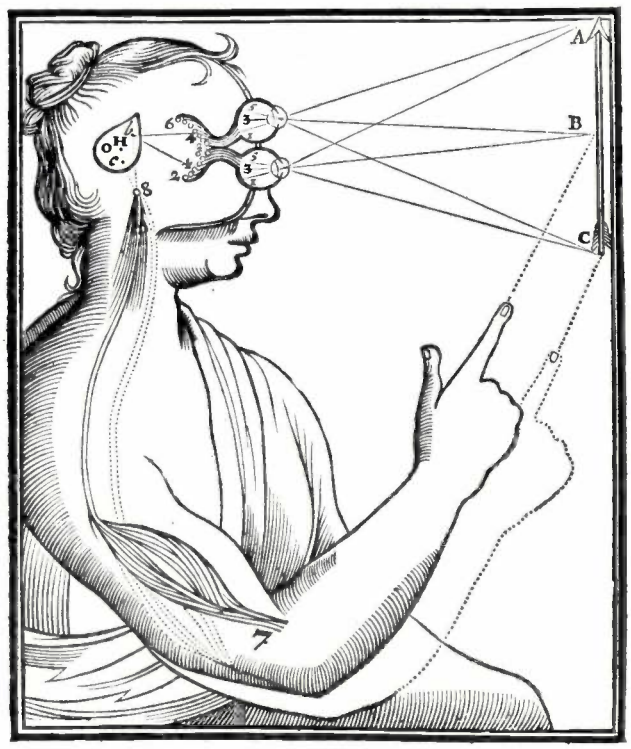


Fig. 2. Diagram from Descartes' Traité de l'homme. Paris 1662, representing the arising of a sensory impression and an active response it evokes. According to Descartes the connection between the two processes was to occur in the *epiphysis cerebri* or pineal gland, an unduplicated part of the brain which he believed to be the seat of the soul.

For the time being, then, there was nothing to take the place of Ptolemy's view of the perception process as a kind of syllogism worked out at infinite speed. Indeed, the syllogism idea survived into the second half of last century, Von Helmholtz pointing to a similarity between perception processes and "unconscious conclusions", which he considered to be a kind of syllogism ("Analogieschlüsse"). Understandably, then, it never occurred to anyone that the time occupied by these very fast processes might be measured. As late as about 1840 the great physiologist Johannes Müller predicted it would never be possible to measure such infinitely short intervals of time. The view continued to be accepted that perception and similar processes were attributable to the infinitely rapid movement of an imponderable (an "animal spirit", for example) or to some special "psychic principle".

Yet astronomers had already noted certain facts that were later considered as evidence of a lag in perception.

The oldest case found in the literature is a communication by the Rev. Maskelyne, who had been placed in charge of the Royal Observatory at Greenwich, and who wrote as follows [3]). "I think it neces-

[2]) It was recently reviewed in an interesting paper, containing extensive references to the literature, by A. C. Crombie, The "animate and sensitive body" in Renaissance science, which he presented at the 10th International Congress on the History of the Sciences, held at Ithaca (U.S.) in 1962 (as yet unpublished).

[3]) Astronomical Observations made at the Royal Observatory at Greenwich, 1799.

sary to mention that my assistant, Mr. David Kinnebrook, who had observed the transit of the stars and planets very well, in agreement with me, all the year 1794, and for great part of the present year, began, from the beginning of August last, to set them down half a second of time later than he should do, according to my observations; and in January of the succeeding year, 1796, he increased his error to 8/10ths of a second. As he had unfortunately continued a considerable time in this error before I noticed it, and did not seem to me likely to get over it, and return to a right method of observing, therefore, though with reluctance, as he was a diligent and useful assistant to me in other respects, I parted with him."

So poor Kinnebrook lost his job, and he would hardly have dreamt that the results of his observations, recognized as erroneous by himself, no doubt, as well as by others, would lead to his name still being bandied about a century and a half later.

Mistakes in the recording of times, as described by Maskelyne, were noted by other astronomers as well.

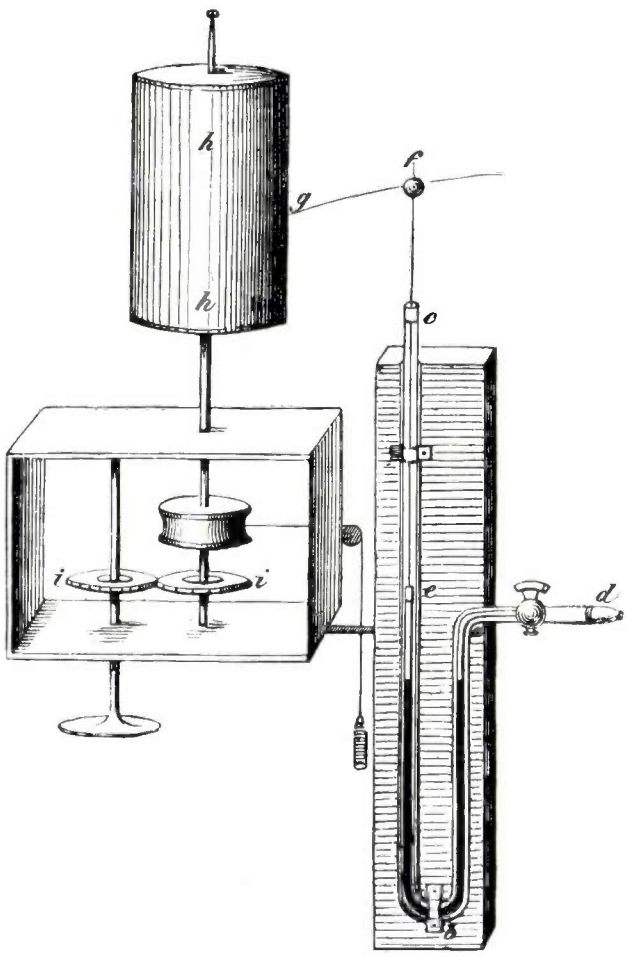


Fig. 3. Ludwig's kymograph, which he designed in 1847 for recording variations in blood pressure. Reproduced from K. Ludwig, Lehrbuch der Physiologie des Menschen, Winter, Heidelberg 1852-1856.

Bessel for example, writing to Gauss, made mention of "some experiments relating to very puzzling discrepancies in absolute times reported by different observers, discrepancies which Maskelyne noticed in 1794 and which I myself have been able to confirm".

Arago attributed the errors in observation to the method that was being employed at Greenwich. The observer saw the star moving slowly through the field of the telescope and, at a certain instant, pass through the cross-wires of the eyepiece. At Greenwich (and elsewhere) the observer had to interpolate that instant between the acoustically perceived (and counted) ticks of a seconds pendulum. Arago adopted a different method, requiring two observers. One was responsible only for watching the star and making a rapid arm movement at the instant of transit. The second had to concentrate on the ticks recurring at one-second intervals and interpolate the observed sudden arm movement between them. Unfortunately the results were even worse than at Greenwich.

Thus astronomers had noticed the reaction time phenomenon, and had started to study it experimentally; they did so as a matter of sheer necessity, for the effect in itself and above all the individual scatter involved (personal equation) prejudiced the required accuracy in the determination of transit times. On the other hand physiologists and psychologists until about 1840 showed no interest in this. During the period 1840 to 1850 this attitude changed, a new generation of great physiologists having appeared who mainly wanted to *measure* quantities. Du Bois-Reymond, a pupil of Johannes Müller, cast doubt on his master's prediction, quoted above, and in 1845 he outlined a method to put it to the test. In 1850 Von Helmholtz put this method into practice. He made use of a *recording instrument* known as the kymograph (*fig. 3*) which had been described a few years before by Ludwig, another of the new physiologists interested in measurement. It had been designed for the purpose of investigating variations in blood pressure [4]), but the same principle was to find numerous applications in physiology, including the experiments performed by Donders, which we shall shortly discuss. The instrument consisted of a vertical, uniformly rotating smoked drum; pressed against the drum was a stylus which underwent vertical displacement in accordance with variations

[4]) K. Ludwig, Beiträge zur Kenntniss des Einflusses der Respirationsbewegungen auf den Blutlauf im Aortensysteme, Müller's Arch. Anat. 1847, 240-302. — Regarding the general significance of Ludwig's apparatus as a stage in the evolution of recording techniques, a continuation of the line of development represented by Ruckert's recording hodometer (circa 1575), Wren's recording barometer (circa 1663), and Watt's well-known indicator, see: Hebbel E. Hoff and L. A. Geddes, The beginnings of graphic recording, Isis **53**, 287-324, 1962 (Part 3).

in the quantity under measurement. Experimenting on a frog, Von Helmholtz stimulated a nerve a short distance in front of the point where it entered the muscle, using the kymograph to record the instants at which stimulation and contraction of the muscle took place. It was found that a finite time elapsed between these instants. Von Helmholtz immediately went a step further and did an experiment in which the nerve was stimulated first at one and then at another point; this yielded two different delay times, from which it was possible to deduce the speed at which the stimulus travelled along the nerve. It turned out to be only 100 feet per second — a figure far from lightning speed; as Donders remarked later on, it was "a speed exceeded by birds in flight, approached by racehorses, and attainable by the human hand if the arm is moved very rapidly". Similar investigations on warm-blooded animals and man produced similar though individually different results. For humans, values varying between 6 m/s and 120 m/s were found, depending on the kind and above all on the diameter of the nerve fibres. Von Helmholtz himself pointed out that man's modest physical dimensions stood him in good stead: because of them, he suffered no great disability from the relatively low speed of nerve conduction (in today's traffic conditions we might think rather differently about that) whereas a whale would be at a decided disadvantage, since a full second would elapse before it felt an injury to its tail, and a defensive movement of the tail would not come until about two seconds after the blow had been struck.

The importance of the kymograph in these investigations should be stressed. It would not have been a very promising procedure to arrange for a human observer to record the instant at which stimulation occurred and the subject reacted, since there would have been some difficulty in deciding whether and in how far the subject or the observer was responsible for the recorded delay times. Indeed, it was no accident that these phenomena were first noticed by astronomers; in the observatory, the reaction times of the "subject" were being measured against the uniform rotation of the earth. With the kymograph, these were measured instead against the uniform rotation of the smoked drum.

There was no need for the investigators of that time to worry about whether the tracing mechanism was introducing extra delays of its own, a question that nowadays is often of considerable importance when one is dealing with recording instruments.

Following Von Helmholtz' findings concerning the velocity of propagation of stimuli in nerves, attention was soon given to a composite but much more important quantity in practice, namely *reaction time*. This was introduced by Hirsch under the name of "physiological time" and defined as the time interval between the instant at which the subject receives the stimulus and the instant at which he indicates by a signal that he has perceived the stimulus [5]). In the terms of Descartes' diagram reproduced at fig. 2, the reaction time is the time taken up by the complete chain of events from the first appearance of the object to the completion of the hand movement (we shall disregard the pointing of the finger to a definite part of the object; this involves, additionally, a much more complicated control process that is dealt with in one of the articles in this issue [6])).

Table I shows some reaction times found by various experimenters for various sense organs. It will be seen that the sense of sight is the slowest to react, at least when the stimuli are fairly strong. For stimuli so weak as to be only just perceptible, all the sensory organs have longer reaction times, and there is scarcely any difference between these.

Table I. Reaction times of various sensory organs in seconds, as measured by various investigators (figures indicating responses to stimuli of threshold strength are averages of the results obtained by a number of investigators).

	Strong stimuli			Threshold stimuli (mean values)
	Donders	Wundt	Von Kries	
Hearing	0.180	0.167	0.120	0.337 ± 0.050
Sight	0.200	0.222	0.193	0.331 ± 0.057
Touch	0.182	0.201	0.171	0.327 ± 0.032

We now come to experiments performed by Donders. Once it had become known that stimulus propagation through nerves and human reactions did not take place "with lightning speed", Donders asked himself the following questions [7]):

Was it possible that thought did not have the infinite speed commonly attributed to it, and would

5) A. Hirsch, Expériences chronoscopiques sur la vitesse des différentes sensations et de la transmission nerveuse, Bull. Soc. Sciences Naturelles Neuchâtel 6 (1861-1864), pp. 100-114. Hirsch was quite explicit about the composite nature of the interval "... that might be called the physiological time for the various senses — hearing, sight and touch. This time comprises three elements which are extremely difficult, if not impossible, to separate, namely (1) the transmission of a sensation to the brain, (2) the action of the brain in, as it were, converting the sensation into an act of will, and (3) the transmission of the will to the motory nerve and the performance of the relevant movement by the muscles".

6) J. M. Westhoff, In search of a measure of perceptual work, Philips tech. Rev. **25**, 56-64, 1963/64.

7) F. C. Donders, Über die Schnelligkeit psychischer Prozesse, Pfl. Arch. Anatomie u. Physiologie 1868, 657-681. — A few years earlier, in 1865, a pupil of Donders, named J. J. de Jaager, had been granted a doctorate for work on the same subject, his thesis (in Dutch) having the title "The physiological time associated with psychic processes". Donders later felt obliged to point out that the work was based on *his* ideas and carried out under *his* guidance.

it be feasible to determine the time taken to conceive an idea or arrive at a decision?

These are questions that directly seek to probe the mysterious link between points *b* and *c* in Descartes' diagram, a link that Descartes himself, who had set out to explain everything in mechanical terms, was prepared to leave aside as mysterious and unknowable. If Donders formulated the questions, it was naturally because he thought there was a chance of answering them. He proposed to do this by expanding the reaction-time experiments. If, he reasoned, a certain though unknown part of the reaction time was taken up by an "actual psychic process", then one could try "to insert new terms relating to psychic function into the process corresponding to the physiological time. I judged that if I investigated the resulting increase in physiological time I could ascertain the duration of the new term".

The new term that Donders proposed to insert concerned the taking of a *decision* as to the signals to be given during the experiment. Suppose the subject receives an electrical stimulus in his right foot and has to report it by pressing a button under his right hand; this would be a reaction time measurement of the kind that was then already familiar, and Donders called it an *a*-reaction. Exactly the same reaction could be measured if a button were pressed by the left hand in response to stimulation of the left foot. But now suppose that the subject has the two buttons to choose from, and that he has to press one or the other to report stimulation of this left or right foot, as the case may be. In a series of experiments of this kind (which Donders called *b*-reaction experiments) the reaction time proved on average to be 1/15th of a second longer than that for the *a*-reaction. The inserted term relating to psychic function, i.e. deciding which button had to be pressed, therefore required a time of 1/15 s.

What we have here referred to as a "decision" was in fact already a composite function; Donders himself was quite aware of that. First it was necessary to distinguish between the received stimuli (was it the right or the left foot that was being stimulated?) and then a choice had to be made between the two push-buttons, the right-hand and the left-hand one. In a further series of experiments Donders tried to separate the two functions, using a different set of stimuli and responses. The tester was now required to call out one of the five syllables ka, ke, ki, ko and ku. In one test run, the subject had to react by repeating the syllable that had been called out (*b*-reaction). Then, in the next run, the subject had to react by repeating "ki" when that particular syllable was called out, and only then. Donders referred to this as a *c*-reaction. He argued that here the subject had to exercise the function of *distinguishing*, but there was no need for him to *select*. He further argued that by also measuring the *a*-reaction, i.e. by arranging for the tester to call out "ki" and the subject to react by repeating it, none of the other syllables being called, it would be possible to determine both the time taken to distinguish between stimuli and the time taken to select the correct response, the first being given by the difference between the reaction times for *c* and *a*, and the second by the difference between the reaction times for *b* and *c*.

Fig. 4 shows the apparatus with which Donders carried out these experiments and which he called a "noematachograph" [8]). This "thought-speed tracer" is still preserved in the Physiological Laboratory of the University of Utrecht. It contains a rotating smoked drum, as did Ludwig's kymograph (but here the drum is mounted horizontally on a threaded axle in order to record along a helical track many times longer than the circumference of the drum), and a diaphragm with a stylus attached to it. Vibrations were set up in the diaphragm by the voice of the tester and that of the subject, and were traced on the drum; only the instant at which vibration started — not the shape of the trace — was of interest. An assistant turned the drum by hand, and a vibrating tuning fork fitted with a stylus produced the required reference trace on the drum.

Donders found average reaction-time values of 197, 285 and 243 ms for reactions *a*, *b* and *c* respectively. It thus appeared that 46 ms was required for distinguishing between the stimuli and 42 ms for the actual choice of response.

What was the significance of this new kind of information about psychic functions? In his own words, Donders was no better able than before to say what thought is, but at least he had measured the time it takes. In his opinion it was only the fact that this was possible that justified going a step beyond the long-existing assumption of a general connection between psychic processes and brain function, and posing questions about individual mental processes. The poor accuracy of the measured times scarcely affected the general correctness of this argument, nor did doubts as to whether the element of choice was really eliminated in the *c*-reaction, and whether times taken for the individual psychic processes can really simply be added together — doubts

[8]) F. C. Donders, Twee werktuigen tot bepaling van den tijd, voor psychische processen benoodigd, Ned. Arch. Geneesen Natuurkunde **3**, 105-109, 1868.

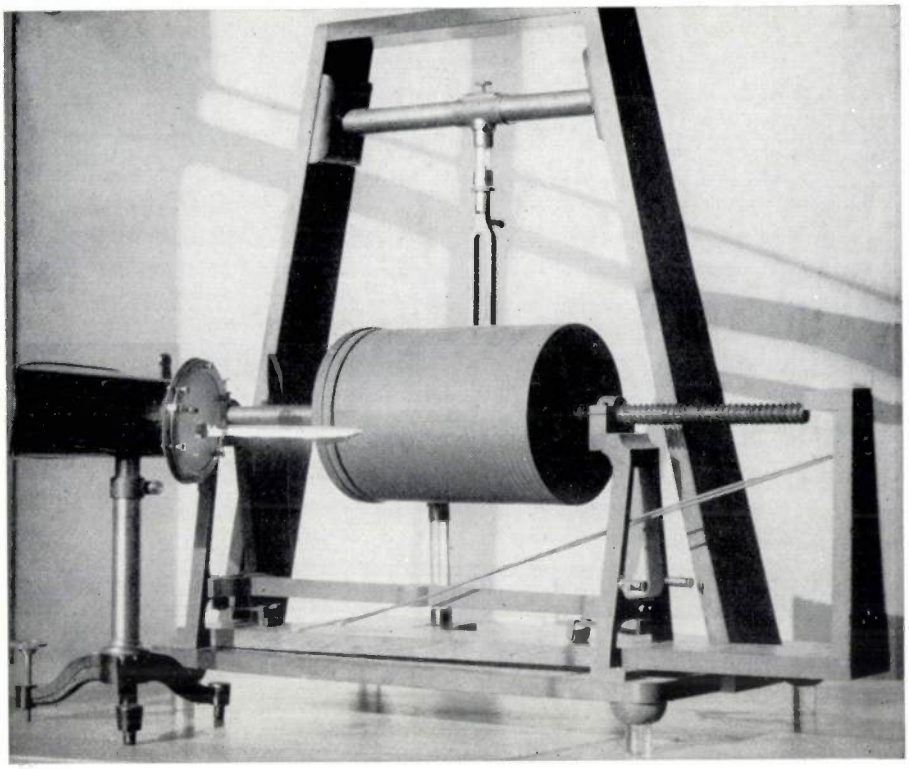


Fig. 4. Donders' noematachograph, which is still preserved in the Physiological Laboratory at Utrecht (photograph reproduced by permission of Prof. Jongbloed, director of that laboratory). In Donders' publication [8]) the apparatus is described in the following words.

"The noematachograph consists of a cylinder, in many respects resembling that of the phonautograph, on which a vibrating tuning fork records the passage of time. Next to this trace is recorded, firstly, the instant at which a stimulus is received, and secondly, the instant at which a signal is given reporting perception of the stimulus.

"Various kinds of stimulus can be employed: the shock produced by breaking the circuit of an induction coil, the making or breaking of a circuit carrying a steady current, a spark or a more intense light flash, transparent letters made visible by a spark behind them, a sound, either originating from a reed struck by a rod projecting beside the cylinder, or from a tuning fork suddenly brought into vibration by a special device, these vibrations being directly recorded, or finally from the human voice, or any other sound, recorded by means of a phonautograph, or better still by a simplified arrangement consisting of a modified König's stethoscope with an elastic membrane stretched over it, this instrument being connected by rubber tubing to two mouthpieces.

"Reactions to the stimulus by various signals is possible:

"*a*. A lever called a key can be depressed to complete the circuit of an electromagnet which brings an armature into motion (a less suitable method because the delay in the response of the magnet may be variable).

"*b*. A tuning fork can be struck, or in certain tests in which the subject necessarily has to distinguish between a large number of stimuli, he can give a vocal signal ....."

which long continued to govern the line of further enquiry pursued by Donders' colleagues. Investigators indeed have gone on posing questions about individual mental processes — as witness many pages in the present issue.

Summary. Well into the 19th century, scientific research on perception was mainly concerned with the *mechanism* of sensory perception. The connection between perception and action was either not enquired into or regarded as an inscrutable secret (Descartes). It was thought that perception processes took place at infinite speed. However, round about the beginning of last century astronomers had recognized the existence of finite reaction times as a very real (and troublesome) phenomenon. In 1850 Von Helmholtz measured the finite speed with which stimuli travelled through nerve fibres. Hirsch introduced the concept of reaction time. Then, about 1865, Donders suggested — and demonstrated — that it was possible, by carrying out reaction time measurements, to determine at least approximately the duration of individual mental processes, such as discrimination and selection, that enter into the connection between perception and action. The "noematachograph", the apparatus used by Donders in these measurements, is described in this article.

VOLUME 25, 1963/64, No. 4 pp. 81-104 Published 24 th March 1964

Philips Technical Review

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

Ir. D. M. Duinker, deputy editor-in-chief of this journal, has now retired. Ir. Duinker joined the company in 1927, working initially in the Research Laboratories in the field of welding and other electrotechnical apparatus. The very first number of Philips Technical Review (Vol. 1, No. 1, p. 11) contained an article by him on "Relay valves as timing devices in seam-welding practice". In 1946 he joined our editorial staff, on which he served for 17 years. He was a very important member, not only through his professional abilities and his energy but also because of his special interest in linguistics. He was keenly aware of the responsibility which rests upon the editors of a technical journal to promote the proper use of language. Of the publications in this journal he was mainly concerned with those dealing with transmitters and transmitting valves, measuring instruments, television and special electronic circuits. Through his gift for exact and lucid exposition he has rendered valuable services to those interested in enlarging their knowledge of these subjects. The editors are sorry to lose him.

AN IMPLOSION-PROOF PICTURE TUBE FOR TELEVISION

by F. de BOER *), P. CIRKEL *), W. F. NIENHUIS *) and C. J. W. PANIS *).

621.397.331.24

In recent decades picture tubes for television have been steadily improved, both in picture quality and design. The article below describes an innovation in envelope design, which makes it possible to dispense with the protective window normally fitted in front of the tube. This offers several advantages.

Television sets normally have a protective screen fitted in front of the picture tube. The atmospheric pressure produces high stresses in the evacuated glass envelope of the picture tube — the bulb. Although of course the bulb is designed so that it can amply withstand these stresses, it may happen in very rare cases that, as a result of external damage, a crack forms which leads to *implosion* of the tube. By implosion is meant the sudden collapse of an evacuated vessel under the influence of the atmospheric pressure, the initial crack propagating and branching out over the whole surface of the bulb before the difference in air pressure has been compensated through the openings so produced. The bulb shatters into fragments which are forced inwards through the pressure difference and then pass on outwards. A crack in the glass that may lead to implosion can be caused by glass fatigue in a superficially damaged part of the bulb, or by a severe knock against the bulb. The first case is referred to as spontaneous implosion.

In recent years more insight has been gained into the phenomenon of *glass fatigue*. It had long been known that the strength of glass is determined to a large extent by the state of the surface. If any part of the surface that is under tensile stress is scratched or otherwise damaged, and if moreover this glass is exposed to moisture and temperature variations, there is a chance that the scratches will gradually

*) Philips Electron Tubes Division, Eindhoven.

deepen, as a result of which the glass is further weakened or becomes "fatigued". In view of the combination of damage and the effects of moisture, this fatigue in the case of an evacuated bulb can only arise in the outside surface, and only if the glass there is under tensile stress. Due to this process it is possible that the strength of the bulb will decrease locally to such an extent as to lead finally to a spontaneous fracture that may result in implosion. In the design and manufacture of the tubes thorough safety factors are of course applied against such a contingency, and in practice a spontaneous implosion is extremely rare.

A *heavy knock* against the bulb can momentarily increase the tensile stresses sufficiently to cause a crack which may be the beginning of an implosion. This case is more difficult to establish because the stresses now are caused not only by the atmospheric pressure but also by the shock waves set up in the bulb.

To protect viewers against the effects of implosion, and to prevent damage to the tube that might lead to an implosion, it has hitherto been the practice to fit a protective window in the television set. In its original form, which is still widely employed, the window is a flat or curved glass plate fitted in the cabinet at a small distance from the face of the tube. This method, however, introduces a space between tube and window where dust can penetrate which is difficult to remove, while picture contrast is reduced by light reflection from the two added interfaces of glass and air. These drawbacks are not found in later developed forms, where the protective window is curved and bonded to the tube face with transparent resin having the same refractive index as the glass.

Drawbacks which all forms of protective window have in common are that they increase the weight of the set and shift the centre of gravity towards the front; the latter is not conducive to the stability of the set, which may be a particularly important point where the cabinet is shallow. Further, a limitation of the protective window is that it is useless to the service engineer working at the back of the set, and also to the person handling the tube during production, while of course it cannot protect the interior of the set itself from damage.

It is evident from the foregoing that a tube which is *intrinsically implosion-proof*, and which therefore makes all protective measures superfluous, would be very attractive. Development research carried out by us along these lines has indeed led to the realization of an implosion-proof picture tube. A description of this work will now be given.

Closer examination of the implosion process

To gain insight into the implosion process it is necessary to know the state of stress in the evacuated bulb. We shall only consider those stresses which are due to the atmospheric pressure, for these are by far the greatest. The strength of the bulb is governed mainly by the tensile stresses, particularly those at the outside surface where fatigue can occur. The stresses in the inside and outside surface of the bulb were already known from earlier measurements with strain gauges; they were found to be greatest in the two planes of symmetry through the axis of the tube. We shall now examine the stresses for one of these planes, *w* in *fig. 1*.

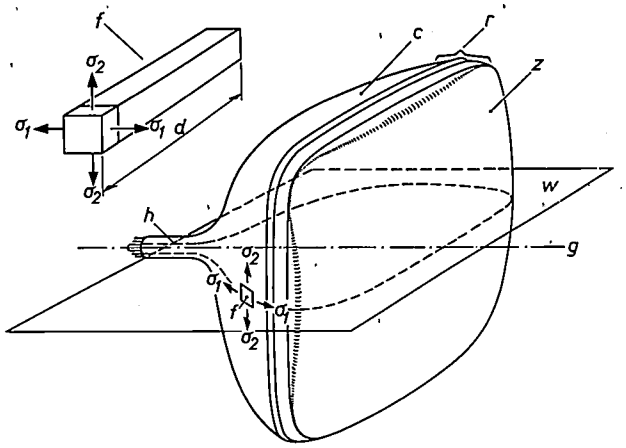


Fig. 1. The strength of an evacuated bulb exposed to atmospheric pressure is chiefly determined by the places where the tangential normal stresses at the outside surface are positive, and thus exercise a tensile force. σ_1 is here the component of the tangential normal stresses in the plane through a given point and the tube axis *g*, such as the plane of symmetry *w* drawn here, and σ_2 is the component perpendicular to it. In the figure σ_1 and σ_2 are represented for a small wall element *f* as stresses at the outer surface. Referring to the bulb, *z* is the face plate, *c* the cone, *h* the neck, *r* the rim and *d* the wall thickness.

For the purpose of analysing the stresses at points in the plane *w*, consider a small wall element, the rectangular parallelepiped *f*, shown enlarged in the inset of fig. 1. For determining the forces that constitute the heaviest load on the glass, the state of this element is sufficiently described by two normal stresses: σ_1 in the plane *w* and σ_2 perpendicular to it. Their magnitude need not be constant over the thickness of the wall; if the glass is in flexure the stress across the wall even changes sign.

In *fig. 2*, which shows the cross-section of the bulb in the plane of symmetry *w*, the principal results of the measurements are presented graphically. For clarity the stresses σ_1 and σ_2 are drawn at different sides of the bulb. The periphery of the bulb is used as the abscissa: the magnitude of these stresses at each point is plotted perpendicular to the periphery, outwards for positive values (tensile) and

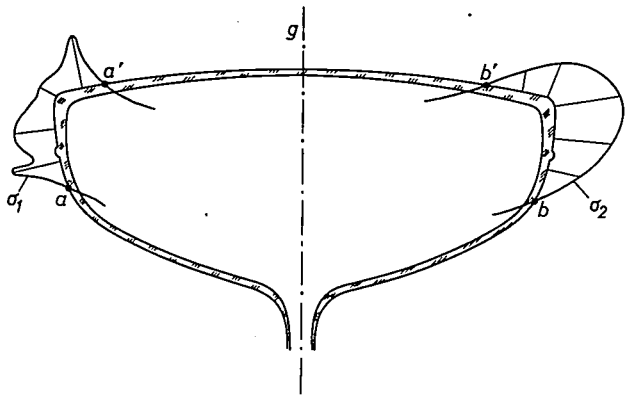


Fig. 2. Bulb cross-section in the plane of symmetry w showing the stresses σ_1 and σ_2 acting in the outer surface, which are represented graphically with the periphery as abscissa. Positive values, plotted outwards, indicate tensile stresses, and negative values, plotted inwards, are compressive stresses. For clarity σ_1 and σ_2 are set out at opposite sides of the bulb. Tensile stresses are seen to occur in the region around the rim of the bulb, between a and a' and b and b'. The compressive stresses are only partially represented.

inwards for negative values (compressive). The graph relates only to stresses on the outside of the bulb. For σ_2 a roughly similar state of stress holds on the inside. For σ_1 the situation is different: between a and a' the glass in the corresponding direction is in flexure, such that a tensile stress prevails on the outside and a compressive stress on the inside. Qualitatively this whole stress pattern also applies in other planes through the axis of the tube, the stress values there being usually smaller than in the planes of symmetry. The greater part of the bulb is thus under compressive stress over the entire wall thickness, and presents few problems as far as strength is concerned; but in the region of the transition between face plate and cone (a to a' or b to b' in fig. 2), the region near the rim or maximum periphery of the tube, tensile stresses prevail [1]). This tensile stress region, or peripheral region, deserves special attention.

As mentioned at the beginning, it is a characteristic of implosion that cracks form over the entire surface of the bulb, and that this happens at such great speed that the resultant fragments experience the full force of the uncompensated difference in air pressure. These two characteristics are bound up with the existence of the tensile stress region at the critical location between face plate and cone. We represent the process as follows. If, through any cause, a crack appears in the peripheral region, the tensile stresses there will widen it rapidly into a fissure. The widening takes place mainly in the direction of the stress σ_2, since the latter is positive over the entire wall thickness and moreover is normally greater on the outside surface than the stress σ_1. As a result the crack also becomes longer and branches out into the adjacent areas — to the face and the cone. The cohesion of the pieces — which initially held together so that the vacuum was largely preserved — is then destroyed. The atmospheric pressure sets the resultant fragments into motion inwards. As a consequence of the kinetic energy thereby imparted to them, the fragments, if they are not stopped by collision on their way, pass on outwards. This assumption has been confirmed by high-speed cine recordings of the implosion process: the cracks are seen to spread over the entire bulb within a few milliseconds before the outward shape is lost and the bulb collapses.

[1]) Tensile stresses are also present in a small region near the transition between cone and neck, but these need not be considered here since a crack in this place never gives rise to implosion.

Prevention of implosion

From the foregoing it is obvious that it should be possible to avoid implosion by simply preventing the above-mentioned widening and propagation of cracks originating in the tensile stress region. To this end we have adopted the method of applying around the peripheral region of the bulb a reinforcement layer, or *band*, which opposes all deformation and expansion of the periphery and thus checks the widening of a crack and stops it spreading (see *fig. 3*). In the normal condition the band is free of stress, and only in the event of glass breakage does it exert forces on the bulb. This method was in fact found to be effective.

A band of this kind should meet the following requirements. Its tensile strength should of course be great enough to rule out any likelihood of the band breaking. Further, the *elongation* resulting from the force to which the band is subjected in the event of glass breakage should be extremely small, the widening of a crack in the glass under the band being equal to this elongation. In the materials concerned here the elastic limit in that case is not exceeded. Within the elastic limit the elongation Δl of an elastic body of length l is given by Hooke's law:

$$\Delta l = \frac{Kl}{ES}.$$

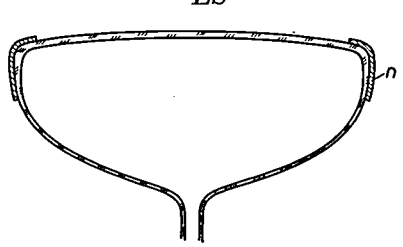


Fig. 3. Prevention of implosion by a reinforcement band n around the tensile stress region of the bulb.

Here E and S are the modulus of elasticity and the cross-section of the band respectively, and K is the tensile force exerted on the band. Increasing the product ES would reduce Δl, but the former may not be desirable owing to the fact that the cross-section is limited by the trend towards tubes of small dimensions and light weight, and a high modulus of elasticity would limit the choice of material.

We shall confine ourselves first to what happens in the direction of the stress σ_2, which is the greater stress and is positive over the entire wall thickness. Where a bulb is reinforced by a band it has been found by experience that a crack that originates in the peripheral region and is perpendicular to σ_2 does not extend much farther than the points where the tensile stress region ends (points b and b' in fig. 2). The force K in this case is the integral of σ_2 over the area of this crack, i.e. of the wall between b and b'. In a 59 cm tube, for example, this force is about 400 kg. For the length l one might at first sight be inclined to put in the entire circumference of the rim of the tube, but in reality l is no greater than the distance between two adjacent corner points (in the case of a 59 cm tube a maximum of 50 cm). This may be understood as follows. Suppose that a crack occurs as indicated in *fig. 4*, and that the force tending to widen it is K. The deformation forces in the band and the frictional forces between the band and the glass are in this case much greater at the corners *1* and *2* than elsewhere: consequently most of the force K is transmitted to the band near these corners. The length l over which the force acts is thus roughly the distance between *1* and *2*. This length can be appreciably reduced, however, by providing good *adhesion* between the glass and the band, adhesion causing the force to act in the immediate region around the crack. In this way we succeeded in reducing l to about 2 cm [2]).

As regards the strength in the direction of σ_1, where the bulb is in flexure, demands are also made on the *stiffness* of the band, for if a crack occurs perpendicularly to σ_1 the band is also required to limit the bending deformation of the bulb. The adhesion between glass and band is here too an important positive factor, since it substantially increases the effect of this stiffness. The influence of stiffness and adhesion in this connection is difficult to express in figures, but it was unmistakable in our experiments.

The total effect of the band in the prevention of implosion is a combination of the three factors mentioned: the product ES, the adhesion and the stiffness. In determining the share of each of these factors there is a certain margin of freedom. By increasing the product ES and the stiffness, for example, one might even dispense with the adhesion.

In the foregoing we have assumed that the crack originates in the peripheral region. The same reasoning also applies, however, if the crack originates elsewhere and penetrates into this region, and it makes little difference whether the real cause is a spontaneous or a forcible fracture.

Construction

As we have seen, the band has to be applied to the region of tensile stress near the rim of the tube. Part of this region, the area of tensile stress in the face plate, is not suitable for covering, for the result would be a reduction of the size of the useful screen area. The question was therefore to what extent the rest of the region had to be reinforced. To answer this question it was necessary to resort to *experiment*.

In the initial experiments, at the beginning of

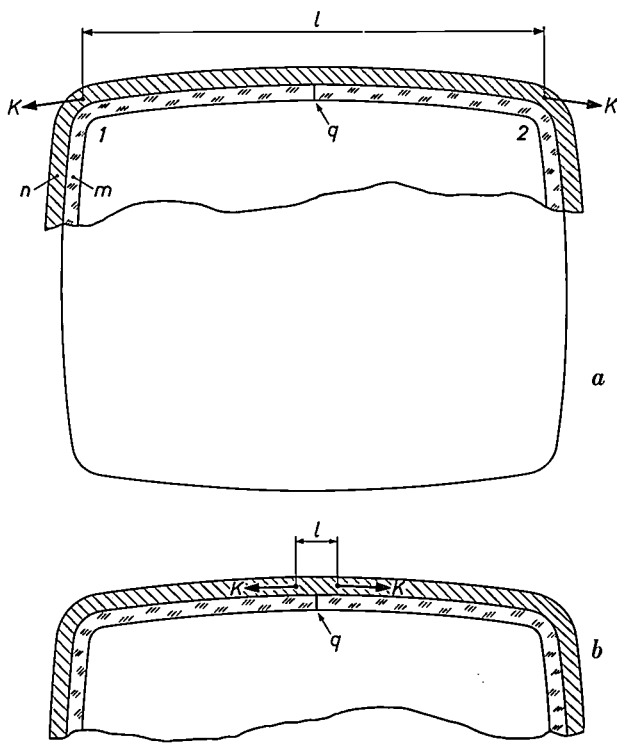


Fig. 4. The peripheral area of a bulb with reinforcement band, in a cross-section perpendicular to the tube axis, illustrating the effect of adhesion on the extent to which the band n limits the widening of a crack q in the glass m. K is the total force acting on the band when the crack is forced open by the stresses σ_2.
a) Without adhesion between bulb and band the force K is largely transmitted to the band at the corner points *1* and *2*, owing to the frictional and deformation forces occurring at those points.
b) With adhesion the effective length l over which the total force acts is only a few centimetres, so that the widening of the crack, which is equal to the elongation of the band over the distance l, is considerably smaller.

[2]) In the case of adhesion the force acts continuously and not at two more or less discrete points. A calculation, taking into account the modulus of elasticity of the adhesion layer, has shown that the continuous action is here equivalent to an effective action at two discrete points roughly 1 cm on either side of the crack.

1960, we chose as the material for the band a polyester resin reinforced with glass fibre [3]). We chose this for the following reasons. Firstly, this material combines high strength and low weight; secondly, it can be applied to a bulb by simple means in any desired form; and thirdly, the mechanical properties of a reinforced layer made from this material can be locally varied simply by varying the layer thickness and the glass-fibre content. For bonding this resin to the bulb a polyvinyl acetate adhesive was used. The manner in which the tubes thus reinforced were tested is described in the next section.

It indeed proved possible to achieve our aim completely without covering any part of the face plate: a strong thick layer on the rim and a thin layer on the entire cone sufficed (*fig. 5a*).

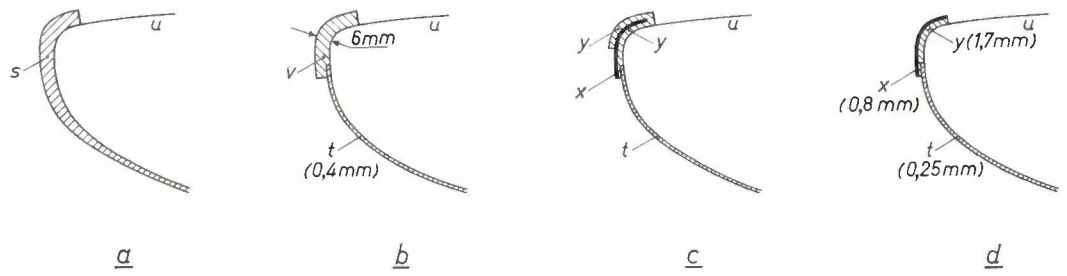


Fig. 5. Different versions of the reinforcement band, in a partial cross-section. The line *u* represents the periphery of the bulb.
a) Experimental version: polyester resin reinforced with glass fibre, *s*.
b) Preform version: the band is in two parts. The part *v* around the rim consists of resin reinforced with a preformed ring of glass fibre. The part *t* around the cone consists of resin reinforced with glass-fibre cloth.
c) Later version: the rim is surrounded by a metal strip *x* embedded in resin *y*; the cone is reinforced as under *b*).
d) Latest version: around the rim is a metal band *x*, with an inside filling of resin *y*; cone reinforced again as under (*b*) and (*c*).

For the purpose of *quantity production* we initially adopted a method in which the rim was reinforced with a preformed glass-fibre ring. The bulb was placed face plate downwards in a jig, leaving a gap between the rim and the jig. The ring was then placed in this intervening space, and the space was further filled with polyester resin. The cone was covered by applying to it a thin layer of glass-fibre cloth onto which the resin was sprayed. Finally a metal strip was fitted around the peripheral band for the purpose of locking in position the lugs by which the tube is secured in the cabinet, as in the case of conventional tubes. The resultant "preform" tube is represented in fig. 5*b*.

In another version we replaced the glass-fibre ring by a metal band, which also locked the fixing lugs in position (fig. 5*c*). This was an improvement in some respects compared with the preform tube: the dimensions were smaller, the elongation was reduced, the outer metal strip was dispensed with and production was simplified.

The *latest version* — in which the tube has meanwhile gone into mass production — is a variant on the latter one, with even smaller dimensions and lower weight. Here the metal band itself serves as casting jig, and thus permanently constitutes at the rim of the bulb the outside of the reinforcement band. Only the small space between the rim and the band is filled with resin. The cone is again covered with the combination of resin and glass-fibre cloth (fig. 5*d*). The weight of this tube is roughly 20% less than that of a conventional tube, including its attachment pieces and the protective window; for a 59 cm tube this means a reduction of 3.4 kg. In a 59 cm tube in this version a rough calculation for the case of the crack perpendicular to σ_2 shows that the crack widens to no more than about 0.01 mm.

The experimental version, the preform type and the latest version are shown in *figs 6*, *7* and *8*.

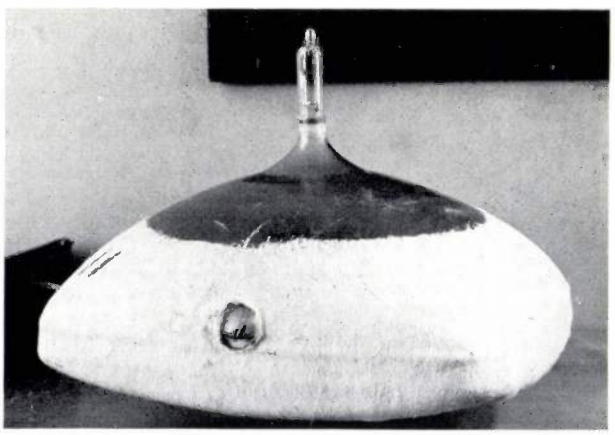
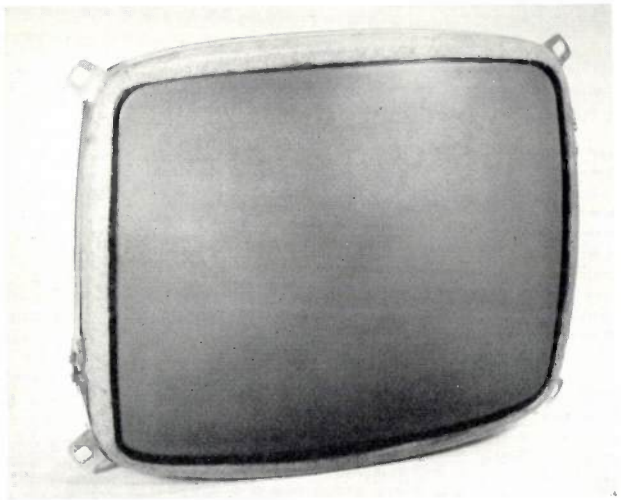


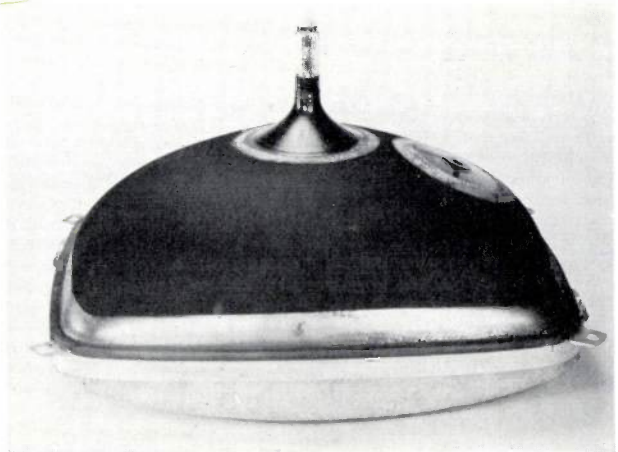
Fig. 6. Bulb with experimental band of polyester resin reinforced with glass fibre.

[3]) We made use of the experience gained with this material in the synthetic materials laboratory of Philips Allied Industries Division. The unsaturated polyester resin which we employ is hardened with the aid of catalysts and accelerators.

Dispensing with the protective window can have an optical consequence on the face plate of the tube. As will be known, some ambient lighting is desirable when looking at a television picture. Some of this outside light will fall on the fluorescent screen, however, and reduce the picture contrast, which of course is not the intention. To limit this effect it is usual to introduce a grey absorption filter in front of the fluorescent screen. Although this also attenuates the light emitted by the screen itself, the light from the outside has to pass through the filter twice (before and after reflection from the fluorescent screen) and is thus more strongly attenuated. This absorption can take place either in the protective window or in the face plate, or in both. Since the absorbent material is homogeneously distributed in the glass, to obtain a uniform filtering action the glass must be of uniform thickness. In the implosion-proof tube the absorption takes place entirely in the face plate. The thickness of the glass therefore must not differ much in the middle from that near the edge.

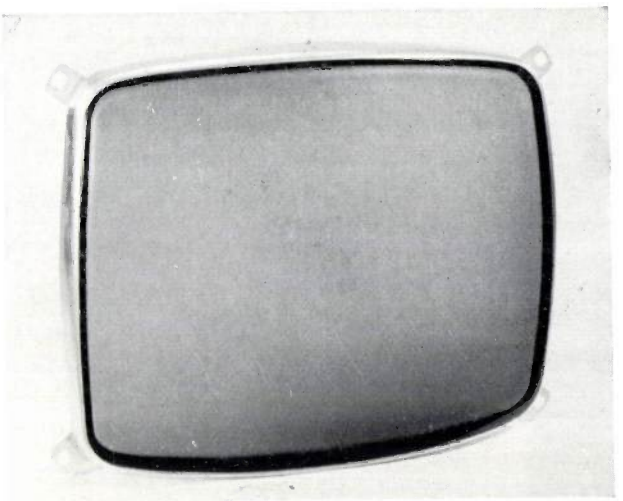


a

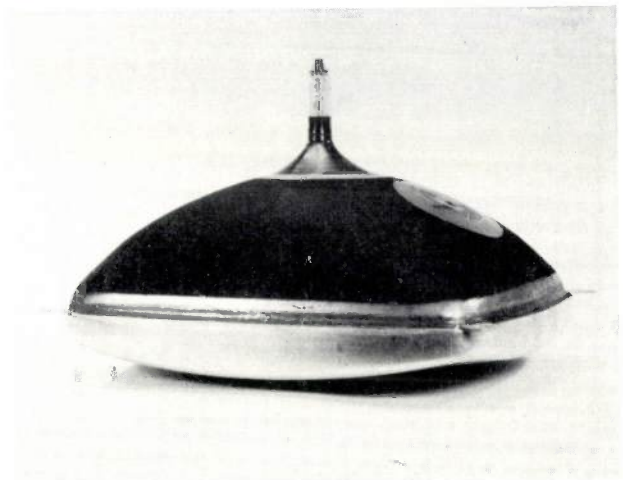


b

Fig. 7. Preform version.
a) Front view. The rim is surrounded by polyester resin reinforced by a preformed glass-fibre ring. The attachment lugs are fixed to the tube by the metal strip subsequently fitted.
b) Rear view. The band around the cone, resin reinforced with glass-fibre cloth, is largely concealed by the outside layer of graphite which, together with a conductive layer on the inside of the cone, forms the EHT smoothing capacitor.



a



b

Fig. 8. The implosion-proof tube now in mass productione Around the rim is a metal band filled with resin; around thscone, under the graphite layer, is resin reinforced with glass-fibre cloth.

Testing the new tube

The reinforced bulbs were tested experimentally in various ways: by simulated spontaneous fracture, by external violence and by fire; the experiments were repeated after artificial ageing. The aim was to investigate whether the safety achieved justified the

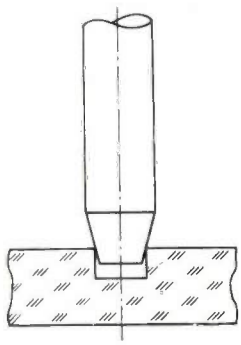


Fig. 9. Simulating the spontaneous fracture process by introducing small surface cracks in the tensile stress region. In the method shown here this is done by tapping lightly with a hammer on a tapered pin, placed in a blind hole drilled into the glass.

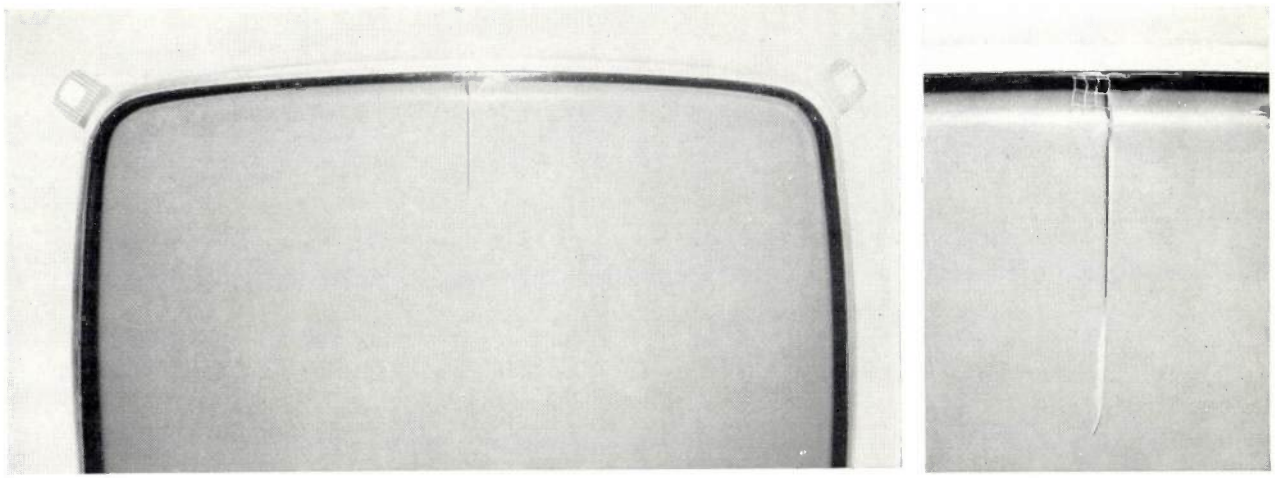


Fig. 10. The new tube with a "spontaneous" crack, simulated by producing surface abrasions near the rim (visible in the detail photograph on the right) and cooling this area rapidly with liquid nitrogen. The plane of fracture is usually slightly twisted. The fissure produced is so slight that it takes about 15 minutes before the air pressure inside the bulb rises to half an atmosphere.

omission of the protective window in all conceivable circumstances, and whether it was in fact also effective for the maintenance engineer and for the set-maker. Many thousands of tubes were subjected to these experiments. The results mentioned below relate to the latest version.

For simulating the *spontaneous* process we introduced small surface cracks in the tensile stress region, using as little energy as possible. Two methods were used. In the first method (punch test), we drilled into the glass a small blind hole, into which a tapered pin was inserted, the pin then being tapped lightly with a hammer (see *fig. 9*). In the second method (thermal-shock test), small surface abrasions were introduced in the glass, and the damaged place was rapidly cooled with liquid nitrogen, so that the resultant thermal stresses deepened the abrasions into cracks. To apply this method to parts under the reinforcement band it was of course necessary to remove a small piece of the band first.

Wherever this is done, it causes no implosion; the cracks remain small in size and number, and the cohesion in the bulb is maintained. *Fig. 10* shows a typical result. There is only one crack, perpendicular to the stresses σ_2.

When there is no protective window there is of course the risk of a direct *knock against the front* by a heavy object. The thick face plate proves to be perfectly capable of withstanding knocks from household objects (brush, broom, vacuum cleaner, table, etc.) as may occur in practice. A greater impact energy, as for example a blow from a hammer or the impact of a thrown bottle or steel ball (which are less likely occurrences in the living room) succeeds in damaging the glass but does not cause implosion. Cracks and perforations may appear in the face plate, but the cohesion is preserved. In some cases glass splinters of 0.1 to 1 gramme may fly from the tube to a distance up to 3 feet, but that is inherent in the normal breakage of glass objects and cannot be attributed to any implosive effect. To obtain more reproducible results we adopted a pendulum method of the kind long used for testing the protective windows of television sets. In this method a steel ball attached to a pendulum is released from a certain height, so that the glass is struck at a required point with a defined impact energy. *Fig. 11* shows a frontal perforation produced in this way. The fact that the cohesion of the glass

Fig. 11. Testing the new tube by the impact of a heavy steel ball on the tensile stress region of the bulb. The result is only a small hole and a few cracks in the face plate. Omission of the protective window is thus entirely warranted. The rapid ingress of air has blown away part of the fluorescent layer, as the photograph shows.

is still preserved, even when cracks are produced all over the face plate by gross mishandling, is attributable to the convex form of the face plate, to the presence of the reinforcement band which prevents expansion of the rim, and to the circumstance that the plane of fracture always shows some twisting as a result of which the broken parts interlock. Consequently the sections, under the influence of the atmospheric pressure, remain in place during the breakage process.

The effect of violence on the *rear of the bulb* was also investigated. A sharp blow with a hammer on the cone or the neck causes the glass to break at the point of impact but does not lead to implosion. We also investigated what happens if, by outside force or spontaneous fracture, the neck together with the electrode system breaks loose and is hurled inwards by atmospheric pressure. We did this by giving a blow to a steel pipe placed concentrically around the neck. The result: superficial damage to the inside of the face plate, at the most a few cracks in it, but not a single glass splinter ejected.

Finally, the tubes were subjected to *heating followed by abrupt cooling*. This was done for the very exceptional case where the apparatus is situated in a burning room and the fire is extinguished with water. For this test we set fire to a television receiver with petrol, and some time later put the fire out with water. The result — again no implosion, only a few cracks in the bulb.

The conclusion of these extensive tests was that implosion has indeed been overcome, even under the most abnormal mishandling of the tube, so that the omission of the protective window is entirely warranted and all the envisaged advantages have been achieved. It seems likely that the exploitation of these advantages will lead to entirely new cabinet designs: as a result of the exceptional strength of the tube, the simplified mounting in the set with fixing lugs which are an integral part of the tube, the omission of the protective window, the reduction of weight and the favourable displacement of the centre of gravity, it will be possible to make the cabinet smaller, shallower and lighter in construction, and even to let the front of the tube project from the cabinet.

Summary. With conventional picture tubes the chance of implosion, although extremely remote, nevertheless exists, and for this reason a protective window is usually fitted in front of the tube. The article describes investigations aimed at *preventing* implosion. The possibility of implosion is attributable to the atmospheric pressure producing tensile stresses around the rim of the bulb, the region between face-plate and cone. The object was achieved by applying a reinforcement band around this region. This band, normally free from stresses, limits the cracks once they have formed, inhibits expansion of the bulb edge, and thus preserves cohesion in the bulb in the event of the glass fracturing. In a version that has now been put into mass production the reinforcement consists of a metal band and polyester resin. Tests show hat the method is entirely effective. The principal advantages are: the protective window is dispensed with, assembly in the cabinet is simpler, the weight and dimensions of the television set are reduced, and new cabinet designs are made possible.

AN EXPERIMENTAL IMAGE-INTENSIFIER TUBE WITH ELECTROSTATIC "ZOOM" OPTICS

by A. W. WOODHEAD *), D. G. TAYLOR *) and P. SCHAGEN *).

621.383.811

The long known optical "zoom" lens has become very important in the last few years. Attempts to create an electron-optical analogue — a system providing electron-optical image-formation with variable magnification — started a decade ago. A very promising result has now been obtained using an electrostatic system based on the concentric-spheres principle.

A system for vision at very low light levels

The special image intensifier tube described in this paper was developed as part of a visual aid for use at extremely low light levels. Such aids have obvious applications e.g. for navigation at sea during dark nights. An experimental version of the complete system, designed in the Mullard Research Laboratories at Salfords, is shown in *fig. 1* and a schematic diagram is given in *fig. 2* [1]). A mirror type

*) Mullard Research Laboratories, Salfords (Surrey), England.

[1]) A proposal for this system was described two years ago: P. Schagen, D. G. Taylor and A. W. Woodhead, An image intensifier system for direct observation at very low light levels, 2nd Symposium on photo-electronic image devices, London, Sept. 1961, published in Advances in Electronics and Electron Physics **16**, 75-83, 1962. Details of the pertaining optical system were published l.c., page 85-89: A. Bouwers, Low brightness photography by image intensification.

is still preserved, even when cracks are produced all over the face plate by gross mishandling, is attributable to the convex form of the face plate, to the presence of the reinforcement band which prevents expansion of the rim, and to the circumstance that the plane of fracture always shows some twisting as a result of which the broken parts interlock. Consequently the sections, under the influence of the atmospheric pressure, remain in place during the breakage process.

The effect of violence on the *rear of the bulb* was also investigated. A sharp blow with a hammer on the cone or the neck causes the glass to break at the point of impact but does not lead to implosion. We also investigated what happens if, by outside force or spontaneous fracture, the neck together with the electrode system breaks loose and is hurled inwards by atmospheric pressure. We did this by giving a blow to a steel pipe placed concentrically around the neck. The result: superficial damage to the inside of the face plate, at the most a few cracks in it, but not a single glass splinter ejected.

Finally, the tubes were subjected to *heating followed by abrupt cooling*. This was done for the very exceptional case where the apparatus is situated in a burning room and the fire is extinguished with water. For this test we set fire to a television receiver with petrol, and some time later put the fire out with water. The result — again no implosion, only a few cracks in the bulb.

The conclusion of these extensive tests was that implosion has indeed been overcome, even under the most abnormal mishandling of the tube, so that the omission of the protective window is entirely warranted and all the envisaged advantages have been achieved. It seems likely that the exploitation of these advantages will lead to entirely new cabinet designs: as a result of the exceptional strength of the tube, the simplified mounting in the set with fixing lugs which are an integral part of the tube, the omission of the protective window, the reduction of weight and the favourable displacement of the centre of gravity, it will be possible to make the cabinet smaller, shallower and lighter in construction, and even to let the front of the tube project from the cabinet.

Summary. With conventional picture tubes the chance of implosion, although extremely remote, nevertheless exists, and for this reason a protective window is usually fitted in front of the tube. The article describes investigations aimed at *preventing* implosion. The possibility of implosion is attributable to the atmospheric pressure producing tensile stresses around the rim of the bulb, the region between face-plate and cone. The object was achieved by applying a reinforcement band around this region. This band, normally free from stresses, limits the cracks once they have formed, inhibits expansion of the bulb edge, and thus preserves cohesion in the bulb in the event of the glass fracturing. In a version that has now been put into mass production the reinforcement consists of a metal band and polyester resin. Tests show hat the method is entirely effective. The principal advantages are: the protective window is dispensed with, assembly in the cabinet is simpler, the weight and dimensions of the television set are reduced, and new cabinet designs are made possible.

AN EXPERIMENTAL IMAGE-INTENSIFIER TUBE WITH ELECTROSTATIC "ZOOM" OPTICS

by A. W. WOODHEAD *), D. G. TAYLOR *) and P. SCHAGEN *).

621.383.811

The long known optical "zoom" lens has become very important in the last few years. Attempts to create an electron-optical analogue — a system providing electron-optical image-formation with variable magnification — started a decade ago. A very promising result has now been obtained using an electrostatic system based on the concentric-spheres principle.

A system for vision at very low light levels

The special image intensifier tube described in this paper was developed as part of a visual aid for use at extremely low light levels. Such aids have obvious applications e.g. for navigation at sea during dark nights. An experimental version of the complete system, designed in the Mullard Research Laboratories at Salfords, is shown in *fig. 1* and a schematic diagram is given in *fig. 2* [1]). A mirror type

*) Mullard Research Laboratories, Salfords (Surrey), England.

[1]) A proposal for this system was described two years ago: P. Schagen, D. G. Taylor and A. W. Woodhead, An image intensifier system for direct observation at very low light levels, 2nd Symposium on photo-electronic image devices, London, Sept. 1961, published in Advances in Electronics and Electron Physics **16**, 75-83, 1962. Details of the pertaining optical system were published l.c., page 85-89: A. Bouwers, Low brightness photography by image intensification.

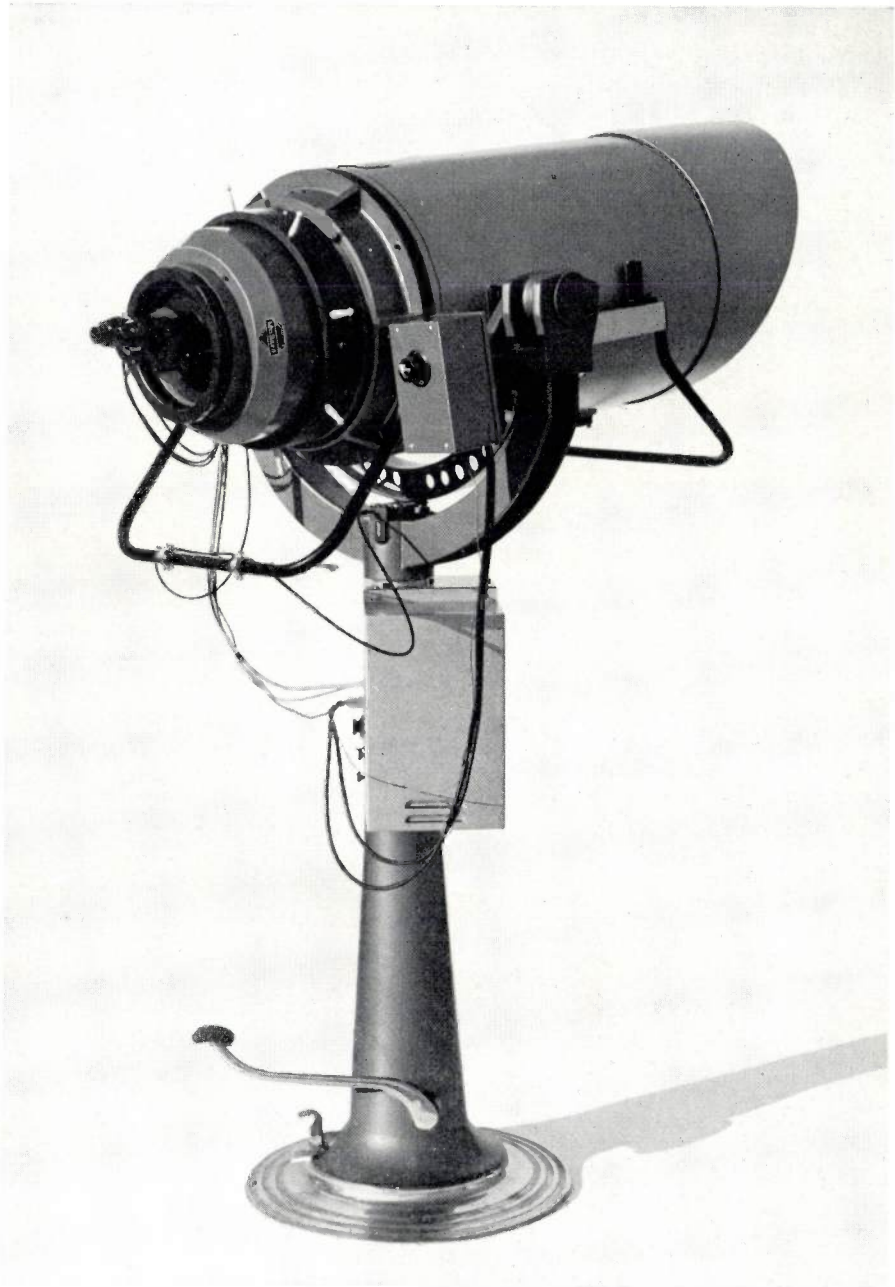


Fig. 1. Complete system for low light level vision developed at Mullard Research Laboratories. The system comprises an image intensifier tube with mirror type optics and a binocular microscope for observing the viewing screen. It is mounted in a clevis ring on a stand easily permitting adjustment of height. On the right side of the instrument is a small box with regulating knob for varying the magnification. A box containing the high voltage supply for the image intensifier is fixed on the stand.

optical system (geometric aperture f : 0.75, effective aperture f : 1.0) focuses an image of the scene to be viewed onto the photocathode of the image intensifier tube. The number of electrons emitted from each unit area of the photocathode of such a tube is proportional to the local illumination. The electron-optical system of the tube accelerates the electrons and focuses them onto the fluorescent viewing screen, where they produce a reduced image of the scene. This is observed through a binocular viewer providing a considerable magnification and possessing a large exit pupil. The brightness of the image observed depends on the acceleration of the electrons (which affects the *lumen gain*) and on the *reduction* in the tube (concentration of the available photons on a small area).

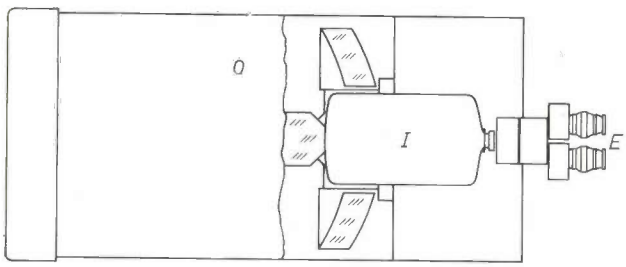


Fig. 2. Schematic diagram of the system. *O* mirror objective of 35 cm focal length and effective relative aperture f : 1.0. *I* image intensifier tube. *E* binocular viewer of magnification 13.5.

Image-intensifier tubes have been described in several articles in this Review. In many cases the principle of intensification is combined with the possibility of wavelength conversion, and very well-known examples of this are the X-ray image intensifier [2]) and the infrared image converter. In tubes of this type, manufactured at present on a fairly large scale, the electron-optical magnification (or reduction, i.e. magnification <1) has so far been constant for a given tube, or nearly so. The image intensifier

[2]) M. C. Teves and T. Tol, Electronic intensification of fluoroscopic images, Philips tech. Rev. **14**, 33-43, 1952/53. — See also a series of articles in this Review, Vol. 17, No. 3, 1955/56 and J. J. C. Hardenberg, An apparatus for cinefluorography with an 11 inch X-ray image intensifier, Philips tech. Rev. **20**, 331-345, 1958/59.

tube described in this article (in which wavelength conversion is only incidental) has the feature that the electron-optical magnification can be very easily and continuously varied over a wide range.

In order to appreciate the need for this facility, consider the improvement in visual perception which an image intensifier system makes possible. This is shown in *fig. 3* for a system which employs a specific combination of optical components. In this figure the smallest angle α which a certain test object with 100% contrast against the background must subtend in order to be detected by the eye is plotted against the brightness level (luminance) L on a double-logarithmic scale. Curve E represents the measured performance [3]) of the unaided eye, after complete adaptation, curves I_1, I_2, I_3 represent the calculated performance of the eye using a particular intensifier system in which the overall angular magnification M (resulting from the objective, the image intensifier tube and the viewer) has been given a number of different values. It is seen that the lowest value, $M = 1$, results in an acuity curve very similar to that of the unaided eye and covering the same range of α but shifted to luminances which are about 1000 times smaller. A higher magnification, $M = 4$, will provide a smaller shift to the left but will result in a simultaneous upward shift. When luminance levels between 10^{-6} and 10^{-8} footlamberts are encountered (1 ftl. $= 3.426 \times 10^{-4}$ stilb), a magnification $M = 4$ clearly will be more useful than $M = 1$, since details subtending smaller angles α are rendered detectable. Similarly $M = 16$ will be more useful for luminances above 10^{-6} ftl. $M = 1$ on the contrary is superior for luminances below 10^{-8} ftl. in order to allow any objects to be detectable at all. By making the overall magnification variable it can therefore be seen that a larger part of the L,α-diagram is made accessible for vision than with a fixed magnification. For the sake of comparison, it may be mentioned that the luminance of a snow field will be about 30×10^{-5} ftl. on a clear but moonless night at our latitudes and at sea-level [4]).

An explanation of the physical facts behind the curves reproduced in fig. 3 (involving among other things the minimum number of photons required for detection and the integration characteristics of the eye) would take us too far. The reader is referred for this to another article by one of the authors [5]).

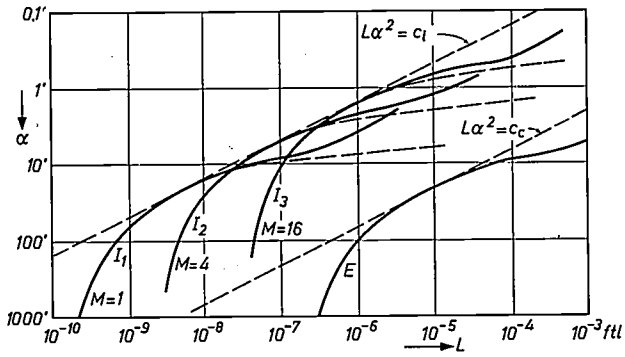


Fig. 3. Measured visual acuity curve E of the unaided eyes after complete adaptation [3]), compared with calculated curves for the eyes aided by an image intensifier system (of the type described but having unlimited resolving power) providing different overall magnifications: $M = 1$, $M = 4$ and $M = 16$. The smallest angle α an object must subtend in order to be detected when viewed against a background of 100% contrast, is plotted against the luminance L.

If the area integration capabilities of the human eye were unlimited, the acuity curve for the unaided eyes would be the straight line $L\alpha^2 = c_c$ and that for the eyes aided by the image intensifier system the line $L\alpha^2 = c_i$ (independent of the selected magnification).

Owing to the limited resolution of the viewing screen (among other things), each of the image intensifier curves flattens out (dotted lines).

An additional advantage of a variable magnification is that the available field of view will increase with decreasing magnification. In cases when the light level is not extremely low, an efficient viewing method will therefore adopt a low magnification with a large field of view for initial viewing and locating objects of interest, while changing to a higher magnification for having a "closer" look at each of these objects.

The experimental tube which has been developed for our system has a lumen gain of about 40 and an electron-optical magnification variable between 0.13 and 0.85, allowing the overall angular magnification of the system to be changed by a factor 6.5, from 2.6 times to 17 times. The total calculated performance curve of the system, based on the choice of optimum magnification in each range of illumination level, is shown in *fig. 4*.

For the sake of comparison, calculated curves for the eye aided by night glasses (7×50) and by a telescope (15×125) are also shown in fig. 4. These curves (N and T) are found by shifting the "unaided" curve E upwards by the magnification factor (7 and 15 resp.) and simultaneously to the *right* by a certain factor in order to account for the loss of light in the optical system (transmission about 70% in both cases). The intensifier system is seen to be superior to both night glasses and telescope, since it enables adequate vision at much lower light levels.

[3]) See: M. H. Pirenne, F. H. C. Marriott and E. F. O'Doherty, Individual differences in night vision efficiency, Brit. Med. Res. Council, Special Report Series No. 294, London 1957.

[4]) See: P. Bouma, Natural illumination and visibility at night, Philips tech. Rev. 5, 296-299, 1940.

[5]) P. Schagen, Electronic aids to night vision, Television Soc. J. 10, 218-228, 1963 (No. 7).

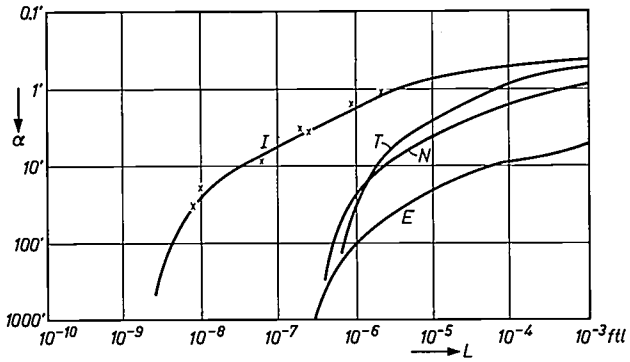


Fig. 4. Theoretical total performance curve I of the system for low light level vision, assuming optimum adjustment of magnification for each level. Points measured with our image intensifying system fit the curve very well. E acuity curve for the unaided eyes, T and N performance curve of the eyes aided by a monocular telescope 15×125 and night glasses 7×50 respectively.

The electrostatic zoom lens

History and general considerations

It should be pointed out that the problem of designing electron-optical systems with a variable magnification is by no means new. In electron microscopes based on magnetic electron lenses it is a well-established technique to change the magnification by adjusting the current passing through the magnet coils [6]). In this process it is easy to keep the picture in focus (in fact, the depth of focus of an electron microscope is extremely large), but the picture rotation inherent to magnetic focusing will change. This does not matter in electron microscopy, but it must be avoided in other applications. A method fulfilling this condition and capable of continuously varying the magnification over a factor of 2 was devised for the Philips image iconoscope (an early type of television camera tube) and reported about ten years ago [7]). This tube was equipped with a focusing system consisting of a combination of electrostatic and magnetic fields. The variable magnification was achieved by dividing the normal focusing coil into three separate sections located at different distances from the photocathode, as illustrated in *fig. 5*. The three degrees of freedom in selecting the current through each section made it possible to choose the required range of image magnification while maintaining picture focus and a constant picture rotation. The same principle has recently been applied to an image converter where a magnification variable by a factor 4 has been reported [8]).

Although a variable magnification is thus attainable with magnetically focused systems, such systems have severe drawbacks for equipment of the kind envisaged here (and, for that matter, for other purposes too): 1) it is rather difficult to obtain a substantial reduction in a tube with magnetic focusing; 2) magnetic focusing leads to increased bulk and weight of the equipment; 3) a high degree of stability in the electrical supplies is required in order to keep the picture in focus.

Electron-optical systems employing purely electrostatic fields do not suffer from such complications. Other difficulties are encountered in this case, which however have been successfully overcome, as will be shown below.

An electrostatic system in which the magnification could be varied has been described by Zworykin and Morton [9]). This was in essence a three-cylinder lens combination. Changing the potential on the centre cylinder varied the magnification and the image was refocused by adjusting the voltage on the low potential cylinder. This arrangement has a limited magnification range and suffers from all the aberrational defects of this type of electron-optics.

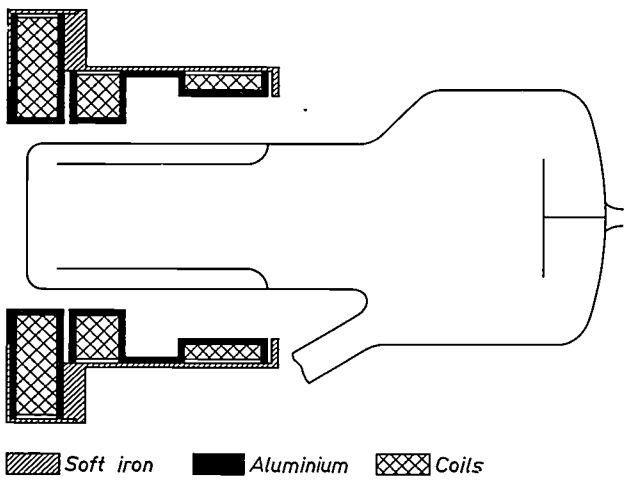


Fig. 5. Electron-optical system of the Philips image iconoscope, an early type of television camera tube, with magnetic focusing. By separately adjusting the currents through the three coil sections the overall magnification could be varied by a factor 2 while maintaining picture focus and picture rotation (cf [7])).

[6]) See J. B. le Poole, A new electron microscope with continuously variable magnification, Philips tech. Rev. **9**, 33-45, 1947/48.

[7]) J. C. Franken and H. Bruining, New developments in the image iconoscope, Philips tech. Rev. **14**, 327-335, 1952/53. See also: P. Schagen, H. Bruining and J. C. Francken, The image iconoscope, a camera tube for television, Philips tech. Rev. **13**, 119-133, 1951/52.

[8]) Tako Ando, Zoom image tubes, Paper presented at the International Television Conference, I. E. E., London 1962. A method of changing the magnification in an image orthicon from 0.84 to 1.4 by using an additional focusing coil and adjusting the strength of the electrostatic lens in the image section of the tube has also been described: S. Miyashiro and Y. Nakayama, Electronic zooming with the image orthicon television pick-up tube, Advances in Electronics and Electron Physics **16**, 2nd Symposium on Photo-electronic image devices 1962, p. 195-211.

[9]) V. K. Zworykin and K, A. Morton, Television, Chapman and Hall, London 1954, 2nd edition, p. 151.

For some time image tubes with electrostatic focusing have been designed based on the *concentric-spheres principle* which results in a much better picture quality than could be obtained with earlier systems [10]). We had reason to believe that one additional converging lens, if incorporated near the screen of a tube employing electron optics based on this principle, would allow the magnification to be varied considerably without undue movement of the image surface. This could be guessed from the results of experiments conducted in the Mullard Research Laboratories with the triode image converter ME 1201, which has been used extensively as an ultrafast photographic shutter [11]), and a range of at least a factor 4 for the magnification could be expected. A basic tube design on these lines is shown in *fig. 6*. When the anode and screen are at the same potential the tube is a conventional triode of the "concentric spheres" type and has a magnification of about 0.85. When the potential of the anode A_1 is reduced, a converging lens is formed between anode and screen and the image is reduced in size. The picture can be kept in focus by the simultaneous adjustment of the focus electrode.

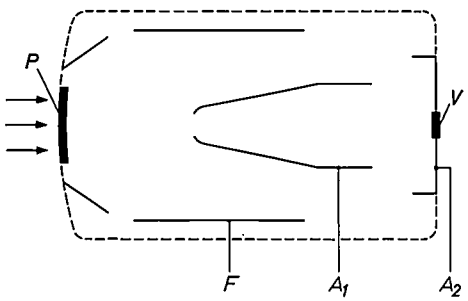


Fig. 6. Basic tube design for an image intensifier with electron-optical system of the concentric spheres type and incorporating an additional electrode A_2 for varying the magnification. P photocathode. A_1 main (spherical) anode. V viewing screen. F focusing electrode.

The actual tube

The practical tube which has sprung from these general considerations is shown in *fig. 7* and *8*. With this tube an even larger magnification range than anticipated is possible, viz, magnification from 0.85 to 0.13, as mentioned in the beginning of this article. Thus, the overall angular magnification of the system can be varied by a factor 6.5. In designing this tube, attention had to be given to two main problems, viz, loss of picture contrast, and loss of definition towards the edges of the picture. The latter problem will be discussed first.

With an electrostatic focusing system based on the concentric spheres principle, resolution will decrease towards the edges of the picture because the image surface is spherical whereas the viewing screen is flat. It would be pointless to adapt the screen to the shape of the image surface, since this would vary with magnification and, moreover a suitable eyepiece could not be designed for such a curvature of the screen. In a well-designed electron-optical system of this type the depth of focus will be large, thus minimizing the whole effect. However, the additional converging lens in the variable magnification tube enhances the curvature of the image plane and increases the loss of definition away from the electron-optical axis.

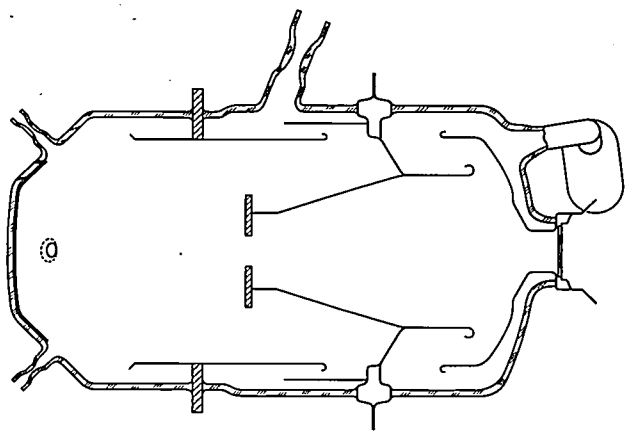


Fig. 7. Simplified cross-section of the tube in which the same elements as in fig. 6 can be found.

A simple solution would be to decrease the radius of curvature of the photocathode, which has the effect of flattening the image surface. In the tube shown in fig. 8, however, the curvature of the cathode was completely determined by the need to match the spherical image surface of the associated mirror optics [1]) (fig. 2), so that this artifice could not be directly employed. A compromise solution has been achieved by making the tube of a larger diameter and shaping the part of the cathode electrode that is outside the actually utilised photo-emissive surface in such a way that the equipotential surfaces near the cathode are more curved (increase of apparent curvature of the cathode). The result is shown in *fig. 9*, where the resolution in the image measured is plotted against distance from the centre. The central resolution is limited by the grain of the viewing screen to about 45 line pairs per mm (for a 100% contrast test pattern), which corresponds to

[10]) P. Schagen, H. Bruining and J. C. Francken, A simple electrostatic electron-optical system with only one voltage, Philips Res. Repts. 7, 119-130, 1952.

[11]) See e.g. J. A. Jenkins and R. A. Chippendale, High speed photography by means of the image converter, Philips tech. Rev. 14, 213-225, 1952/53.

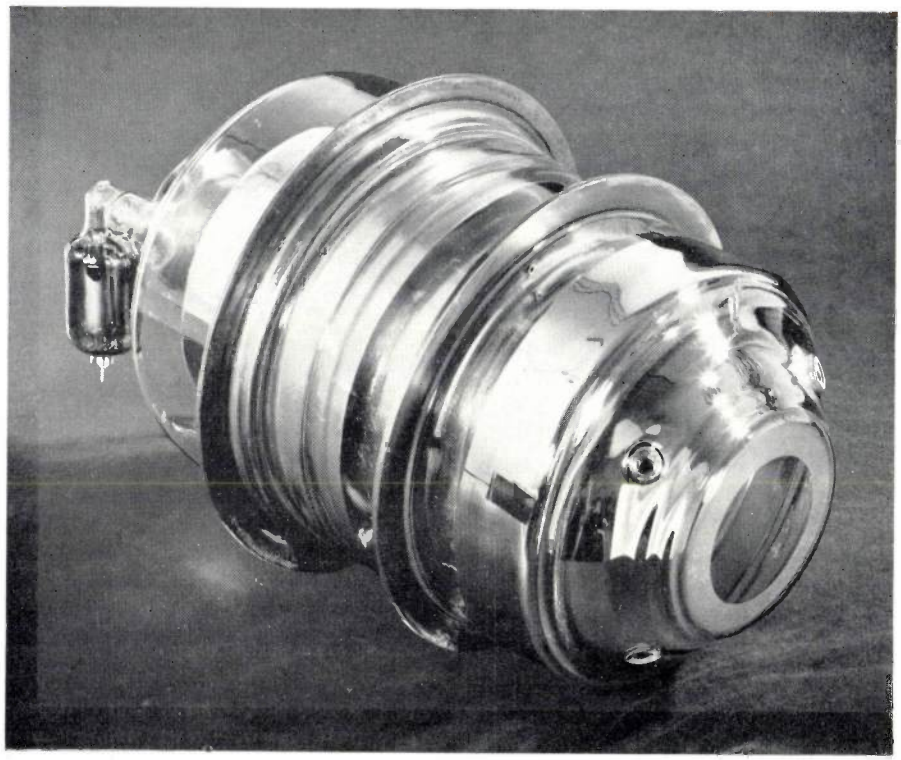


Fig. 8. Image intensifier tube for the low light level vision system developed at Mullard Research Laboratories. To the right the window with photocathode (6 cm diameter). The metal rings forming part of the tube envelope provide means for mounting the tube and at the same time for connecting the various electrodes to their respective supplies. To the left a getter pump for maintaining the required vacuum is visible.

about 3.5× the resolution of a 625 line television picture. *Fig. 10* shows a series of photographs taken from the viewing screen of the tube at different magnifications.

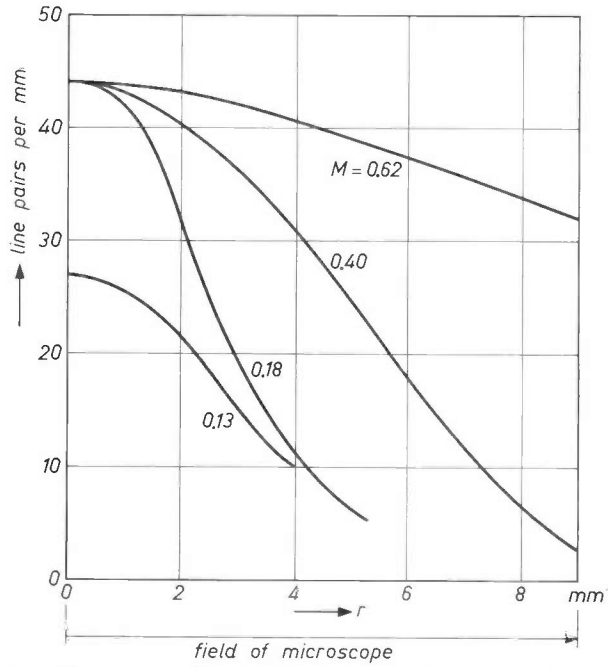


Fig. 9. Resolution of the image on the viewing screen plotted against the distance from the centre, for different magnifications.

A loss of picture contrast may be caused by spurious illumination of the fluorescent screen, which can have several origins. Stray electrons in the tube can cause the build-up of charges on the insulating wall surfaces of the tube. Radiation can then be emitted when breakdown occurs between charged areas, so that the wall surface can feed light back to the photocathode. This effect, which may also occur in X-ray image intensifiers and infrared image converters, is sufficiently reduced by suitable shaping of the electrodes, preventing this radiation from reaching the cathode. Other possible causes of background brightness are thermionic emission by the cathode and ion bombardment of the cathode releasing secondary electrons. The background brightness of the viewing screen due to these effects is less than 10^{-6} ftl when the tube is operated at 25 kV anode voltage. Finally, a serious loss of picture contrast can also be caused by light being transmitted by the photocathode (which cannot be made completely opaque) and reflected within the tube back to the cathode. In order to cut down this effect, the electrodes are coated with a layer having low reflectivity. Owing to all these precautions the contrast reproduction of the tube is exceptionally good.

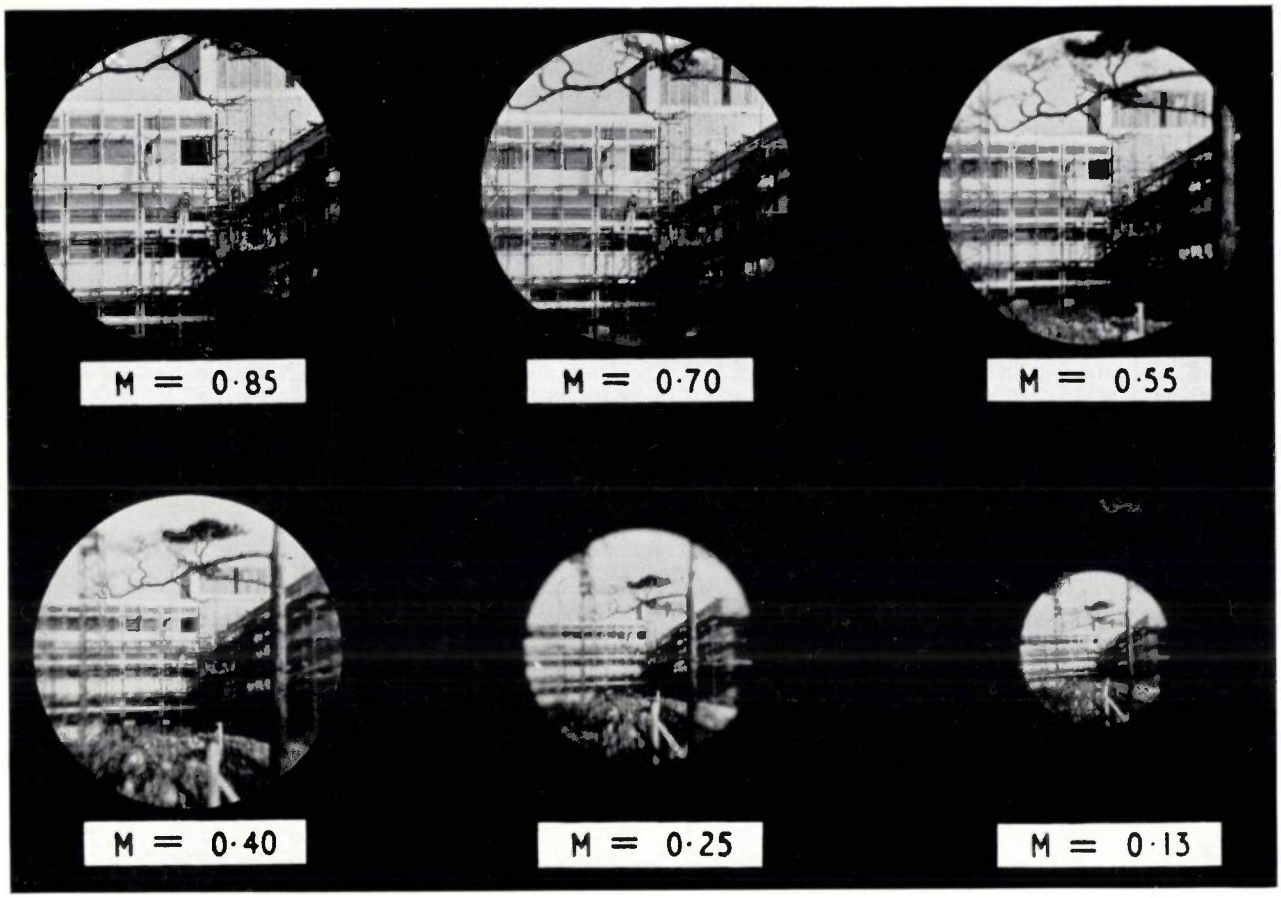


Fig. 10. Photographs taken from the viewing screen at different magnifications.

A few words about the control of magnification should be added. The variation in magnification as a function of the anode potential with respect to the photocathode is shown in *fig. 11*. In the same graph the potential required for the focus electrode to keep the picture in focus is plotted. It is possible to design a tube in which the latter potential is very small or even zero over a large part of the range. This would simplify the control of magnification. The shape of the focus potential curve, as that in fig. 11, however, is extremely sensitive to small changes in the photocathode to anode distance. The danger would exist that slight deviations from the nominal tube dimensions which are to be expected in manufacture, could change the focus potential characteristic considerably and even result in a curve which takes *negative* values over some portion of the range. This would lead to undesirable complications in the design of the power supply.

The present tube, therefore, has deliberately been designed for comparatively large positive focus potentials (fig. 11). Control of magnification is nevertheless made very easy by using a simple cam to couple the drive of two potentiometers, regulating the anode and focus-electrode voltages. A continuously variable picture size over the whole magnification range is thus obtained by turning a single knob.

Final remarks

The tube described was designed to suit the optical system which was necessary for the particular system of low-light-level vision illustrated in fig. 1.

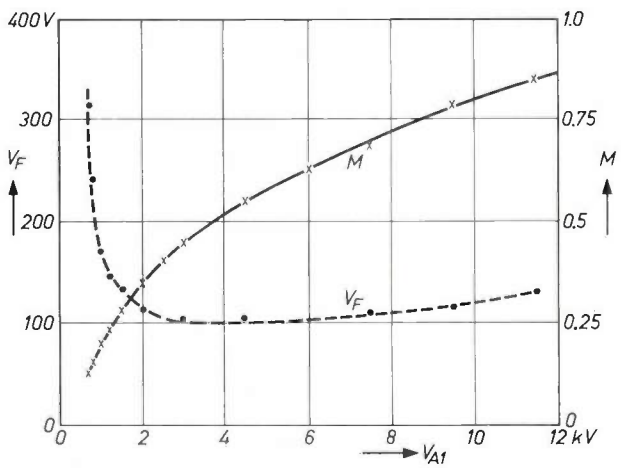


Fig. 11. Magnification M by the image intensifier tube, plotted against the voltage V_{A1}. The dotted curve represents the potential V_F required at the focus electrode in order to keep the image in focus.

A similar design for achieving variable magnification can, however, be applied to any image tube which employs electrostatic focusing, including those tubes which have the additional function of wavelength conversion. An example of an instrument of this type in which a variable magnification may be extremely useful is an ultra-violet microscope which employs an image converter as its final viewing element. The technique could also be valuable in the well-known X-ray image intensifier used for medical purposes. Finally, an electron-optical "zoom lens" of the kind described may even replace its optical equivalent, common nowadays, in applications where the facility to control the magnification by electrical rather than mechanical means would be important.

In some of these applications a greater latitude of choice of objective optics (if any) may exist, allowing an increase in the curvature of the photocathode. This will enable the image surface to be sufficiently flattened without having to increase the diameter of the tube much beyond that of the useful area of the photocathode.

Summary. A system for vision at low light levels employing a special image intensifier tube has been developed at Mullard Research Laboratories. The tube, which contains an electrostatic electron-optical system of the concentric spheres type, produces a reduced image on the viewing screen and provides a lumen gain of about 40. Owing to the addition of an extra electrode near the screen, the reduction can be varied between 1 : 0.85 and 1 : 0.13 by simply adjusting the anode voltage. The variation of the overall angular magnification by a factor 6.5 makes it possible to select the optimum acuity curve in a very broad range of light levels, permitting adequate vision at luminances down to 10^{-9} footlambert (i.e. several thousand times less than the luminance of a snow field on a clear but moonless night).

A VIBRATING CAPACITOR DRIVEN BY A HIGH-FREQUENCY ELECTRIC FIELD

by A. G. van NIE *) and J. J. ZAALBERG van ZELST *).

621.317.723

For many years the vibrating capacitor has formed the basis of vibrating-plate electrometers for converting a DC voltage, charge or current into an AC voltage which can be amplified by a normal amplifier. The article describes a new type of vibrating capacitor having various advantages over other types and which has recently been put into production.

Principle of the vibrating-plate electrometer

Electrometers are instruments used for measuring electric charges, DC voltages and very weak DC currents. A characteristic feature is their very high input resistance, which is in the order of thousands of megohms. In most kinds of electrometer the DC signal to be measured is converted into a proportionate alternating voltage; this is amplified by valves or transistors and, after rectification, produces a deflection on a meter.

A familiar device for converting a small direct voltage into an alternating voltage is the *vibrating capacitor*. The principle — which also underlies the condenser microphone — is represented in *fig. 1a*. A vibrating capacitor consists of two plates, one of which is stationary and the other (the earthed plate in fig. 1*a*) is kept in vibration, so that the capacitance C_v varies periodically with time t:

$$C_v = \frac{C_{v0}}{1+\hat{a}\cos pt}.$$

Here C_{v0} is the capacitance of the capacitor with both plates stationary, $\hat{a}$ the relative amplitude of the vibrating plate and p the angular frequency. The vibrating capacitor is connected to the source

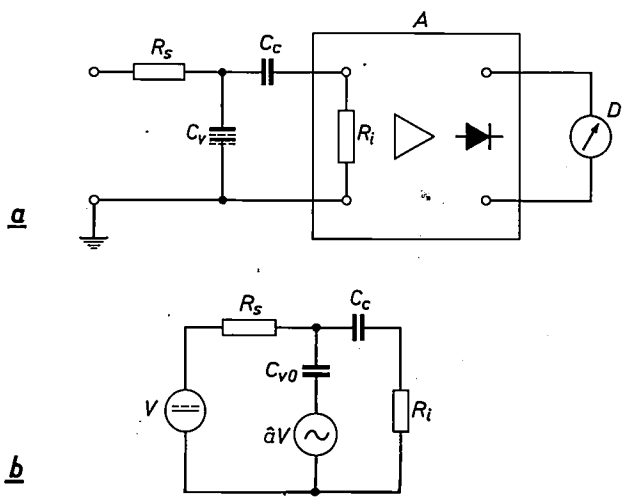


Fig. 1. *a*) Basic diagram of a vibrating-plate electrometer. C_v vibrating capacitor. C_c coupling capacitor. *A* AC voltage amplifier and rectifier. R_i input resistance of amplifier. *D* moving coil meter. R_s high series resistance.
b) Equivalent circuit, in which the vibrating capacitor is approximately represented by a fixed capacitor C_{v0} in series with an AC voltage source, whose voltage $\hat{a}V$ is proportional to the measured DC voltage V.

*) Philips Research Laboratories, Eindhoven.

A similar design for achieving variable magnification can, however, be applied to any image tube which employs electrostatic focusing, including those tubes which have the additional function of wavelength conversion. An example of an instrument of this type in which a variable magnification may be extremely useful is an ultra-violet microscope which employs an image converter as its final viewing element. The technique could also be valuable in the well-known X-ray image intensifier used for medical purposes. Finally, an electron-optical "zoom lens" of the kind described may even replace its optical equivalent, common nowadays, in applications where the facility to control the magnification by electrical rather than mechanical means would be important.

In some of these applications a greater latitude of choice of objective optics (if any) may exist, allowing an increase in the curvature of the photocathode. This will enable the image surface to be sufficiently flattened without having to increase the diameter of the tube much beyond that of the useful area of the photocathode.

Summary. A system for vision at low light levels employing a special image intensifier tube has been developed at Mullard Research Laboratories. The tube, which contains an electrostatic electron-optical system of the concentric spheres type, produces a reduced image on the viewing screen and provides a lumen gain of about 40. Owing to the addition of an extra electrode near the screen, the reduction can be varied between 1 : 0.85 and 1 : 0.13 by simply adjusting the anode voltage. The variation of the overall angular magnification by a factor 6.5 makes it possible to select the optimum acuity curve in a very broad range of light levels, permitting adequate vision at luminances down to 10^{-9} footlambert (i.e. several thousand times less than the luminance of a snow field on a clear but moonless night).

A VIBRATING CAPACITOR DRIVEN BY A HIGH-FREQUENCY ELECTRIC FIELD

by A. G. van NIE *) and J. J. ZAALBERG van ZELST *).

621.317.723

For many years the vibrating capacitor has formed the basis of vibrating-plate electrometers for converting a DC voltage, charge or current into an AC voltage which can be amplified by a normal amplifier. The article describes a new type of vibrating capacitor having various advantages over other types and which has recently been put into production.

Principle of the vibrating-plate electrometer

Electrometers are instruments used for measuring electric charges, DC voltages and very weak DC currents. A characteristic feature is their very high input resistance, which is in the order of thousands of megohms. In most kinds of electrometer the DC signal to be measured is converted into a proportionate alternating voltage; this is amplified by valves or transistors and, after rectification, produces a deflection on a meter.

A familiar device for converting a small direct voltage into an alternating voltage is the *vibrating capacitor*. The principle — which also underlies the condenser microphone — is represented in *fig. 1a*. A vibrating capacitor consists of two plates, one of which is stationary and the other (the earthed plate in fig. 1*a*) is kept in vibration, so that the capacitance C_v varies periodically with time t:

$$C_v = \frac{C_{v0}}{1+\hat{a}\cos pt}.$$

Here C_{v0} is the capacitance of the capacitor with both plates stationary, $\hat{a}$ the relative amplitude of the vibrating plate and p the angular frequency. The vibrating capacitor is connected to the source

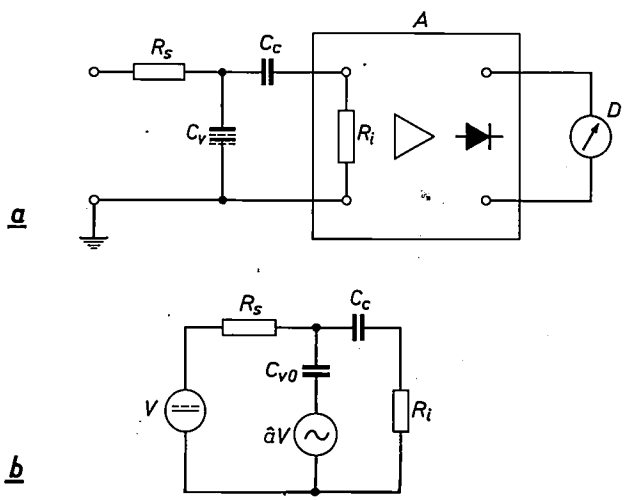


Fig. 1. *a*) Basic diagram of a vibrating-plate electrometer. C_v vibrating capacitor. C_c coupling capacitor. *A* AC voltage amplifier and rectifier. R_i input resistance of amplifier. *D* moving coil meter. R_s high series resistance.
b) Equivalent circuit, in which the vibrating capacitor is approximately represented by a fixed capacitor C_{v0} in series with an AC voltage source, whose voltage $\hat{a}V$ is proportional to the measured DC voltage V.

*) Philips Research Laboratories, Eindhoven.

of the DC signal via a series resistance R_s. A coupling capacitor C_c blocks the flow of direct current through the input resistance R_i of the amplifier. The vibrating capacitor is thus charged to the potential of the DC signal to be measured. If the charge q of the capacitor is constant, the voltage v_v across the capacitor is given by:

$$v_v = \frac{q}{C_{v0}} (1 + \hat{a} \cos pt) . \quad . . . . (1)$$

The alternating component $v_\sim$ of v_v is in this case therefore proportional to the DC voltage component $V = q/C_{v0}$ to be measured:

$$v_\sim = \hat{a} V.$$

To a good approximation this still applies when the condition q = constant is not entirely fulfilled. This is always the case in practice, because part of the charge q varies with the frequency $p/2\pi$. This can be understood from the approximate representation in fig. 1*b*, where the vibrating capacitor is replaced by the fixed capacitor C_{v0} in series with an alternating voltage source $\hat{a}V$; the latter delivers an alternating current, which means that the charge varies.

The alternating current flows partly through C_c and R_i, partly through R_s and the DC voltage source. If the conditions are chosen so that

$$pC_cR_s \gg 1, \quad . . . . . (2)$$

then only a negligible fraction of the alternating current flows through R_s, most of it thus flowing through R_i.

Given perfect insulation of the vibrating and the coupling capacitor, the electrometer draws no energy from the DC voltage source (except that needed for charging the capacitors). Energy is, however, delivered to the amplifier; this energy is drawn by the vibrating capacitor from the system that keeps the one plate in vibration. The vibrating capacitor can thus be considered as a *parametric amplifier*.

The better the insulation of the two capacitors, the more sensitive the measurement (provided at least the noise sets no limit to the sensitivity). In this case the resistance R_i can have any value.

Principle of the new vibrating capacitor

In an earlier form of vibrating capacitor the plates were open to the atmosphere and the one plate was driven by an electrodynamic loudspeaker system [1] [2]. The present article deals with a new design in which the vibrating capacitor with its drive system — now capacitive — is contained inside an evacuated bulb (*fig. 2*). This effectively protects it against dust, moisture and other atmospheric influences, which is very much to the benefit of the insulation resistance. Since the materials employed are capable of withstanding high temperatures, the component parts are thoroughly outgassed while pumping. This makes the contact potentials — which as mentioned below, can cause the zero point to drift — very much less subject to variation, and thus much easier to compensate. A new drive system, by means of a high-frequency electric field, has the virtue of causing no interference in the measuring circuit. Owing to the absence of air friction, the drive requires exceptionally little power. The noise level, moreover, is particularly low.

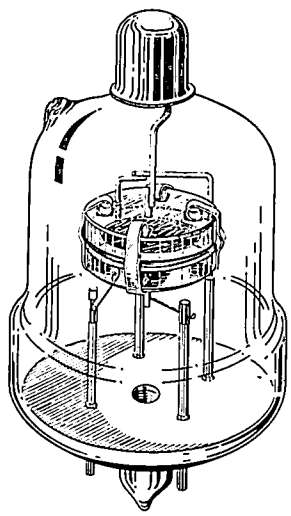


Fig. 2. The new vibrating capacitor, now produced by Philips' X-ray and Medical Apparatus Division. The vibrating capacitor is mounted in an evacuated bulb, which protects it from dust, moisture and other atmospheric influences. Roughly true size.

A transistor circuit has been designed both for the amplifier and for the drive. The amplifier will not be dealt with in this article, but a brief description will be given of the oscillator for the drive.

Fig. 3a shows a schematic cross-section of the new vibrating capacitor. A round glass membrane M, 0.135 mm thick, is clamped at its rim between two glass insulators, I_1 and I_2. The middle portion of the insulators is hollow ground, and coated with a layer of metal, as are both faces of the membrane. These layers, indicated in the figure by a thick line, serve as electrodes (E_1, E'_1, E_2, E_3). Electrodes E_1 and E_2 constitute the vibrating capacitor C_v, which is circuited as shown in fig. 1*a*. A second, similar capacitor, C_d, is formed by electrodes E'_1 and E_3; these are used in the manner described below for causing the membrane to vibrate at its natural frequency.

The hollow-ground parts of the insulators are encircled by a groove. This lengthens the leakage path and thus increases the insulation resistance between the electrodes E_2 and E_3 on the one hand and the earthed membrane electrodes E_1 and E'_1 on the

[1] C. Dorsman, A *p*H-meter with a very high input resistance Philips tech. Rev. **7**, 24-32, 1942.

[2] J. van Hengel and W. J. Oosterkamp, A direct-reading dynamic electrometer, Philips techn. Rev. **10**, 338-346, 1948/49.

other. In this way an insulation resistance of the order of 10^{14} ohm is obtained.

Fig. 3*b* shows how the coupling capacitor is incorporated in the system.

The drive

The membrane is kept in vibration by the fact that the capacitor C_d forms part of an oscillator O (fig. 3). This maintains a high-frequency voltage (frequency about 1 Mc/s) across C_d which is modulated in amplitude with a fundamental frequency automatically equal to the natural frequency of the membrane (approx. 6 kc/s).

Assuming for simplicity that the modulation is sinusoidal, then the voltage u_d across C_d can be represented by:

$$u_d = (1 + m \cos pt)\, U_0 \cos \omega t = \\ = U_0 \cos \omega t + \tfrac{1}{2} m U_0 \{\cos(\omega + p)t + \cos(\omega - p)t\}. \quad \ldots \quad (3)$$

Here U_0 is the amplitude and $\omega/2\pi$ the frequency of the unmodulated voltage, $p/2\pi$ is the natural frequency of the membrane, and m the depth of modulation. The electrical attractive force F on the membrane is proportional to u_d^2. It follows from (3), after squaring and trigonometrical manipulation, that the force $F(t)$ consists of the following groups of components:

1) A constant component,

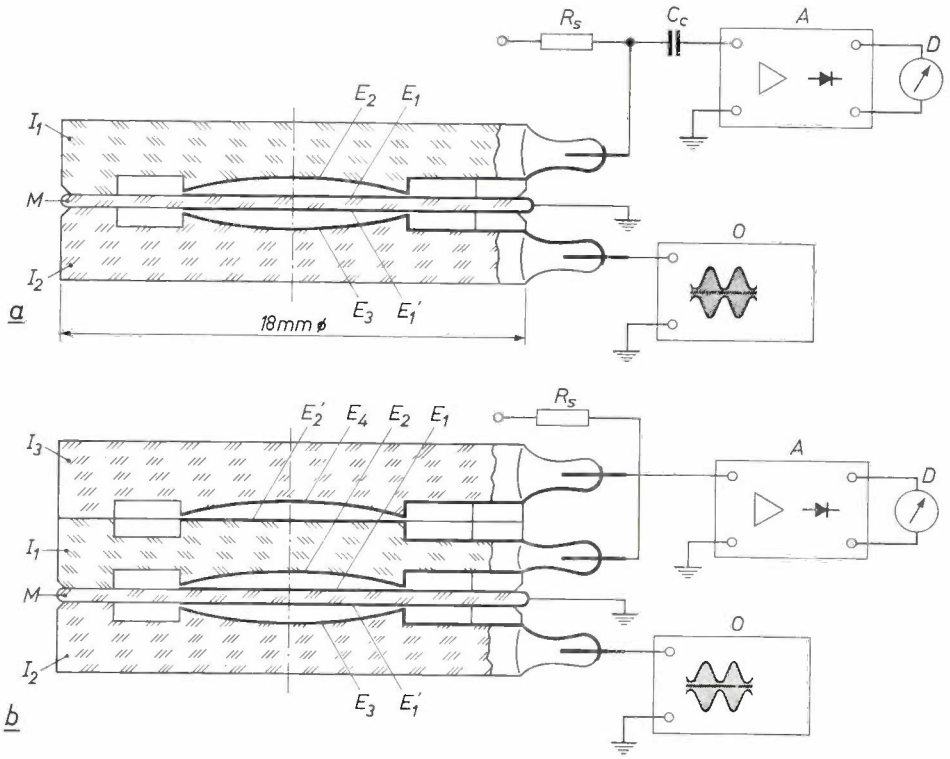


Fig. 3. *a*) Schematic cross-section of the vibrating capacitor. M round glass membrane (0.135 mm thick) the edge of which is clamped between glass insulators I_1 and I_2. The thick line indicates metal layers that serve as electrodes (E_1, E_1', E_2, E_3). Electrodes E_1 and E_2 form the vibrating capacitor, which is connected as shown in fig. 1*a* to the components R_s, C_c, A and D. Electrodes E_1' and E_3 form a second capacitor. This is part of an oscillator O, which maintains a high frequency voltage (1 Mc/s) between E_1' and E_3; the voltage is amplitude-modulated at the natural frequency of the membrane (6 kc/s). The pulsed electric attractive force keeps the membrane in vibration.
b) Here insulator I_1 has an electrode (E_2') at the top also which is connected to E_2; an insulator I_3 with electrode E_4 has been added. E_2' and E_4 together form the coupling capacitor (C_c).
For the sake of clarity the curvature of the hollow-ground middle section of the insulators is shown greatly exaggerated.

2) Low-frequency components with the angular frequencies p and $2p$,
3) High-frequency components with the angular frequencies 2ω, $2\omega \pm p$ and $2\omega \pm 2p$.

Since the membrane shows a sharp resonance at the frequency $p/2\pi$, the component having this frequency is the only one of importance for the drive, and this is also the frequency of the alternating voltage into which the measured quantity is to be converted. As seen from (3) the voltage u_d, on the other hand, contains only high frequency compo-

nents. Parasitic components with these high frequencies, in the region of 1 Mc/s, can therefore be kept out of the measuring circuit by simple means. This is a considerable advantage compared with the old type of vibrating capacitor, which was driven by a current having *the same* frequency as that of the alternating voltage into which the measured quantity was converted.

As can be seen from the oscillogram in *fig. 4*, the modulation of the high-frequency oscillator voltage is far from sinusoidal. This does not detract, however, from the principle described.

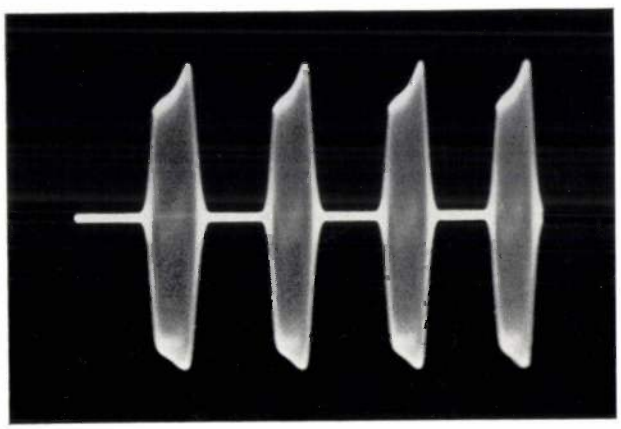


Fig. 4. Oscillogram of the voltage u_d of the oscillator O in fig. 3. The oscillator oscillates at certain intervals with a frequency of 1 Mc/s. The amplitude modulation is thus not sinusoidal but there is a periodic interruption. The repetition frequency is equal to the natural frequency of the membrane (approx. 6 kc/s).

Oscillator circuit

The oscillator circuit used is shown in *fig. 5a*. The circuit oscillates (at about 1 Mc/s) when the fed back alternating voltage V_b between base and emitter of the transistor Tr has the proper magnitude and phase. When the latter is the case, V_b is called positive. This voltage is taken from a bridge circuit consisting of a centre-tapped coil S_2 and capacitors C_1 and C_d. Of these C_d is formed by the electrodes E_1' and E_3 (fig. 3). We provisionally assume that the membrane is not yet in vibration, i.e. that C_d has a fixed capacitance (C_{d0}). If C_{d0} and C_1 are equal in magnitude, the bridge is balanced ($V_b = 0$). Provided the differences between the two capacitances are small, V_b is nearly proportional to $C_{d0} - C_1$ (fig. 5*b*). C_1 is adjusted ($> C_{d0}$) to give V_b the exact positive value that is just great enough to cause the circuit to oscillate at constant amplitude. When the membrane vibrates at its natural frequency, so that C_d becomes alternately larger and smaller, then V_b, according to fig. 5*b*, will likewise alternately increase and decrease and so too will the amplitude of the vibration.

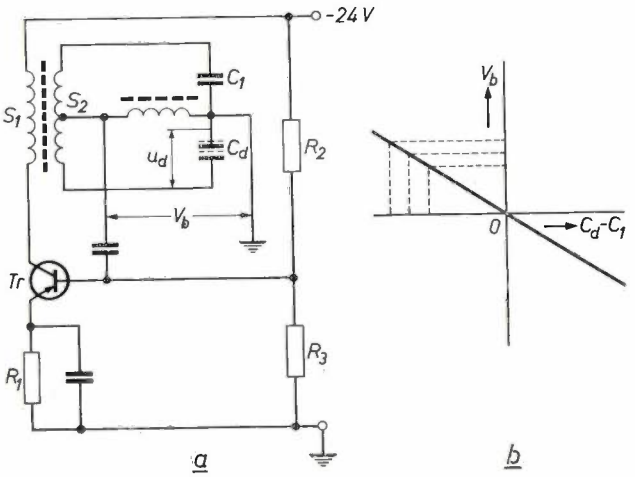


Fig. 5. *a*) Basic diagram of the oscillator. Tr transistor type OC 44. Of the coupled coils S_1 and S_2 the later is centre-tapped to form a bridge circuit with capacitors C_1 and C_d; C_d is formed by the electrodes E_1' and E_3 (fig. 3).
b) The diagonal voltage V_b in (*a*) as a function of the capacitance difference $C_d - C_1$. Capacitor C_1 is adjusted so that V_b has the magnitude and phase required for oscillation at constant amplitude. When the membrane is in vibration, C_d varies periodically and the high-frequency vibration is modulated in amplitude.

Since the membrane would touch the other electrode if the amplitude were excessive, and since the amplitude depends on the maximum peak value of the oscillator voltage, it is necessary to keep the latter below a certain limit. The operating point is therefore so chosen (class B-C) that with increasing voltage amplitude U_0 the effective current amplification factor α'_{eff} increases, first gradually and then rapidly when a certain value of U_0 is reached (*fig. 6*); α'_{eff} is understood to be the ratio of the fundamental component of the collector current to the alternating base current. The oscillation condition — loop gain $= 1$ — is fulfilled at that amplitude at which $\alpha'_{eff} = 1/\beta$; here $1/\beta$ is the fraction of the oscillator voltage which is fed back to the base via the oscillator network, and β is the feedback factor. If the resistance R_1 and the ratio $R_2 : R_3$ (fig. 5*a*) are

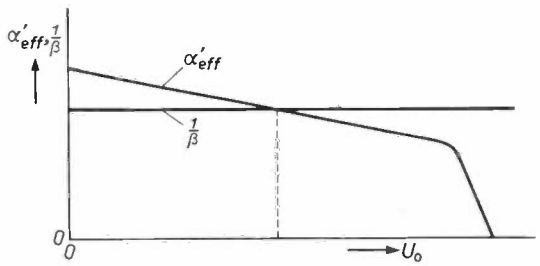


Fig. 6. Effective current amplification factor α'_{eff} of the transistor circuit in the oscillator and $1/\beta$ (β = feedback factor) as a function of the amplitude U_0 of the high-frequency voltage. The oscillation condition is satisfied at the amplitude at which $\alpha'_{eff} = 1/\beta$. If the resistances R_1, R_2 and R_3 (see fig. 5*a*) are given the appropriate values, the α'_{eff} characteristic can be made to intersect the horizontal line $1/\beta$ at a suitable amplitude U_0. The steeply descending part of the characteristic limits the maximum amplitude.

properly chosen, the α'_{eff} characteristic in fig. 6 can be given a slope such that the point $\alpha'_{eff} = 1/\beta$ comes to lie at a suitable average value of U_0 and that the bend in the characteristic ensures effective limiting of the maximum amplitude — without the operating point becoming critical.

Fig. 7a-e shows that the phase differences between the various quantities are exactly as required to maintain the oscillation.

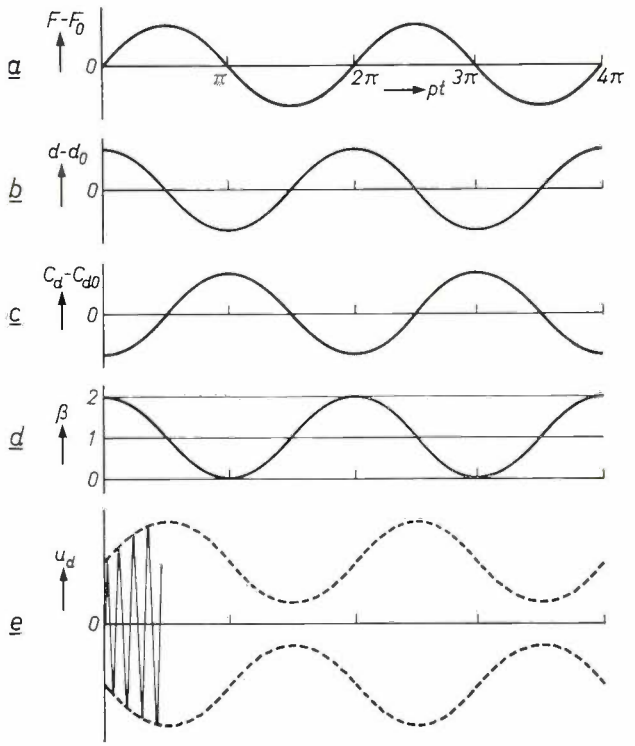


Fig. 7. The following are plotted as a function of pt:
a) the alternating component with frequency $p/2\pi$ of the electric attractive force F between the electrodes of the drive capacitor (C_d),
b) the alternating component of the distance d between the electrodes,
c) the alternating component of the capacitance C_d,
d) the feedback factor β,
e) the amplitude modulated oscillator voltage u_d.
(The subscript *0* in F_0 and d_0 denotes the quiescent state.)
Since the membrane is in resonance, $d - d_0$ lags 90° in phase behind $F - F_0$. The capacitance varies in antiphase with the distance d, and β varies in antiphase with capacitance C_d (cf. fig. 5*b*). When $\beta > 1$, the amplitude of u_d increases, and when $\beta < 1$ the amplitude decreases. The amplitude variation (*e*) is seen to be in phase with the variation of the attractive force (*a*).

Equivalent noise resistance

The sensitivity of the electrometer is primarily governed by the signal-to-noise ratio at the input of the amplifier. As known from the literature [3]), it is useful in this respect to connect between the input terminals of the amplifier a coil (which may be the primary of a transformer) tuned to the frequency of the vibrating plate. We decided on a transformer because it can be matched, thus minimizing the noise factor. Since, with the design employed, the frequency $p/2\pi$ is relatively high (6 kc/s), a low self-inductance is sufficient at the primary of the transformer.

Fig. 8a gives the equivalent circuit for the case where a direct current I is to be measured. This is conducted through a measuring resistance R_m. The noise of R_m is represented by the noise voltage

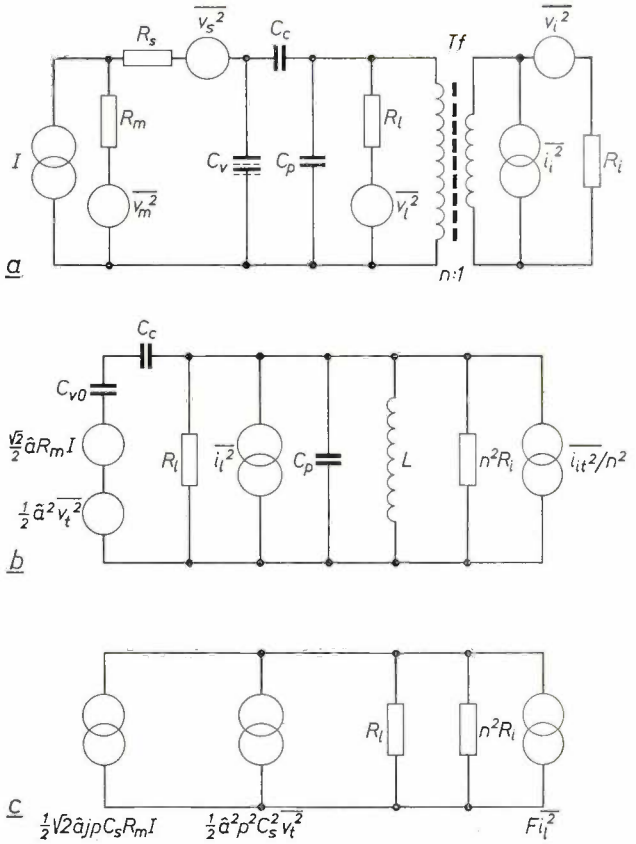


Fig. 8. *a*) First equivalent circuit of electrometer. The direct current I to be measured flows through the measuring resistance R_m. The amplifier has an input transformer *Tf* with a turns ratio $n:1$ and stray capacitance C_p: the resistance R_l represents the transformer losses. R_s, C_v, C_c and R_i have the same meaning as in fig. 1*a*.
The noise of resistors R_m, R_s and R_l is taken into account by the noise voltage sources v_m^2, v_s^2 and v_l^2, the noise of the amplifier by the noise voltage source v_i^2 and the noise current source i_i^2.
b) Second equivalent circuit. The vibrating capacitor converts the direct current I to be measured into an alternating voltage source $\frac{1}{2}\sqrt{2}\,\hat{a}R_mI$ (r.m.s. value). The noise voltage sources v_m^2 and v_s^2 of (*a*) are combined to form a single noise-voltage source v_t^2. The noise voltage source v_l^2 is replaced by a noise current source i_l^2. The resistance R_i on the secondary side is replaced by n^2R_i at the primary side, and the noise sources i_i^2 and v_i^2 at the secondary side by a single noise current source i_{it}^2/n^2 at the primary.
c) Third equivalent circuit. The voltage sources $\frac{1}{2}\sqrt{2}\,\hat{a}\,R_mI$ and $\frac{1}{2}\hat{a}^2\overline{v_t^2}$ in (*b*) are transformed into equivalent current sources, the noise current sources i_l^2 and i_{it}^2/n^2 in (*b*) are replaced by a single noise current source $F\,i_l^2$, and the inductance L — by the addition of capacitance — is tuned to the frequency $p/2\pi$.

[3]) F. A. Muller, Het meten van stromen en spanningen met de trilplaatcondensator, Thesis Amsterdam 1951, p. 43 et seq.

source $\overline{v^2_m}$. The other elements that produce noise are the series resistance R_s, the parallel resistance R_l representing the transformer losses, and the transistor amplifier. The noise-voltage sources $\overline{v_s^2}$, $\overline{v_l^2}$ and $\overline{v_i^2}$ and the noise current source $\overline{i_i^2}$ [4]) represent their respective contributions [4]).

With the exception of R_m, all noise sources in this diagram are considered to be collectively replaced by one equivalent noise resistance R_{eq}. The total noise power is thus split into a part $4kTBR_m$ and a part $4kTBR_{eq}$, where k is Boltzmann's constant, T the absolute temperature of the resistances and B the bandwith. In the section in small print below, the following expression is derived for R_{eq}:

$$R_{eq} = R_s + \frac{F}{\frac{1}{2}\hat{a}^2 p^2 C_s^2 R_l} . \quad . \quad . \quad . \quad . \quad (4)$$

Here C_s is the capacitance of C_{v0} and C_c in series: $C_s = C_{v0}C_c/(C_{v0} + C_c)$, and F is the noise factor of the amplifier; F becomes minimal at one particular turns ratio $n:1$ of the transformer. It is assumed that — in line with reality — the primary inductance L of the transformer is tuned to the frequency $p/2\pi$; for this purpose a capacitance C_e is added to the capacitance $C_p + C_s$ already present (C_p being the stray capacitance) to give:

$$p^2 L(C_p + C_s + C_e) = 1 . \quad . \quad . \quad (5)$$

Derivation of eq. (4)

The calculation of R_{eq} can be considerably simplified if we consider that in the case of *selective rectification* there are at the output of the amplifier only two frequency bands in which the noise contributions can be of importance: a band around the frequency $p/2\pi$ and a band beginning at the frequency 0; the noise contributions in the latter band are converted by the vibrating capacitor to the first. In the highly selective synchronous rectification, which is generally used in vibrating-plate electrometers [5]), both bands are no wider than about 1 c/s. Moreover, some noise sources contribute to only one of these bands, the reason being that the transformer coils constitute a virtual short-circuit to very low frequencies, so that the noise from the sources $\overline{v_l^2}$, $\overline{v_i^2}$ and $\overline{i_i^2}$ may be neglected in the band beginning at the frequency 0. The only contributions concerned are thus the following:

Frequency band	Noise sources
From 0 tot $\frac{1}{2}$ c/s	$\overline{v_m^2}$ and $\overline{v_s^2}$.
From $\frac{p}{2\pi} - \frac{1}{2}$ to $\frac{p}{2\pi} + \frac{1}{2}$ c/s	$\overline{v_m^2}$, $\overline{v_s^2}$, $\overline{v_l^2}$ (of these three $\overline{v_l^2}$ is the most important, since $R_l < R_m + R_s$), $\overline{v_i^2}$ and $\overline{i_i^2}$.

To derive equation (4) we transform the circuit of fig. 8*a* into a parallel arrangement of current sources and noiseless resistors. This transformation proceeds in two steps. First we simplify fig. 8*a* to fig. 8*b*. Here $\frac{1}{2}\sqrt{2}\,\hat{a}R_m I$ is the alternating voltage into which the vibrating capacitor converts the measured direct current. The noise of R_m and R_s at 0-$\frac{1}{2}$ c/s is summarized by $\frac{1}{2}\hat{a}^2\overline{v_t^2}$. The input impedance R_i transformed to the primary side is n^2R_i and $\overline{i_{it}^2}/n^2$ appears in the place of the input noise sources $\overline{i_i^2}$ and $\overline{v_i^2}$.

According to the definition, the noise factor F is given by:

$$F = \frac{\overline{i_l^2} + \overline{i_{it}^2}/n^2}{\overline{i_l^2}} .$$

This yields:

$$\overline{i_l^2} + \overline{i_{it}^2}/n^2 = F\,\overline{i_l^2}.$$

The second step consists among other things in replacing the parallel noise-current sources $\overline{i_l^2}$ and $\overline{i_{it}^2}/n^2$ by one single noise-current source $F\,\overline{i_l^2}$, as is done in fig. 8*c*. The magnitude of F, when the turns ratio n is optimally chosen, can be found from the literature if the data of the amplifier input are known [6]).

The voltage sources $\frac{1}{2}\sqrt{2}\hat{a}R_m I$ and $\frac{1}{2}\hat{a}\overline{v_t^2}$ in fig. 8*b* are replaced in fig. 8*c* by the equivalent current sources $\frac{1}{2}\sqrt{2}\,\hat{a}jpC_sR_mI$ and $\frac{1}{2}\hat{a}^2p^2C_s^2\overline{v_t^2}$ respectively. Further, by the addition of capacitance C_e in accordance with equation (5), the inductance L is tuned to the frequency $p/2\pi$, so that L and all C's have disappeared from the diagram. The transformation is now complete; what remains is the wanted parallel arrangement of current sources and resistances.

The signal-to-noise ratio S/N can be read from fig. 8*c*:

$$\frac{S}{N} = \frac{\frac{1}{2}\hat{a}^2p^2C_s^2R_m^2I^2}{\frac{1}{2}\hat{a}^2p^2C_s^2\overline{v_t^2} + F\overline{i_l^2}} . \quad . \quad . \quad . \quad . \quad (6)$$

The noise power N can be split into a part due to the measuring resistance R_m, and a part due to the rest of the circuit, represented by the equivalent noise resistance R_m:

$$N = \tfrac{1}{2}\hat{a}^2p^2C_s^2\overline{v_t^2} + F\,\overline{i_l^2} =$$

$$= 4kTB\left\{\tfrac{1}{2}\hat{a}^2p^2C_s^2(R_m + R_s) + \frac{F}{R_l}\right\} =$$

$$= 2kTB\hat{a}^2p^2C_s^2\left(R_m + R_s + \frac{F}{\frac{1}{2}\hat{a}^2p^2C_s^2R_l}\right) . \quad . \quad . \quad (7)$$

This leads to:

$$R_{eq} = R_s + \frac{F}{\frac{1}{2}\hat{a}^2p^2C_s^2R_l},$$

which is eq. (4).

From eq. (4) we may conclude in the first place that to obtain minimum noise (minimum R_{eq}) we must aim at:

a) the lowest possible value of R_s,
b) a minimum noise factor F of the amplifier — which is obvious — (for which purpose n must be optimally chosen),
c) minimum transformer losses (maximum R_l).

The other quantities in (4), which relate to the vibrating capacitor, indicate that the relative amplitude $\hat{a}$, the natural frequency $p/2\pi$ and the capacitance $C_s(\approx \frac{1}{2}C_{v0} \approx \frac{1}{2}C_c)$ should be made as high as possible. It is particularly important to raise the natural frequency, because this not only reduces the

[4]) To account for the noise of a four-terminal network (here the transistor amplifier) it is necessary to introduce both a noise-voltage and a noise current source. See A. G. Th. Becking, H. Groendijk and K. S. Knol, The noise factor of four-terminal networks, Philips Res. Repts. **10**, 349-357, 1955.

[5]) See e.g. the article in [2]), p. 340 et seq.

[6]) K. Hinrichs and B. B. Weekes, "Squarved" input stages for low-level transistor amplifiers, I.R.E. Wescon Convention Record 1958, Part 2, 104-114.

term $2F/\hat{a}^2p^2C_s^2R_l$ but also allows R_s to be made smaller without jeopardizing the condition (2): $pC_{v0}R_s \gg 1$. The thicker the membrane the higher the vibration frequency it can have, but this at the same time requires a high electric field to drive it. The permissible relative amplitude $\hat{a}$ depends not only on the absolute amplitude but also on the precision with which the distance from the membrane to the fixed electrode can be controlled. The construction shown in fig. 3 is particularly favourable in this respect. The extent to which C_{v0} can be increased is limited by the time constant $R_m(C_{v0}+C_c) \approx 2R_mC_{v0}$, which should not make the electrometer response too slow at the very high values (e.g. 10^{11} ohms) that R_m must be given for the purpose of measuring very weak currents.

In a particular case the following numerical values broadly apply to the vibrating capacitor described: $F = 2$, $\hat{a} = 0.30$, $p = 2\pi\times 6000$ rad/s, $C_s = 20$ pF, $R_l = 1$ MΩ, $R_s = 10$ MΩ. After inserting these values in (4) we find: $R_{eq} = 10 + 80 = 90$ MΩ.

Fig. 9 shows a recording of the output voltage of the electrometer. The recording was taken with the input short-circuited, so that the resistor R_s ($= 10$ MΩ) was shunted across the vibrating capacitor. The noise is so weak as to be scarcely perceptible. The slow variation in the meter deflection will be discussed presently.

In the measurement of weak currents with the aid of a measuring resistance R_m, the signal-to-noise ratio S/N is given by:

$$\frac{S}{N} = \frac{R_m}{R_m + R_{eq}} \frac{R_m I^2}{4kTB}. \quad \ldots \quad (8)$$

Equation (8) is found by substituting the right-hand side of (7) for the denominator in the right-hand side of (6) and putting $R_s + F/\frac{1}{2}\hat{a}^2p^2C_s^2R_l$ equal to R_{eq} in accordance with (4).

At $R_m = 10^9$ ohm, R_{eq} is already negligible compared with R_m. Eq. (8) then reduces to:

$$\frac{S}{N} = \frac{I^2R_m}{4kTB}.$$

As far as is permitted by the signal-to-noise ratio, it is generally favourable to keep the value R_m as low as possible: this value can then be measured all the more accurately and it is all the less sensitive to moisture and dust.

For measuring low *voltages* it is even more desirable to keep R_{eq} small. In voltage measurements R_m is replaced by the given internal resistance of the voltage source. With $R_{eq} = 10^8$/ohms, measurements are possible on voltage sources whose internal resistance is of the order of 10^8 ohms or higher. The new vibrating capacitor therefore links up well with the electronic DC voltmeter type GM 6020, for which 10^8 ohms is roughly the upper limit of the source resistance.

Drift of zero point

Two principal requirements which a good vibrating capacitor is expected to meet — high insulation resistance and low equivalent noise resistance — have already been dealt with sufficiently in the foregoing. We now come to a third requirement, which is the *minimum drift of the zero point.*

Zero point drift has various causes. In the first place there are *contact potentials*, which occur whenever there is contact between dissimilar conductors and produce a deflection on the electrometer with the input short-circuited. This effect is generally compensated by means of a correction potentiometer connected to a constant DC voltage source, e.g. a standard element or a zener diode. The better the design succeeds in keeping the resultant contact potential low and above all constant, the less frequently will it be necessary to adjust the correction voltage.

Since the contact potentials vary with temperature, the zero point is in general dependent on the ambient temperature.

A second cause of disturbance is the *electrical "after-effect"* of the insulating material used in the vibrating capacitor. After exposure to an electric field many insulating materials develop a new potential difference at their surface as a result of the diffusion of charge carriers towards the surface. This effect can occur in a vibrating capacitor that forms part of an automatic control system: a disturbance of a magnitude sufficient to cause the system to work outside its control range may give rise to an abnormally high input voltage in the electrometer; if its insulators are subject to the after-effect mentioned, it may be quite some time before the normal state is restored.

As will appear below, the contact potentials and electrical after-effect have been taken into account in the design of our vibrating capacitor and the choice of materials.

Details of construction

Various insulating materials have been investigated in our laboratory for electrical after-effect [7]). Of the eligible materials, hard glass and quartz glass showed the least after-effect. Hard glass is used for

[7]) By H. A. Oele, at the time with this laboratory.

Time ← → | 1 hour | ←

Fig. 9

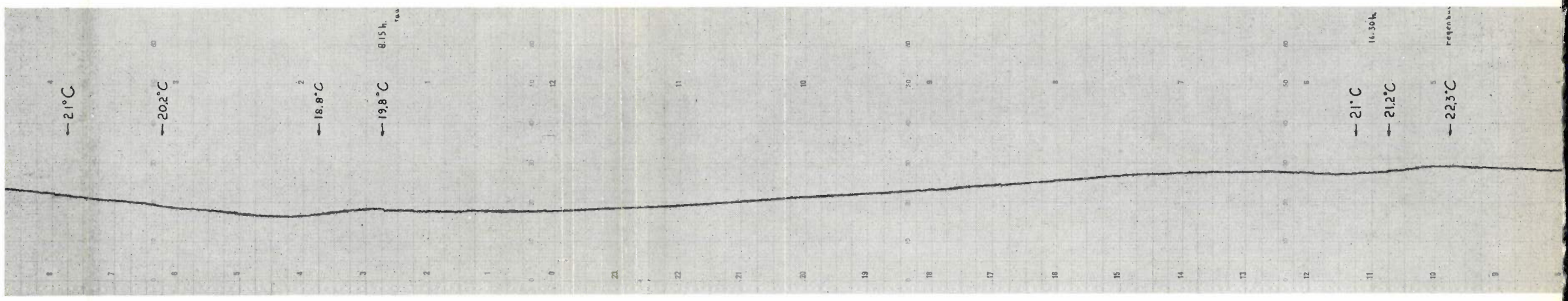


Fig. 11*a*

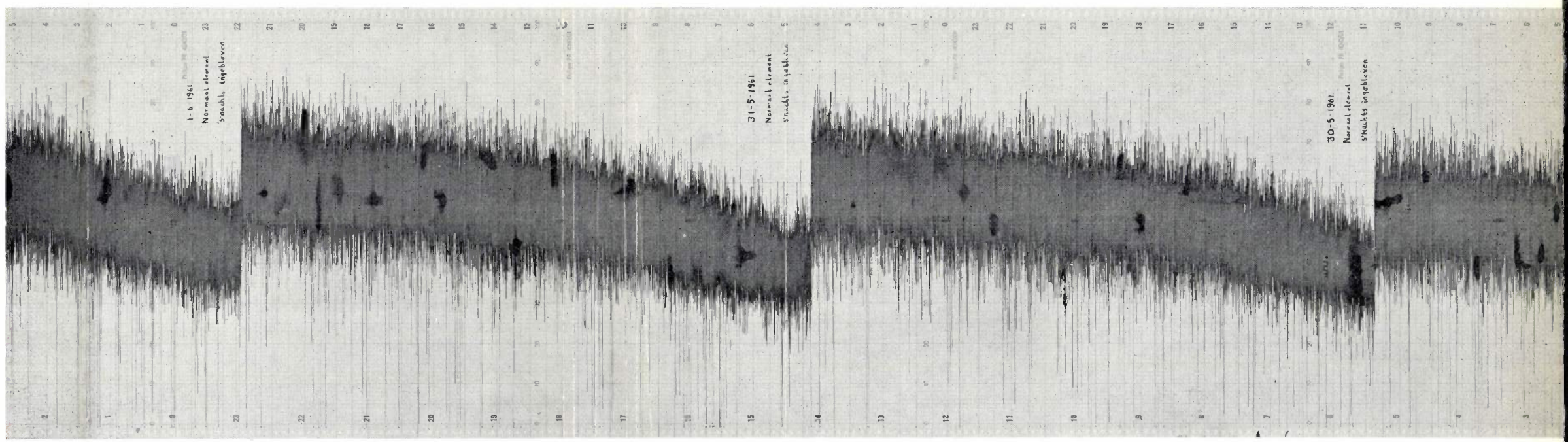
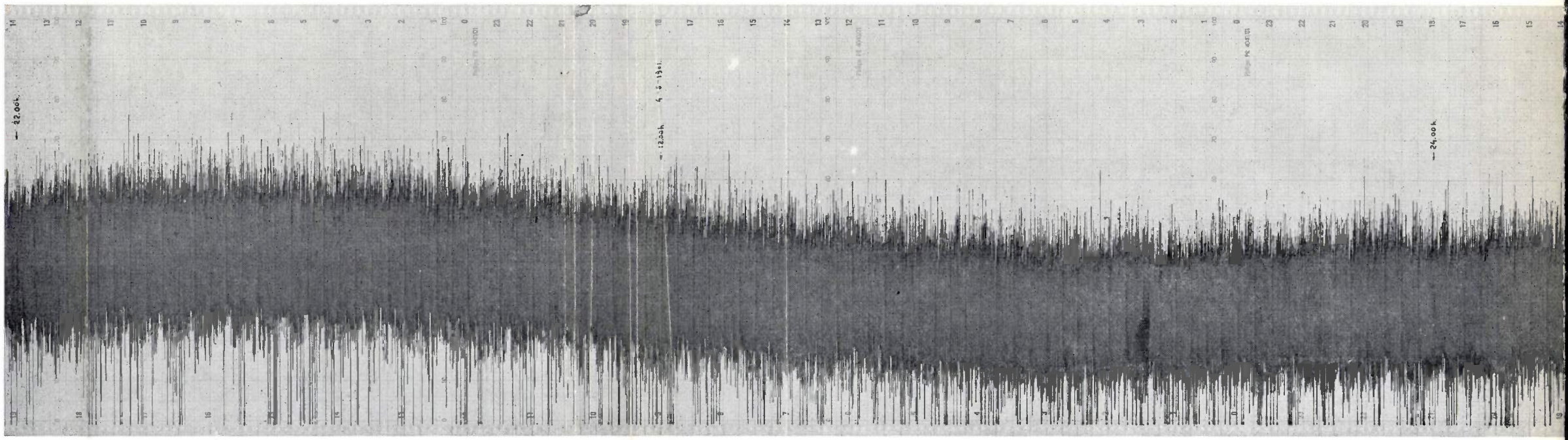


Fig. 11*b*



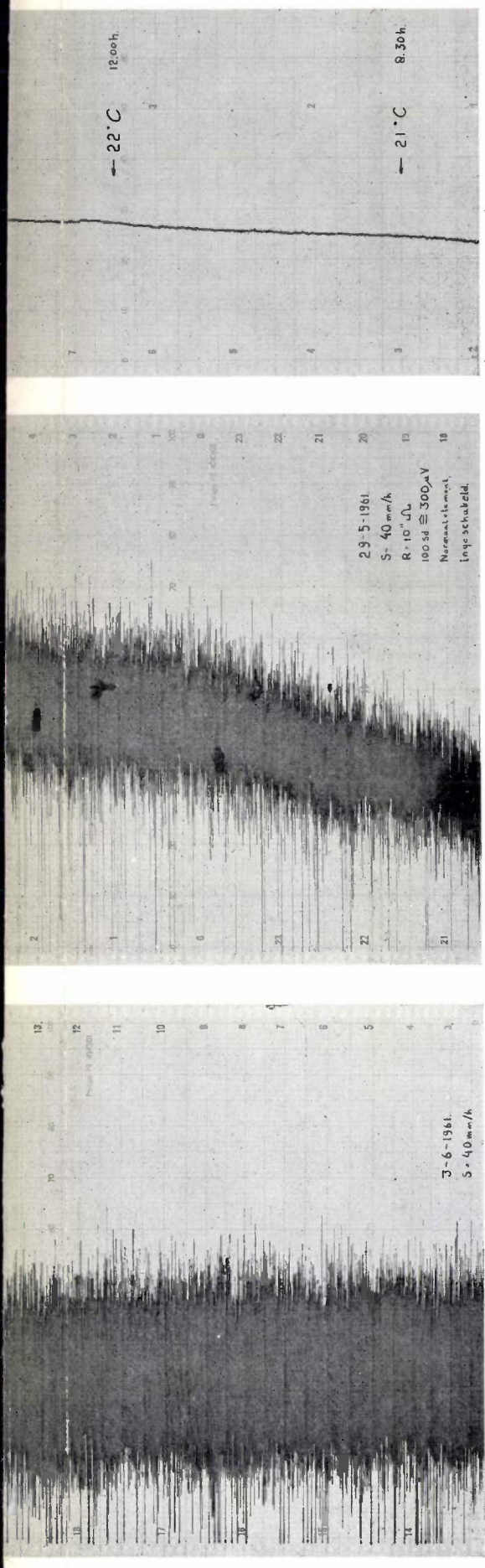


Fig. 9. Recording (during a period of more than 36 hours) of the output voltage with short-circuited input (R_s = 100 MΩ parallel with C_v). The noise is scarcely visible. The slight drift of the zero point is due to fluctuation of the ambient temperature and amounts to approximately 15 μV/°C.

the membrane and also for the insulators from which the vibrating capacitor is built up.

The *insulators* (I_1 and I_2 in fig. 3*a*) are made from centreless-ground hard-glass discs 3 mm thick and 18 mm in diameter (*fig. 10a*). In each disc a concentric groove is cut (G, fig. 10*b*) which lengthens the leakage path and also facilitates the machining operations.

In the border remaining outside the groove a radial slot is now sawn (H, fig. 10*c*) and on the rim of the disc, close to the slot, a glass bead P with a tungsten contact pin W is sealed [8]). The slot later serves for passing through the connection between the electrode and the contact pin.

Next, one face of the insulator is hollow-ground (fig. 10*d*) with a radius of curvature of more than 1.5 m. Finally, the border outside the groove G is ground again (fig. 10*e*). The deepest point of the concave middle section is now at a nominal 16 μm below the surface of the ground border, with a tolerance — thanks to the high precision of the grinding technique — as close as ±1 μm.

The membrane consists of hard glass, made from the same batch as the associated insulators so as to avoid differences in coefficient of expansion due to varying compositions. A disc 1 mm thick and 18 mm in diameter is polished on one face and this face is cemented to a flat plate. The other face is now ground with abrasive powder until the thickness of the disc is 0.135±0.0025 mm. The deviation from a plane-parallel surface is less than 2 μm. The resonant frequency of this membrane, at the given clamping diameter and a clamping force of 15 newton, is 6 kc/s ± about 10%.

[8]) In a later construction (fig. 2) the glass beads have been dispensed with.

Fig. 11. Recording made while the input was connected to a measuring resistance R_m of 10^{11} ohms; this explains the much stronger noise than in fig. 9.
a) Switched on nine hours a day four days in succession, switched off at night. The drift is mainly due to the temperature increase when the apparatus was first switched on. The first day the drift was somewhat greater because the standard element had been out of use for a considerable time. The following nights it was left switched on.
b) Left switched on for 32 hours after the temperature was stabilized. The drift is due to fluctuation in the ambient temperature and amounts, as in fig. 9, to about 15 μV/°C.

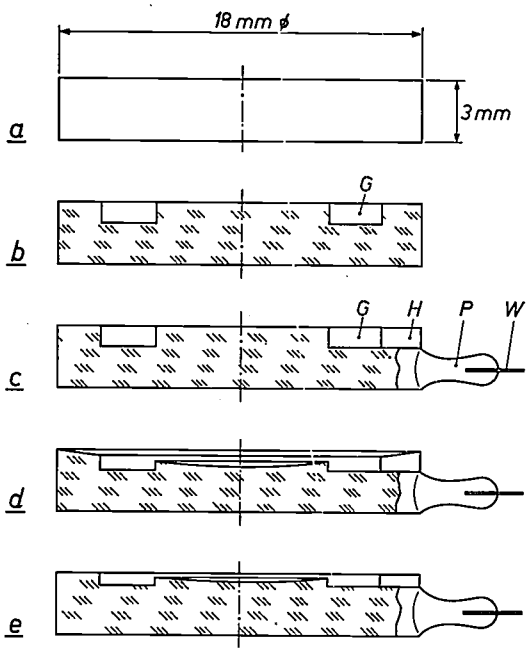


Fig. 10. Stages in the fabrication of one of the insulators (I_1 or I_2, fig. 3*a*).
a) Centreless ground hard-glass disc.
b) A concentric groove *G* is cut.
c) In the border a radial slot *H* is sawn, and on the rim a glass bead *P* with tungsten contact pin *W* is fused [8]).
d) One face is hollow- ground (radius of curvature about 1.5 m; the curvature is exaggerated here for the purpose of illustration).
e) The border outside the groove *G* is again flat-ground. The deepest point of the concave middle section is 16±1 μm below the surface of the ground border.

The hollow-ground part of the insulators and both faces of the membrane are now coated, by cathode sputtering, with the metal layers that serve as electrodes. The metal used is tantalum [9]). After sputtering, the layers are oxidized in an ozone atmosphere, so that a thin homogeneous oxide film forms on them, which is chemically and mechanically highly resistant and has exceptionally low absorption and adsorption. The constancy of the contact potentials is further improved by prolonged heating of the components during evacuation (down to a residual pressure of 10^{-5} to 10^{-6} torr.)

The greatest possible symmetry is observed in the sequence in which the various metals join on to one another in the system. Thus the contact potential variations caused by temperature fluctuations largely compensate each other. The result can be seen by glancing again at fig. 9. This recording of the output voltage extended over a period of 36 hours, during which the ambient temperature varied 3.5 °C; the output voltage returned to the input varied during this time by no more than 50 μV, representing a zero-point deviation of only about 15 μV per °C. It is noteworthy that, apart from this slight temperature effect, no other perceptible deviation occurred.

Fig. 11 shows two recordings, both of which relate to the case where the input was connected to a measuring resistance $R_m = 10^{11}$ ohms. The noise, now clearly visible, is attributable to this high resistance (1000 times greater than the resistance R_s involved in fig. 9). Half (150 μV) of the full deflection in fig. 11*a* and *b* corresponds at $R_m = 10^{11}$ ohm to a current of 1.5 times 10^{-15} A (about 10^4 electrons per second).

For the recording in fig. 11*a* the electrometer was switched on nine hours a day for four days in succession. The zero-point drift was due, apart from to the fluctuation of the ambient temperature, to warming up as a result of switching on, and amounted to about 60 μV. On the first day the drift was somewhat greater, the reason being that the correction-voltage source used for compensating the contact potential had then not been in operation for some considerable time. During the following nights this voltage source was left switched on.

The recording in fig. 11*b* shows the drift of the zero point during 32 hours in which the electrometer was kept continuously switched on after its temperature had become steady. The drift here was 60 μV, the fluctuation of the ambient temperature 4 °C. Here again, then, the zero point drifted 15 μV per °C, indicating that the electrical after-effect could not be noticed.

Summary. Description of a new vibrating capacitor for electrometers. A thin glass membrane is clamped between two glass insulators, the middle section of which is hollow-ground. These middle sections and both faces of the membrane are coated with a layer of tantalum. The layers constitute two capacitors, the capacitance of which varies periodically when the membrane vibrates. One capacitor is the actual vibrating capacitor, the other serves for the capacitive drive of the membrane and forms part of the oscillator, The latter generates a high-frequency voltage (1 Mc/s) which is amplitude-modulated at the natural frequency of the membrane (6 kc/s). Since the frequencies in the spectrum of the modulated voltage are much higher than the frequency of the alternating voltage into which the measured DC signal is converted, they can cause no interference. The relatively high natural frequency of 6 kc/s favours a low noise-level.

The two capacitors are contained inside an evacuated bulb. The contact potentials are low and exceptionally constant, and therefore easily compensated. Variation of the ambient temperature causes a zero-point drift of no more than about 15 μV/°C. For use in conjunction with the new vibrating capacitor an amplifier and an oscillator (for the drive) have been designed, both using transistors.

[9]) Proposed by J. H. J. Lorteije, of this laboratory.

RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF THE PHILIPS LABORATORIES AND FACTORIES

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

3066: K. H. Hanewald, M. P. Rappoldt and J. R. Roborgh: The antirachitic activity of previtamin D_3 (Rec. Trav. chim. Pays-Bas **80**, 1003-1014, 1961, No. 9/10).

3067: J. G. van Pelt: Determination of molecular weights with a semi-micro-ebulliometer (Rec. Trav. chim. Pays-Bas **80**, 1023-1028, 1961, No. 9/10).

3068: F. J. Mulder and K. J. Keuning: Spectrophotometric assay of α-tocopherol (Rec. Trav. chim. Pays-Bas **80**, 1029-1039, 1961, No. 9/10).

3069: B. G. van den Bos, C. J. Schoot, M. J. Koopmans and J. Meltzer: Investigations on pesticidal phosphorus compounds, IV. N-bis(dimethylamido)phosphoryl heterocycli (Rec. Trav. chim. Pays-Bas **80**, 1040-1047, 1961, No. 9/10).

3070: P. Westerhof and A. Smit: Investigations on sterols, XX. The synthesis and properties of 8α,10α-progesterone and 8α,10α-testosterone (Rec. Trav. chim. Pays-Bas **80**, 1048-1056, 1961, No. 9/10).

3071: J. H. Uhlenbroek: Preparation of diaryl sulphides (Rec. Trav. chim. Pays-Bas **80**, 1057-1065, 1961, No. 9/10).

3072: P. A. van Zwieten, J. A. van Velthuijsen and H. O. Huisman: Synthesis and physiological properties of some heterocyclic-aromatic sulphides and sulphones, I. Synthesis of some aryl-pyridyl sulphides (Rec. Trav. chim. Pays-Bas **80**, 1066-1074, 1961, No. 9/10).

3073: H. Koopman: Investigations on herbicides, IV. The synthesis of 2,6-dichlorobenzonitrile (Rec. Trav. chim. Pays-Bas **80**, 1075-1083, 1961, No. 9/10).

3074: C. W. Pluijgers and G. J. M. van der Kerk: Plant growth-regulating activity of S-carboxymethyl-N,N-dimethyldithiocarbamate and related compounds (Rec. Trav. chim. Pays-Bas **80**, 1089-1100, 1961, No. 9/10).

3075: J. L. M. A. Schlatmann and E. Havinga: Studies on vitamin D and related compounds, XVI. Synthesis of model compounds for the study of the previtamin D $\rightleftarrows$ vitamin D interconversion (Rec. Trav. chim. Pays-Bas **80**, 1101-1114, 1961, No. 9/10).

3076: P. H. van Leeuwen and H. O. Huisman: Investigations in the vitamin A series, V. Some aspects of dehydration of polyenic alcohols (Rec. Trav. chim. Pays-Bas **80**, 1115-1125, 1961, No. 9/10).

3077: J. Meltzer: Evaluation of the activity of some diphenyl compounds on winter eggs of the fruit tree red spider (Nature **192**, 474-475, 1961, No. 4801).

3078: J. Davidse: Modified N.T.S.C. colour T.V. signal for single-gun display systems (Electron. Technol. **38**, 388-392, 1961, No. 11).

3079: M. van Tol: Simple equation links stabilizing techniques (Control Engng. **8**, No. 12, 91, 1961).

3080: Th. G. Schut and W. J. Oosterkamp: Methoden zur Bildspeicherung (Automatik **6**, 487-490, 1961, No. 12). (Methods of image storage; in German.)

3081: C. J. M. Rooymans: The crystal structure of $LiScO_2$ (Z. anorg. allgem. Chemie **313**, 234-235, 1961, No. 3/4).

3082: J. Meltzer: Insecticidal and acaricidal properties of "Wepsyn" (Meded. Landbouwhogeschool Opzoekingsstat. Gent **26**, 1429-1434, 1961, No. 3).

3083: H. Rinia: Beschouwingen over een nieuw axiaal leger (Verslag gewone Vergad. Afd. Natuurk. Kon. Ned. Akad. Wet. **70**, 144, 1961, No. 10). (Some notes on a new axial bearing; in Dutch.)

3084: J. A. W. van Laar: Blistering of painted steel, I, II, III (Paint Varn. Prodn. **51**: No. 8, 31-37 + 88; No. 9, 49-52; No. 11, 41-44 + 97; 1961).

3085: J. S. C. Wessels: Reduction of dinitrophenol by chloroplasts (Biological structure and function, Proc. 1st IUB/IUBS int. Symp., Stockholm 1960, Vol. II, pp. 443-447, Academic Press, London 1961).

3086: J. B. de Boer and T. van Oosterom: Flight operational evaluation of approach and runway lighting (Ingenieur **73**, L 29-L 44 and L 45-L 54, 1961, Nos. 49 and 51).

3087: D. Kleis: Grondslagen en praktijk van toespreekinrichtingen, akoestische gezichtspunten (T. Ned. Radiogenootschap **26**, 191-216, 1961, No. 5/6). (Principles and practice of public address systems, acoustic considerations; in Dutch.)

3088: B. de Bruin: Grondslagen en praktijk van toespreekinrichtingen, elektrische gezichtspunten (T. Ned. Radiogenootschap **26**, 217-226, 1961, No. 5/6). (Principles and practice of public address systems, electrical considerations; in Dutch.)

VOLUME 25, 1963/64, No. 5 pp. 105-132 Published 27th May 1964

Philips Technical Review

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

OPERATIONS RESEARCH

by W. F. SCHALKWIJK *).

65.012.122

Since the second world war a branch of science now called Operations Research — but which in fact originated long before then — has flourished and accordingly become a focal point of general interest. In the following article an attempt is made, mainly through discussion of some typical examples, to give an impression of this branch of science.

There are two reasons why it is hard to give a definition of "Operations Research". In the first place it is, generally speaking, difficult to define a science clearly, particularly if it is still developing. And there is the additional factor that Operations Research is closely related to other branches of science, which have existed longer but beside which it nevertheless occupies an independent place. These other branches of science include business economics, mathematical statistics, cybernetics ("study of controls") and the investigations designated as "human engineering". Operations Research has in common with all these activities that it is a "study of control or guidance", dealing with the way in which organizations, large or small, should be controlled. The features distinguishing it from those branches of science are not easy to summarize, however: rather, a whole article is required, such as the present one. In the remainder of the article the name "Operations Research" will be abbreviated to the usual "O.R.".

With statistics O.R. has in common that in it mathematics, and especially probability theory, occupies a central position. It differs from statistics in that it analyses and compares possibilities which can be chosen at will, with the object of selecting the most economical. For this reason O.R. is sometimes called "besliskunde" (study of decisions) in the Netherlands. This term is not considered to define O.R. satisfactorily, however. O.R. is, rather, the scientific research which precedes decision making.

*) Philips' Research Laboratories, Eindhoven.

In the subsequent paragraphs we shall first say something about the origin and development of O.R. This will be illustrated by a specific example. Next we shall discuss, in somewhat more detail, three components of O.R. Finally, this will be followed by a brief survey of the remaining subjects belonging to the field of activities of this science.

Origin and development of O.R.

The advent of O.R. as a specialized branch of science to which many research workers devote their time was about the beginning of the second world war, but investigations in certain spheres now included in O.R. were conducted earlier. The research into waiting-times, for example, now an important component of O.R., started as long ago as 1909 with the study of telephone communications by the Dane Erlang [1]). In 1922 the American Camp published the formula for the optimum economical size of series in stock and production control, named after him [2]), and in 1916 the Englishman Lanchester wrote his book "Aircraft in warfare — the dawn of the fourth arm", which has become well known [3]). We shall pursue the first two of these subjects in the next two chapters. At this point we should like to consider Lanchester's work somewhat more closely,

1) A. K. Erlang, Nyt Tidsskrift for Matematik **B 20**, 33, 1909.
2) W. E. Camp, Determining the production order quantity, Management Engng. **2**, 17-18, 1922.
3) F. W. Lanchester, Aircraft in warfare — the dawn of the fourth arm, Constable, London 1916. Also see P. M. Morse and G. E. Kimball, Methods of Operations Research, Chapman & Hall, London 1951, Chapter 4.

since it strikingly illustrates the fact that O.R. can lead to results which may certainly be described as surprising.

One of the situations examined by Lanchester was the following. There are two groups x and y of objects, say, aircraft, firing at each other, and we assume that the decrease in each group per unit of time is proportional to the size of the other group. The problem is to calculate how x and y decrease during the entire course of the battle. From the above assumption follow the two elementary differential equations:

$$\frac{dx}{dt} = -a_1 y\,, \quad \frac{dy}{dt} = -a_2 x\,.$$

It is easy to eliminate dt from these equations. If, moreover, we suppose $a_1 = a_2$, then we get the simple equation:

$$x\,dx = y\,dy\,,$$

of which the solution is:

$$x^2 - y^2 = x_0^2 - y_0^2\,.$$

Here x_0 and y_0 are the initial sizes of the two groups. Let us now assume that at the beginning of the "operations" side x_0 were the more numerous ($x_0 > y_0$); then at the end, when y has become zero, the remaining size of x will be:

$$x_e = \sqrt{x_0^2 - y_0^2}.$$

The ratio of the total losses is then found to be:

$$\frac{\Delta y}{\Delta x} = \frac{y_0}{x_0 - x_e} = \frac{x_0}{y_0} + \sqrt{\left(\frac{x_0}{y_0}\right)^2 - 1}\,.$$

We see from this that not only the relative but also the absolute losses of the weaker side are greater. Take as an example $x_0 = 200$ and $y_0 = 100$. The strength of x will then still be 173 units when y has been reduced to zero. This shows the great importance of superior numbers very clearly indeed. This was known previously, of course, but it is O.R. which has proved quantitatively how important these effects can become. In the second world war the insight gained by means of O.R. played an important part. Sailing in convoys and operating with large concentrations of submarines were based on such insight.

The example taken from Lanchester demonstrates a general feature of the modus operandi of O.R., namely that a complex process is approximated or simulated by a simpler mathematical form which may be assumed to bear sufficient resemblance to that process. The actual process showed, in this example, the fundamental characteristic that probabilities play an essential role. In industrial processes often a more complex type of probability is involved than a simple, constant probability, such as that assumed in the example given. A well-known, more complicated, example of this kind is found in telephone communications. Suppose there is a small telephone exchange where, on an average, one call a minute is put through, either automatically or not. At an average duration of four minutes per call it might, perhaps, be thought possible to make do with a capacity of four simultaneous calls. In fact the capacity will have to be much greater if subscribers are not frequently to be kept waiting for their connections. Not only the fluctuating duration of calls but also their irregular arrivals, in particular, cause these difficulties. We shall return to this presently, when discussing the problem of waiting-times.

In dealing with problems in which statistical fluctuations play a part the theory of the stochastic processes occupies a central position. The name is related to the Greek word "stochasmos", which means guessing, or aiming. In that theory variables or systems which can successively assume a number of values or states are considered; the probability of occurrence of one of these states may then depend on the preceding states. In the case of a telephone exchange this state may be the number of calls taking place at a given moment.

Since mathematical treatment of the stochastic processes soon becomes very complicated, attempts are sometimes made to imitate (simulate) them by means of experiments. Hence an electronic computer can be programmed to make a selection, periodically, from a number of admitted possibilities (intervals, processing-times, etc.). In this way the process is reproduced, as it were. If the experiment is repeated a number of times, an idea as to what will happen in practice can be formed. Simulation of stochastic processes is called the Monte Carlo method. This name has been taken from a similar method, which was employed in 1943 for simulating a nuclear-physics process and which bore the code name Monte Carlo.

The sphere of O.R. includes, in addition to these processes, numerous problems for which a solution can be found by using geometrical or topological methods. We shall give a simple example in a subsequent chapter, where linear programming is dealt with. Generally speaking, we are concerned here with organizing an industrial activity as favourably as possible (planning). In particular, the *sequence* of certain operations may then be an important consideration.

We shall now pass to a more concrete treatment of some branches of O.R.

Waiting-times [4])

Many industrial and other activities take the form of a more or less regular flow of orders ("jobs"), which arrive at a certain point for processing. Examples are: calls received in a telephone exchange, orders in a workshop, ships to be discharged in a port, traffic having to pass a crossroads, etc. A processing point of this kind is generally called a "station" (or, sometimes, a "service point"). As we have already seen, there are two continuously fluctuating times which determine progress, viz the *intervals* at which the jobs come in and the *processing-times* ("service times") required.

The course of the arrival of the jobs can now be represented in a diagram, as indicated in *fig. 1*. On

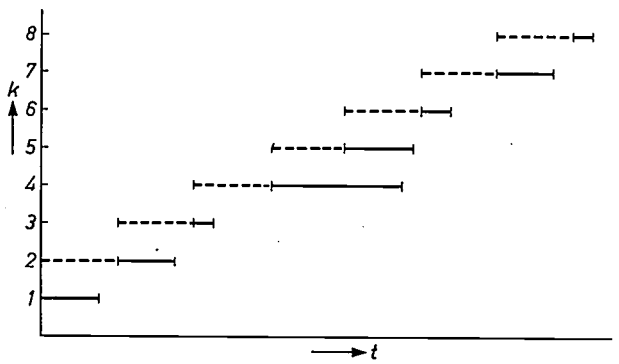


Fig. 1. Diagram of the arrival of orders ("jobs", with serial number k) at constant intervals (broken lines), but with various processing-times (continuous lines).

the vertical axis the consecutive number k of the orders coming in is shown, and both the interval and the processing-time are plotted horizontally. For simplicity it has been assumed here that the intervals, indicated by broken lines, are identical. In this particular case we immediately observe that if all the processing-times are shorter than the intervals, waiting will never be necessary ($k = 1$, 2 and 3 in the diagram). Each succeeding job does not arrive until its predecessor has been finished. Where a job takes longer to complete, however, such as No. 4, the next one arrives while No. 4 is still in hand. This means that No. 5 has to wait. As a result so much delay may occur that No. 6 as well, and possibly even No. 7, will have to wait. Not until the arrival of a few more jobs taking a short time will the waiting come to an end.

It is not hard to realize that the necessity for waiting (or queuing) may arise even where the *average* processing-time is shorter than the (constant) interval. Tbis is obviously due to the circumstance that the processing-times vary in length, i.e. show a *variance*. It will likewise be clear that the problem of waiting-times will become still more pressing if the intervals, too, show a variance. We shall now pursue this subject a little further.

It is at least approximately true that in practice the procedure concerned is usually one in which the moment of arrival of an order is entirely independent of that of the preceding or succeeding orders. If it is assumed, nevertheless, that the *average* duration of the intervals, viewed over successive long periods, is constant, then the distribution of the intervals over all possible values is found to be an exponential quantity. This exponential function is a special case of a more general distribution function known as a *Poisson* distribution [5]). In queuing theory this case is therefore sometimes called the Poisson arrival of the jobs.

Suppose that an average of m orders per unit of time, e.g. per hour, come in (the average interval is then $1/m$). If, now, the probability that, after the arrival of an order (job), the next arrival will fall in the interval between t and $t + \mathrm{d}t$ is called $p(t)\mathrm{d}t$, this probability is given, according to Poisson, by:

$$p(t)\mathrm{d}t = m\,\mathrm{e}^{-mt}\,\mathrm{d}t\,. \qquad \ldots\ldots \quad (1)$$

The function $p(t)/m = \mathrm{e}^{-mt}$ has been drawn in *fig. 2*.

Let us take as an example $m = 10$ arrivals, on an average, *per hour*. According to equation (1) the probability that, following an arrival, the next one

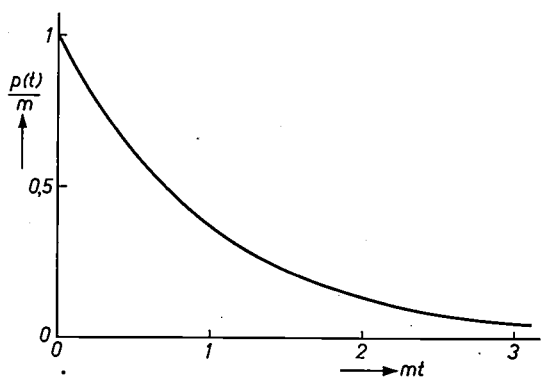


Fig. 2. For the occurrence of m random events per unit of time the probability $p(t)\mathrm{d}t$ that, starting from an arbitrary moment, the next event will take place in the interval between t and $t + \mathrm{d}t$ is given by $m\mathrm{e}^{-mt}\mathrm{d}t$. This distribution of probability is a special case of a more general distribution known as a Poisson distribution. The graph shows that *short* intervals between the events are the more likely. That the average interval $\bar{t} = 1/m$ is nevertheless rather long is connected with the relatively large contribution to $\bar{t}$ made by the fairly rare, very long intervals. It is a consequence of this distribution of intervals that randomness often gives the impression of systematic group-formation.

[4]) A concise survey of this field will be found in D. R. Cox and W. L. Smith, Queues, Methuen, London 1961.

[5]) See, for example, W. Feller, An introduction to probability theory and its applications, Wiley, New York 1961, Vol. I, Chapter 17.

will fall *in the first minute* is: 10/60 = 17%. The high degree of probability of a rapid succession of two arrivals is indeed a surprising result. This becomes still clearer if, on the basis of equation (1), the probability of a following arrival in, say, the *seventh* minute is calculated ($mt = 1$): the probability will then be found already to have decreased to 6%. From this we observe that in the case of random or Poisson arrival there is an apparent tendency towards "group-formation" in the arrival of jobs. That is a well-known phenomenon, which necessarily leads to allowance of liberal capacities for telephone exchanges, workshops, etc., if it is desired to keep the waiting-times short.

The general Poisson distribution indicates the degree of probability that in an interval t the number of orders coming in will be k. This probability is:

$$P(t,k) = \frac{e^{-mt}(mt)^k}{k!}.$$

If, in this equation, it is assumed that $k = 0$, then $P(t,0) = e^{-mt}$. In order to find the probability given by (1), we must first determine the degree of probability of just one arrival in a very short interval dt. This is: $P(dt,1) = m dt$, as was to be expected. According to the rule of composite probability, the probability that the following arrival will fall in the interval between t and $t + dt$ will then be:

$$p(t)dt = P(t,0)\, P(dt,1) = m\, e^{-mt}\, dt.$$

This holds for any period t, starting at a completely arbitrary moment, thus also from the time of each arrival. In this way the exponential distribution (1) has been derived from Poisson's general formula.

The aforementioned necessity will become even more obvious if we now combine an arrival of the jobs according to Poisson with a random distribution of the duration of the processing-times. In that case the average waiting-time $\bar{w}$ is given by the formula of Pollaczek-Khintchine [6]):

$$\frac{\bar{w}}{\bar{t}} = \frac{\varrho}{2(1-\varrho)}\left(1 + \frac{\sigma^2}{\bar{t}^2}\right), \quad \ldots \quad (2)$$

in which $\bar{t}$ is the average processing-time, ϱ the "utilization factor" of the station, i.e. the ratio between $\bar{t}$ and the average interval $(1/m)$ between the arrivals, and σ the variance in the processing-times. This variance is defined by:

$$\sigma^2 = \int_0^\infty \varphi(t)\,(t-\bar{t})^2\, dt,$$

in which $\varphi(t)dt$ is the fraction of the processing-times lying between t and $t + dt$.

[6]) See, for example, D. G. Kendall, Some problems in the theory of queues, J. Roy. Statist. Soc. **B 13**, 151-173, 1951.

In agreement with what has already been discussed, equation (2) shows that if the variance σ in the processing-time increases, so does the average waiting-time. From the relation follows, in addition, the remarkable conclusion that for a 100% degree of loading of the station ($\varrho \equiv \bar{t}m = 1$) the length of the waiting-times will become infinite. Hence here, once again, the need for overcapacity of the processing facility is apparent. In workshops this overcapacity can, if necessary, be profitably employed for work which is not urgent ("filler jobs").

A third important conclusion can be drawn from formula (2), in particular for an *exponential* distribution of the processing-times. If it is assumed that $\bar{t} = 1/n$, then for a distribution of this kind $\varphi(t)dt = n\, e^{-nt}\, dt$ applies in view of (1), and from the definition of σ^2 it follows accordingly that $\sigma = \bar{t}$. Hence formula (2) changes into the simpler form:

$$\frac{\bar{w}}{\bar{t}} = \frac{\varrho}{1-\varrho},$$

and it will now be seen that, for a fixed value of the utilization factor ϱ, the average waiting-time becomes proportional to the average processing-time.

This insight immediately permits us to assess the well-known possibility of combining two queues, each at its individual station, into one queue served by two stations which are "connected in parallel": the job first in line is handled by the first station to become available. For large values of ϱ, in which case there is usually a long queue, it can then be proved that the two stations function approximately as one, also having an exponential distribution of the processing-times but with *half* of the average processing-time. Since in this case the simpler form of equation (2) again applies, we observe that now the average waiting-time has also been halved. From this it can be concluded that a single large workshop will operate faster than two smaller ones having half of the equipment each. This is an example of the advantage of *concentration*. When the example borrowed from Lanchester was discussed, we already saw something of the kind. Such concentration or centralization is one of the fundamentals of modern business organization. Here limitations are imposed by factors other than those considered in the foregoing. Human peculiarities, for instance, may play a part; we have in mind what is known as Parkinson's law!

A slightly different example of the theory is that in which a mechanic has the task of maintaining and, if necessary, repairing a number of machines. It may then happen that a second machine breaks down while the mechanic is still occupied with a repair job.

With the aid of equation (2) it is now possible to estimate or calculate how long, on an average, the machines will be idle. This will also depend, of course, on the number and type of the machines. In this way the desirability of engaging a second mechanic can be investigated quantitatively.

Another form of waiting-time occurs where the orders come in irregularly, in the manner discussed, but all have to wait until a certain moment, in order to be processed *simultaneously*. This is encountered in all kinds of transport facilities: train, bus, aircraft, post, etc. Accordingly this type of waiting-time is referred to as *platform* or *stock* waiting-time.

Finally, a third type of waiting-time is found in cases where the service concerned is not given until a fixed quantity (batch) of orders are on hand. All the orders then have to wait until the last one has come in. This situation occurs in the conveyance of parties by motor coaches, for example. Hence the name *motor-coach* waiting-time. Another example is the delay occurring in the publication of books and magazines. For waiting-times of this kind, too, mathematical calculations can be derived.

Stock control [7])

In the foregoing we saw that in the case of fluctuating arrival of orders long waiting-times can be avoided by providing more stations or, if this can only be done to a limited extent, through provision of an overcapacity. The latter is, of course, rather uneconomical. In the case of manufacturing or repairing *identical* products or components, building-up of stocks is a suitable method of avoiding long waiting-times. This is, of course, impossible in the establishment of telephone connections, for example. Overcapacity is then imperative. Generally speaking, however, stocks will indeed be built up in the case of the manufacture of radio sets, incandescent lamps, etc. Such stocks, which serve to obviate waiting-times should fluctuations in demand occur, are called *buffer stocks*. They can serve to absorb both rapid statistical fluctuations (e.g. Poisson arrival) and slow fluctuations (e.g. seasonal influences).

There is, however, an entirely different reason why stocks are often built up in industry. This is connected with the fact that usually more than one product is manufactured in a factory or workshop. Each time there is a change-over from one product to another time losses and other losses occur. Machines have to be reset, there may be more rejects at first and, moreover, the change-over usually involves administrative charges. All these extra charges are either fully or almost independent of the size of the series to be manufactured. Such setting-, resetting- or initial charges are often called the *fixed costs*. Hence both this and the remaining, *variable*, costs always relate to *the whole series*, produced without interruption. (The two designations may sometimes cause confusion, since in relation to the cost *per unit of product* they would have to be chosen exactly the other way round!)

In order to keep the fixed cost relatively low, production of large series will be preferred, i.e. large stocks will be built up intermittently. (Such stocks, which in principle are not buffer stocks, are called *series-size stocks*). This cannot be carried too far, however: keeping stocks in hand itself entails costs. Stores have to be built and maintained, and the number of personnel required increases. In addition, the stocks represent noninterest-bearing capital. Added to this, there is the possibility of deterioration, unforeseen falling-off of demand or even unsaleability of the product. Generally speaking, very large stocks must be regarded as uneconomical. Hence it is probable that there will be *optimum sizes of series*, for which the over-all cost per unit produced reaches a certain minimum.

We should like to work this out in slightly greater detail, using somewhat idealized suppositions. For this purpose we introduce the following symbols:

F = total fixed cost per series when production is started;

C_i = storage cost per unit of product per unit of time;

D = the quantity of the product that is demanded per unit of time, briefly "the demand";

Q = the series size.

For simplicity let us assume that the demand D for the product is constant. The problem, then, is to express the optimum size of series, corresponding to the minimum total cost per unit of product, in F, C_i and D.

The *variable* cost per series, corresponding exactly to the cost per unit of product which remains constant (materials, piece rate, etc.), can be disregarded in the calculation. We are concerned only with the additional production cost per unit of product and the average storage cost per unit of product. For the former we can immediately write F/Q; in order to express the average storage cost in these terms we shall consider the fluctuation of the stock, which is represented in *fig. 3*. In this calculation it is assumed that the production time is very short compared to the maximum storage time Q/D and that, to avoid waiting-times, a new quantity Q is manufactured as

[7]) C. W. Churchman, R. L. Ackoff and E. L. Arnoff, Introduction to Operations Research, Part IV, Wiley, New York 1957.

soon as the stock has been exhausted. The average storage time will be half of the maximum, i.e. $Q/2D$, the average storage cost is thus $C_iQ/2D$ and for the total additional production cost and storage cost per unit of product we find the expression:

$$\frac{F}{Q}+\frac{C_iQ}{2D}.$$

This expression as a function of Q has the well-known form $ax + b/x$. It has a minimum for a certain value of x, or Q, which can easily be determined by differentiation. Hence for the optimum series size:

$$Q^* = \sqrt{\frac{2DF}{C_i}}. \quad \ldots\ldots \quad (3)$$

This formula was derived by Camp as long ago as 1922, but aroused little interest at the time [2]).

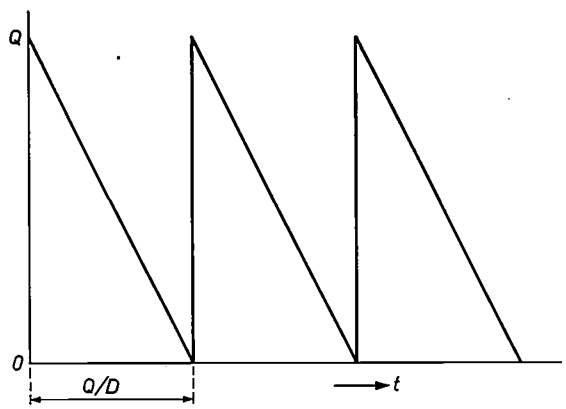


Fig. 3. The fluctuation of the stock of a product, as a function of the time, where a constant demand D is met from the stock and a series of a constant size Q is immediately manufactured every time the stock has been exhausted.

Both for derivation of the formula and for fig. 3 simplifying assumptions were made. Generally speaking, manufacture of a subsequent series will start before the entire stock has been used up, for example. This is desirable if the risk of being unable to supply at all at some time is to be avoided. Nevertheless equation (3) gives a sufficiently clear picture of the most efficient method.

As an approximation the formula can also be used for the case of platform waiting-time, already mentioned, e.g. in the literal sense with regard to the organization of rail transport. For that purpose the time coordinate in fig. 3 must be supposed to run from right to left. Q^* then corresponds to the *most economical train capacity*, D to the number of passengers per hour, F to the fixed cost of running a train and, finally, C_i to the expense incurred owing to loss of time, the need for waiting-rooms, etc., per passenger and per hour. This shows how in O.R. apparently quite different problems can be dealt with on the basis of the same model.

A certain similarity exists between the problem of the optimum series size and that of the most favourable switch-over time for *traffic lights*. In this connection also, some measure of reference can be made to fixed and variable losses of time. A certain time is required for setting the waiting line of vehicles in motion, for example. If the light changes rapidly from red to green and back to red again, the traffic cannot get going and the waiting-time becomes infinitely long. Should a very long switch-over time be introduced, however, the average waiting-time again becomes very long. Hence it is obvious that, dependent on the density of traffic, there will be an optimum switch-over time, in which the average waiting-time is shortest.

Linear programming [8])

In the foregoing we saw that O.R. makes it possible to determine the most economical value for a certain quantity which can be chosen at will (e.g. a series size). The case dealt with is an example of a process which may be more commonly referred to as *optimization*. Usually this process is concerned with the simultaneous optimization of more than one quantity, and we shall now give an example based on a case which may occur in practice and is easy to view as a whole. At the same time this illustrates a general method which has become well known under the name *linear programming* [9]).

We shall consider a factory manufacturing a relatively small quantity of a product every year, the demand for which shows a marked seasonal dependence. We shall assume that the normal annual productive capacity equals the total demand in respect of one year. The product being of large size, storage for a few months is expensive, and it may therefore be more economical to work overtime during certain periods, though that too entails additional expense. The possibility of overtime, which we shall again call overcapacity, is also limited, however.

The problem, then, is to determine to what degree overtime should be worked, and at what time of the year, to ensure that the total extra cost of overtime and storage is reduced to the minimum. Since this research soon becomes complicated, *Table I* gives a survey of the important data for a comparatively simple case. In the table the year is divided into periods of four months (terms). This seems somewhat unusual; the reason for it is to limit the number of variables to two, as will presently be apparent.

[8]) See the book mentioned in [7]), Part V.

[9]) The method was recently discussed 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 (No. 4/5).

Table I. Data relating to the production problem.

Term number	1 (Jan.-Apr.)	2 (May-Aug.)	3 (Sept.-Dec.)
Demand (turnover)	6	9	15
Normal productive capacity	10	10	10
Overcapacity	3	3	3

Let the additional cost entailed by overtime be a per unit of product and the cost of storage in the event of transfer to a succeeding term b per unit of product. Since no more than 13 units of product can be manufactured per term, in order to deliver 15 units of product in the third term *at least* 2 units of product will have to be made in the preceding terms. It is likewise clear, of course, that one of them will be manufactured in the second term, this involving no overtime. Hence we can write:

Term 1 — Production $= 6 + x$;
Term 2 — Production $= 9 + 1 + y$;
Term 3 — Production $= 15 - (1 + y) - x$,

to which the following restrictions will apply:

$$\left.\begin{aligned} 0 \leqslant x \leqslant 4; \\ 0 \leqslant y \leqslant 3; \\ 1 \leqslant x + y \leqslant 4. \end{aligned}\right\} \quad \ldots\ldots\ldots \quad (4)$$

The first restriction shows that it is pointless to work overtime in the first term, there being sufficient overcapacity in the second term; the second restriction relates to the extent of that overcapacity; the third restriction means that — naturally — at least 10 units of product will be made in the third term and that, as stated, at least two have to come from the preceding terms.

The cost of overtime will be:

$$a\{y + (4 - x - y)\} = a(4 - x),$$

and the cost of storage (or of transfer) is given by:

$$b(2x + y + 1) = 2bx + by + b.$$

Hence the total additional cost is:

$$\begin{aligned} E &= x(2b - a) + by + 4a + b \\ &= b\{(2 - a/b)x + y + 1 + 4a/b\}. \quad \ldots \quad (5) \end{aligned}$$

According to the problem set, we must so determine x and y that the cost function E, given by equation (5), is reduced to the minimum, with due regard to the restrictions (4). Since both (5) and (4) contain only linear functions of the unknown quantities x and y, the designation "linear programming" will be understandable.

In an x-y plane the restrictions (4) fix an admissible field which, being bounded by straight lines, has the form of a polygon. The boundary lines are given by the equations: $x = 0$, $x = 4$, $y = 0$, $y = 3$, $x + y = 1$ and $x + y = 4$. The polygon has been drawn in *fig. 4*.

Let us now first consider the special case $a/b = 2$, in which (5) is a function of y only. Evidently the extra cost E will then be minimum for $y = 0$. It follows from the third restriction that x may, according to choice, take one of the values 1, 2, 3 and 4. These points (x,y), all of which are equally favourable, lie on the bottom, horizontal boundary line of the polygon drawn in fig. 4. In practice this means that no overtime is worked in the second term ($y = 0$) and that it is immaterial whether overtime is worked in the third term or this production takes place in the first term. The relation $a = 2b$ expresses the fact that the storage cost in the latter case equals the cost of overtime for production in term 3.

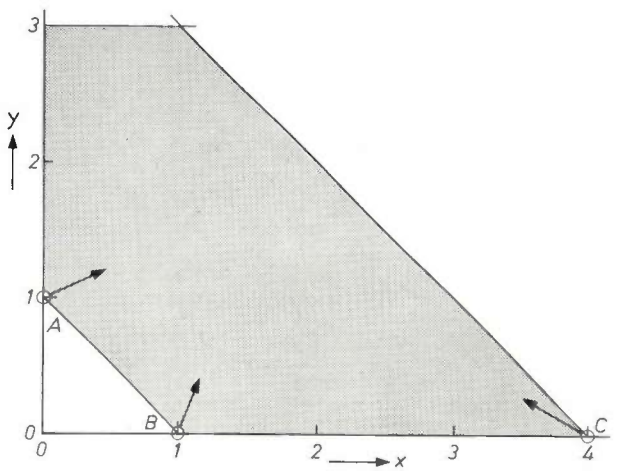


Fig. 4. Graph illustrating the process of linear programming. The aim is to distribute a certain production over three periods with various demands such that the sum of the additional cost of overtime (a per product) and that of maintaining a stock (b per product per period) is a minimum.
In the x-y plane (in which x and y can assume whole numerical values only) x and y represent the numbers of units of product in excess of 6 and 10, respectively, manufactured in the periods 1 and 2. Points $A(0,1)$, $B(1,0)$ and $C(4,0)$, all of which lie on the boundary of the "admissible field", correspond to minimum extra cost in the respective cases $0 < a/b < 1$, $1 < a/b < 2$ and $a/b > 2$.

In order to deal in the most rational way with the general case, with arbitrary values for a and b, we make use of an elementary result from the theory of linear functions, viz that the value of the expression $px + qy + r$ increases most rapidly for progress along a line whose slope (dy/dx) is q/p. This direction is perpendicular to the lines $px + qy + r =$ constant. Applied to the cost function (5), this means that the cost rises most rapidly if we vary x and y so that

$$\frac{dy}{dx} = \frac{1}{2 - a/b}. \quad \ldots\ldots \quad (6)$$

In this connection we can make a distinction between the following cases:

A) The cost of overtime a is lower than the storage cost b, with the result that $0 < a/b < 1$ applies. The direction given by (6) then makes an angle narrower than 45° with the x axis. Point A (0,1) indicated in fig. 4 is then the point of minimum additional cost, since within the admissible field it has an extreme position in relation to this direction. The arrow drawn at A shows the direction of the sharpest rise in cost.

B) The ratio between a and b is such that $1 < a/b < 2$ applies. The direction given by (6) now makes an angle of between 45° and 90° with the x axis. Similarly it can be seen that point B (1,0) is the one with the minimum extra cost.

C) Finally, the case $a/b > 2$ remains. The right-hand term of (6) then becomes negative, which indicates that the direction of the greatest increase in cost makes an angle wider than 90° with the x axis. Point C (4,0) now corresponds to optimum working-results.

The results are summarized in *Table II*: the most economical production per term is given as a function of the relation a/b between cost of overtime and cost of storage.

Table II. Optimum distribution of the production over the three terms, for the three cases.

Term number	1 (Jan.-Apr.)	2 (May-Aug.)	3 (Sept.-Dec.)
Case A ($0 < a/b < 1$)	6	11	13
Case B ($1 < a/b < 2$)	7	10	13
Case C ($2 < a/b$)	10	10	10

We have already seen that in the special case $a = 2b$, point B or C can be taken in fig. 4 according to choice, or alternatively, one of the two intermediate points (2,0) and (3,0). Something of the kind applies to the special case $a = b$, in which one of the points A and B can be chosen.

The example discussed shows how even in a simple case of business economics fairly complicated mathematical methods are involved, although the one dealt with here is of an elementary nature. In more complex cases with, say, more than two independent variables, the method will be more difficult and less convenient. Nevertheless, for these cases methods of ascertaining, in a finite number of steps, the most economical process are also available (Simplex method; transport method) [8] [9].

Treatment of all linear-programming problems is split up into two parts: analysis of the problem (i.e. introduction of the unknown quantities and indication of the restrictions) and the actual optimization (i.e. finding the most economical distribution or process). In the example considered we have seen that the second part — generally speaking, the more difficult — is a mathematical problem. This explains why, with regard to problems of this kind, reference is sometimes made to "mathematical programming" — unfortunately a misleading term at times, when it is remembered that electronic computers are often used to solve these problems and where "programming" has an entirely different meaning.

Other branches of O.R.

In those examples from the large field of O.R. that have been discussed certain simplifications were made, e.g. that the average time between two orders received at one station remains constant. We can likewise formulate this by saying that in fact only components of, or isolated phenomena in, industrial organizations were considered. In actual practice interactions or connections between the individual processes will, generally speaking, be present. A connection of this kind may be inherent in the whole system, but it may also be applied deliberately. A stock that is growing too large can be diminished by means of a reduction in price, for instance. Another example is that personnel can be transferred from a vacant station to one with a long queue. Hence, by taking deliberate action the management can make an organization operate more economically. The attendant phenomena are related to similar ones that were already known in control technique and in cybernetics. Here the term "industrial dynamics" [10] is sometimes employed. Undesirable and unforeseen fluctuations in industrial activities can be detected and counteracted by these means.

A branch of O.R. that must not be left undiscussed in an introductory article is *game theory* [11]. It deals mainly with the tactics a player can adopt in a game of chance (stakes, bids, etc.) and which may largely affect the result of the game. The object is to optimize tactics, whereby those of an opponent can either be taken into account or not. Such investigations may be regarded as a further development of the classic problems relating to games of chance, which have been well known from ancient times as an important subject of the calculus of probability. Those classic problems include the question of the

[10] J. W. Forrester, Industrial dynamics, Wiley, New York 1961.

[11] See, for example, J. D. Williams, The compleat strategyst, McGraw-Hill, New York 1954; Melvin Dolsher, Games of strategy: theory and applications, Prentice Hall, Eaglewood Cliffs 1961.

profitability of gambling-establishments (casinos). The strong position of an establishment of this kind, such as the one at Monte Carlo, is explained by the large capital at their command: generally the individual players do not have enough cash to endure a period of reverses, and, if they do have sufficient, limitation of the maximum stake allowed prevents them from taking advantage of it. Such insight fits in well with the foregoing explanation regarding the significance of superior numbers and of economic concentration. Optimization of tactics demands considerably more complicated mathematical aids than the classic problems referred to, however. At the same time the results obtained, where the underlying "rules of play" are suitably interpreted, have a far more general scope: applications of game theory to economic and military operations have already attracted attention.

All this still does not exhaust the fields covered by O.R. Without being able to aim at completeness we would mention a few more. An extension of the theory of waiting-times is arrived at if the possibility (important in practice) of affecting the production cycles by means of an optimum system of priorities is taken into account. This leads to entirely new theories. The planning-methods that have recently emerged for very complex projects (PERT, "network planning") likewise form a separate branch. Another extensive branch in which O.R. is making more and more contributions is the theory of communications of every kind. Here the theory of waiting-times of course plays a leading part, as indicated a few times in the foregoing, but many other effects also come into play. Finally, it may be observed that O.R. and subjects such as econometrics are drawing closer and closer together. Methods such as the aforementioned mathematical programming already constitute an accepted part of econometrics.

In the foregoing an attempt has been made, with the aid of examples worked out more or less in detail, to give an outline of Operations Research. Its main task is to discover, following analysis of the industrial activity, the most economical processes. Moreover, to sum up, it may be said that O.R. is a mathematically oriented branch of science, which links together long-existing forms of industrial and controlling activity. This aspect makes O.R. fit in well with every pursuit of unity of science.

Summary. Operations Research, which has flourished in the course of the past twenty years, is closely allied to business economics on the one hand and mathematical statistics on the other. The statistical element is introduced because of the fact that O.R. is always concerned with social activity subject to fluctuations, such as road traffic. Since the purpose of O.R. is not only to analyse but also to indicate optimum methods of working, it becomes at the same time a "study of control or guidance". Following an outline of the history of O.R., its modus operandi is illustrated with the aid of more or less detailed examples. These were taken from the theory of waiting-times, from stock-control theory and from the method known as linear programming. Reference is also made to game theory. By means of this consideration of O.R. an attempt has been made to show that its characteristic feature is the establishment of connections between branches of science which, formerly, were more or less isolated.

A TUBULAR FLUORESCENT LAMP WITH INCREASED LUMINOUS EFFICIENCY

by H. J. J. van BOORT *).

621.327.534.15

During the twenty years of its existence, the tubular fluorescent ("TL") lamp has been considerably improved, including an increase of its luminous efficiency from about 40 to about 70 lm/W. Applications of "TL" light were originally limited by the peculiar colour, but this has also improved considerably with the availability of new phosphors. The development of the "TL" lamps has now reached a stage where no further revolutionary improvements are to be expected; but, as appears from this article, refinements are still possible at a number of points. Improvements in one of the aspects of the "TL" lamp, viz the gas discharge, have recently made it possible to increase the luminous efficiency to 75 lm/W.

The development of the "TL" lamp

The first "TL" lamps, which were marketed about 1940, had a light output of about 40 lm/W. This was so much higher than the corresponding value for the incandescent lamp that the unusual colour and colour rendering, which differed both from that of daylight and from that of the incandescent lamp, were tolerated for the moment. Since then, "TL" lamps have undergone many improvements, most of them due to an increase in the quality of the phosphor covering the wall of the lamp, in which the ultraviolet light from the gas discharge is converted into visible light. The quality of the powder in this respect is determined by the colour and colour rendering of the emitted light, and also by the *quantum efficiency*.

By quantum efficiency we understand the ratio of the number of quanta emitted by the powder to the number absorbed by the powder. At present, the most commonly used phosphor is a *calcium halophosphate* activated by manganese and antimony, whose quantum efficiency has been gradually increased during the past twenty years to about 0.70. For this purpose, it was necessary to develop a method of preparation which gave a powder of sufficient purity, with the activators incorporated in the calcium halophosphate in the right amounts and with the right valency [1]). A high quantum efficiency is however not enough to guarantee the "TL" lamp a high luminous efficiency: part of the ultraviolet radiation is not absorbed by the powder, but is reflected back towards the discharge, where it is partially converted into heat. The total amount of reflected radiation increases with decreasing particle size of the powder, i.e. with greater reflecting surface [2]).

*) Philips Lighting Division, Development Department, Roosendaal, Netherlands.

1) W. L. Wanmaker, A. H. Hoekstra and M. G. A. Tak, Philips Res. Repts **10**, 11, 1955.

2) J. L. Ouweltjes, Elektrizitätsverwertung **33**, 293, 1958.

The luminous efficiency was thus also considerably improved when it was found possible to remove the fine grains from the powder by means of a "*hydrocyclone*". This device (*fig. 1*) consists mainly of a small conical vessel *A* (height e.g. 10 cm) with a side tube *B* through which the powder, mixed with a large amount of water, is pumped tangentially into the vessel. Under the influence of a complicated

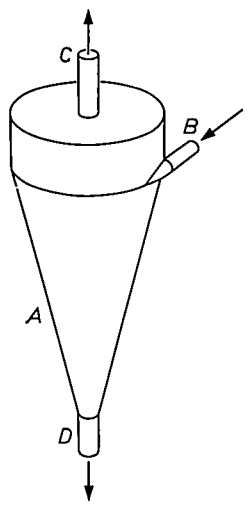


Fig. 1. Hydrocylone, used to remove the fine particles from the phosphor. The powder, mixed with a large amount of water, is pumped through the tube *B* into the vessel *A* (height e.g. 10 cm). The particles are separated into two fractions by a complicated interplay of forces in the body of water rotating in the vessel. Particles with less than a certain diameter leave the vessel through tube *C*, while the rest are led off through tube *D*.

interplay of forces in the rotating body of water, the particles of less than a certain size (which depends on the dimensions of the vessel) are driven to the middle of the vessel and are carried off by the stream of liquid through the axial tube *C*. The larger particles leave the vessel at the bottom with a smaller amount of water, through the tube *D*.

The possibility of changing the colour of the emitted light by slight changes in the composition of the phosphor has been made use of in producing a number of types of "TL" lamps, each of which satisfies certain requirements. The most important types are the "daylight" lamp, which can be used during the day-time to reinforce the daylight, the "warm white" lamp, which can be used together with incandescent

lamps, and the "white" lamp, whose colour is intermediate between those of the other two. Apart from the colour, the spectral composition of the light plays an important role, as this determines the colour rendering, i.e. the extent to which the colours of all sorts of objects look natural in the light in question [3]). The requirements for a good colour rendering are to a certain extent in opposition to those for a high luminous efficiency. The lamps on the market at present can be divided into two categories, the "*standard*" lamps and the "*de luxe*" lamps, in each of which the three colour types "daylight", "white" and "warm white" are represented. In the standard lamps, not too much stress is laid on the colour rendering, so that the *luminous efficiency* can be made *as high as possible*. In the "de luxe" lamps, on the other hand, some of the luminous efficiency is sacrificed in the interests of a very good colour rendering.

In contrast to the almost continual development of the phosphors, the gas discharge in which the ultraviolet radiation is produced has altered but little. Now that there has been of recent years a tendency to increase the luminous efficiency of "TL" lamps as much as possible, we found it advisible to investigate carefully whether the gas discharge offered any further possibilities of increasing the luminous efficiency.

It was found that this is indeed so: a gas discharge not only produces ultraviolet radiation but also a certain amount of heat, and this amount is more than is needed to maintain the discharge. We have found it possible to take a number of measures which reduce the excess heat produced, with the result that the luminous efficiency has been increased by about 4%. These measures will be applied first of all in the white standard lamp, whose efficiency has already been increased by about 3% by a recent improvement in the phosphor. The luminous efficiency of this lamp has thus been increased by about 7% in all, to 75 lm/W. (This figure refers to the lamp alone, i.e. not to the lamp with ballast.)

We shall now discuss in detail the various measures taken to limit the heat losses. Since some insight into the operation of the gas discharge in the "TL" lamp is necessary for this purpose, we shall begin with a brief description of this.

The gas discharge

The voltage across a 40 W "TL" lamp which is burning steadily is about 100 V. The way in which this voltage is distributed over the length of the lamp is shown in *fig. 2*. Relatively large voltages are concentrated within a few mm of the electrodes: the cathode fall of 5 V in the region KB and the anode fall of 8 V in CA. In the region BC, the "positive column", the potential increases linearly, so that the field strength is constant here.

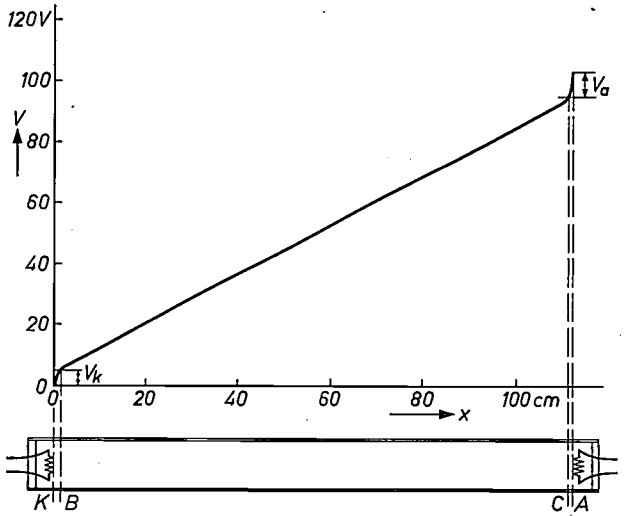


Fig. 2. The variation of the potential V in a 40 W "TL" lamp with the distance x to the cathode K. A is the anode. In the range BC (the positive column), where the desired mercury radiation is produced, the variation is linear so that the field strength is independent of x. Relatively large voltages are concentrated in the regions KB and CA: 5 V at the cathode (cathode fall) and 8 V at the anode (anode fall). The "TL" lamp shown under the graph is not drawn to scale.

The electric current through the lamp (about 0.4 A) is mainly carried by the free electrons emitted by the hot cathode and moving from there to the anode. A certain contribution is also made by the positive mercury ions, which move relatively slowly in the opposite direction. Despite the accelerating action of the electric field, the average velocity of the electrons in the positive column is constant. This means that an electron loses on the average as much energy in its collisions with the atoms of the gas filling — which consists practically entirely (99.8%) of e.g. argon, the rest being mercury — as it gains from the field in the interval between two collisions. These collisions may be elastic or non-elastic. The elastic collisions cause the temperature of the gas to rise until equilibrium is established with the loss of heat to the surroundings. This situation is similar to that of the conduction of electricity in metals, where the electrons are continually being retarded by the atoms of the lattice and thus transfer part of their energy to the lattice in the form of heat. Contrary to conduction in metals, conduction in gases gives rise to the *neutralization of charge carriers*, the electrons and positive ions diffusing to the walls of

[3]) See e.g. A. A. Kruithof and J. L. Ouweltjes, Colour and colour rendering of tubular fluorescent lamps, Philips tech. Rev. 18, 249-260, 1956/57.

the lamp and there recombining to form neutral atoms. This loss is made up for in the positive column by the continual ionization of mercury atoms as a result of the above-mentioned non-elastic collisions between mercury atoms and fast electrons. Since the number of ionizations per second depends strongly on the velocity of the electrons, equilibrium between the loss of charge carriers by recombination and the formation of new ones by ionization is only achieved at a very definite value of the field strength. The discharge can thus only be stable when the potential distribution in the positive column has adjusted itself so that the field strength has the desired value. In the lamp in question, the field strength is 0.8 V/cm.

Although the above-mentioned processes (charge transport, emission of heat, recombination and ionization) are essential for the maintenance of the discharge, they are only incidental to the main purpose, which is the excitation of the ultraviolet mercury radiation (2537 Å) by means of non-elastic collisions between electrons and mercury atoms. We may wonder in this connection why the lamp contains so much argon when it is the mercury radiation we are concerned with. Now, of course, it is possible to bring about a gas discharge in an atmosphere composed entirely of mercury, but when the mercury is present at the pressure needed to ensure a reasonable yield of 2537-Å quanta, a large part of the radiation produced is lost again by *self-absorption*. By self-absorption we understand in the present case that a 2537-Å quantum emitted by a mercury atom may be intercepted by another mercury atom on its way to the wall of the lamp, exciting the latter. This quantum may of course be re-emitted, but there is a certain chance that it will be converted into heat in one way or another, and in that case it no longer contributes to the production of light. If we reduce the mercury pressure, the chance of self-absorption is reduced, but the mean free path of the electrons is increased. If the mercury pressure is chosen to have a suitable value with respect to self-absorption, the mean free path is so great that the electrons only meet a few atoms between cathode and anode, so that the excitation falls off to practically nil. The addition of argon reduces the mean free path, without causing the electrons to lose an appreciable part of their energy to the argon by non-elastic collisions: there are very few electrons which have enough energy to excite or ionize an argon atom. As a result of the presence of the argon atoms, the electrons follow zigzag paths whose total length is hundreds of times greater than the distance between cathode and anode, so that the chance of collisions between electrons and mercury atoms is much increased. Not too much argon must be added, however, since the electrons do lose a certain part of their energy each time they collide elastically with an argon atom. It is true that this part is very small, but when the number of collisions becomes very large an appreciable amount of energy is still lost as heat to the surroundings. Usual values of the partial pressures of argon and mercury in the steadily burning lamp are about 2.8 and about 0.006 torr respectively.

We have already mentioned that the potential in the positive column varies linearly. This indicates that there is *no space charge* in this part of the lamp, in other words that the number of electrons per cm^3 is equal to the number of positive ions. (Since the average velocity of the ions is hundreds of times smaller than that of the electrons, the ions do not make an appreciable contribution to the electric current.) In each of the regions KB and CA (fig. 2), however, there *is a space charge* of ions near the cathode and electrons near the anode, and thus there is a greater potential drop (cathode and anode fall respectively) and a stronger electric field than in the column. At the cathode, this strong field accelerates the ions so that on arrival at the cathode surface they can give up enough energy to keep the cathode at the temperature required for the emission of electrons. The cathode is coated with an emitting material, e.g. a mixture of BaO, CaO and SrO, which ensures a low value of the work function of the electrons, so that a relatively low temperature (about 1250 °K) is sufficient. When the lamp is burning steadily, the potential distribution and thus the ion velocity near the cathode has adjusted itself so that the temperature is reached at which the emission of electrons is just sufficient to maintain the discharge. The cathode fall thus has a useful function. The anode fall, on the other hand, only reduces the luminous efficiency: the strong field near the anode accelerates the electrons towards this electrode, so that when they give up their kinetic energy to the anode it becomes unnecessarily hot.

It should have become clear from this description that the conduction of electric current through the gas is brought about by an extremely complicated interplay of all sorts of factors. We have mentioned, for example, three equilibria: between the disappearance and production of charge carriers, between the heat absorbed and evolved by the gas and between the ion bombardment and electron emission of the cathode. The lamp can only burn steadily if these three equilibria, and others which we have not mentioned, are maintained and moreover suitably in-

terrelated. If a "TL" lamp is connected directly to a normal voltage source, this steady state is never reached: the current just keeps on increasing until some part of the lamp breaks down. A "TL" lamp must therefore always be used in series with a current limiter, which may consist of a coil of sufficient self-inductance when the lamp is fed from the AC mains. Naturally, with an AC supply, the electrodes play the part of cathode and anode alternately.

The mercury radiation coming from the discharge is largely absorbed by the fluorescent powder on the wall of the lamp, and thus converted into visible light. This aspect of the "TL" lamp — which is an important one in fact — will not be discussed any further in this article.

Increasing the luminous efficiency

Reduction of the pressure of the gas filling

Fig. 3 shows the variation of the luminous efficiency of a "TL" lamp as a function of the pressure of the gas filling, at a constant mercury pressure.

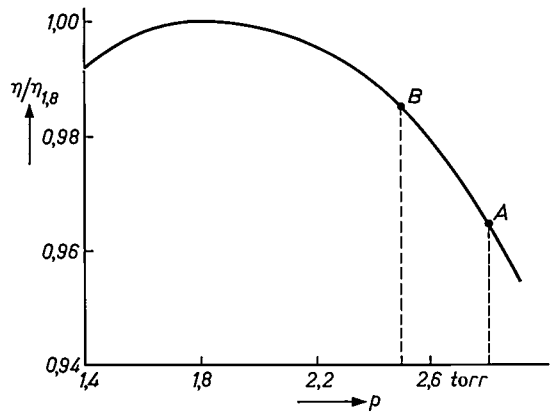


Fig. 3. The luminous efficiency η, expressed as a fraction of the maximum value $\eta_{1.8}$, as a function of the pressure p of the argon filling. By reducing the pressure from the normal value of 2.8 torr (point A) to 2.5 torr (B), the luminous efficiency has been increased by about 2%.

The efficiency has a maximum at about 1.8 torr. To the right of this point, the efficiency falls because the electrons lose an increasing amount of their energy to the gas filling as a result of elastic collisions. The decrease in efficiency with decreasing pressure, to the left of the maximum, is the result of the increase in the mean free path. Both effects have been touched on in the previous section.

It will be seen that the pressure used until now (2.8 torr) does not correspond to the maximum possible luminous efficiency. This is connected with a subsidiary function of the gas filling, not mentioned so far, viz limiting the mean free path of the *ions in the cathode fall*. Without this limitation, the ions would strike the cathode with such high velocities as to cause violent sputtering of the emitting material. Under these conditions, the lamp would have a very short life. The choice of the argon pressure is thus strongly dependent on how resistent the cathode is to ion bombardment. Until recently, it was not possible to reduce this pressure below 2.8 torr for this reason, while in the past a much higher argon pressure had to be used. We are now able to improve the situation still further by use of an electrode construction which can contain more emitting material and moreover retain this material better, so that the protection of the gas filling is no longer so necessary (for further details of this new electrode construction, see the next section). This makes it possible to reduce the pressure of the gas filling to 2.5 torr, thus increasing the luminous efficiency by about 2%. This electrode would in fact be able to stand the ion bombardment at even lower pressures, e.g. 2.2 torr, but then other undesirable phenomena arise whose discussion would take us beyond the scope of this article.

Ring round the electrode coil

As we have already mentioned, potential differences of about 5 and 8 V are concentrated near the electrodes of a burning "TL" lamp. In the cathode fall, $5 \times 0.4 = 2$ W of electrical power is converted into heat. Most of this is used to keep the cathode at emission temperature; the rest is lost. The power dissipated at the anode, $8 \times 0.4 = 3.2$ W, does not contribute at all to the maintenance of the discharge and must thus be considered as a pure loss. Now it has been found quite empirically that these electrode losses are reduced by placing a *metal ring* round each electrode, insulated from the latter (*fig. 4*). The increase in the luminous efficiency associated with this reduction in the electrode losses amounts to about 2%. Very little is known about the influence of the ring on the discharge, so that the real cause of this improvement in the efficiency cannot be stated. Nor can we explain an undesirable side-effect, viz, the somewhat more rapid consumption of the emitting material in the presence of the ring. Because of this latter effect, the ring has so far only been used in rare cases. It is now possible to use this ring in normal "TL" lamps, if we also make use of the improved electrode mentioned above. The life of the lamp then continues to meet the demands made on it.

This ring also catches the sputtering products from the cathode, so that these are no longer deposited on the ends of the lamp. The advantage of this is that the *blackening* which used to occur is now

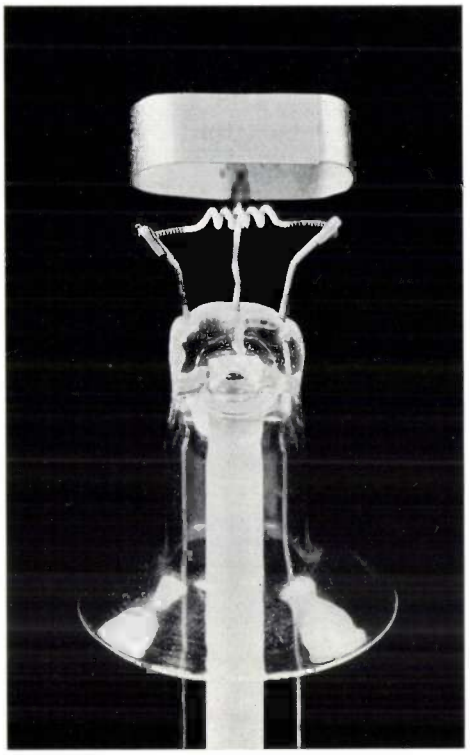


Fig. 4. If the electrode coil is surrounded by an insulated flat metal ring, the "electrode losses" are found to decrease. The luminous efficiency thus increases by about 2%. — In order to make the coil itself visible in the photo, the ring has been bent backwards. Magnification about 1.5×.

avoided, so that the lamp retains its original brightness over its whole length throughout its life.

Adjusting the power consumption

The measures described above reduced certain losses: the loss of heat to the surroundings and the electrode losses. This means that the new lamp, used in conjunction with a standardized ballast, will give more or less the *same* amount of light as the old model, but will need *less power* to produce this much light. It has in fact been found that the power consumption is now 38.6 W instead of 40 W.

We would like to bring the power consumption back to 40 W, without changing the dimensions of the lamp. This can be done by replacing some of the argon gas filling by *neon* [4]). Neon atoms are much smaller and lighter than argon atoms, so that when some of the argon is replaced by neon the mercury ions are hindered less in their movement. As a result of this the mercury ions diffuse in greater numbers to the wall of the lamp where, as we have seen, they combine with electrons to form neutral atoms, so more charge carriers are lost in this way. This extra loss is compensated for by a slight shift in the equilibria described above, which causes the field strength in the positive column to increase somewhat, so that the number of ionizations increases. This increase in the field strength causes the voltage across the lamp as a whole to rise, and when the partial pressure of neon is suitably chosen (0.7 torr) the power consumption once again becomes 40 W. Fortunately, at this partial pressure the presence of the neon has no adverse effect on the luminous efficiency (at higher pressures it does have an adverse effect).

The greater mobility of the mercury ions, which as we have just mentioned causes the extra loss of charge carriers, also causes the bombardment of the cathode by mercury ions to be more violent in the argon-neon mixture than in pure argon at the same pressure, which leads to still stronger sputtering of the emitting material. The new electrode is however so great an improvement on the old that this reduction in the life due to the partial replacement of argon by neon is also compensated for. In the section which now follows we shall describe the construction and properties of this electrode.

Modification of the electrode construction

While the electrode constructions used until now have been based on a "coiled coil" of tungsten wire, the new one is based on the "*triple coil*". This form of electrode, which has been known for some time [5]), has been dimensioned by us so that it can be used in a 40 W "TL" lamp.

To make this electrode, a tungsten wire (diameter 65 μm) and a molybdenum wire (diameter 110 μm) are placed side by side and a very thin (18 μm) tungsten wire is wound round both of them (*fig.* 5). This compound "wire" is wound into a coiled coil with the aid of a molybdenum mandrel and a steel

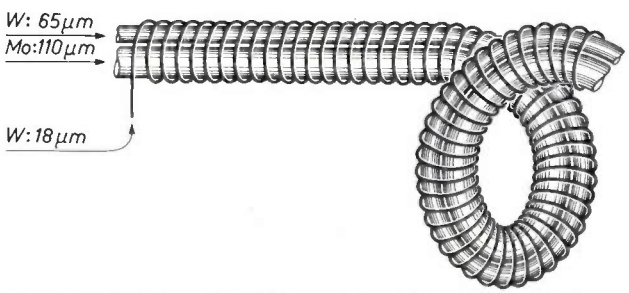


Fig. 5. A "triply coiled" filament is obtained by winding a compound "wire" into a coiled coil. The compound wire consists of a tungsten wire (diameter 65 μm) and a molybdenum wire (110 μm) placed side by side, and a very thin (18 μm) tungsten wire wound round both of them. The molybdenum wire is then dissolved away in a strong acid, leaving the thin tungsten wire coiled very loosely round the thick one.

[4]) This method of increasing the power consumption of a gas discharge, which has been known for a long time, was also used for a tubular fluorescent lamp by D. D. Hinman and R. S. Fox, Illum. Engng. **56**, 222, 1961.

[5]) J. O. Aicher, U. S. Patent No. 2 306 925.

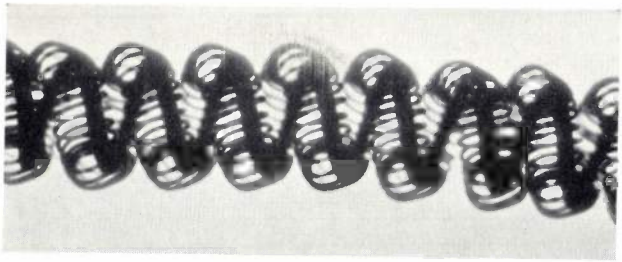


Fig. 6. Photo of part of a triple coil not yet coated with emitting material. Magnification 10×. (This picture shows a straight part of the coil.)

mandrel. The steel mandrel is then withdrawn and the two molybdenum wires are dissolved in strong acid (*fig. 6*). The name triple coil refers to the thin tungsten wire that is coiled *three times*. This thin tungsten wire wound loosely round the thicker wire (the primary tungsten mandrel) ensures that the emitting material applied to the electrode stays in place better than on a single wire. Moreover, if we make the diameters of the tungsten and molybdenum primary mandrels large enough, a triple coil can contain about 50% more emitting material than a normal coiled coil. Although the ion bombardment of the cathode is much heavier in the new lamp than in the old one, these favourable properties of the triple coil give it at least as long a life.

The diameter of the tungsten core wire has been chosen so large because the discharge current flows through part of this wire, giving rise to Joule losses which reduce the luminous efficiency and which we would therefore like to limit as much as possible. The reason for this is that the ends of the coil cannot be covered with emitting material, since material there would not be heated sufficiently during the degassing of the lamp. While burning the discharge is now concentrated just at the edge of the emitting layer next to the pole P_1 through which the current leaves the lamp after flowing through the intervening uncovered length of core wire (*fig. 7*). The Joule losses thus caused are the more undesirable because they gradually increase during the life of the lamp: sputtering of emitting material from the place where the discharge is concentrated causes the discharge to move gradually further from the above-mentioned pole, as a result of which the luminous efficiency of the lamp shows a gradual though slight decrease.

The choice of a thick core wire thus reduces both the original Joule losses and their gradual increase. The thickness is limited by a subsidiary condition which must also be satisfied. In most circuits for a "TL" lamp, the electrodes are pre-heated during ignition by passing a current through the coil. The resistance per unit length of the electrode coil must therefore be sufficient to allow the electrode to reach the desired temperature within the allowed time.

By choosing the diameter of the core wire as large as possible, we have ensured that the Joule losses in the coil are no greater than in a coiled coil. The triple coil introduced to maintain the life of the lamp does not therefore prevent us from taking full advantage of the measures to improve the luminous efficiency and to adjust the power consumption. These measures, together with the improvement of the quality of the phosphor mentioned at the start of this article, have allowed the luminous flux of the 40 W white standard "TL" lamp to reach a value of 3000 lm.

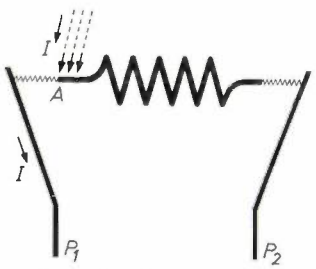


Fig. 7. In a new electrode, the discharge is concentrated at the junction A between the part of the coil coated with emitting material and the uncoated part (see also fig. 4) at the side of the pole P_1 through which the discharge current I leaves the lamp. This is because the current through a burning lamp causes a potential gradient along the electrode coil, and the bombardment by positive ions is concentrated at the spot with the lowest potential where emitting material is still to be found. The pole P_2 is only connected into the circuit when a current has to pass through the coil for pre-heating it.

Summary. During the last few years the development of the "TL" lamp has been mainly confined to the phosphors. Careful investigation of the other aspect of this lamp, the gas discharge, has shown that this too offers a number of possibilities of increasing the luminous efficiency. The loss of heat to the surroundings and the electrode losses are reduced respectively by lowering the pressure of the gas filling and by surrounding the electrode coils by an insulated metal ring. Thanks to these reduced losses, a lamp which in previous form used 40 W for a given luminous flux now uses 38.6 W for the same luminous flux. The power consumption can be brought back to 40 W without reducing the luminous efficiency by replacing part of the argon gas filling by neon. All these three measures increase the cathode sputtering, but the influence of this effect on the life of the lamp can be suppressed by replacing the normal coiled-coil electrodes by "triple coils", suitably dimensioned for use in a 40 W "TL" lamp. Such a coil retains the emitting material better, and can moreover contain 50% more of this material than a coiled coil. If these changes are combined with a recent improvement in the fluorescent powder in a 40 W white standard "TL" lamp, they allow the luminous efficiency of this lamp to reach the value of 75 lm/W.

PHYSICAL PRINCIPLES OF PHOTOCONDUCTIVITY

by L. HEIJNE *).

537.312.5:537.311.4

I. BASIC CONCEPTS; CONTACTS ON SEMICONDUCTORS

Interest in photoconductivity has increased considerably in recent years. This is due not only to the numerous technical applications of photoconductive materials but also to the fact that the photoconductive effect has proved to be a useful aid to the study of the properties of semiconductors. A series of three articles in this journal will deal at some length with the general aspects of photoconductivity. It will be shown how the phenomena concerned can be explained in terms of the band theory of solids. Special attention will be paid to the manner in which the photoconductive properties depend on the lifetime of the charge carriers, on the recombination mechanism, on the presence of various types of impurity centres, on the nature of the contacts and on the possible occurrence of space charge.

The article below, the first of the series, deals with the most important basic concepts, and discusses the influence of fitting contacts on a photoconductor.

Introduction

As the name implies, photoconductivity is the effect whereby the electrical conductivity of a solid changes under the action of light. This effect, which was first observed in 1873 on selenium [1]) and which subsequently found application in the well-known selenium cells, has again become increasingly prominent in the past 15 years. There are two reasons for this. In the first place, advances in the preparation of very pure substances and the deeper insight into the electronic processes taking place in solids have made it possible to manufacture in a reproducible way photoconductive cells (photoresistors) of high sensitivity, suitable for use in diverse spectral regions. In the second place, measurements of photoconductivity and of phenomena associated with it have yielded useful information on many important properties of semiconductors or insulators, such as the lifetime and mobility of charge carriers, the "depth" of impurity centres, and so on.

Some practical aspects of photoconductivity have been dealt with in earlier articles in this journal, examples being the application in a television camera tube [2]) or in a solid-state image intensifier [3]), and the preparation and properties of photoconducting cells of cadmium sulphide [4]) and of indium antimonide [5]).

In three articles, of which this is the first, we shall examine some general and fundamental aspects of photoconduction, and attempt to illustrate the physical background of various well-known properties of photoconductors. Among the subjects considered will be the sensitivity of photoconductors and their speed of response, the dependence of the photocurrent on the wavelength of the incident light, the influence of temperature, and the consequences of the fact that a piece of photoconducting material can only be incorporated in a circuit by providing it with metal contacts. The relation between photoconductivity and luminescence will also emerge. No account, however, will be taken of processes associated with chemical changes in the substance, such as those occurring in materials used for photographic emulsions [6]).

The present article will be concerned with some basic concepts and the influence of contacts; the second will deal with the influence of impurity centres on the lifetime of the charge carriers and on the speed of response (for cases with one and with more than one type of impurity centre), and in the third some special effects will be discussed.

Our considerations will be based on the insight into the properties of semiconductors which has been gained in recent years from the study of the transistor materials germanium and silicon. In the series, particularly in this first article, the treatment of the various subjects will often begin by considering semiconductors in general, after which they will be extended or confined to the case where the material also shows photoconductivity.

*) Philips Research Laboratories, Eindhoven.

1) W. Smith, Nature **7**, 303, 1873.
2) Philips tech. Rev. **16**, 23, 1954/55 and **24**, 57, 1962/63.
3) Philips tech. Rev. **19**, 1, 1957/58.
4) Philips tech. Rev. **20**, 277, 1958/59.
5) Philips tech. Rev. **22**, 217, 1960/61.
6) For a more comprehensive treatment of photoconductivity, see: R. H. Bube, Photoconductivity of solids, Wiley, New York 1960, or: T. S. Moss, Photoconductivity in the elements, Butterworths, London 1952. A review article by F. Stöckmann appeared in Z. angew. Phys. **11**, 68, 1959.

Energy levels of the electrons in a solid; the band scheme

Just as electrons in isolated atoms cannot possess any arbitrary energy, nor can they when the atoms are united to form a crystalline solid. Whereas in an isolated atom, however, the energy of an electron can have only a few discrete values, in a solid the possible energy levels are very numerous. Usually these levels fall into groups, called the *allowed energy bands*, which in a perfectly pure crystal are separated by zones in which there are no energy levels at all: the *forbidden zones*. The differences in energy between the neighbouring levels in an allowed energy band are extremely small, and such a band may broadly be regarded as a continuum of possible energy values. According to Pauli's exclusion principle, however, the number of electrons contained in a band cannot be greater than the number of levels of which the band is made up. If these numbers are equal, the band is completely filled. In that case there can be no electron movement and therefore no electric conduction. In an insulator the bands are either completely filled or completely empty; a metal contains one band which is only partially filled.

Electrons occupying a level in a partially filled band are able, in general, to take up a small amount of energy from an applied electric field, and thus to move freely. This band is therefore termed the *conduction band;* the next band below is called the *valence band.*

In semiconductors the occupation of energy levels is nearly the same as in insulators but now either some levels at the bottom of the conduction band are occupied or some at the top of the valence band are empty (or both conditions occur at the same time). In the first case the electron can, as described, move freely under the action of an applied field; this case is referred to as N-type conductivity. In the other case the conduction can best be described by considering not the motion of the electrons in the valence band but that of the few vacant places which an electron might still have occupied, called *holes*. This case is referred to as hole conduction or P-type conductivity, since the holes behave as positive charge carriers.

The energy states of the charge carriers are conventionally represented in what is termed a *band scheme*, a diagram in which the energy of the electrons is set out vertically. *Fig. 1* shows the band schemes of an insulator and of a metal.

As will later be explained, the occupation of the energy levels is not entirely independent of temperature. It is interesting to consider the case of a solid in which the conduction band is entirely empty at absolute zero. If the temperature is now increased, electrons are able, by acquiring thermal energy, to rise from the valence band into the conduction band and to remain there for some time. Such a transition can take place more easily the smaller is the width of the forbidden zone, called the *energy* or *band gap*, and the higher is the temperature. The substance which, at absolute zero, was an insulator has now become at finite temperature an *intrinsic* semiconductor, that is to say the electrons contained in the conduction band all originate from the valence band. The latter band therefore contains an equal number of holes. Disregarding small spontaneous fluctuations ("noise"), in thermal equilibrium the density of electrons and holes is constant because, on average, as many conduction electrons continuously recombine with holes as are newly generated.

The equilibrium densities of the electrons and holes are no longer equal if the crystal contains *impurity centres*, e.g. foreign atoms. Some of the various functions of these impurity centres will be dealt with later in this article. Principal among them are the properties of "donating" and "accepting" electrons. There are *donor centres*, i.e. impurity centres that can give up electrons (often an atom containing more valence electrons than the atom of the parent lattice whose place it takes), and *acceptor centres*, which take up electrons or supply holes (e.g. atoms with fewer valence electrons). An electron bound to such a centre may possess an energy which lies in the forbidden zone of the pure crystal. An energy value of this kind is generally called an "impurity level" or, more specifically, a donor or acceptor level. If the energy level of an electron bound to a donor centre

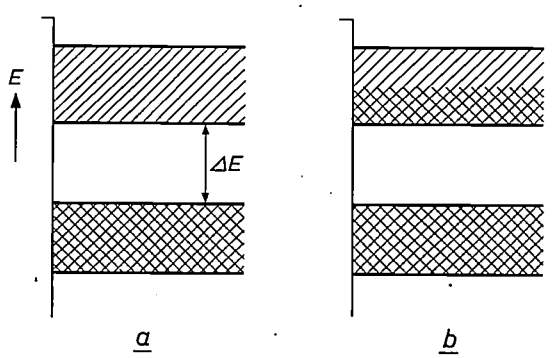


Fig. 1. Energy band scheme of an insulator (*a*) and of a metal (*b*). The electron energy E is set out vertically; the top of the ordinate axis gives the energy of an electron which is just outside the material. The horizontal coordinate has no physical meaning here, but may be used to represent a space coordinate. Of the various allowed bands those shown are the highest of the entirely filled bands (cross-hatched), called the *valence band*, and the *conduction band* immediately above it. The forbidden zone in between, of width ΔE, is called the *energy* or *band gap*. In an insulator the conduction band is completely empty (single hatching), in a metal about half filled.

lies just below the conduction band, very little thermal energy will be sufficient to raise such an electron into the conduction band; at room temperature such a donor level is as a rule unoccupied. If the impurity level lies deeper, thermal excitation will seldom occur. Similar reasoning applies to impurity centres which supply holes; the energy difference between the level and the valence band must be taken into account here. Semiconductors in which the charge carriers primarily originate from impurity centres are called *extrinsic* semiconductors [7]). The band scheme of a semiconductor with impurity centres can be seen in *fig. 2*.

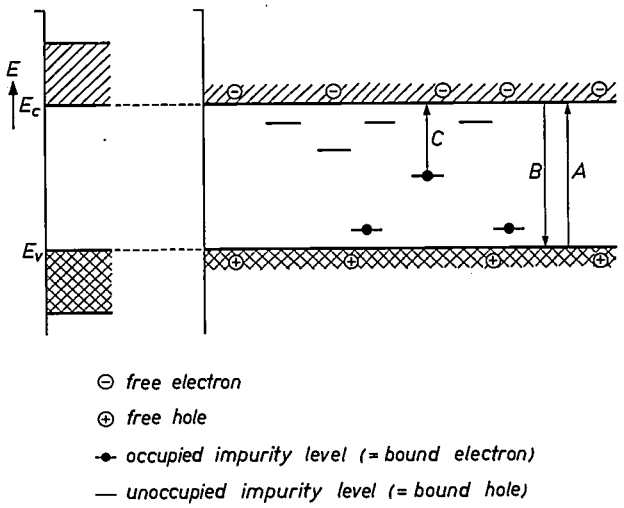


Fig. 2. Band scheme of a semiconductor containing impurity centres. The drawing indicates only the lower limit of the conduction band ($E = E_c$) and the upper limit of the valence band ($E = E_v$); see the sketch on the left. Energy levels corresponding to impurity centres are conventionally indicated by horizontal dashes, representing symbolically that electrons or holes occupying such an energy level are physically localized. Relatively high impurity levels are largely unoccupied, low impurity levels are largely filled.

The arrows A and B represent respectively a generation (excitation) and a recombination; C represents an excitation from an impurity level. The difference compared with an insulator is not expressed by the hatching but by indicating a few electrons at the bottom of the conduction band and/or a few holes at the top of the valence band.

On statistical grounds it can be shown that in thermal equilibrium the product of the concentrations of electrons and holes is independent of the presence of impurity centres and simply a function of the temperature and of the energy gap.

Finally it may be remarked that impurity centres can have a marked effect on the speed with which equilibrium is restored or a new steady state sets in after a disturbance. This point will be dealt with at some length in the second article of this series.

[7]) The physics of semiconductors is dealt with extensively by W. Shockley in: Electrons and holes in semiconductors, Van Nostrand, New York 1950, and by E. Spenke in: Electronic semiconductors, McGraw-Hill, New York 1958.

A profound analogy exists between the "spectrum" of the allowed energy values of the band scheme and that of the discrete energy levels that can be occupied by the electrons in a single atom: each band corresponds to one such discrete energy level. This may be understood in qualitative terms as follows. We first consider a crystal whose atoms are so far apart that they have no influence on each other. For the electrons in this crystal the diagram of allowed energy values would be identical with the atomic. If the atoms are brought closer together, interaction becomes perceptible: the electrons are no longer solely in the field of their own atom but are also influenced by the fields of the other atoms. As a result each energy level splits up into numerous components. The energy difference of the sublevels increases with decreasing interatomic spacing, but in an absolute sense it is always extremely small. Since moreover the number of sublevels is very large, the aggregate of sublevels originating from one level can be regarded as a quasi-continuous region of allowed energy values.

Photoconductivity

The transition of an electron from a low level to a higher one can be brought about not only by thermal excitation but also by the absorption of a light quantum in the solid (internal photoelectric effect). Although the same applies equally to X-ray or gamma quanta, we shall confine ourselves in the following to the quanta of infrared, visible and ultraviolet radiation. Plainly, light quanta possessing an energy lower than the energy gap will not be able to cause a transition from the full to the empty band. In the optical absorption spectrum of the substance this appears as an *absorption edge*, i.e. an abrupt discontinuity in the absorption spectrum at a definite wavelength: light with a shorter wavelength (greater quantum energy) is strongly absorbed, light with a longer wavelength is transmitted (*fig. 3*). At longer wavelengths than that of the absorption edge, however, absorption bands can still occur as a result of the excitation of electrons from impurity levels (fig. 2, arrow C).

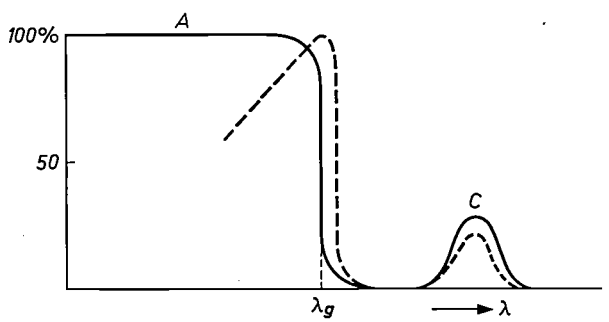


Fig. 3. Light absorption (solid line) and photocurrent (dotted line) as a function of the wavelength of the light incident on a photoconductor. The maximum value is put at 100% for both. The region A of the absorption spectrum corresponds to the excitation of electrons out of the valence band (transition A in fig. 2); the quantum energy of light of wavelength λ_g, at which the absorption edge occurs in the spectrum, is equal to the energy gap $E_c - E_v$. The peak C corresponds to excitation of electrons out of an impurity centre, the energy level of which lies in the forbidden zone (transition C in fig. 2).

In consequence of these excitation processes, which take place when the substance is exposed to radiation, the concentration of free electrons and holes is increased, causing an increase in electrical conductivity. This is the phenomenon of photoconductivity. Usually the photocurrent is strongest for light of roughly the wavelength of the absorption edge, and at longer wavelengths corresponding peaks appear as in the absorption curve as a result of excitation from impurity levels (see the dashed line in fig. 3).

The decrease of photoconductivity when the wavelength becomes considerably shorter than that of the absorption edge is a surface effect. This will be dealt with in the third article.

Fermi-Dirac distribution function

The probability $f(E)$ that an electron will occupy a certain quantum state of energy E is given by the Fermi-Dirac distribution function:

$$f(E) = \frac{1}{1 + \exp\,[(E-E_F)/kT]}. \quad . \quad (I,1)$$

In the quantum statistics of particles subjected to Pauli's exclusion principle this expression takes the place of the well-known Boltzmann distribution [8]). The quantity k is Boltzmann's constant and T is the absolute temperature; E_F is the Fermi limit or Fermi energy, which to a first approximation is independent of temperature. From (I,1) it can be deduced that, at absolute zero ($T = 0$), all levels for which $E<E_F$ will be occupied: $f(E) = 1$, and all others unoccupied: $f(E) = 0$. At higher temperatures the situation is scarcely any different. Although the value of f does not now change abruptly from 1 to 0 but changes gradually, the change largely takes place in a narrow energy region around $E = E_F$. The width of this region is greater the higher the value of T (*fig. 4a*). For $E = E_F$ the occupation probability $f(E)$ is now 0.5.

Formula (I,1) is wholly valid for the occupation of energy levels in solids, on the understanding of course that an electron can have no forbidden energy value. In metals E_F lies roughly in the middle of the conduction band (fig. 4*b*). In semiconductors and insulators E_F lies in the forbidden zone between the valence and the conduction band (fig. 4*c*). In nonmetals, at not very elevated temperatures, the entire region of E values within which $f(E)$ rapidly decreases is well within the forbidden zone. This, then, confirms the earlier assertion that in a semiconductor or insulator the conduction band is empty and the valence band completely filled.

Quite apart from calculating the occupation of energy levels, E_F is important in another connection. In a closed system in thermal equilibrium there can (by definition) be only one value of E_F. If two crystals of different Fermi energy are brought into contact with each other, processes take place at their junction such that, after equilibrium is restored, E_F is identical in both crystals, which now of course form a single closed system. In the band scheme of these crystals in contact with each other the hori-

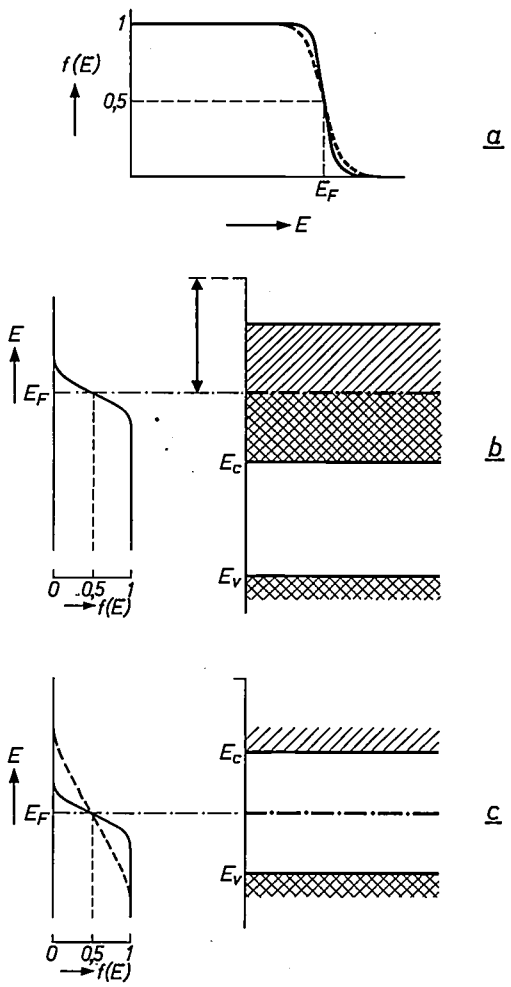


Fig. 4. *a*) The occupation $f(E)$ of the energy levels in accordance with the Fermi-Dirac distribution function. Except in a narrow transition region around $E = E_F$, all levels with $E < E_F$ are occupied by electrons and all levels with $E > E_F$ are unoccupied. The transition region is wider the higher the temperature (dashed line).
b) Band scheme of a metal. The Fermi level here lies roughly in the middle of the conduction band. The distance indicated by the double-headed arrow represents the *thermionic work function*.
c) Band scheme of an insulator. The Fermi level lies in the forbidden zone. If the energy gap $E_c - E_v$ is not too wide, at high temperature the transition region of the function $f(E)$ has sufficient width (see dashed line) for a few electrons to occupy the bottom levels of the conduction band. The insulator has then become an intrinsic semiconductor.

[8]) For a treatment of quantum statistics see e.g. R. C. Tolman, Statistical mechanics with applications to physics and chemistry, Chem. Catalog Co., New York 1927, or L. D. Landau and E.M. Lifshitz, Statistical physics, Pergamon Press, London 1958. See also J. Volger, Solid-state research at low temperatures, part I, Philips tech. Rev. 22, 190-195, 1960/61, especially p. 192.

zontal line marking the energy $E = E_F$ must be at the same height on both sides of the vertical line representing the junction. This directly indicates the mutual situation of the bands, since the *relative* situation of E_F in each crystal remains unaffected. In the following we shall frequently use this rule.

Since the value of E_F, as may be deduced from the foregoing, depends on the situation of the energy levels and the number of electrons to be contained, in an N-type semiconductor (excess donor centres) E_F will be higher than in the pure material, and in a P-type semiconductor (excess acceptor centres) E_F will be lower.

The energy level occupation of an *illuminated* semiconductor, i.e. one in which the electron concentration in the conduction band and the hole concentration in the valence band have been artificially increased, may also in some cases approximately be described by a distribution function of the form of (I,1), provided another value, called the quasi Fermi energy, is substituted for E_F. The value of this energy is *not* the same for holes and electrons. We must therefore reckon with *two* quasi Fermi levels, the one for electrons being above the ordinary Fermi level, and the one for holes below it.

The energy E_F corresponds to the (electro)chemical potential of electrons known in thermodynamics. In a closed system which is in thermal equilibrium, this potential is everywhere identical.

It may be pointed out that the Fermi-Dirac distribution function (I,1) holds for a "gas" of free electrons. The electrons moving in the periodic potential field of a crystal may be formally treated as free provided we substitute pseudo or "effective" quantities for various physical quantities needed to describe their behaviour. This applies e.g. to the mass, the density of states, etc.

With the aid of (I,1) one can easily prove that, as mentioned above, the product of the respective concentrations n and p of conduction electrons and holes depends only on the energy gap $(E_c - E_v)$ and the temperature. If the top of the valence band contains N_v levels per cm³ — this being the *effective* density of states just referred to — and the bottom of the conduction band contains N_c, then provided E_F is not too close to E_v or E_c (i.e. when $n \ll N_c$ and $p \ll N_v$) we may write:

$$n = N_c \exp\left[-(E_c - E_F)/kT\right], \quad \ldots \quad \text{(I, 2a)}$$

and

$$p = N_v \exp\left[-(E_F - E_v)/kT\right]. \quad \ldots \quad \text{(I,2b)}$$

From this it follows that:

$$np \sim N_c N_v \exp\left[-(E_c - E_v)/kT\right]. \quad \ldots \quad \text{(I, 3)}$$

The quantity N_c is given by the relation:

$$N_c = 2\left(\frac{2\pi\, m_n^*\, kT}{h^2}\right)^{3/2}. \quad \ldots \quad \text{(I, 4)}$$

Here h is the Planck constant, and m_n^* the effective mass of an electron in the conduction band. If we substitute m_p^* (the effective mass of a hole) for m_n^*, we obtain the expression for N_v. Since these effective masses are constant, np indeed depends only on $(E_c - E_v)$ and T.

Lifetime and mobility of charge carriers

An electron freed by irradiation or thermal excitation will recombine with a hole after a certain time. The average time τ which elapses between the moments of generation and recombination is termed the lifetime of the charge carrier. It is easy to see that the increase Δn in the steady state concentration of the charge carriers caused by constant irradiation will be greater the longer is τ. The lifetime τ therefore determines the sensitivity of a photoconductor, i.e. the extent to which the conductivity changes when the intensity of illumination is varied (a more accurate definition is not needed for our purposes).

The relation between τ and Δn can quickly be found when it is considered that, in a steady state, the number of charge carriers generated per unit time must equal the number that recombine in that time. Let G be the number of electrons freed per unit volume and per second. The average number of free electrons recombining per second is equal to $\Delta n/\tau$. We therefore find

$$G = \Delta n/\tau, \text{ or } \Delta n = G\tau. \quad \ldots \quad \text{(I, 5)}$$

The extra concentration is thus directly proportional to τ. A similar formula applies to holes. It may be noted here that τ is not always constant but sometimes may itself depend on the intensity of illumination. This point will be dealt with in article II. For the moment we can disregard this complication.

If a block of the semiconducting material is provided with two electrodes and we apply a voltage, the free electrons will move with a mean velocity v which is proportional to the field strength F produced in the material. Expressed as a formula:

$$v = -\mu F. \quad \ldots \quad \text{(I, 6)}$$

The proportionality factor μ is called the mobility of the electrons; the minus sign expresses that the electrons move in the opposite direction to the field. The current density j_f of the photocurrent, which is equal to $-\Delta n e v$, follows from formulae (I,5) and (I,6):

$$j_f = -\Delta n e v = \mu \tau e F G \quad \ldots \quad \text{(I, 7)}$$

(e being the absolute value of the charge of the electron). The magnitude of the share of electrons in the photocurrent is therefore determined not only by the field strength and light intensity but also by the product $\mu\tau$; this quantity is characteristic of the substance. It should be remembered here, however, that μ sometimes depends on the field strength, just as τ may depend on the light intensity.

To distinguish we shall give quantities relating to electrons the suffix n, and those relating to holes the suffix p.

A current of charge carriers can be produced by a concentration gradient as well as by an applied field. Such a *diffusion current*, which is a fundamental feature of the behaviour of semiconductor diodes and transistors [3]), is proportional to the magnitude of the concentration gradient. The proportionality factor, the diffusion coefficient D, is related to the mobility of the charge carriers by Einstein's relation:

$$D = \frac{kT}{e}\mu. \quad \ldots\ldots \quad (I, 8)$$

If we consider an electron current flowing in one particular direction (x), the complete current equation is thus:

$$\Delta j_{n,x} = \Delta n\, e\, \mu_n F_x + eD_n \frac{d(\Delta n)}{dx}. \quad (I, 9)$$

The first term on the right-hand side corresponds to the field component, the second to the diffusion component.

The influence of metal contacts on the current-voltage characteristic of a semiconductor

In the foregoing we have tacitly assumed that when a number of charge carriers at a certain location in the semiconductor move e.g. to the right, an equal number can flow from the left. In the bulk of a homogeneous semiconductor of not too small dimensions this is indeed the case, but it is often not true close to a contact. Consequently the contacts fitted to a piece of a semiconducting material in order to apply a voltage, for example, can have a considerable effect on the current-voltage characteristic of the whole. This applies to illuminated and non-illuminated semiconductors. In the following we shall deal first with non-illuminated semiconductors, after which we shall indicate in how far the theory remains applicable when the substance is illuminated.

Depending on their influence on the current-voltage characteristic of the combination of semiconductor and contacts a distinction is made in practice between a) contacts that do not affect the characteristic of the device — these are generally called "ohmic contacts" — and b) contacts that do. The latter contacts generally pass the current in one direction better than in the other, and are therefore termed *rectifying* or *blocking* contacts. Some combinations exist in which the characteristic, although symmetrical, is not linear. It may be noted that this classification is purely phenomenological and that the behaviour of a contact may depend on the strength of the current.

The behaviour of a combination of semiconductor and contacts is determined by what takes place in the layer of the semiconductor which is directly adjacent to the metal acting as the contact. In the following we shall therefore understand a "contact" to be the metal plus the contiguous layer of the semiconductor. We shall now proceed to examine the processes taking place in the contact, and in doing so establish a criterion for classification based on more fundamental physical aspects.

In general the semiconductor and the metal possess different thermionic work functions (see fig. 4), in other words their E_F values differ. Consequently, when they are brought into contact with each other, electrons flow from the one substance to the other, so that both become charged. At the boundary between them an electrical double layer is therefore formed. The metal contains a very large number of mobile charge carriers per unit volume, and therefore its charge may be present as virtually a pure surface charge. In the semiconductor, however, the charge can only be formed by the supply or removal of a very limited number of mobile charge carriers (per unit volume) and consequently this should take place over an appreciable depth in order to supply the charge required for the double layer. Making the contact, then, gives rise to a *space charge* at the junction.

Fig. 5a shows the band scheme of the contact formed by a metal with a *non-illuminated* N-type semiconductor whose thermionic work function is *smaller* than that of the metal, i.e. whose Fermi level is higher. When the contact is established, electrons therefore flow from the semiconductor to the metal. The positive space charge in the semiconductor is formed by the ionized donor centres whose positive charge in the boundary layer is no longer compensated; the concentration of free electrons in this layer is extremely small.

The presence of a space charge in the boundary layer of the semiconductor appears in the energy band scheme as a curvature of the bands in this layer. (In this case, then, we in fact regard the horizontal coordinate as the distance to the boundary.) For as soon as equilibrium has been restored, the Fermi level in the metal and in the semiconductor must be at the same height while, at the same time, at the boundary plane no change must have taken place in the distance between the Fermi level of the metal and the bottom of the conduction band of the semiconductor, which is sometimes called the work func-

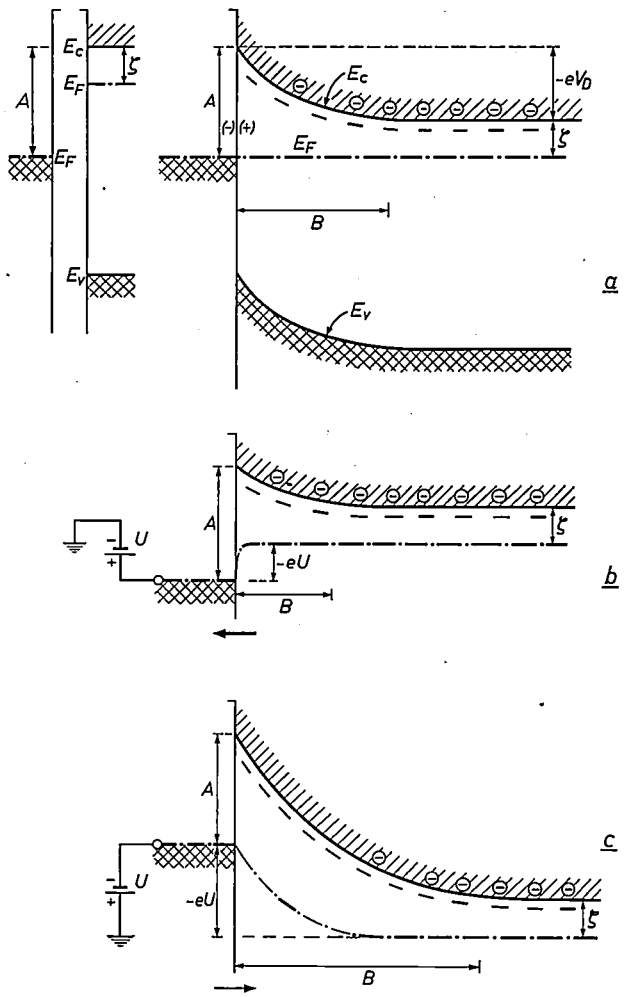


Fig. 5. *a*) Band scheme of the contact between a metal and a semiconductor for the case where the thermionic work function of the semiconductor is smaller than that of the metal, i.e. where the Fermi level in the semiconductor is higher before the contact is made (see sketch on the left). When the contact is established the Fermi levels are brought to the same height, but the limits of the bands remain unchanged in their relative position at the boundary plane: electrons pass from the semiconductor to the metal — as a result of the difference in E_F value — thus forming an electrical double layer which maintains this situation. The positive charge in the semiconductor is due to free electrons flowing away and leaving behind bound holes. The result is a space charge (of thickness B). The bands in the charge region are curved. Outside this region $E_c - E_F$ again has the value ζ which is characteristic of the pure material. The negative charge in the metal may be regarded as a surface charge; the curvature of the bands of the metal may therefore be approximated by a step function. In the boundary layer the free electrons are faced with a potential barrier of the magnitude V_D. (The energy difference A is sometimes referred to as the height of the barrier.)
b) Band scheme of the same contact after the application of a voltage U to make the metal positive. The potential barrier is lowered by an amount eU, and the region of positive space charge is narrowed (smaller B). In this case the contact is biased in the forward direction and passes the current.
c) The same, but with the voltage reversed (metal negative). The barrier is now higher and B larger. In this case the contact blocks the passage of the current. In the boundary layer the Fermi level is drawn thinly to indicate that in a large part of this layer there is no thermal equilibrium; the thin line represents the quasi Fermi energy of the electrons.

tion A of metal to semiconductor. The value of A is determined solely by the nature of the contact materials, and to a first approximation is independent of the concentration of electrons (which is always small) in the semiconductor. At a considerable distance from the boundary, on the other hand, the energy difference ζ between the conduction band and the Fermi level in the semiconductor has the value that holds for this material without contacts. As can be seen, with a contact of this type there is a potential barrier of height V_D and a width B. The energy $-eV_D$ is equal to the difference in E_F value mentioned at the beginning. The quantity V_D is also referred to as the diffusion potential.

The curvature of the bands in the neighbourhood of a contact depends on the space-charge density ϱ. Since the electrostatic potential V and the potentials $-E_c/e$ and $-E_v/e$ differ only by a constant amount, the form of the bands is identical with that of V. This form may be derived from Poisson's equation which, in one dimension, reads:

$$\frac{d^2V}{dx^2} = -\frac{\varrho}{\varepsilon}. \quad \ldots\ldots\ldots \quad (I,10)$$

Here ε is the dielectric constant of the medium. Marked curvatures, i.e. narrow barriers, can only occur if a high space-charge density can be built up. Very marked curvatures can occur at a metal surface; the potential can here be considered as making a jump. In a strongly doped semiconductor (i.e. with numerous donor or acceptor levels) narrow, steep barriers will be found, and in a doped semiconductor broad, slightly sloping barriers.

The fact that a contact of the type in fig. 5*a* can function as a rectifier when a voltage U is applied may be illustrated in qualitative terms as follows. As we have seen, the concentration of free electrons in the boundary layer is extremely low, so that the resistance of this layer is high. It may well be that this resistance entirely governs the resistance of the whole device. If the voltage source is connected such that the metal is positive with respect to the semiconductor (fig. 5*b*) the boundary layer is made thinner. The most depleted part thereby disappears — the E_c curve remains identical with that of fig. 5*a* but the boundary lies more to the right — and the resistance therefore becomes smaller, the more so the greater the value of U. In this case, then, the current will increase sharply with U. If we reverse the polarity of the voltage source (fig. 5*c*) the boundary layer becomes thicker and its resistance greater, here again the more so the greater the value of U. The current consequently increases only slightly with U, and when U is high the current is almost constant.

Calculations of the characteristic of a rectifying contact are given in an appendix to this article. One calculation is based on *diode theory*, which is applicable to the special case where the width B of the bar-

rier is small compared with the mean free path of electrons having sufficient kinetic energy to surmount the potential barrier, but not so small as to give rise to the tunnel effect. The other calculation is based on *diffusion theory*. In both cases it is shown that the density j_n of the electron current can be described by a relation of the form:

$$j_n = \text{const.} \exp[eU/kT - 1]. \quad . \quad \text{(I, 11)}$$

The constant, however, is not the same in the two cases.

Fig. 6 gives the band scheme of a contact where the *metal* has the lower work function. In this case the metal "injects" electrons into the semiconductor, thus giving rise to a *negative* space charge in the boundary layer. Where the semiconductor is N-type the electron concentration in the boundary layer is therefore greater here than elsewhere (and the same thus applies to the conductivity). Since there is no potential barrier, such an "injecting" contact can

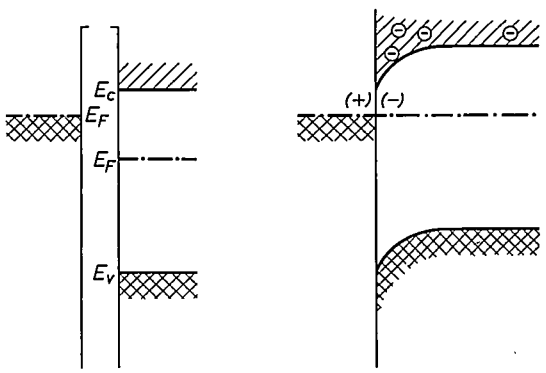


Fig. 6. Band scheme of the contact of a metal and an N-type semiconductor which has a *higher* thermionic work function (lower Fermi level) than the metal. The sketch on the left shows the situation before the contact is made. In the semiconductor a negative space charge is formed, which is concentrated in a thin layer. There is no potential barrier.

have no blocking action, although of course its resistance is not completely independent of the direction of the current.

If the metal and semiconductor have the same work function, which will very seldom be the case in practice, the contact is termed "neutral" (*fig. 7*). In such a contact, even when an external field is applied, there is no field distortion, so that the current-voltage characteristic is governed entirely by the homogeneous semiconductor and is therefore linear (it obeys Ohm's law).

The injecting contact, too, more or less obeys Ohm's law, and it does so more closely the smaller is the width of the better conducting junction layer (roughly at the order of 1 μm) compared with the length of the semiconductor. If this length, however, is itself very small (thin layer) the region whose conductivity has been increased by injection may cover the entire layer. When a voltage is applied a relatively strong current then flows, which increases more than proportionally with the voltage (superlinear behaviour). This current is limited only by the space charge formed by the migrating charge carriers.

The characteristic of the contact of a metal and an *illuminated* semiconductor can be calculated in the same manner as the non-illuminated case if the occupation of the levels in the conduction band, although mainly due to the photoconductive effect, can be described in terms of the Fermi-Dirac distribution function. This is only possible if we confine our considerations to one kind of charge carrier; in that case, of course, E_F in the equation must be replaced by the above-mentioned quasi-Fermi energy for the relevant charge carrier.

Under strong illumination the occupation of the levels may undergo such a marked change as to alter the type of the contacts. If the occupation can in fact be desceribed in terms of a quasi-Fermi energy, the reason is immediately evident: the illumination has shifted the quasi-Fermi level beyond the E_F value of the metal.

In an illuminated semiconductor or insulator the current can only be as strong as follows from formula (I,7), if the "replenishment" of charge carriers by the contact presents no difficulties. As we have seen, this calls for contacts that will inject carriers and which moreover do not change their type, even under the strongest illumination and at the highest field strength. It is often impossible, however, to find materials with the appropriate work function; moreover, the work function also depends to a great extent on the state of the surface. Using an N-type photoconductor, reliable injecting contacts can often be obtained by employing a metal which acts as a donor when incorporated in the semiconductor. This can be done by diffusion, resulting in a thin N-type layer which is highly conductive (symbol N^+)

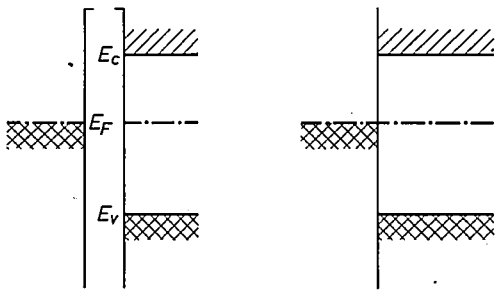


Fig. 7. Band scheme of a metal and a semiconductor with *identical* thermionic work functions (neutral contact). The establishment of the contact causes no change either in the metal or in the semiconductor.

and which forms with the rest of the photoconductor an N^+-N junction. This junction has virtually the same electrical properties as an injecting contact. Its band scheme is shown in *fig. 8*. As can be seen, it indeed closely resembles that of an injecting contact (fig. 6). A difference is the presence of a barrier at the boundary of the metal and the N^+ layer. Owing to the high concentration of donors in the N^+ layer, however, this barrier is very narrow. It is therefore possible that quite a large number of the electrons whose kinetic energy is lower than the height of this barrier can pass it by means of the tunnel effect. To a first approximation, therefore, the barrier may be treated as non-existent.

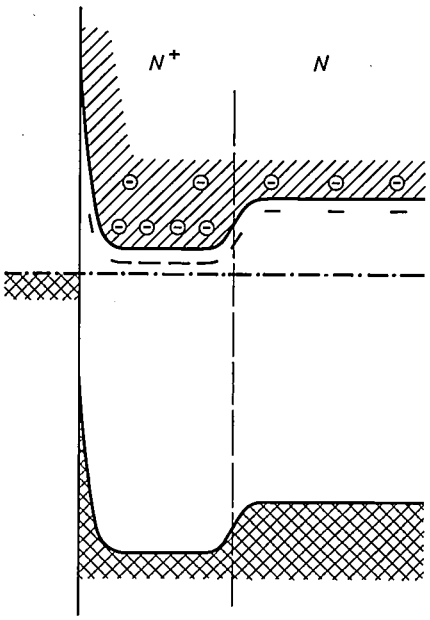


Fig. 8. Band scheme of an injecting contact, produced from a blocking contact by causing atoms to diffuse from the metal into the semiconductor. If these atoms can act as donors, a large concentration of free electrons is obtained in the boundary layer. At the position of the vertical dashed line the boundary layer forms an injecting contact (an N^+-N junction) with the rest of the semiconductor and with the metal it forms a contact which, although blocking, has such a narrow potential barrier that relatively low-energy electrons can easily pass it by means of the tunnel effect.

The behaviour of a photoconductor with blocking contacts

We shall now consider the behaviour of an illuminated photoconductor provided with contacts which we assume to be fully blocking, i.e. from which no carriers, either electrons from the cathode or holes from the anode, can enter the substance. We further assume that in the non-illuminated state the substance is an insulator. We start with the case where both kinds of carriers possess good mobility.

If we apply a voltage to such a device the charge carriers freed by the photoeffect move in the direction of the electrode concerned and a current starts to flow (*fig. 9a*). If the applied voltage is relatively low, a substantial number of the carriers will recombine on the way and will thus fail to reach the electrode. The current is then smaller than corresponds to the number of charge carriers generated per second. In proportion as the applied voltage is raised, the carriers will on an average need less time to flow from the place where they were generated to the electrode, reducing the chance of recombination. When the field strength is still further increased, all charge carriers freed by the light quanta will be extracted from the photoconductor before they have a chance to recombine. A further increase of the voltage will then no longer give rise to a higher photocurrent: the photocurrent is then saturated (see fig. 9*b*). In this situation, which is analogous to that in a vacuum photocell, the number of charge carriers flowing through a cross-sectional area of the external circuit is equal to the number of carriers generated in the same period by the light quanta. The ratio η of these two numbers is usually termed the *quantum yield* or quantum efficiency. For the situation described here the quantum yield is therefore at the most equal to unity. It can be deduced that:

$$\eta = \frac{\tau_n}{T_n} + \frac{\tau_p}{T_p}, \quad \ldots \quad \text{(I, 12)}$$

where T is the "carrier transit time", i.e. the time taken by a charge carrier, if it is not trapped, to migrate from one contact to the other.

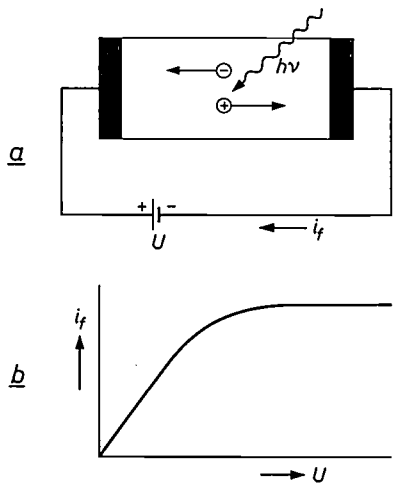


Fig. 9. *a*) Schematic representation of the flow of the photocurrent i_f in a semiconductor in which both kinds of charge carriers have adequate mobility and which is provided with completely blocking contacts.
b) The photocurrent i_f, under constant illumination, as a function of the applied voltage U. When U is small there is considerable recombination; when U is high nearly all charge carriers reach the electrode concerned. The current is then independent of U (saturation) and corresponds to the number of charge carriers freed per second by the light quanta (quantum yield = 1).

Since the two kinds of charge carriers are separated by the field, the field in the semiconductor may be modified by space charge. This effect, which is more marked the greater the light intensity and the lower the applied voltage, reduces the photocurrent. As a result, when the supply voltage is not particularly high, the relation between the photocurrent and the light intensity can become sublinear.

As just remarked, the magnitude of the photocurrent is determined by the degree of recombination, or in other words by the average distance which the charge carriers can cover during their lifetime τ. Since their average velocity is μF this distance — termed the "Schubweg" [9]) — can be defined by

$$S_n = \mu_n \tau_n F,$$
$$S_p = \mu_p \tau_p F.$$

In a sample with a cross-sectional area of 1 cm^2, and where the distance between the electrodes is d cm, Gd pairs of charge carriers are generated per second. If these travel on an average a distance S_n or S_p respectively, then when $S_{n,p} \ll d$, the current flowing in the external circuit is given by:

$$j = j_n + j_p = G\,d\,e\left(\frac{S_n}{d} + \frac{S_p}{d}\right).$$

The quantum yield η is therefore:

$$\eta = \frac{j}{eGd} = \frac{S_n}{d} + \frac{S_p}{d}. \quad \ldots\ldots\ldots \quad (I,13)$$

Introducing the transit time

$$T_{n,p} = \frac{d}{\mu_{n,p} F},$$

we arrive at the above formula (I,12):

$$\eta = \frac{\tau_n}{T_n} + \frac{\tau_p}{T_p}.$$

In the case under consideration the total Schubweg $(S_n + S_p)$ of a pair of charge carriers can never be greater than d; from (I,13) we therefore see once again that the quantum yield is at the most equal to unity. In how far this value is reached at a given field strength and electrode spacing again depends on the product $\mu\tau$.

Up to now we have considered cases in which both kinds of charge carrier had considerable mobility. We shall now examine what happens, in the presence of blocking contacts, when one type of carrier is virtually immobile, e.g. the holes. When a voltage is applied in this case, the electrons freed by irradiation will migrate towards the anode, but the holes formed scarcely move at all, giving rise to an excess of holes particularly in the neighbourhood of the cathode. The continued excitation causes this surplus to increase, so that a strong space charge is finally formed. The electric field therefore tends to concentrate near the cathode while the rest of the substance becomes field-free (see *fig. 10*). In this part, charge carriers recombine without having moved. These charge carriers therefore make no contribution to the photocurrent. The holes formed in the cathode region, if the field strength there finally becomes very high, can nevertheless reach the cathode and a stationary state sets in. The photocurrent, however,

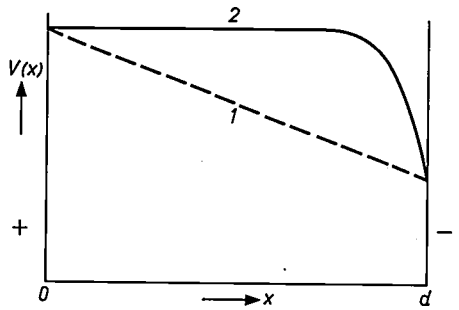


Fig. 10. Potential distribution in a non-illuminated (curve *1*) and in an illuminated (curve *2*) semiconductor with blocking contacts (distance between contacts d) where the holes have low mobility. The curvature of curve *2* is due to the occurrence of a positive space charge at the cathode: there the field strength is very high, at the anode it is zero.

is small because only the light absorbed in the thin cathode layer makes any contribution to it. If a fairly strong photocurrent is required from substances in which one type of charge carrier is virtually immobile, then contacts must be used that are relatively easily capable of replenishing the charge carriers which are being drawn out of the photoconductor by the field, e.g. injecting contacts. In the following, final section we shall see however that blocking contacts (see fig. 5) having a relatively low barrier (poorly blocking contacts) may also replenisb charge carriers.

Replenishing contacts

We shall now further examine the case just touched upon, where a semiconductor is provided with rectifying contacts and where only one of the two types of charge carriers possesses good mobility. For simplicity we assumed above that the blocking action at the electrodes was absolute, in other words that no single charge carrier from a contact could penetrate the semiconductor. In practice, however, some penetration will take place at the cathode in the case described, the field strength at the cathode being very high owing to the space charge. If a contact is used which has a fairly low barrier, it may even happen that the entry of electrons begins "earlier" than the extraction of the relatively immobile holes. In this sense "earlier", depending on the circumstances, may mean at a lower applied voltage or at lower light intensity. If voltage and light intensity have a value at which the effect can occur, it may well happen that during the build-up of the

[9]) This name is due to Gudden and Pohl, who did considerable research on photoconductivity in the twenties. See e.g. B. Gudden and R. Pohl, Z. Phys. **16**, 170, 1923, or B. Gudden, Lichtelektrische Erscheinungen, Springer, Berlin 1928.

space charge the effect does in fact, in terms of time, become appreciable earlier than the extraction of the holes. When equilibrium is reached, we then have a situation in which, under the influence of the space charge, the electrons freed by the light and which are taken up by the anode are replenished by the negative electrode, despite the fact that the contacts belongs to the type shown in fig. 5. Where a neutral or an injecting contact is used, a very small space charge will be sufficient to maintain replenishing in this way. (The electric field then divides itself more or less homogenously and the current-voltage characteristic approximately obeys Ohm's law.)

A special aspect of the replenishment is that the lifetime of an electron has not formally ended when it has reached the anode and has been taken up there, the reason being that a new electron has taken its place at the cathode. In the relation $\eta = \tau/T_n$, which can be derived from (I,12) by postulating $T_p = \infty$, the lifetime τ can therefore formally become longer than the transit time T_n. (The value of τ follows from the consideration that it must be equal to the lifetime of the holes, which are immobile.) Consequently the quantum yield η can be greater than unity. For this reason η is sometimes referred to as the current amplification factor. The formula (I,7) mentioned at the beginning is applicable to this case.

The mechanism described above, where one of the two types of charge carrier is immobile thus making the contact for the other type an injecting contact, explains the high sensitivity obtainable with a good photoconductor, such as CdS. The migration of holes in these substances is hampered by capture in impurity centres which are known as traps. The nature of these traps is such that the probability of trapping a conduction electron is very small, resulting in a long lifetime and hence in a high amplification factor. The nature and effects of traps will be dealt with at greater length in the second article.

The complicated and apparently irregular behaviour of photoconductors, arising from the space-charge effects produced when the contacts used are poorly injecting and not completely blocking prompted research workers in the twenties to distinguish between a primary and a secondary photocurrent [10]). In cases where the quantum efficiency had a saturation value of unity the current was regarded as identical with the primary current. A quantum efficiency greater than unity was accounted for by assuming the presence of an additional "secondary" current in this case. Since it is now clear that a quantum efficiency greater than unity is due to replenishment of electrons by contacts which is by no means a secondary effect, there is no longer any reason to make this distinction.

[10]) See the article by Gudden and Pohl in reference [9]).

True secondary effects do occur in substances showing ionic conduction. These, however, belong to the category disregarded here, ionic conduction being bound up with chemical changes.

Appendix: Calculating the characteristic of a blocking contact

1) We first consider the case where B (fig. 5) is small compared with the mean free path of electrons that have sufficient kinetic energy to pass the potential barrier at the junction; this may be the case if the donor concentration N is very high. The boundary layer can then be left out of consideration and *diode theory*, as it is called, is then applicable. We further assume for convenience that the potential gradient outside the boundary layer is small, so that the potential drop U takes place almost entirely inside the boundary layer. We may then treat the metal and the semiconductor as reservoirs which are both in thermal equilibrium and which exchange electrons. The minimum energy which an electron must possess in order to change from one reservoir to the other differs for the two directions, however, by an amount eU.

The density j_{ms} of the partial electron current from metal to semiconductor may therefore be written (omitting the suffix n):

$$j_{ms} = j_0 \exp\,[-A/kT],$$

and the density in the opposite direction:

$$j_{sm} = j_0 \exp\,[(-A+eU)/kT].$$

The density j of the net current is therefore:

$$j = j_0 \exp\frac{-A}{kT}\left(\exp\frac{eU}{kT}-1\right). \quad \ldots \quad \text{(I, 14)}$$

If we connect the voltage source so that eU is positive — the metal is then positive with respect to the semiconductor — the form between brackets may become large with the respect to unity ($kT/e = 1/40$ volt at $T = 300$ °K). With the voltage source connected in the opposite polarity, and when $|eU| \gg kT$, formula (I,14) becomes:

$$j_{block} \approx j_0 \exp\,(-A/kT).$$

This indicates that the current, if the contact is a blocking one, is indeed relatively weak and almost independent of the applied voltage.

2) If the width B of the barrier is large compared with the mean free path of the electrons, the calculation becomes more intricate (*diffusion theory*). The formulae now needed, in addition to the current equation (I,9), in order to calculate the characteristic of a rectifying contact, can be derived from Poisson's equation (I,10) given above. If we may assume a) that the concentration n of the electrons in the boundary layer is small enough to be neglected with respect to that of the donors (N) so that $\varrho = Ne$, and b) that N does not depend on the place coordinate x, i.e. the distance to the boundary plane, we can then find the field strength F, which is equal to $-\mathrm{d}V/\mathrm{d}x$, by a single integration of (I,10). For the region in which the energy bands are curved ($x<B$) we may write:

$$F = -\frac{Ne}{\varepsilon}(B-x). \quad \ldots \quad \text{(I, 15)}$$

The variation of V with x can be found by a further integration. Suitably choosing the zero point of V, this gives:

$$V = -\frac{Ne}{2\varepsilon}(B-x)^2. \quad \ldots \quad \text{(I, 16)}$$

The curve is thus parabolic. The vertex of the parabola is at the point $x = B$. For larger x, V is constant and equal to zero.

Since V must have the value $-V_D$ at the position $x = 0$ (see fig. 5), we can derive from (I,16) an expression for B:

$$B = \sqrt{\frac{2\varepsilon V_D}{eN}}. \quad \ldots\ldots \text{(I, 17a)}$$

When an external voltage U is applied (and provided still that $n \ll N$) the expression becomes:

$$B = \sqrt{\frac{2\varepsilon (V_D - U)}{eN}}. \quad \ldots \text{(I, 17b)}$$

It follows from (I,15) and (I,17b) that the value $F(0)$ of the field strength at the boundary plane is given by:

$$F(0) = -\sqrt{\frac{2eN(V_D - U)}{\varepsilon}}.$$

If now the external voltage is so applied that the contact tends to block the passage of current (U negative), then provided U is not too small we can deduce the current density directly from (I,9) by neglecting the diffusion term and inserting for the electron concentration $n(r)$ and the field strength $F(r)$ the values applicable at the boundary ($x = 0$). In this way (again dropping the suffix n) we find:

$$j(U) \approx -n(0)\, e\mu \sqrt{\frac{2eN(V_D - U)}{\varepsilon}}. \quad \ldots \text{(I,18)}$$

It is evident from this formula that j is only slightly dependent on U.

When the voltage is so applied that the contact becomes conductive (U positive), then the boundary layer, compared with the situation at $U = 0$, is much less depleted in charge carriers, giving rise to an undisturbed Boltzmann equilibrium even very close to the boundary plane. We may then write:

$$n(x) = n(\infty) \exp[-e\{V(x) - U\}/kT].$$

The density of the diffusion current j_{diff} thus becomes:

$$j_{diff} = eD \left(\frac{dn(x)}{dx}\right)_{x=0} =$$

$$= -n(\infty) \frac{e}{kT} \frac{dV(0)}{dx} eD \exp\left[-e\{V(0) - U\}/kT\right],$$

which, with the aid of (I,8), can be reduced to:

$$j_{diff} = n(0)\mu F(0)\, e \exp(eU/kT).$$

As regards the field current we assume that it is in no way changed by the application of U (in reality it decreases slightly). The density of the total electron current is then found to be:

$$j = n(0)\mu F(0)\, e\{\exp eU/kT - 1\}.$$

This expression has the same form as (I,14), but differs in the proportionality factor between j and the exponential function. An expression of this form with yet another proportionality factor applies to the characteristic of a P-N junction [11]). The latter formula will be used in the third article of this series.

[11]) See e.g. M. Beun and L. J. Tummers, Philips tech. Rev., to be published.

Summary. This first article of a series of three on the physical principles of photoconductivity deals first of all with some elementary concepts and hypotheses, such as the energy band scheme; the difference in this respect between metals, semiconductors and insulators; intrinsic and extrinsic semiconductors; donor and acceptor centres; electron and hole conduction; the lifetime and mobility of charge carriers; and the Fermi-Dirac distribution function for the occupation of energy levels. It is explained how the absorption of light quanta causes electrons to be raised from the valence band or from impurity centres into the conduction band, resulting in photoconductivity. Similar considerations apply to holes. Finally, a fairly extensive treatment is given of the theory of semiconductors fitted with metal contacts. Their properties are determined by the relative situations of the Fermi levels of the semiconductor and that of the contact metal. If the Fermi level of an N-type semiconductor (before a contact is made) is higher than that of the metal, a boundary layer forms in the semiconductor which is greatly depleted in charge carriers and exhibits a rectifying action. In the other case "injecting" contacts are obtained. The device then behaves as a pure resistance, provided the current is not too strong. The photocurrent through a semiconductor with blocking contacts is no higher than corresponds to the number of charge carriers freed by the light quanta. This value is reached when the magnitude of the applied voltage is such that the transit time of the electrons is small compared with their lifetime. If the contacts are of the "replenishing" type, i.e. inject charge carriers or have a poor blocking action, the photocurrent can be very much stronger; it is this that underlies e.g. the high sensitivity of CdS photoresistors.

RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF THE PHILIPS LABORATORIES AND FACTORIES

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

3089: G. D. Rieck: The effect of temperature and deformation on the recrystallization of doped tungsten wires (Acta metallurgica **9**, 825-834, 1961, No. 9).

3090: L. Heijne: Contact influence on the photoconductivity of lead oxide (Phys. Chem. Solids **22**, 207-212, 1961).

3091: J. J. A. Ploos van Amstel: Some methods of producing stable transistors (Comm. Colloque int. sur les dispositifs à semiconducteurs, Paris 1961, Vol. II, pp. 716-724, publ. Chiron, Paris).

3092: A. Baelde, H. Groendijk and M. T. Vlaardingerbroek: Reduction of the noise figure of

an amplifier by a negative conductance in its input circuit (J. Electronics and Control **11**, 177-187, Sept. 1961, No. 3).

3093: W. Albers: Thermal conversion of germanium (J. Electronics and Control **10**, 197-206, March 1961, No. 3).

3094: J. D. Fast: Frottement intérieur des métaux (Métaux, Corr., Industr. **36**, 383-398 and 431-453, 1961, Nos. 435 and 436). (Internal friction in metals; in French.)

3095: A. Claassen: Methodes en problemen in de anorganische sporenanalyse (Chem. Weekbl. **58**, 33-38, 1962, No. 4). (Methods and problems in inorganic trace analysis; in Dutch.)

3096: M. Koedam and A. A. Kruithof: Transmission of the visible mercury triplet by the low-pressure mercury-argon discharge; concentration of the 6^3P states (Physica **28**, 80-100, 1962, No. 1).

3097: J. Bloem and J. C. van Vessem: Etching Ge with mixtures of HF-H_2O_2-H_2O (J. Electrochem. Soc. **109**, 33-36, 1962, No. 1).

3098: J. S. van Wieringen: Magnetic resonance in semiconductors (Progr. Semicond. **6**, 199-231, 1962).

3099: H. C. Hamaker: On multiple regression analysis (Statistica neerl. **16**, 31-56, 1962, No. 1)

3100: W. J. Oosterkamp and Th. G. Schut: Magnetische Festlegung von Röntgenbildern (First Int. Congress on medical photography and cinematography, Düsseldorf 1960, pp. 96-99, Thieme, Stuttgart 1962). (Magnetic recording of X-ray images; in German.)

3101: H. Bouma: Size of the static pupil as a function of wavelength and luminosity of the light incident on the human eye (Nature **193**, 690-691, 1962, No. 4816).

3102: P. A. H. Hart: On cyclotron-wave noise reduction (Proc. Inst. Radio Engrs. **50**, 227-228, 1962, No. 2).

3103: Th. J. van Kessel, F. L. H. M. Stumpers and J. M. A. Uyen: A method for obtaining compatible single-sideband modulation (E.B.U. Rev. Part A, No. 71, 12-19, 1962).

3104: H. J. M. Moonen: Het bepalen van bestelniveaus wanneer afname en levertijd gamma-verdeeld resp. normaal-verdeeld zijn (Statistica neerl. **16**, 113-120, 1962, No. 1). (Determination of re-order levels when demand has a gamma distribution and delivery time a normal distribution; in Dutch.)

3105: J. F. Schouten, J. W. H. Kalsbeek and F. F. Leopold: On the evaluation of perceptual and mental load (Ergonomics **5**, 251-260, 1962, No. 1).

3106: P. Clausing: On the molecular flow with Langmuirian adsorption of the molecules on the wall of the tube; a correction (Physica **28**, 298-302, 1962, No. 3).

3107: O. Bosgra and J. H. G. Roerink: Parental immunity in fowl pox and serum neutralization (T. Diergeneesk. **87**, 106-112, 1962, No. 2).

3108: J. Cornelissen and A. L. Zijlstra: The strength of glass rods as a result of various treatments (Symp. Résist. mécan. du verre, Florence 1961, pp. 337-358, Union Scientifique Continentale du Verre, Charleroi 1962).

3109: A. L. Zijlstra and J. de Groot: Surface condition and strength of glass objects (as **3108**, pp. 359-376).

3110: Y. Haven and A. Kats: Hydrogen in α-quartz (Silicates industriels **27**, 137-140, 1962, No. 3).

3111: J. J. Engelsman: Enkele elektrochemische methodes in de sporenanalyse (Chem. Weekblad **58**, 113-115, 1962, No. 11). (Some electrochemical methods in trace analysis; in Dutch.)

3112: H. Bremmer: On the theory of wave propagation through a concentrically stratified troposphere with a smooth profile, II. Expansion of the rigorous solution (J. Res. Nat. Bur. Stand. **66D**, 31-52, 1962, No. 1).

3113: W. L. Wanmaker and H. L. Spier: Luminescence of copper-activated orthophosphates of the type $ABPO_4$ (A = Ca, Sr, or Ba and B = Li, Na, or K) (J. Electrochem. Soc. **109**, 109-114, 1962, No. 2).

3114: J. A. Kok: Diëlektrische verliezen in heterogene diëlektrica (Ingenieur **74**, Ch 16-Ch 20, 1962, No. 8). (Dielectric losses in heterogeneous dielectrics; in Dutch.)

3115: P. A. H. Hart and G. H. Plantinga: Millimetre-wave noise of a plasma (Proc. 5th int. Conf. on ionization phenomena in gases, Munich 1961, Vol. I, pp. 492-499, North-Holland Publ. Co., Amsterdam 1962).

3116*: B. Okkerse: Preparation of semiconductor materials (Handbook of semiconductor electronics, editor L. P. Hunter, 2nd ed., pp. 6-3 to 6-31, McGraw-Hill, New York 1962).

3117: J. H. Stuy: Inactivation of transforming deoxyribonucleic acid by nitrous acid (Biochem. biophys. Res. Comm. **6**, 328-333, 1961, No. 5).

3118: J. H. Stuy: Studies on the mechanism of radiation inactivation of micro-organisms, IX. Mechanism of the ultraviolet-induced inactivation of transforming deoxyribonucleic acid (Photochem. Photobiol. **1**, 41-48, Jan./March 1962).

3119: G. W. van Oosterhout and C. J. Klomp: On the effect of grinding upon the magnetic properties of magnetite and zinc ferrite (Appl. sci. Res. **B 9**, 288-296, 1962, No. 4/5).

3120: N. W. H. Addink and L. J. P. Frank: Zinc content of hair from the head of carcinoma patients (Nature **193**, 1190-1191, 1962, No. 4821).

VOLUME 25, 1963/64, No. 6/7 pp. 133-180 Published 7th July 1964

Philips Technical Review

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

THE "PLUMBICON", A NEW TELEVISION CAMERA TUBE

by E. F. de HAAN *), A. van der DRIFT *) and P. P. M. SCHAMPERS *).

621.397.331.222

*In this article a description is given of the "Plumbicon" **), a new type of television pick-up tube. The use of photoconducting material as a light detector and the construction of the tube are such that the "Plumbicon" can in many ways be considered as a type of vidicon, with which it has in common simple construction and easy operation. The less favourable properties of the conventional vidicon, however, are absent: the picture quality and the speed of response are also excellent at low levels of illumination. With this new tube, which equals or surpasses the existing pick-up tubes in all respects important in broadcast television, particularly good results are obtained when using it in cameras for colour television. This is because the "Plumbicon" meets to a high degree the requirement that the signal supplied by one picture element depends solely on the amount of light falling on it — so it does not depend on the position of the picture element, or on its history, or on the situation in the neighbouring elements.*

Principles and construction of the "Plumbicon"

In view of the requirements imposed by television broadcasting, it has been necessary till now to use either an image iconoscope or an image orthicon for direct broadcasts. These were the only types of pick-up tube with sufficient resolving power and speed to respond adequately to the rapidly changing details of the broadcast scene. The vidicons, which far surpass these types in regard to simplicity and ease of operation, were unsuitable because the picture they supply to the receiver is too uneven at low light levels — a consequence of local differences in dark current; also, at low light levels vidicons have too slow a response [1]). Their employment has been limited to applications where a high level of illumination is possible, as in film scanning.

The new camera tube we describe in this article, the "Plumbicon" (*fig. 1*), possesses the good properties of both classes of pick-up tube referred to above, and even surpasses them in certain respects: the "Plumbicon" combines small size, simple construction and easy operation with having a low dark current, high sensitivity, high speed of response, and good resolution. The life of the "Plumbicon" is no shorter than that of other studio-quality tubes. It also affords striking advantages particularly in colour television and in X-ray television set-ups.

Broadly speaking, the electrode layout of the "Plumbicon" is like that of the vidicon, and the mode of functioning is much the same (*fig. 2*). A glass plate is coated with a thin transparent conducting layer of SnO_2, on which is deposited a thin layer of photoconducting material, which in the "Plumbicon" consists of lead monoxide (PbO). The scene to be transmitted is projected via the glass substrate and the SnO_2 layer onto the PbO. A beam of slow electrons strikes the other side of the PbO layer. The SnO_2 layer, known as the signal plate, carries a potential of about +30 V with respect to the cathode of the electron gun. The side of the photoconducting layer facing the gun has roughly the same potential as the cathode when the layer is not illuminated; lighted areas periodically attain a potential of a few volts higher.

Although the "Plumbicon" may be considered in

*) Philips Research Laboratories, Eindhoven.

**) A short article about this tube appeared in the preceding volume: E.F. de Haan, Philips tech. Rev. 24, 57, 1962/63.

[1]) A description of the types of pick-up tube most used up to now may be found in D.G. Fink, Television engineering, McGraw-Hill, New York 1960.

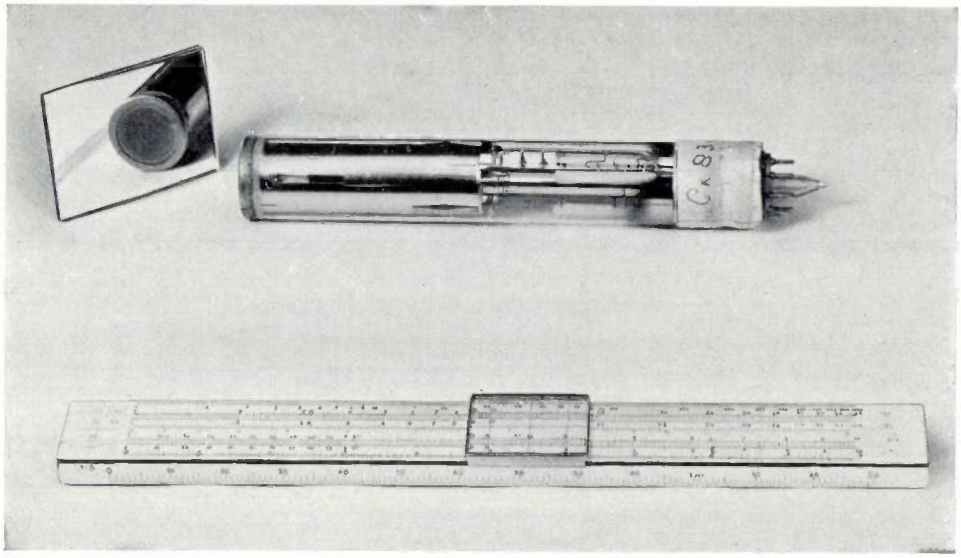


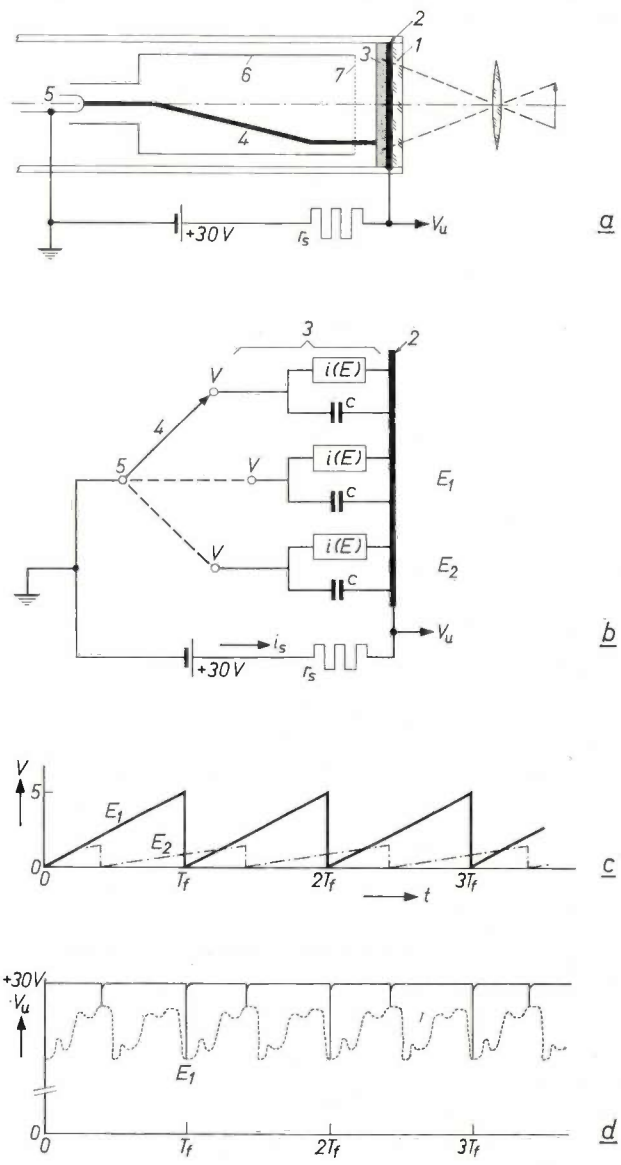
Fig. 1. A pick-up tube of the "Plumbicon" type. The tube is in the form of a cylinder having a diameter of 3 cm and a length of 19 cm. The target diameter is 2 cm.

Fig. 2. *a*) Electrode layout of the "Plumbicon" (schematic). In the front of the tube (on the right) is a glass window *1*, on the inside of which have been applied, in that order, a transparent, conductive SnO_2 layer *2* and a photoconducting layer of PbO constituting the target *3*. An image of the scene is projected on the target, the other side of which is scanned by the electron beam *4*. The beam electrons are supplied by gun *5* and accelerated by anode *6*. A mesh screen *7* has been fitted to the front of the anode in order to make the field between target and anode more uniform. The anode is at a potential of about +300 V, the SnO_2 layer (signal plate) is at about +30 V, both with respect to the gun cathode. As will be explained, when the tube is in operation the potential V of the free surface of the PbO target fluctuates within an interval ΔV of only a few volts, the lower limit of which is approximately equal to the cathode potential of the gun.

b) Equivalent circuit of the "Plumbicon" explaining the functioning of the tube. The electrode numbering is the same as in (*a*). The signal plate is on the right. The photoconducting target can be regarded as being made up of a large number of capacitors c, each of which is in parallel with a current source supplying a current i_f whose magnitude depends solely on E, the intensity of illumination; these capacitors represent "picture elements". The beam can be regarded as a multi-way switch that connects each of the picture elements in turn to the negative terminal of the battery supplying the voltage on the signal plate. Thus when a picture element is brought into circuit, the potential V on its free surface falls abruptly to zero (the cathode is earthed). During the remainder of the frame period T_f — the time elapsing before the electron beam returns to the same picture element, which is usually 1/25th of a second — c partially discharges owing to the flow of current i_f; V therefore rises. The increase ΔV in V that takes place between successive scans is greater when the picture element is more strongly illuminated because i_f is then greater. A charging current i_c, proportional to ΔV, flows each time the beam completes the circuit containing an element, and this current causes a corresponding difference of potential to arise across signal resistor r_s. Consequently the potential V_u varies and an output signal is obtained. If E is changing rapidly, ΔV will have a value roughly corresponding to the mean of the intensity of illumination on the element during the frame period in question.

c) The variation over time t in the potential V on the free surface of two picture elements, one exposed to a high intensity of illumination E_1 and the other to a low intensity of illumination E_2. Since, under normal operating conditions, i_f does not depend on the difference of potential $V_u - V$, between scans the variation in V is linear with respect to time.

d) The variation in V_u over time t, corresponding to (*c*). When the contributions of *all* the picture elements are taken into account, then V_u varies e.g. as shown by the dotted line.



many ways to be a kind of vidicon, there is a characteristic difference however between the "Plumbicon" and the present vidicons, and this concerns the photoconducting layer. Not only has this layer been made from a different photoconductor — Sb_2S_3 and sometimes As_2S_3 or Se have so far been used in ordinary vidicons — but what is more important, the PbO layer together with the SnO_2 layer form a unit consisting of three sublayers, each of differing conduction type. The inner sublayer consists of almost pure PbO, which is an intrinsic semiconductor. The PbO in the layer struck by the electrons is doped to make it a *P*-type semiconductor (hitherto not possible with Sb_2S_3). The SnO_2 signal plate is strongly *N*-type. The contact between the PbO and the SnO_2 may also give rise to a thin *N*-type layer in the PbO. The *P*-type and *N*-type layers are relatively thin, so that the inner (intrinsic) layer, the *I* layer, takes up most of the overall thickness of the PbO layer. For simplicity it is assumed in the following that the *N*-type PbO layer is always present [2]).

Examination under the electron microscope shows the PbO layer to be porous in structure; it is built up of crystallites having dimensions of about $1.0 \times 1.0 \times 0.1\,\mu m$. The filling factor ranges from 30% to 50%. The dimensions of the crystals are small compared with the line spacing (20 μm). They are therefore too small to be detrimental to the resolving power. The overall thickness of the photoconducting layer ranges from 10 to 20 μm.

The fact that the tube has many favourable properties for television broadcasting is thanks to the multilayer structure of its photoconducting target. Its ability to satisfy two of the main desiderata, namely low dark current and high sensitivity, is easily explained. When the tube is in operation its photoconducting layer — in contrast with a conventional vidicon — constitutes a reverse-biased diode. The dark current is the (small) inverse current through this diode. The tube owes its high sensitivity to the *I* layer sandwiched between the *P* and *N* layers. Conduction electrons and holes generated by light cannot contribute to a photo-current unless they originate in a region where a relatively high field-strength prevails. If the diode in question were simply a *P-N* device, the requisite field-strength would only be available in the immediate vicinity of the junction, and a large proportion of the charge-carriers generated in the PbO would be ineffective. In the "Plumbicon", however, there is a high field-strength throughout the *I* layer, in consequence of the fact that this is a relatively poor conductor, and since the PbO layer consists almost entirely of *I* material, almost all the charge-carriers generated in the PbO contribute to the photo-current [3]).

In principle, an excessive dark current can be compensated by electrical means. In practice, however, electrical compensation is scarcely ever adopted: unless the dark current has the same value within close limits at all points on the screen, compensation leads to an uneven image signal which is undesirable and in colour television is quite unacceptable. An additional difficulty is that the dark current is strongly dependent on temperature. If, on the other hand, the dark current has a very low value (say 10^{-9} A, or about 1% of the signal current), local variations, even though quite large (the values differing by a factor of 2, for example) will not perceptibly affect the uniformity of the image.

In most other important respects — spectral characteristic, definition, speed of response and service life — the favourable properties of the tube are mainly dependent on a suitable choice of the parameters governing the properties of the sublayers, such as their thickness, the doping substance and its concentration, and so on. Although some of the demands of studio use give rise to conflicting requirements, it has not been necessary to make compromises between the various parameters. On the contrary, there is so much play that certain properties can be varied quite widely without interference with the others. This makes it possible to manufacture camera tubes of the "Plumbicon" type with properties very closely fitted to the demands of a given application. Left out of discussion are the very large variations encountered when the choice of photoconducting material is not restricted to PbO, but when e.g. PbS is added [4]).

Various points mentioned above will now be elaborated [5]). First however the relevant physical and chemical properties of PbO will be discussed; some observations will be made on the process of deposition of the PbO layer, and a brief account given of the way in which the potential of the free surface of the photoconducting layer varies during a frame period.

The properties of PbO; deposition of the photoconducting layer

Two modifications of PbO are known: the red (tetragonal) modification, which is stable at temperatures below 488 °C, and the yellow (orthorhombic),

[2]) The characteristics of the "Plumbicon" are compared with those of other tubes in an article by A.G. van Doorn and S.L.Tan, shortly to be published in this Review.

[3]) For a succinct explanation of the physical principles underlying photoconductivity see e.g. L. Heijne, Philips tech. Rev. 25, 120-131, 1963/64 (No.5).

[4]) See the article by E.F. de Haan, F.M. Klaassen and P.P.M. Schampers, shortly to be published in this Review.

[5]) See also L. Heijne, thesis Amsterdam 1960, Ch. 8.

which is stable at higher temperatures. The red PbO is built up from "sandwiches" consisting of a plane occupied by O atoms, on either side of which is a plane occupied by half as many Pb atoms. These sandwiches have a thickness of 2.38 Å, and the spacing between the O planes is 4.99 Å.

The structure of the yellow modification differs considerably from that of the red. This too is built up from sandwiches, with Pb atoms on the outside, but the "filling" is more complicated. Accordingly, the thickness of the sandwich is rather greater (2.72 Å). Various investigators have demonstrated that oxygen can be inserted between the sandwiches (especially those of red PbO) without drastically modifying the crystal structure of the compound; in other words, departures from stoichiometric ratio are possible. It is in virtue of this important property, among others, that PbO can be turned into either an *N*-type or a *P*-type semiconductor. It has been found that PbO becomes a *P*-type semiconductor when an excess of oxygen is present, or when it has been doped with Tl, Cu or Ag. The compound becomes an *N*-type semiconductor when an excess of lead is present, or when it has been doped with Bi.

The energy gap ΔE of the forbidden zone between the valency band and the conduction band, the quantity that determines the upper limit of the range of wavelengths within which photo-excitation can occur, is 2.0 eV for red PbO and 2.7 eV for the yellow modification. The band gap of red PbO gives rise to a cut-off wavelength of about 6200 Å, that of yellow PbO gives about 4500 Å.

The production of the PbO layer in a tube of the "Plumbicon" type is by vapour deposition. PbO contained in a small platinum crucible is evaporated by inductive heating, and condenses on a window previously coated with SnO_2. The crucible temperature is held at about 900 °C; at this temperature evaporation proceeds at a reasonable rate. During the deposition process the window is likewise kept, within fairly close limits, at a certain temperature. A high temperature is especially undesirable because the crystals of the deposited PbO become too large to give the required image resolution. Deposition takes place not in vacuum, but in a certain gas atmosphere.

Owing to the considerable difference in the cut-off wavelengths of red and yellow PbO, the red-yellow ratio has an important effect on the spectral response of the "Plumbicon". An X-ray diffraction study, in which the diffraction pattern of PbO layers deposited by the normal process was compared with those of certain mixtures of yellow and red PbO powders, has revealed that these layers consisted of about 90% of the red modification and about 10% of the yellow. At vapour pressures lower than the normal one, the proportion of yellow PbO was greater. Moreover, the red crystals were found to have a preferred direction.

Potential of the target surface; stabilization

As already stated, the free surface of the PbO target has a potential V that varies within a range close to the potential of the cathode. At the instant the scanning beam leaves an element of this surface, the potential of this element is roughly equal to that of the cathode; in the time elapsing before the beam returns, V rises by a few volts only. The actual range over which the surface potential varies is determined by the requirement that in the steady state, the amount of (negative) charge removed per unit time, in consequence of the flow of photo-current, must be equal to that supplied by the electron beam. (Since the negative charge is supplied intermittently, this equality is only valid if an interval of time is considered containing an integral number of frame periods.)

Fig.3 shows, schematically, how i_a, the net current flowing to a surface element of a PbO target under electron bombardement, depends on the potential V of the target surface. The various curves refer to different values of beam current, i.e. the current formed by the electrons leaving the electron gun. When V is the same as the cathode potential (which is zero), the deceleration of the electrons in the final part of their path is equal to their previous acceleration, with the result that only a few of them actually reach the target. The *net* flow of current to the target is even smaller, owing to secondary emission; it is on account of increasing secondary

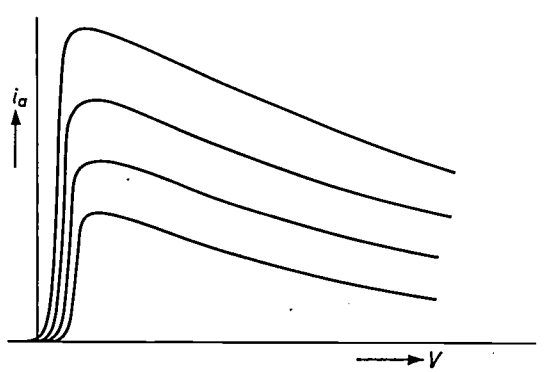


Fig.3. The net flow of current i_a to a surface element of a PbO target under continuous electron bombardment as a function of the potential V of the target surface (the cathode of the electron gun is assumed to have zero potential). As V increases, i_a also initially increases but subsequently falls off on account of increasing secondary emission. The above curves, which are schematic, relate to various values of beam current (i.e. the current constituted by the electrons leaving the gun).

emission that the net current to the target, after attaining a maximum for a certain potential value, falls off again as V is raised further. *Fig. 4* gives an overall impression of the way the photo-current that flows through the target layers, and so removes negative charge from its free surface, depends on the potential difference U between the two faces of the PbO

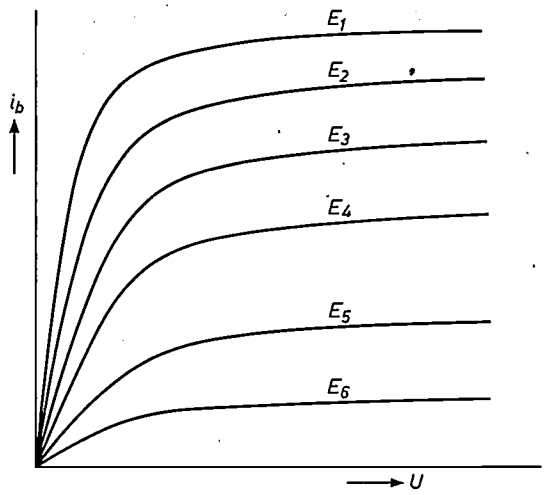


Fig. 4. The variation of i_b, the photo-current that removes (negative) charge from an elementary area of the free target surface, with the potential difference U across the target. The curves, which are schematic, hold for different values E_i of the illumination. The stronger the illumination, the heavier is the current i_b.

target ($U = V_u - V$; see fig. 2). The plotted quantity i_b is again the value of the current per surface element.

In *fig. 5* one of the curves of fig. 3 has been combined with the family of curves in fig. 4. That, in fact, corresponds to the situation prevailing in the tube: the intensity of illumination may assume almost any value, but unless altered from outside the beam current always remains the same. For reasons that will shortly become clear, the ordinate values for the curves representing the current flowing through the target layers have been multiplied by a factor N representing the number of surface elements into which the surface is understood to be divided. Accordingly, the quantities plotted are i_a and Ni_b.

Consider first the imaginary case in which the beam electrons continuously bombard the same spot on the target surface, i.e. a spot whose area is $1/N$th the total target area (this is what we call a picture element), and in which a constant photo-current Ni_b is flowing at this spot. The potential V of the spot then assumes a value given by the projection on the abscissa of one intersection point of the i_a curve with the relevant Ni_b curve: the loss of charge is then exactly balanced by the rate of supply. The points of intersection representing stable states have been marked in fig. 5 with a dot. It will be noted that these points occur both on the rising and on the falling branch of the i_a curve. In the former case the value of V is close to zero, and the voltage U across the target is therefore roughly equal to V_u. The points of intersection on the falling branch represent states in which the value of V is close to the signal-plate potential, and U is very small.

Let us now consider the true situation, in which the beam scans the target and the current supplying a given picture element only flows for $1/N$th of the time. On account of the surge-like character of the charge-supply process V does not assume a steady value; it can however be said that the range over which V varies must extend on either side of one of the stable-state values referred to above. Since the range of variation ΔV is normally small compared with V_u, these stable-state values nevertheless give a fairly good indication of the operating conditions that are possible in the tube: there are again *two* possibilities when the tube is operational. Operating conditions such that $V \approx V_u$, U accordingly being small, are extremely unfavourable, however: in the first place the equilibrium value of i_a differs very little with different intensities of illumination, with consequent loss of contrast, but apart from that, special forms of sluggish response are liable to occur at low U values.

Therefore in practice the beam-current value is chosen high enough for an intersection to be available on the *rising* branch of the i_a curve even at the highest intensity of illumination likely to occur on the PbO target during a broadcast. In this way a third, but no less important, disturbing effect is

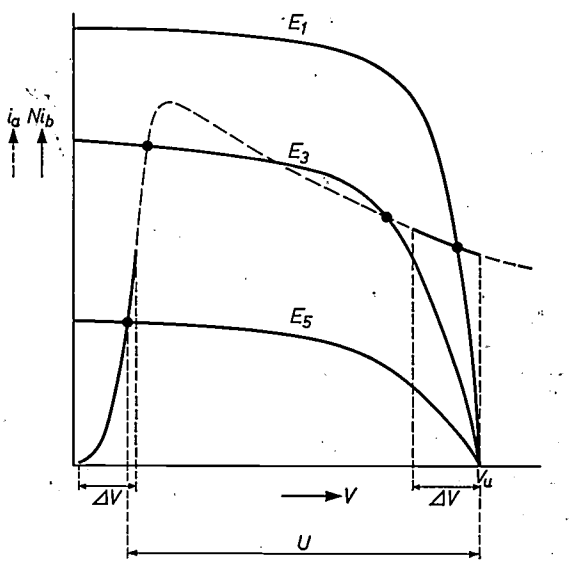


Fig. 5. To explain the fact that the interval ΔV can fluctuate in the neighbourhood of the cathode potential of the electron gun ($V \approx 0$), as well as in the neighbourhood of the potential of the signal plate ($V \approx V_u$).

avoided that occurs when one surface element has a potential close to V_u and the other a potential close to zero. In these circumstances the element with the higher potential will start to attract electrons towards it when the beam approaches, and continue to do so after the beam should have passed. In consequence, lighter-coloured parts of the scene will be "blown up", i.e. appear bigger than they really are.

The fact that V moves within an interval ΔV, contained between the abscissa values of the intersections of the i_a and Ni_b curves, can be proved as follows. During the short time T_p in which the beam passes a surface element, then for the (negative) charge:

$$dQ = (i_a - i_b)dt \approx i_a dt = c\,dV .$$

Here dQ is the charge supplied in time dt and c is the capacitance of the layer per surface element. It follows from this that $dV/i_a = dt/c$. In the time T_p, V decreases from V_2 to V_1 and so:

$$\int_{V_2}^{V_1} \frac{dV}{i_a} = \frac{T_p}{c}.$$

For the potential increase following from the flow of the discharge (photo-) current, we have by analogy:

$$\int_{V_1}^{V_2} -\frac{dV}{i_b} = \frac{T_f}{c},$$

or, as $T_f = NT_p$:

$$\int_{V_2}^{V_1} \frac{dV}{Ni_b} = \frac{T_p}{c}.$$

If $1/i_a$ and $1/Ni_b$ are both set as functions of V, then the areas under the curves between the ordinates $V = V_1$ and $V = V_2$ are equal. This implies that the curves must intersect at least once in the interval $V_1 - V_2$. So the same holds for the curves for i_a and Ni_b.

The target considered as a *P-I-N* diode

To obtain the smallest possible dark current and the greatest sensitivity possible in combination with it, then theoretically one must use a layer of an intrinsic photoconductor fitted with two contacts one of which, when current is flowing in a given direction, will hinder electron supply and the alter the hole supply ("blocking contacts"; see *fig.6*). In the "Plumbicon", the P layer acts as the contact hindering the entry of electrons, and the obstacle to the entry of holes is formed by the SnO_2 or by the PbO immediately adjoining it, when this region of the target has become an N-type conductor [6]). The fact that blocking contacts have been attached to the photoconducting layer in the "Plumbicon" constitutes one of the most striking differences with the conventional vidicons.

[6]) See F.A. Kröger, G. Diemer and H.A. Klasens, Phys. Rev. **103**, 279, 1956.

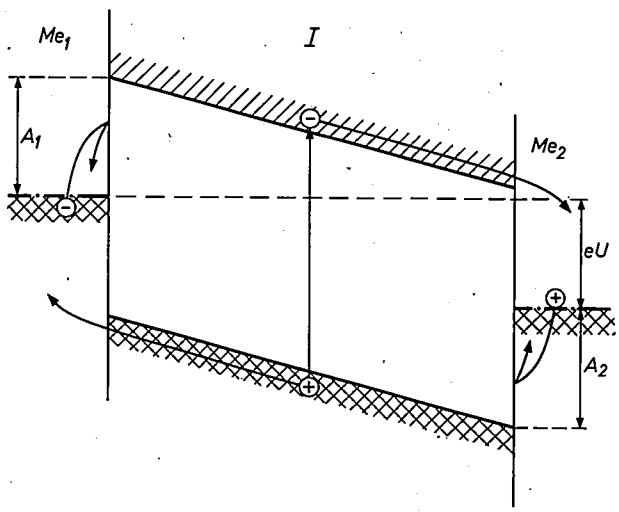


Fig. 6. Energy-band diagram (greatly simplified) of an intrinsic semiconductor (I) fitted with two metal contacts (Me_1 and Me_2). Me_1 has such a high work function (A_1) that electrons contained in it have very little chance of crossing the interface and entering the conduction band of the semiconductor. In the same way holes arriving from the metal Me_2 are unable to reach the valence band. On the other hand the charge-carriers that the incident light has liberated in the semiconductor, having moved over to the appropriate contact under the action of the electric field due to applied voltage U, encounter no obstacle hindering entry to that contact. (For simplicity, A_1 and A_2 have been chosen equal to $\frac{1}{2}\Delta E$.)

The energy-band diagram of a *P-I-N* diode is indicated in *figs.7a* and *b*. Fig.7*a* shows the details and fig.7*b* is a simplified diagram. It differs from the drawings in fig.6 in that the bands in the I layer are curved. This curvature can be explained as follows. In reality, the I layer can never be completely free from impurities; donor and acceptor centres in certain number are always present. The diagrams apply to the case where both concentrations are relatively high and roughly equal. When the concentrations are low the curved parts merge and the band diagram has the appearance indicated in fig.7*c*. Fig.7*d* shows how this diagram alters when a voltage is applied to the diode. The central layer (I), having by far the greatest resistance, shows the steepest fall-off in potential. However, because of the curvature of the bands the field-strength is not everywhere the same. *Fig.8* shows how the potential variation in the middle layer differs from that of fig.7 when the donor and acceptor concentrations are unequal, so that the layer is either slightly P-type or slightly N-type. In the former case the steepest fall-off in potential (and the higher field-strength) is on the N-contact side; in the latter case, the steepest fall-off is on the P-contact side. The practical significance of all this will be discussed below.

Fig. 9 shows the band diagram of a single contact, along with the more important quantities that allow the shape of the potential barrier to be described,

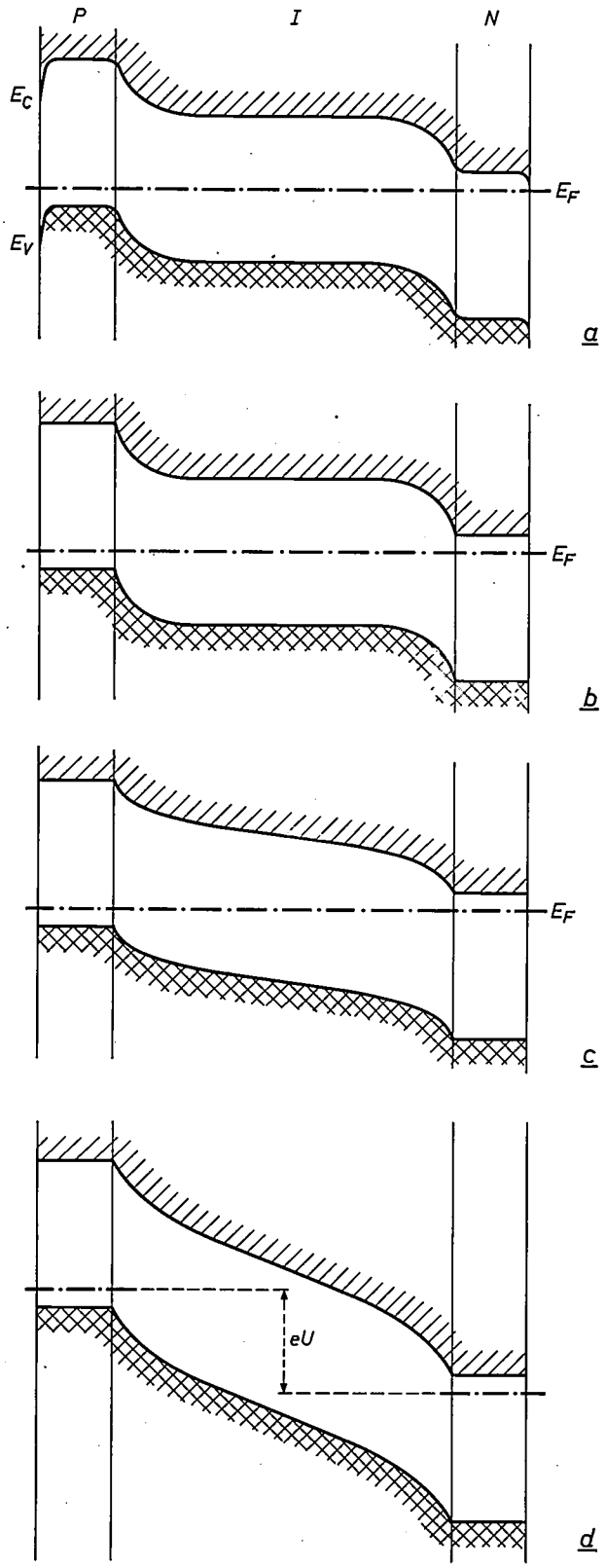


Fig. 7. *a*) Energy-band diagram of *P-I-N* diode. The curvature shown by the bands in the *I* layer is due to the fact that in practice this layer can never be completely free of impurities. In the case illustrated here, the concentration of impurities is relatively high. Diagram (*b*) is a simplified version of (*a*), the marked curvature exhibited in the two junctions having been converted into an abrupt change of slope. Diagram (*c*) is similar to (*b*) but relates to a diode with an *I* layer of such a high purity that the bands are nowhere horizontal and the curved portions meet. Diagram (*d*) shows how (*c*) is modified when a voltage *U* is applied to the diode. It must be made plain that this band diagram is itself a simplification since it implies that the target material has a homogeneous structure while in reality the PbO is made up of a large number of small crystallites. However, the diagram is quite adequate for the purpose of explaining the action of the diode and its more important properties.

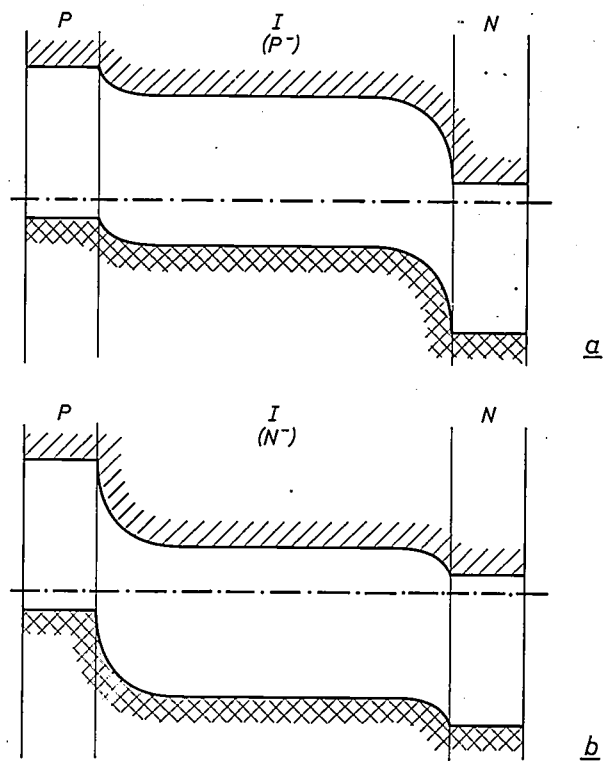


Fig. 8. Band diagrams simplified in the same way as fig. 7*b* and relating to *P-I-N* diodes whose *I* layers are slightly *P*-type (*a*) and slightly *N*-type (*b*). In these cases there is a fairly steep fall-off of potential close to one contact and a region with a gradual potential fall-off (and consequently a very low field-strength) near the other.

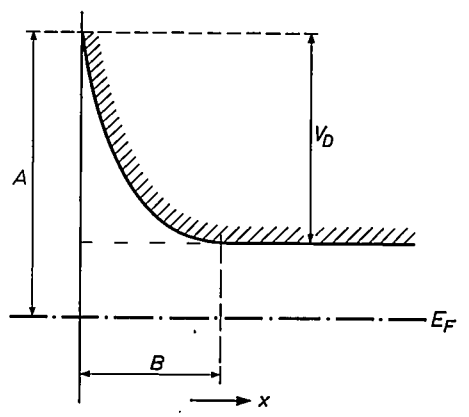


Fig. 9. Near the interface between two regions of different conduction type a potential barrier arises which can be described in terms of the work function *A*, the diffusion potential V_D, and the width *B* of the barrier [7]).

namely the work function (barrier height) A, the width B of the region in which the bands are curved, and the diffusion potential V_D [7]). B is given by the formula

$$B = \sqrt{\frac{2\varepsilon(V_D + U)}{eN_D}}, \quad \ldots \quad (1)$$

and the capacitance c' per unit area by

$$c' = \frac{\varepsilon}{B} = \sqrt{\frac{\varepsilon\, eN_D}{2(V_D + U)}}. \quad \ldots \quad (2)$$

Here ε and e are the dielectric constant and the absolute value of the electric charge respectively, N_D is the (small) donor concentration in the I region, and U is the applied voltage. Further, for the field-strength F in the barrier,

$$F = \frac{eN_D(x-B)}{\varepsilon}, \quad \ldots \quad (3)$$

in which x is the distance to the junction. Use is made of these formulae in the following section.

The dark current

As we have seen, the dark current is mainly determined by the inverse current through the P-I and I-N junctions. We shall now consider the magnitude of this current, taking only one of the contacts and only the kind of charge-carrier obstructed by that contact — the P-I junction, say, and the electron current. For such a contact, when reversed biased, the density j_n of the electron current is given by

$$j_n(U) = n(0)\, e\mu_n\, F(0) = -n(0)\, e\mu_n \sqrt{\frac{2eN_D(V_D+U)}{\varepsilon}}, \quad \ldots \quad (4)$$

where μ_n is the mobility of the electrons and $n(0)$ is the electron concentration in the boundary plane. The value of $n(0)$ depends on the absolute temperature T, the work function A and the constant N_c (sometimes called the effective density of the states in the conduction band), being connected by the formula

$$n(0) = N_c\, e^{-A/kT} \quad \ldots \quad (5)$$

Given the values of U, T, N_D, ε and μ_n, one can use (4) and (5) to calculate the minimum value A must have for the dark current to be smaller than, say, 10^{-8} A (roughly 10% of the signal current; this corresponds, since the target area is a good 3 cm², to a current density of $j_n \approx 3\times10^{-9}$ A/cm²). The results of a number of such calculations have been collected in *Table I*. These are based on the assumption that $U = 50$ V, $T = 300$ °K and $\varepsilon = 12\,\varepsilon_0$ (for non-porous PbO, $\varepsilon = 26\,\varepsilon_0$).

Table I. Minimum values of A (eV), the relative work function of a contact on a photoconducting layer, appropriate to certain combinations of mobility μ_n (cm²/Vs) and donor concentration N_D (cm⁻³). The minimum values are obtained from the requirement that the dark current should not exceed 10^{-8} A when a voltage of 50 V is applied to the diode. (The relative dielectric constant is assumed to be 12, the absolute temperature 300 °K, and the area of the layer 3 cm². The value of V_D has been put equal to zero.)

μ_n \ N_D	10^{14}	10^{15}	10^{16}	10^{17}	10^{18}
1	0.79	0.81	0.84	0.87	0.90
10	0.84	0.87	0.90	0.93	0.96
100	0.90	0.93	0.96	0.99	1.02

As can be seen, over a wide range of μ_n and N_D values, the required value of A lies around 0.9 eV (it should be noted that the N_D value of 10^{14} cm⁻³ is for very pure material, that of 10^{18} cm⁻³ for heavily doped material). We can infer from fig. 7*b* that the band gap ΔE must be equal to or greater than A; in other words, the band gap too must be at least about 0.9 eV. At higher temperatures a bigger work function and a wider band gap will be required (see eq. 2); but smaller A and ΔE values suffice at lower temperatures. Or expressed the other way round: the use of a material with a band gap of less than 0.9 eV is not impossible but then cooling is required to get a sufficiently low dark current. It will now be clear why an extremely low dark current can be achieved with a layer of PbO to which blocking contacts have been attached: the band gap of red PbO is no less than 2.0 eV. It can be seen in *fig. 10* how the dark current in the "Plumbicon" varies with the potential difference across the target.

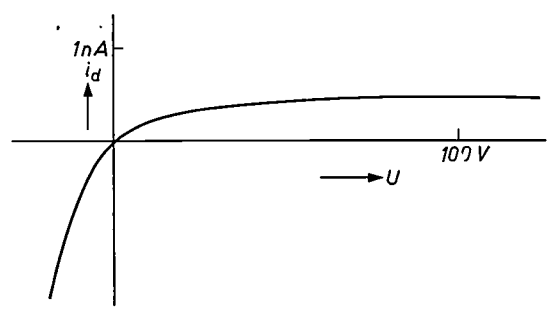


Fig. 10. The dark current i_d as a function of the potential difference U across the target. At no value of U likely to occur in practice does the dark current exceed 0.5×10^{-9} A. As might be expected, the curve resembles a diode characteristic. (In fact, the signal-plate potential V_u has been plotted along the horizontal axis; in practice V_u is nearly equal to U. The same holds for figs. 11 and 12.)

[7]) See the article cited under [3]). Since the thermal work function solid-vacuum does not appear in this article, for simplicity the thermal work function contact-semiconductor A will be referred to as the work function.

In the foregoing it has been tacitly assumed that there is negligible thermal generation of charge-carriers in the *I* region. Calculation confirms that this effect can, in practice, safely be neglected. (Only in the most unfavourable case, viz, that where the impurity levels lie at about half-height in the forbidden zone, can too strong a dark current be obtained when using a material with a band gap of 0.9 eV.)

Sensitivity

We shall now discuss the way in which the photo-current flowing through the PbO depends on the potential difference between the *P* and *N* layers, on the character of the light and on the intensity of illumination, and examine how these relations are affected by the thickness and other characteristics of the sublayers composing the target. *Fig.11* is a graph of the photo-current i_f versus the applied voltage *U* perfectly practicable, the breakdown voltage of the PbO target being so high that a value of 50 V, say, can safely be chosen for the applied voltage.

The reason why saturation is attained so quickly will now be explained. As is well known, a photo-current saturates when the transit time, the time the charge-carriers take to reach a contact, is shorter than their mean life. The transit time depends not only on the applied voltage but also on the field pattern. In an *I* layer with a relatively high concentration of impurities, zones in which the energy bands are curved are rather narrow and in the middle of the layer there is a region of extremely low field-strength (see fig. 7 and formula 1). In these circumstances the transit time is very large and most charge-carriers recombine before reaching a contact. It is not until the applied voltage is raised so high that the zones just referred to extend pretty well throught

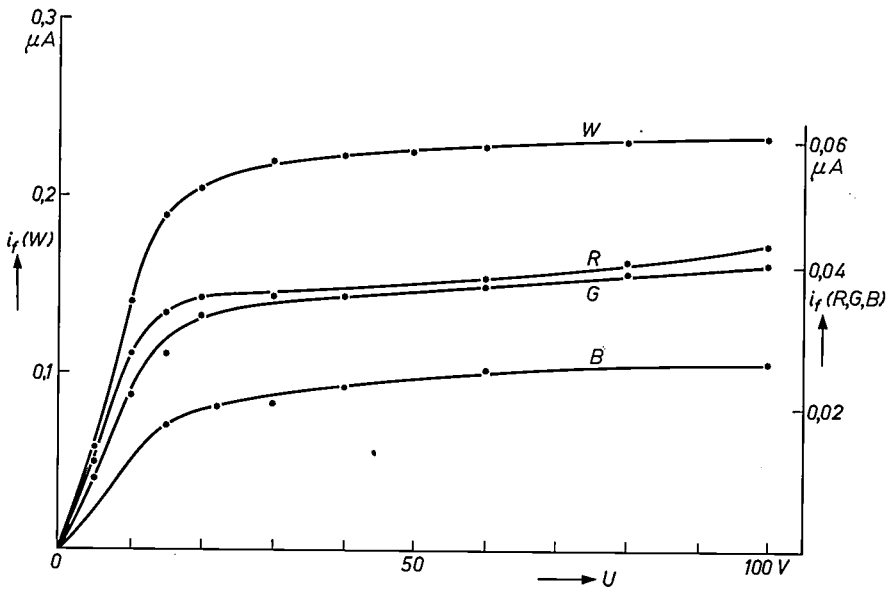


Fig.11. The variation of the photo-current i_f as a function of the signal-platevoltage V_u for different kinds of light, the intensity of illumination remaining the same (measured on a randomly selected tube). Curve *W*, which is to be read in conjunction with the i_f scale on the left, relates to white light with a colour temperature of 2640 °K. Curves *R*, *G* and *B*, to be read in conjunction with the i_f scale on the right, relate to red, green and blue light respectively; these latter were obtained by placing coloured filters in the path of the light, no other detail of the experimental set-up being altered. The filters in question have transmission characteristics similar to those of the red, green and blue filters used in a colour-television camera. (They are Gevaert R 586, G 537 and B 488 respectively.) For all four kinds of light, i_f already attains its saturation value at *U* values of the order of some tens of volts. The curves are slightly concave at the extreme low-voltage end.

for four kinds of light — red, green, blue and white. It will be seen that rising initially with *U*, the curves subsequently flatten out. This saturation, combined with the fact that the quantum efficiency is close to one, is an indication that the contacts are acting as required, and not supplying additional charge-carriers.

It will also be seen that for all kinds of light, the saturation value of i_f is attained at a *U* value that is out the thickness of the layer, that the field-strength is high enough for all the charge-carriers liberated by the incident light to be able to contribute to the flow of current to the fullest extent. Now, in the "Plumbicon" the donor and acceptor concentrations in the *I* layer are so small that the band curvature extends through the whole thickness of the layer, even when the applied voltage is very small.

Moreover the impurity that is inevitably present is of a type that least affects the properties of the tube. To a very limited extent the middle sublayer is a P-type semiconductor. The region where field-strength is lowest and where, accordingly, most of the recombination takes place — from now on we shall call it the "field-free" region — is therefore to be found in that part of the I layer which lies next to the P layer (cf. fig.7*a*). Here the field-free region does much less harm than if it were close to the N layer (fig.7*b*), owing to the way in which PbO absorbs light. Red light is absorbed rather gradually as it passes through the target, but most of the blue is absorbed in the first 5 μm of the target thickness. In a target containing a wide field-free region immediately behind the N layer, charge-carriers generated by blue light make no contribution whatsoever to the photo-current. It is true that the field-free region shrinks as the applied voltage is increased, but those parts of it which lie immediately behind the window, and which absorb the greatest amount of blue light, are the last to be affected. In these circumstances the characteristic for blue light changes at the low-voltage end (*fig. 12*), and a far higher value of applied voltage is required for saturation than when the central layer has P-conducting properties and the field-free region is on the gun side.

From the point of view of service life, it is even an advantage for the I layer to be somewhat P-type, as is shown later.

In approximation, the field-strength can be equated with U/d in cases where the curvature of the bands is only slight, d being the thickness of the layer. The velocity of the charge-carriers is then $\mu U/d$, and for the transit time to be smaller than the mean life τ of the charge-carriers it is required that $d^2/\mu U < \tau$. For a U value of 10 V and a layer thickness of 10 μm the requirement becomes $\mu\tau > 10^{-7}$ cm²/V.

The way in which i_f, the photo-current flowing through the PbO layer, depends on the incident luminous flux L may be found represented graphically in *fig. 13*. Plotted on log-log paper, this function appears as a straight line; that is to say, it can be expressed by a formula of the type $i_f \propto L^\gamma$. The value of the exponent is close to unity. Generally, γ is found to have a value between 0.8 and 1.0, which means that the photo-current is more or less proportional to the incident luminous flux; a single value, uniform at all light levels, can therefore be quoted for the sensitivity of the "Plumbicon" [8]).

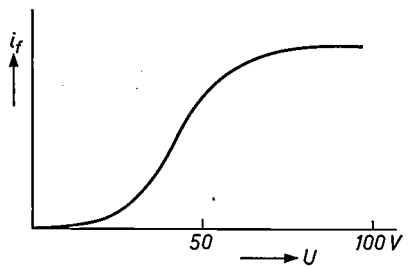


Fig. 12. When the I layer is too strongly N-type, the low-voltage end of the i_f,U characteristic will be depressed, and a high applied voltage will be required for i_f to attain its saturation value. This effect is strongest for blue light.

The significance of this will be further discussed in the article quoted in [2]). For the tube whose characteristic appears in fig. 13, this value is 210 μA/lm. (For the conventional vidicons one value only cannot be quoted; γ differs widely from unity — roughly $\gamma = 0.5$ — and is even dependent on L.)

[8]) Also when i_f is below the saturation value the photo-current is still proportional to the incident luminous flux, and from this fact we can probably conclude that the recombination is monomolecular.

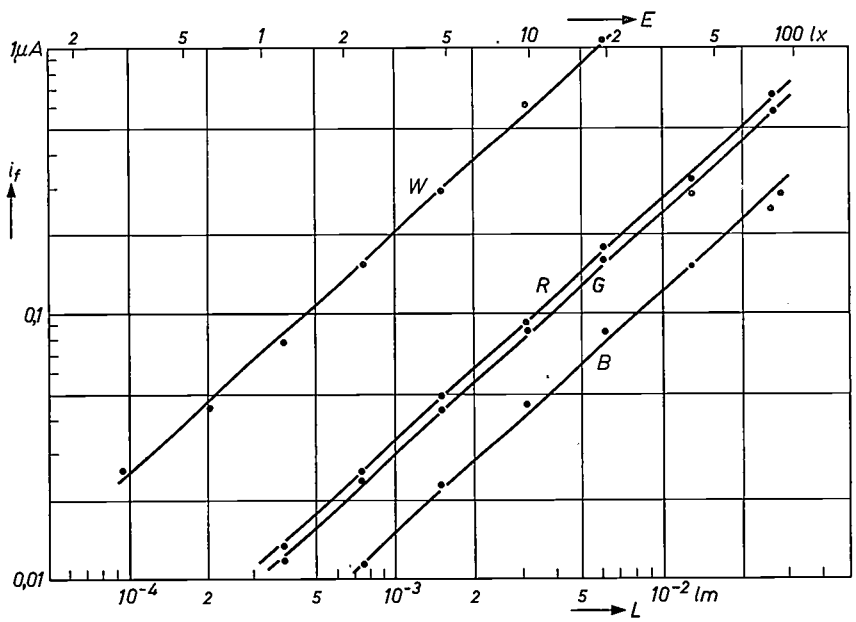


Fig. 13. The measured variation in photo-current i_f as a function of luminous flux L (or illumination E) for white, red, green and blue light (curves W, R, G and B), the applied voltage having been kept at a fixed value. In contrast to conventional vidicons, in the "Plumbicon" L and i_f are roughly proportional ($\gamma \approx 1$) for all kinds of light. Curves R, G and B were obtained in the same manner as the corresponding curves in fig.11. They have been plotted against the luminous-flux values for white light, i.e. the flux incident on the tube before the filter was interposed.

It has been found that tubes made in the same way show very little spread in sensitivity, the widest variation being of the order of 10 μA/lm. Another point of interest is that higher sensitivity values can be achieved, where necessary, e.g. 400 μA/lm, by modifying the deposition process or by depositing a thicker layer of PbO.

When sensitivity to different kinds of light is measured, γ still has a value around unity.

The spectral sensitivity distribution

It has already been pointed out that PbO does not absorb different kinds of light to the same degree. The shorter the wavelength, the higher is the degree of absorption; roughly speaking, blue light is almost completely absorbed after passing through the first 5 μm of PbO, but quite a high proportion of red light passes right through the target (*fig.14*). Clearly, then, the spectral sensitivity of the target can be varied between rather wide limits: a higher relative sensitivity to red light can be obtained by making the *I* layer thicker; maximum sensitivity to blue can be obtained by making the *N*-type layer as thin as possible, and by ensuring (for the reasons explained above) that there is no field-free region in the part of the *I* layer directly adjoining the *N* contact.

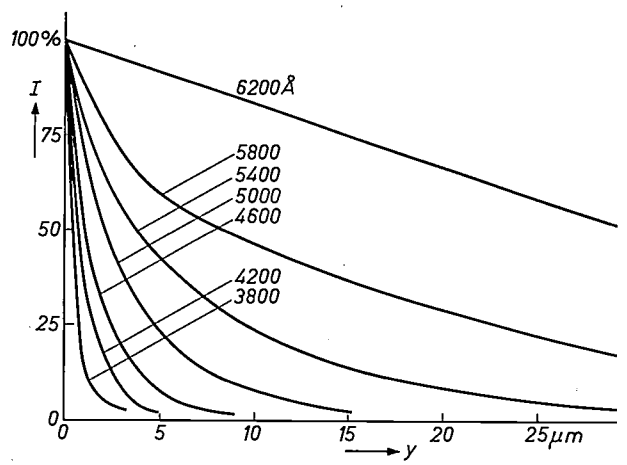


Fig. 14. The absorption of monochromatic light of different wavelength by PbO. The variation of the intensity *I* is plotted against the distance *y* the light has travelled.

The upper limit to the range of wavelengths within which the tube is sensitive is roughly the same as that for red PbO. A band gap of 2.0 eV corresponds, as already mentioned, to a cut-off at 6200Å. The lower limit depends on the thickness of the *N* layer and the distribution of potential in the neighbouring part of the *I* layer. (The upper limit can be shifted to considerably longer wavelengths, without cooling being required, by making the target of a material having a smaller band gap. See the article quoted in [4].)

Fig. 15 shows the spectral sensitivity distribution of two tubes whose PbO targets were intentionally made in a different way. The dashed line represents the response of the human eye. Curve *1* is that of a tube whose PbO target was made by the standard process. By modifying the process to reduce sensitivity to blue — and possibly increase sensitivity to red — the peak of the characteristic can be shifted towards the right. It is even possible to shift it to the right-hand side of the response curve of the eye (curve *2*).

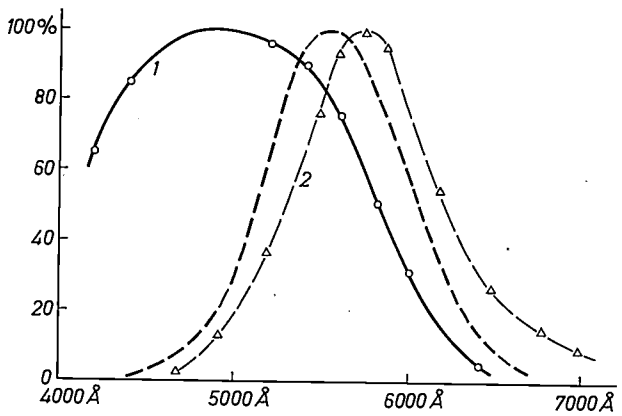


Fig. 15. The spectral sensitivity distribution of two tubes of the "Plumbicon" type. The thick dashed line represents the spectral sensitivity of the eye. Curve *1* is that of a tube whose PbO target was made by the standard process. The peak of the curve can, if desired, be shifted to the right (curve *2*), to the other side of the eye sensitivity curve.

Since the (standard) "Plumbicon" has a spectral sensitivity distribution much closer to that of the human eye than the corresponding characteristics of the ordinary vidicon (Sb_2S_3) or the image orthicon (Ag-Bi-O-Cs), it requires no filters when used in monochrome television, yet gives a far better gradation of colours than the tubes just named. Also for *colour* television the somewhat smaller sensitivity to red of the "Plumbicon", compared to the human eye, does not appear in practice to be a serious objection [2]. In such a case, however, a tube with greater sensitivity to red offers a solution [4].

Resolution

A light-to-dark transition in the image of the scene projected on the window is not reproduced in the video signal with exactly the same abruptness. Apart from the properties of the electronic equipment to which the pick-up tube is connected, and the properties of the electron beam — factors we shall not be discussing here — this is due to the fact that the corresponding transition in the charge image formed on the free surface of the photoconducting

layer between two scans is less sharp. Two effects are responsible for this: in the first place a certain amount of the light is scattered in the target, and in the second place some transport of charge takes place in the PbO parallel to the target surface (cross-conduction). The latter effect can be subdivided into cross-conduction in the *P* layer and cross-conduction in the *I* layer.

It was found that the lack of definition caused by blue light is much less than that caused by red. If the target thickness is reduced, blurring of red outlines becomes less but that of blue remains the same. The reason will be clear if it is remembered that blue light does not penetrate so far into the target as does red (*fig.16*).

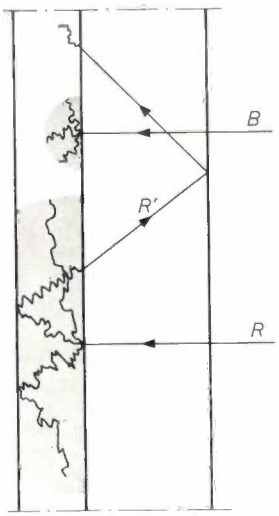


Fig. 16. PbO has a lower absorption for red light than for blue, and consequently the scattering of red light has a more adverse effect than that of blue on the resolving power of the PbO target. The shaded areas indicate, schematically, the extent of the zones penetrated by the diffused light originating from one narrow beam (*R* and *B* resp.).
A certain amount of diffused light escapes from the target by way of the front face of the target, but is then reflected back by the front face of the glass window; it is again the less rapidly absorbed red light (*R'*) that is mainly responsible for the contribution of this effect to the loss of definition.

Consequently from the standpoint of definition it is desirable to make the target as thin as possible. As we have seen, this will involve a reduction in sensitivity to red light; furthermore, the capacitance of the target will be increased and, for reasons which will be explained below, only a limited increase in target capacitance can be accepted. Thus there is not complete freedom of choice of the target thickness. (The difficulty disappears if the PbO is replaced by a basic material with a much greater absorptivity for red light [4].)

Cross-conduction in the *P* layer can naturally be limited by making the layer as thin as possible; and further, the less the doping of the layer, the less cross-conduction there will be. However, as we have seen, if the layer is too lightly doped the dark current will be excessive. Also, as will be explained below, from the viewpoint of service life it is advantageous to make the *P* layer thick and dope it heavily. Accordingly, here again the choice is not entirely free.

Cross-conduction occurs in the *I* layer when, owing to insufficient purity of the material, this layer contains a "field-free" zone. In such a zone the charge-carriers are liable to fan out instead of crossing the layer by the most direct route. This does not matter if the charge-carriers in question are electrons, since the point where these arrive on the signal plate has no bearing on the resolving power. But the place where the holes arrive is of course important: a hole that has not crossed the target by the most direct route alters the distribution of potential on the free target surface in a way that does not correspond to the light pattern of the broadcast scene. So from the viewpoint of resolving power also, it is desirable that the *I* layer be as pure as possible.

Before the values of resolving power relevant to the "Plumbicon" are quoted, first a word about how this resolving power is expressed in figures. Suppose that a pattern like that in the upper half of *fig.17* is being projected on the screen of the tube. The pattern consists of alternate vertical light and dark stripes of the same width. In some parts of the screen the width of the stripes is such that 20 light-and-dark pairs would completely fill the picture height (in the language of the television engineer this is called 40 "lines"); elsewhere this number is 200 (400

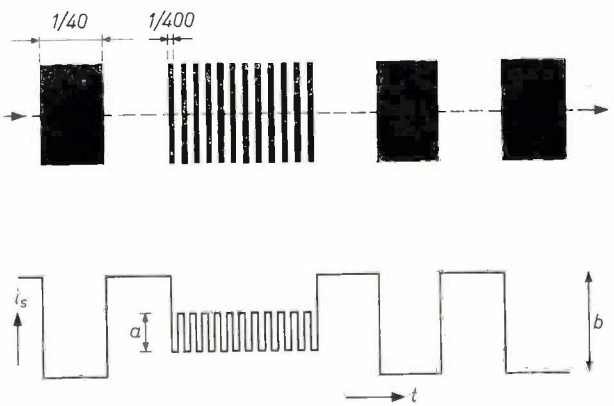


Fig. 17. Explanation of how the resolving power of a television camera tube is expressed in figures. A pattern of alternate light and dark vertical stripes, some having a breadth of 1/40th of the picture height and others a breadth of 1/400th of the picture height, is projected on the screen. This pattern is scanned in the direction of the dashed line. The broad and the narrow stripes give rise to alternating voltages with fundamental frequencies of 0.5 Mc/s and 5 Mc/s respectively. The ratio between amplitudes *a* and *b*, expressed as a percentage, and known as modulation depth, provides the required measure for resolving power.

"lines"). If an electron beam scans the corresponding charge image in the direction of the broken line, the signal current (fig.17, lower half) will have the form of an alternating current with fundamental frequencies of 0.5 and 5 Mc/s respectively (in the case of a 625-line system with a frame period of 1/25th second). Parts of the signal current corresponding to the broad dark stripes will have approximately the dark-current value, but the narrow dark stripes will give rise to higher current values. Parts of the signal current arising from the broad light stripes will have the same value as if the window were illuminated over its whole surface; the narrow stripes will yield lower current values. Let the letters a and b denote the difference between the light and dark values of i_f in the region of fast and slow alternations respectively. The ratio a/b expressed as a percentage, and known as the modulation depth, is commonly adopted as a measure of resolving power.

A more detailed impression of the tube properties is obtained if, instead of restricting the measurement of a to a pattern with 400 lines per picture height, the variation of a/b is investigated when a number of patterns with different stripe breadths are used. An example of this kind of measurement, done on a random-selected tube of the "Plumbicon" type, is shown in *fig.18*. It will be noted that at 400 lines the tube under investigation had an a/b ratio of about 45%. In general a/b is found to be not less than about 35%.

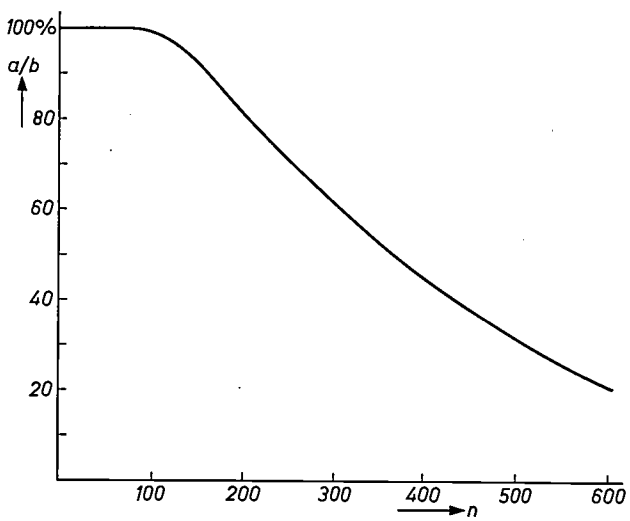


Fig. 18. A detailed description of resolving power can be obtained by plotting the variation in a/b against the breadth of the narrow stripes in the pattern appearing in fig. 17, the stripe breadth being expressed as the number n of black-white pairs of such stripes that would be required to fill one picture height. The figure shows the results of measurements done with white light on a random-selected tube of the "Plumbicon" type having a PbO layer 20 μm thick. As is inevitable, the ancillary equipment causes some interference; so normally such measurements give a figure that is lower than corresponds to the real performance of the tube. By adopting a very careful procedure, the difference is here reduced to a minimum.

Speed of response

When a sudden change occurs in the luminous flux incident on the photosensitive layer of a pick-up tube, the signal current does not immediately reach its new equilibrium value. When the target of a tube undergoes the variations of illumination shown in *fig.19a*, in the signal current the forms of response

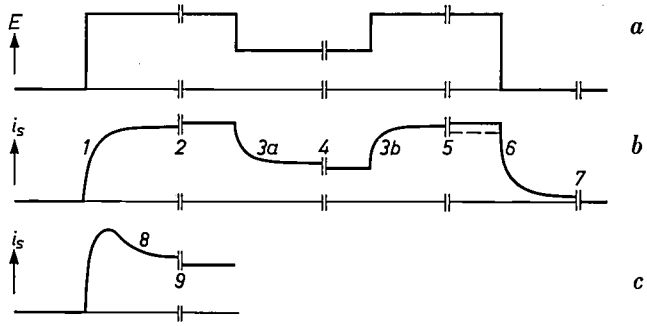


Fig. 19. The various forms of delayed response which, in principle, a "Plumbicon" may exhibit. In a good tube they are too slight to cause any trouble, and some are absent altogether. Diagram (*a*) shows the programme of varying intensity of illumination presented to the tube: the light is switched on, decreased in intensity after an appreciable time has elapsed, then brought back to its full intensity and finally switched off. The various forms of delayed response are represented in diagram (*b*). They are: 1) black-white inertia, 2) black-white trailing (a slight increase over a long period), 3a) and 3b) intermediate inertia, 4) intermediate trailing, 5) persistence (which may also manifest itself in *attenuation* of the signal), 6) white-black inertia, and 7) white-black trailing. Instead of effects (1) and (2) the tube may exhibit, see diagram (*c*): 8) fatigue and 9) fatigue trailing.

illustrated in fig.19*b* may be encountered. The names that we have given these inertial effects may be found in the caption to that figure.

From the practical point of view one of the most important forms of response is "intermediate inertia". (Inertia effect occurring when the light intensity changes from white to grey or grey to white.) Black-white inertia only becomes noticeable under the most unfavourable conditions, and white-black trailing can be compensated electrically. Intermediate inertia is always fully manifest in the received picture, however, and the same applies to persistence and fatigue effects.

All the above forms of response can be regarded as resultants of two components, that are due to 1) the electron beam being incapable of supplying an unlimited amount of charge during the brief time in which it is directed on a given picture element; this component is accordingly called *beam-current-* or *discharge lag*, and 2) the presence of traps in the *I* layer. In the "Plumbicon" the contribution made

by both these factors have been reduced to a satisfactorily low level.

Both the discharge lag and that due to the presence of traps decrease with increasing target voltage U, although it should be observed that no further appreciable increase in the speed of response is achieved above a certain value of the applied voltage. The traces in *fig. 20* show this effect. It will be noticed that the highest speed of response is attained at the relatively low U value of 30 V. This speed of response is very satisfactory; the change in signal level resulting from a transition from strong to less strong illumination (and involving intermediate inertia) is 95% complete after only 3/50ths of a second.

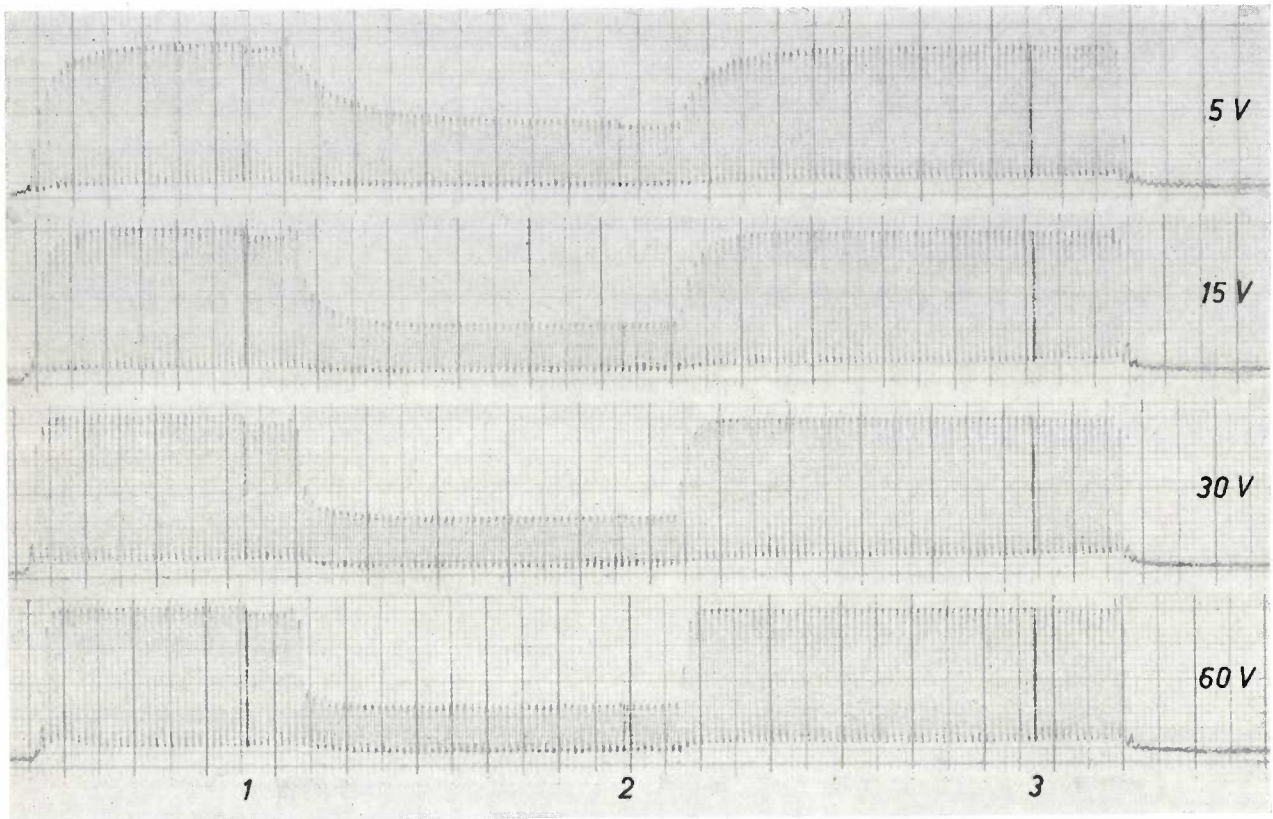


Fig. 20. The response of a randomly selected "Plumbicon" tube at four values of U, the voltage across the PbO target. At the relatively low U value of 30 V the response of the tube is already quite fast, and there is scarcely any persistence. The traces were obtained by illuminating the top half of the screen only, so that the signal current consisted of a train of pulses spaced at intervals of 1/50 s (*half* a frame period because of the interlaced scanning). The programme of illumination was equal to that of fig. 19 and began in each case with a dark period lasting one minute. At the points *1*, *2* and *3* the reading was interrupted for 10, 10 and 30 seconds respectively.

The discharge lag

When we were discussing how the potential of the free surface of the PbO target adjusts itself to an equilibrium value, we found that the current flowing towards that surface was dependent on its potential V (fig. 3). The deflection and subsequent fall-off in the curve of current versus potential was explained as a result of an increase in secondary emission. The initial rise (V low) is due to the fact that the electrons leaving the cathode do not all have the same velocity in the axial direction, the axial components of velocity having a quasi-Maxwell distribution. To a good approximation, the rising portion of the curve can be described by an expression of the form $i_a = ae^{b\Phi}$, where a is a constant connected with the beam-current and b is inversely proportional to the cathode temperature; Φ differs from V by a small constant amount.

Over the range of illumination values within which V remains small and the above formula accordingly remains valid, the discharge lag has the following properties:

1) It is independent of the beam-current.
2) It decreases with decreasing target capacitance C.
3) When the light intensity drops, the higher the new intensity the shorter the discharge lag.
4) The lower the cathode temperature, the shorter the lag.

Property (2) explains why the discharge lag initially decreases as voltage U is increased. At low U values the I layer contains a field-free region which,

in effect, forms part of the adjoining contact. This region becomes narrower and narrower as U increases. Consequently the distance between the capacitor "plates" becomes greater and the target capacitance decreases, giving a shorter discharge lag; see formula (2).

At higher intensities of illumination, corresponding to the region in which the i_a versus V curve starts to change direction, (1) ceases to apply: the discharge lag decreases with increasing beam-current when the light level is high. Property (2), relating to the low target capacitance desirable, is retained at high intensities of illumination; properties (3) and (4), for obvious reasons, are not.

In practice, for the decrease of the discharge lag, only the decrease of the target capacitance can be considered. If a time constant shorter than about 1/25th of a second is required for the response to a change from dark to light (to a moderate intensity of illumination giving rise to a photo-current of 10^{-8} A, say), then the target capacitance must not exceed about 2000 pF. Also at high intensities of illumination the discharge lag will continue to be negligibly short, provided it is possible to keep the target capacitance below 2000 pF.

These requirements can be fulfilled very comfortably with a target having the *P-I-N* structure described above. For example, suppose that the PbO layer is 10 μm thick and that it has an I layer so pure that it contains no field-free region at the U values employed in practice — when discussing sensitivity we saw that this degree of purity can actually be attained — then the target will have a capacitance ranging from 1000 to 1500 pF.

The four properties of the discharge lag at relatively low illumination levels, as listed above, can be inferred from the following. As before, we shall first assume that the beam is supplying charge *continuously*. The behaviour of V, the potential of the free surface of the target, is in accordance (cf. figs. 2*b*, 3, 4 and 5) with the differential equation

$$\frac{dV}{dt} = \frac{1}{C}(i_f - i_a).$$

(The whole target is imagined to be evenly illuminated; but the above equation applies equally to a single picture element, the capacitance C and the (negative) charge supply then being smaller in the same proportion.) Further:

$$i_a = ae^{b\Phi}.$$

Now suppose that a change in illumination abruptly raises the photo-current i_f from the value i_1 to the value i_2; the problem is to ascertain how i_a, initially equal to i_1, arrives at its new value of i_2.

On solving the above set of equations and eliminating V, we obtain:

$$i_a(t) = \frac{i_2}{1-(1-i_2/i_1)\exp(-tbi_2/C)}. \quad \ldots \quad (6)$$

The four properties can be directly inferred from this formula. It can be shown, namely, that (6) remains valid when the fact that the beam supplies charge in surges is taken into account.

Just as there is, as we have seen, an upper limit to the capacitance of the target, there is also a lower one: to prevent the oncoming electrons from being excessively deflected by the charges on the target, the potential V must not vary over too wide a range during a frame period. Suppose for example that a limit of 10 V is placed on the variation of V and that a photo-current of up to 10^{-7} A is required, then C must not be smaller than about 800 pF.

The capacitance of the PbO layer in the "Plumbicon" has, it appears, a value such that on the one hand the lag is sufficiently short and on the other hand the influence of the electron beam on neighbouring picture elements can be neglected. The signal given by a picture element depends solely upon the intensity of the incident light, and is independent of its position, of its history, and of the situation in the surrounding picture elements. This is why the tube is so very suitable for colour television [2]).

Lag due to traps

Let us now go a little more deeply into the type of inertial response caused by traps. It is known from photoconductor theory [3]) that the presence of traps does not in the first instance affect the relationship between intensity of illumination and the steady-state electron concentration in the conduction band, or similar relationship; it does however have a bearing on the *speed* with which a new situation supervenes when the intensity of illumination is altered. Generally there are many more electrons in the traps than in the conduction band, the respective concentrations being c_t and c_c (much the same thing applies to the holes). If the illumination E is increased, with a consequent increase in c_c, the c_t/c_c ratio must nevertheless remain the same, and this implies a large absolute increase in the number of trapped electrons. Initially the demand is largely supplied by electrons liberated by the light incident on the target; c_c cannot therefore jump directly to the value corresponding to the change in E. The c_t/c_c ratio is proportional to N_t, the concentration of traps; the greater N_t is, then, the greater will be the (absolute) deviation of c_c at a given time after the increase in E. Similar reasoning applies if E is reduced: when this happens a large number of filled traps have to "dry out", and the higher the concentration of such centres, the greater will be the absolute deviation of c_c.

Since the concentration in the conduction band

falls off with increasing U, it is understandable that these inertial effects should become less noticeable at higher values of applied voltage.

Apart from these *direct* consequences of the presence of traps, which will not be further analysed, the traps have an *indirect* effect which at low U values may be clearly manifest in the response of the tube. The capture of charge-carriers in traps can modify the curvature of the energy bands in the I layer so drastically that the characteristics of the tube are affected. Any such modification is purely temporary, of course, beginning subsequent to a change in E and proceeding at the same rate as the establishment of a new value of space charge (c_c+c_t), so that for the observer it has the character of an inertial effect. We shall now look a little more deeply into this effect; for simplicity we take as an example the phenomena of *fatigue* and *black-white inertia*. To obviate misunderstanding it should be pointed out that in the discussion which follows, cases will only be considered in which the applied voltage U is chosen so small that a field-free region is present in the I layer.

Fatigue, a slow decay in photo-current following an abrupt increase, occurs because charge-carriers are trapped in such a way that the field-free region expands. Here a distinction must be made between cases in which the I layer is slightly N-type, and those in which it is slightly P-type.

As we have seen, if the I layer is slightly N-type there will be a field-free zone on the N-layer side when U is small. In consequence of hole capture this zone widens (see *fig. 21a*) with the result that the sensitivity of the target is modified, particularly its sensitivity to blue.

If the I layer is slightly P-type there will be a field-free zone on the P-contact side, which widens owing to electron capture (fig. 21*b*). This affects the tube's sensitivity to red light.

Under normal operating conditions U is high enough to eliminate the field-free zone or at least to reduce it to such small dimensions that the effects just described are of little importance.

In the case of *black-white inertia* there is, in addition to the direct effect of charge-carrier capture, a side-effect that is the opposite of that just discussed; here the field-free zone *shrinks*, causing the sensitivity of the target to increase (see *fig. 22*).

It will be clear that the properties of a tube in regard to speed of response allow conclusions to be drawn about the nature of the I layer. If for example the tube exhibits both fatigue and black-white inertia, this is an indication that both kinds of charge-carrier are being trapped, though the probabilities of capture are different for the two types.

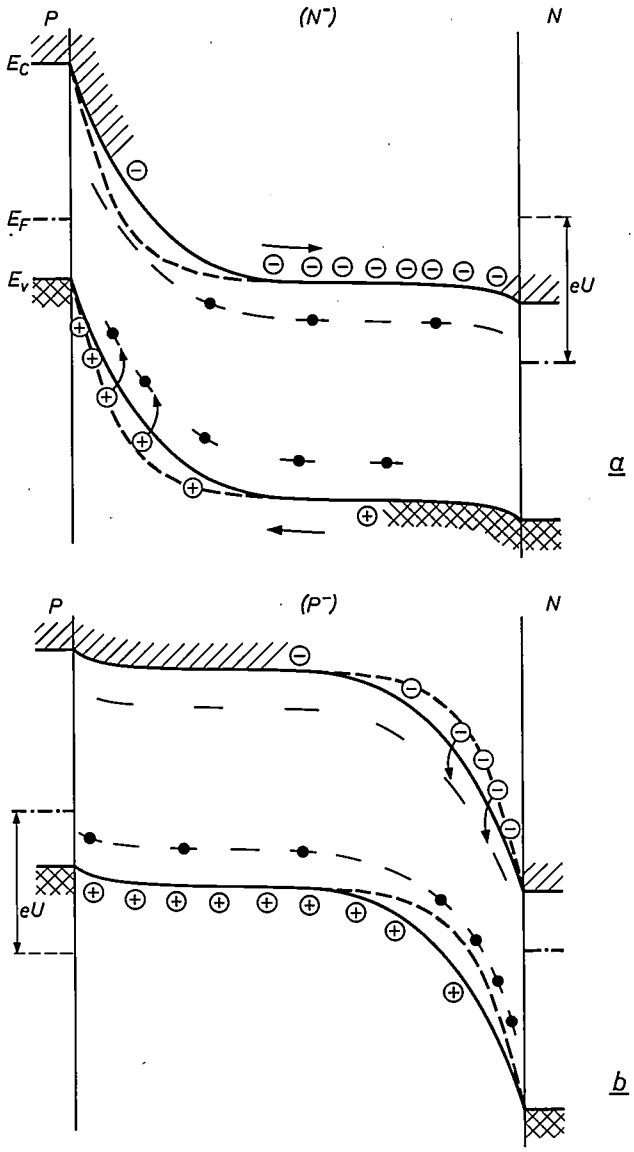


Fig. 21. To explain "fatigue".

Finally, it must be pointed out that there is yet another form of delayed response which is likewise a consequence of a change in the curvature of the energy bands but, in contrary, becomes more pronounced as U increases. This effect occurs in old tubes when, on account of long use, the P contact ceases to have an adequate blocking action. In such a case hole capture lowers the height A of the barrier to a point such that it no longer prevents the flow of current. So long as the target is illuminated and a relatively heavy photo-current continues to flow, this component will not be noticed; but it persists when the incident light has been removed, until the captured holes have left the traps. This "stimulated dark current" (one type of white-black trailing) is quite unacceptable and its occurrence is an indication that the tube has reached the end of its

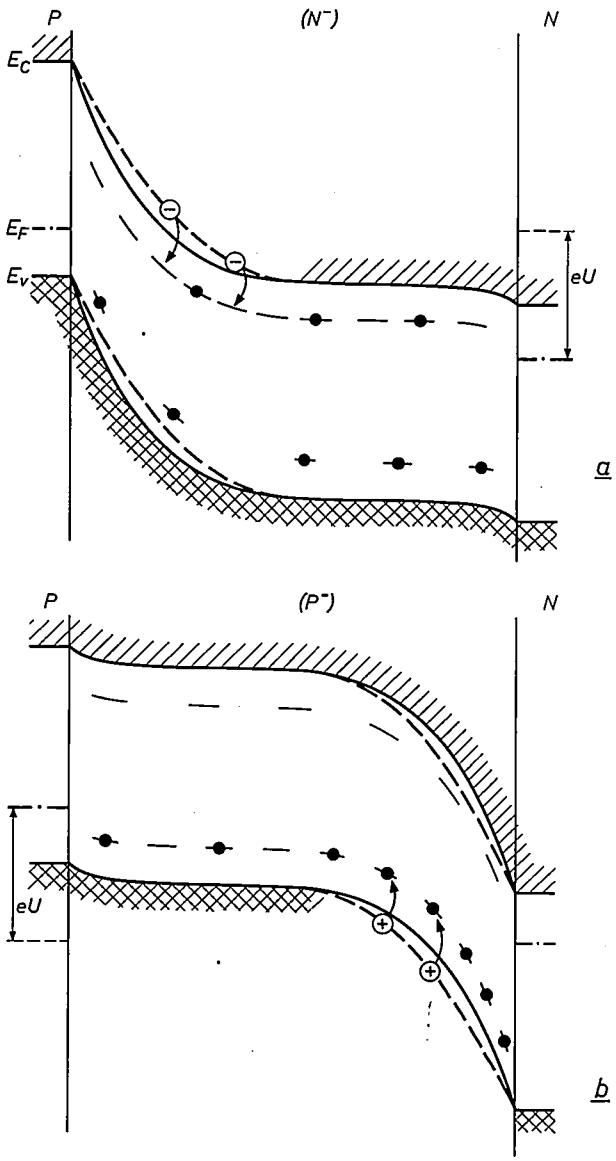


Fig. 22. To explain "black-white trailing".

service life. In the next section we shall see why the P contact deteriorates in this way.

Constancy of tube properties; service life

Any variation in the properties of a tube of the "Plumbicon" type in course of time is the result of the changes that take place in the PbO target. We shall review the most important of these changes and, amongst other things, explain how they ultimately make the tube unserviceable.

The changes in question are caused in the first instance by the diffusion of excess oxygen in the PbO target. As a result, small irregularities in oxygen concentration in the P and I layers are to some extent evened out. In addition, the transition between the P and I layer becomes less sharp. The evening-out of irregularities in the I layer naturally gives rise to a more uniform fall-off of potential through the layer, and hence to greater sensitivity and a faster response. More important than internal oxygen migration — at least from the viewpoint of service life — is the *overall loss* of oxygen from the PbO target. Oxygen in gaseous form cannot remain free for long enough to build up any pressure within the tube; it immediately combines with a barium getter, or with residual gases, and in these circumstances the PbO slowly decomposes. The P layer, because it is at the free surface side of the target, is the one most affected. But owing to the porous nature of the target the regions more remote from the free surface, and in particular the I layer, gradually lose oxygen too. The oxygen loss of the P layer is accelerated by ion bombardment while the tube is in operation. Further, it is possible that the P layer loses some oxygen by electrolysis within the PbO.

One consequence of this removal of oxygen is that the P layer loses some of its P-type conductivity; the same applies to the I layer which, it will be remembered, is also to some extent P-type. It will now be clear why, for long service life too, it is advisable to make the I layer slightly P-type. In consequence of oxygen loss a truly intrinsic I layer would gradually become N-type, which as we have seen would not be altogether desirable.

The decrease in the P-conductivity of the P layer is no disadvantage in the first instance; from one point of view it is even the reverse. The decrease of conductivity also cuts down cross-conduction, of course, so improving the resolving power of the target. In the long run, however, the height of the barrier is reduced to such a point that the tube starts to exhibit the form of sluggish response ("stimulated dark current") discussed at the end of the preceding section; at a later stage the normal dark current also becomes excessive. When this happens, the tube has come to the end of its service life.

Very occasionally a tube becomes unserviceable because of the sudden appearance of a white speck in the received picture. The cause is again an excessive dark current, but here it is restricted to a small part of the target area. Speckling will be discussed in the next section.

Fig. 23 is a plot of a number of tube properties with respect to operating time. It will be noted that the overall life of the tube falls into two distinct parts. The earlier period is one of rapid change; a sort of "forming" process takes place. During this time the tube remains in the factory. The tube is ready for use as soon as it has entered the second phase of its life. From then on the quantities determining serviceability remain more or less constant for a

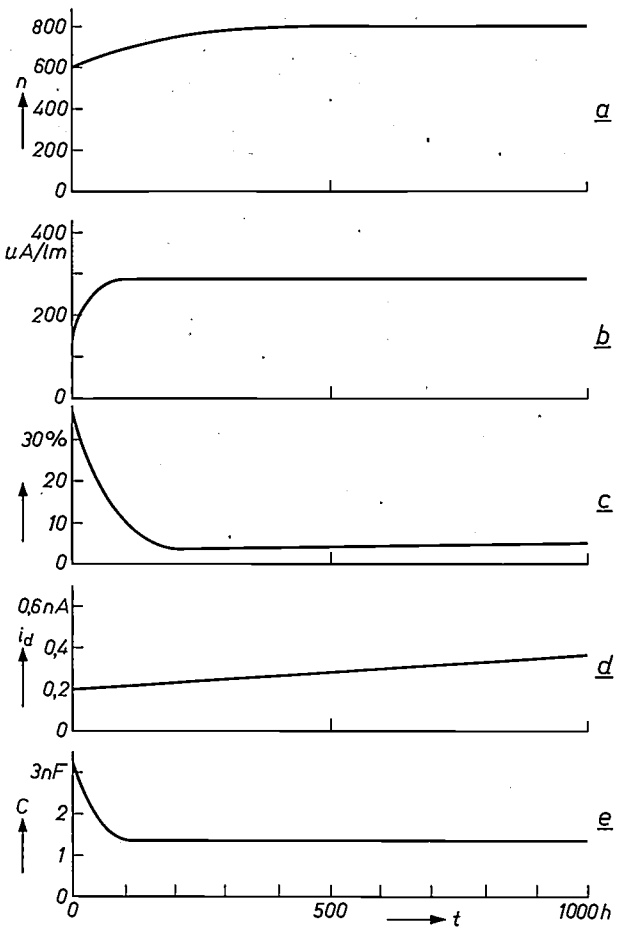


Fig. 23. Changes in various tube characteristics and in the capacitance of the PbO target, measured on a randomly selected tube of the "Plumbicon" type, during the first 1000 hours in operation. *a*) Resolving power, shown as the number of lines that can just be resolved by the eye (the corresponding *a*/*b* ratio — see fig. 18 — is roughly 5%). *b*) The sensitivity of the tube. *c*) The intermediate inertia, expressed as the proportion of the required change in signal current that has still to take place after a lapse of 3/50 s following an abrupt change in illumination. *d*) The dark current. *e*) The capacitance *C* of the PbO target.

Initially, the characteristics of the tube under test changed considerably — a normal feature of the "Plumbicon"; the tubes remain in the factory during this early phase of their life. Thereafter their characteristics show a high degree of stability, as can be seen from the diagrams.

lengthy time (sensitivity and resolving power) or increase extremely slowly (dark current and the various forms of sluggish response). The most direct evidence of the changes that take place in the target, early on in the life of the tube, is obtainable from the curve of target capacitance as a function of time (fig. 23*e*). Since the target capacitance is inversely proportional to the effective thickness (see eq. 2), the variation in effective target thickness during the life of the tube can be determined directly from capacitance measurements. By the same means it is also possible to investigate how target behaviour is affected by changes in U or in the intensity of illumination.

Speckling

It has been stated that the PbO target can be regarded as a large flat *P-I-N* diode; alternatively, since there is little cross-conduction, it can also be viewed as an aggregate of small diodes lying side by side corresponding to the picture elements. To obtain a good picture at the receiver it is necessary that all these little diodes should be roughly similar. If there is one picture element with properties differing greatly from those of the others, it will be visible as a speck in the received picture. Depending on the nature of the defect, the speck may be light or dark, sharp or diffuse, and of constant or variable luminous intensity. If the brightness variation is periodic the speck is said to "twinkle". We shall now briefly review the commonest defects and the kinds of speck they give rise to.

1) Regions of high local *P*-conductivity occur in the *I* layer, such that the fall-off in potential is very steep near the *N* contact, with the result that the barrier is extremely narrow. Thus a strong dark current flows and a white speck is produced. This strong dark current can be the cause of an "avalanche effect" owing to the very high field-strength, or of the tunnel effect: if the barrier is very small, a hole in a conduction-band level of the *N* region can be moved by the tunnel effect to the *I* region, and there appear in an equally high level of the valence band (see *fig. 24*). With regard to the avalanche effect it is very conceivable that the dark current can periodically vary in strength. In this case the speck "twinkles".
2) A similar situation arises if parts of the *I* layer have a relatively high *N*-conductivity; but now it is the *P*-layer side that is affected. The dark current resulting from the tunnel effect is now not a hole but an electron current.
3) Inadequate doping of a part of the *P* layer. This part of the target will have a much shorter life than the rest. A heavy dark current will soon start to flow there, producing a white speck in the received picture and making the tube unserviceable.
4) Excessive thickness of the *P* layer in a part of the target (for example, because it has been doped excessively). Accordingly, the *I* layer is then too thin in the affected target area, and here the sensitivity — particularly sensitivity to red — is less than it is elsewhere in the target. The result is a dark speck in the received picture.
5) Excessive thickness of a part of the *N* layer. The result is again a local loss of sensitivity giving rise to a dark speck, but it is now the blue sensitivity that is mainly affected.

The defects listed above are essentially irregularities in the multilayer structure of the target; in addition, defects of a purely mechanical nature — due for example to the presence of a dust particle — are also possible.

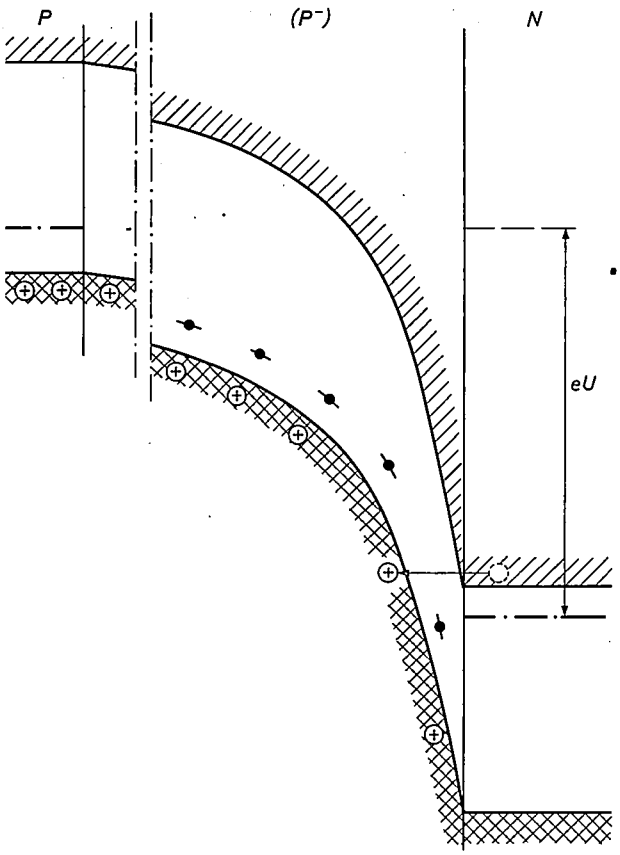


Fig. 24. To explain the occurrence of a strong dark current of holes, when the inner sublayer contains a limited region of high *P*-conductivity. The barrier can then be so narrow that tunnelling is possible (see arrow).

By arranging for fabrication of the tube under carefully controlled conditions in dust-free rooms, each part being subjected to rigid inspection, it has been possible to eliminate most of the causes of these defects, and the PbO targets now being produced are accordingly of consistently high quality in regard to freedom from speckling.

Summarizing the properties of camera tubes of the "Plumbicon" type, as discussed above, we can say first of all that the multilayer structure of the target leads to a very small dark current and to a much higher sensitivity than is obtainable with a simple *P-N* junction. In addition, it is possible by suitably choosing the individual layer parameters to give the tube an excellent spectral sensitivity distribution, a very good resolving power and a high speed of response. The "Plumbicon" has made it a practical proposition to build small, lightweight television cameras with very good properties.

We have also seen that there is scope for a fairly wide range of variations on the basic design: some properties, such as the spectral sensitivity distribution, can be modified quite drastically without severe interference with the others. A further wide range of possibilities becomes available when the basic material PbO is abandoned in favour of PbO-containing material having a band gap smaller than 2.0 eV [4]).

In virtue of all these opportunities for varying the characteristics of the tube it is possible, for example, to make pick-up tubes of the "Plumbicon" type that possess a very high resolving power together with high sensitivity, qualities desirable in medical applications when the tube is used for radiological examination; or to make tubes combining high sensitivity with long service life and a reasonable speed of response, the qualities desirable in industrial closed-circuit applications and in traffic control at night; or to make tubes which, without any sacrifice of studio requirements for equipment to be used in live broadcast television, are eminently suitable for incorporation in colour-TV cameras.

Summary. The "Plumbicon", a new television pick-up tube, is a kind of vidicon whose photoconducting target is built up of micro-crystalline PbO. This PbO layer consists of a *P*-type sublayer at the gun side, an intrinsic sublayer (*I*) and possibly a thin *N*-type sublayer next to the signal plate. The signal plate is made from *N*-type SnO_2. The PbO layer and the signal plate form a unit having the properties of a *P-I-N* diode: the *P*-type sublayer hinders the entrance of electrons to the *I* layer, the signal plate (or the *N*-type sublayer) hindering the holes. When the tube is in operation this *P-I-N* diode is reverse-biased; the dark current (i.e. the inverse current through the diode) is therefore very small. The sensitivity of the tube, which is determined by the thickness of the *N* and *I* layers and by the distribution of potential through the *I* layer, can exceed 200 μA/lm; its spectral sensitivity can be matched to the human eye more closely than is the case for existing camera tubes, but this characteristic can be modified within wide limits. The thinner the *N*-type sublayer, the greater is the sensitivity of the target to blue light. Its overall sensitivity increases with the thickness of the *I* layer. Gamma is close to unity. Resolving power and speed of response are excellent (the depth of modulation is is about 35% at 400 lines; response time approx. 3/50th of a second). The "Plumbicon" has a longer service life than other studio-quality tubes. On all points important in television broadcasting the "Plumbicon" equals or betters existing tubes, and above all when it is employed for colour TV. The tube's more important properties can be drastically modified without prejudicing the others, and because of this it is possible to make versions that are specially suited to widely divergent applications. A further wide range of designs becomes available if the basic target material is replaced by PbO-containing material having a smaller band gap (minimum 0.9 eV).

MEASURING THE LIGHT-INTENSITY DISTRIBUTION IN THE SPOT ON A CATHODE-RAY-TUBE SCREEN

535.247.4:621.385.832

For judging the quality of the electron gun of a cathode-ray tube the diameter of the focus on the fluorescent screen is an important quantity. This diameter is defined with the aid of the diametrical distribution of the light intensity in the spot — theoretically a Gaussian curve — as that width within which the intensity is greater than 1/e of the peak value. It is therefore important to be able to measure the intensity distribution. For this purpose various set-ups are used in the Philips laboratories and elsewhere. Although our set-up contains no essential elements that are not used by others, it may still be useful to describe it here and to mention some general practical experience which we and others have gained over the years in the design of set-ups for such measurements.

Fig. 1a shows the principle of our set-up. The gun to be investigated is mounted in a testing tube *A*. The non-deflected beam is focused on the fluorescent screen, and an optically magnified image of the spot *F* is made to pass over a slit *S* and a photomultiplier *B* behind the slit. Scanning with a slit instead of a hole is possible because we assume the beam to have circular symmetry (see fig. 1*b*) [1]). The advantage is that more light falls on the photomultiplier, and moreover the often troublesome graininess of the fluorescent screen is averaged out. A pen recorder *Reg* connected to the photomultiplier traces a curve from which the intensity distribution in the spot can be calculated.

There are various methods of scanning the spot image. In one of them the slit is moved by means of a micrometer screw [2]), and in another the spot is projected via a mirror onto the slit, the mirror being rotated [3]). In our set-up (*fig. 2*) the electron beam is slightly deflected with the aid of a coil, the slit remaining stationary.

If a beam with a voltage of 15 kV and a current of 0.1 to 1 mA were to bombard the fluorescent screen continuously, the latter would very soon be damaged. For this reason the beam is intermittently suppressed by using pulse technique, so that electrons strike the screen only during a fraction 10^{-4} to 10^{-5} of the time. The pulses may, for example, be applied to the Wehnelt cylinder of the gun (*W* in *fig. 3*). The simplest procedure is then to give the peaks of the pulses during which the electron beam is passed a fixed potential, e.g. earth potential. The beam current can then be adjusted by means of the cathode voltage.

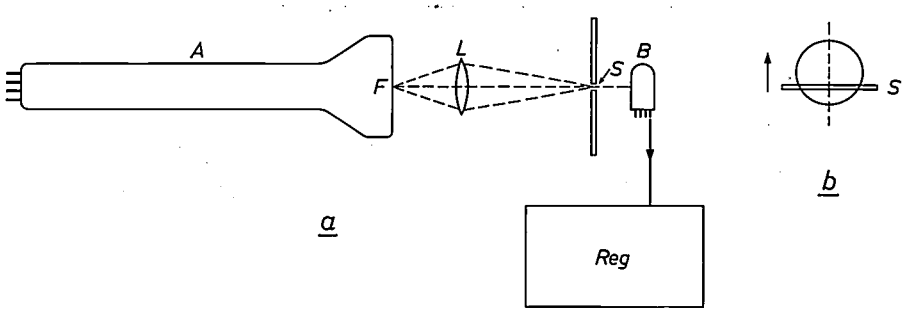


Fig. 1. *a*) Diagram of the set-up. The electron gun under investigation is mounted in a testing tube *A*. A magnified image of the focused spot *F* is produced on a screen by the lens *L*. The slit *S* cuts a narrow strip from the image (see *b*), the light from which falls on a photomultiplier *B*. The image is scanned by the relative movement of slit and image (see arrow in *b*). *Reg* pen recorder.

Even small movements of the spot in relation to the slit will cause considerable variations in the output voltage of the photomultiplier. Unwanted movements can be caused by external magnetic interference and by hum and other voltage fluctuations in the power-supply apparatus. External magnetic interference is suppressed by magnetically screening the whole cathode-ray tube with mu-metal (1 mm thick). Voltage fluctuations are avoided by

[1]) A slit is also suitable for measuring an elliptical spot in a simple way, provided the slit is parallel to one of the axes of the ellipse.

[2]) A. Ciuciura, Mullard Applications Research Laboratory, Mitcham, England. See Mullard tech. Commun. 5, 141-158, 1960.

[3]) R. R. Bathelt, Philips Electron Tubes Division, Eindhoven (not published).

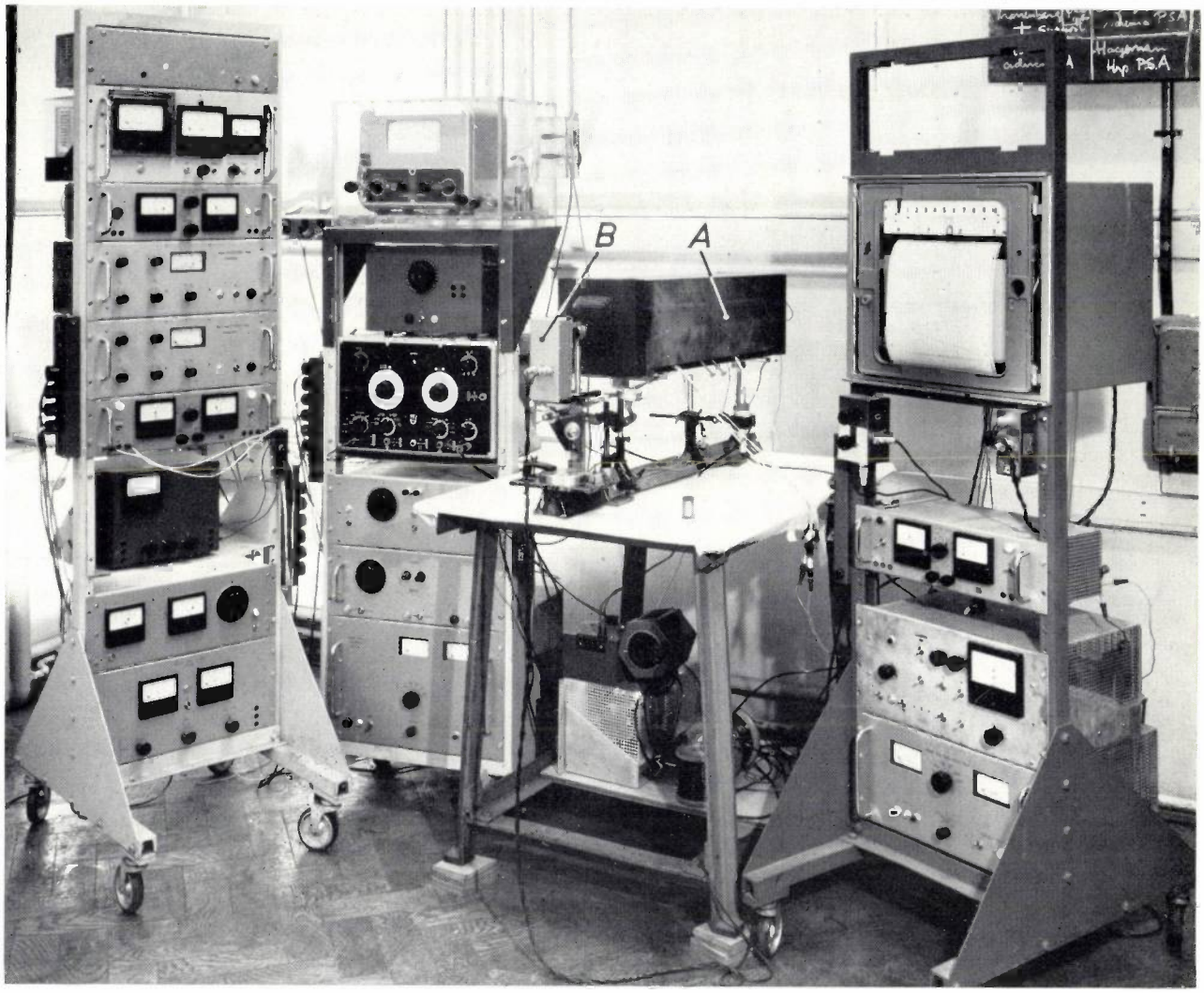


Fig. 2. Measuring set-up as in the diagram shown in fig. 1, used in the Philips Research Laboratories. The spot on the screen is scanned by slightly deflecting the beam. The testing tube is enclosed in a magnetically screened can of mu-metal (*A*). The photomultiplier is at *B*. The light-intensity distribution in the spot is approximately Gaussian (see the recording).

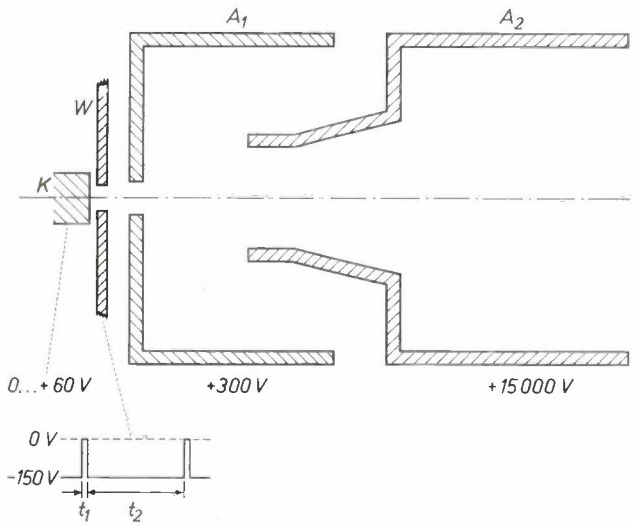


Fig. 3. Axial cross-section of the electron gun of a picture tube. *K* cathode. A_1, A_2 anodes. During the measurement the beam is intermittently suppressed by applying pulses to the Wehnelt cylinder *W*; $t_1/t_2 = 10^{-4}$ to 10^{-5}.

carefully stabilizing the power-supply apparatus. The influence of hum and of external interference synchronous with the mains can be eliminated by synchronizing with the mains frequency the pulses on the Wehnelt cylinder. Every time the beam appears the electrodes then have the same potential with respect to the cathode in spite of the hum. As in a television set, the HT of 15 kV is obtained by rectifying an alternating voltage having a frequency of about 17 kc/s. This frequency is arranged to be a whole multiple of the mains frequency, so that the 17 kc/s ripple of the HT can also do no harm.

In tubes with electrostatic focusing the focusing electrodes are combined with the electron gun to form a single system, and can therefore be used in the measurement without additional measures. If the gun under investigation is to be used in a tube with magnetic focusing, we then focus by means of a coil. To avoid image aberrations, it is necessary in

that case to ensure that the symmetry axis of the focusing coil coincides with the axis of the beam. For that purpose the coil is rigidly fixed to the neck of the tube by means of an adjustable holder. If the tube is accidentally moved the coil (*1* in *fig. 4*) will move with it, so that the coil setting is not affected. The coil holder consists of two rings (*2*, *2'*) which are each clamped to one side of the coil. In each ring a phosphor-bronze leafspring (*3*) presses the tube neck against two adjusting screws (*4*, *4'*), with which the rings can be aligned. By aligning the two rings the axis of the coil can be made to coincide with the axis of the unfocused beam, which can be verified as follows.

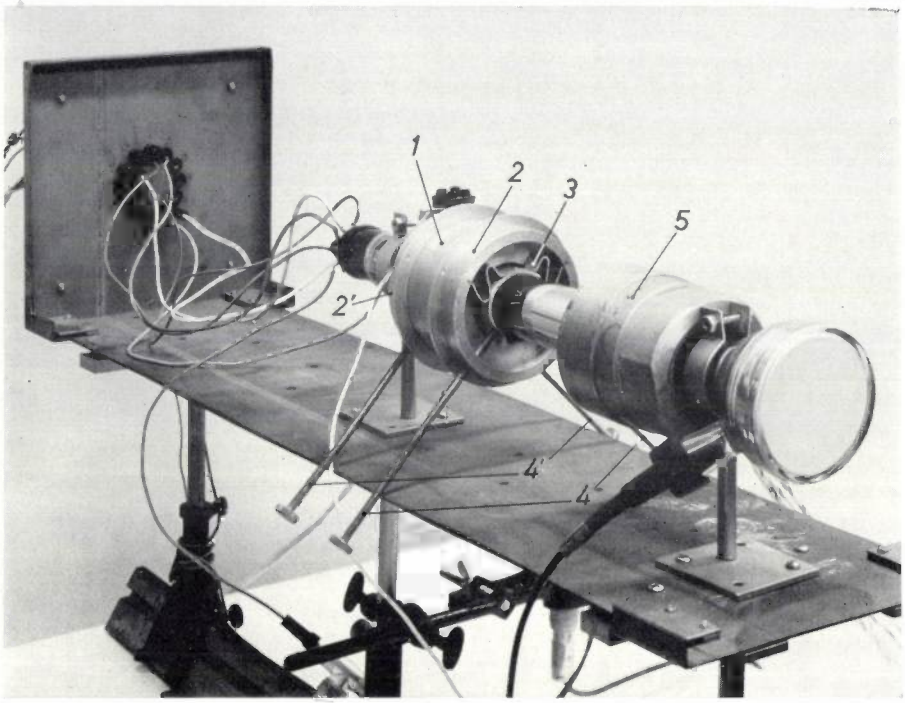
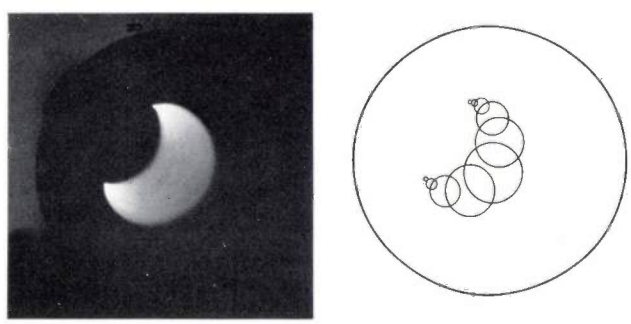


Fig. 4. The testing tube with its coils; screening can, optical system and photomultiplier have been removed. *1* focusing coil. *2*, *2'* clamping rings. *3* leafspring (ring *2'* contains an identical leafspring). *4*, *4'* adjusting screws. *5* coil used to deflect the beam slightly for scanning the spot.

If, after having focused the beam by passing a given current through the coil, we pass a current of equal magnitude but opposite direction through the focusing coil, the beam is again focused on the screen: in the expression for the focal length the magnetic induction on the coil axis occurs only in even powers, so that on the induction changing sign the focal length remains unchanged. If, however, the coil is out of alignment, the focus will not lie at the same point on the screen as the first time, the electron paths in the magnetic field having been turned through an equal but opposite angle. By turning the adjusting screws, the two spots can be made to coincide. This can happen, however, at any point of the screen, so that we need one more alignment mark. If we make the two spots coincide in the middle of the fluorescent area produced by the unfocused beam, we can be sure that the coil axis and the axis of the unfocused beam will also coincide. To this end we pass through the coil an alternating current whose amplitude is equal to the focusing current. We then see the two spots on the screen at the same time, joined by a series of intermediate positions in which the beam is not focused (*fig. 5a*, *b*). It is now a simple matter to adjust the



a *b*

Fig. 5. *a*) Example of image that can appear on the screen when the non-aligned focusing coil is carrying an alternating current whose amplitude is equal to the focusing current (comma and asymmetrical forms can also occur, and also self-intersecting half-moons; see e.g. T. Soller *et al.*, Cathode ray tube displays, McGraw-Hill, New York 1948, p. 105, fig. 3.16). *b*) How the image in (*a*) is produced.

The coil is aligned when the two focal points (ends of half-moon) are made to *coincide* in the *middle* of the fluorescent area of the unfocused beam. This is done by turning the four adjusting screws on the coil holder.

coil accurately with the four screws. If we connect a capacitor of appropriate value in series with the coil, this adjustment can be carried out at a relatively low alternating voltage.

Finally, we shall give an example of a test using the set-up described, relating to a certain type of electron gun for a television picture tube [4]). Calculations and measurements were made of a quantity Q that can be used as a figure of merit and which is given by:

$$Q = \frac{2r_i D}{L},$$

where $2r_i$ is the beam diameter at the position of the deflection coil of the picture tube with the beam focused, D the diameter of the focus on the fluorescent screen and L the distance between deflection coil and screen. Q is virtually independent of the positions of focusing coil and deflection coil, so that it gives some information about the quality of the *gun* (the smaller Q the better the gun). *Fig. 6* shows the percentage deviation of the experimental value Q_{exp} of Q of the gun under investigation from the theoretical value Q_{th}. In determining the value Q_{exp} the quantity D was measured with the set-up described above.

In order to measure the diameter $2r_i$ as well, we mounted a plane grid of parallel wires of known equal spacing in the testing tube, in the plane of the deflection coil in the picture tube for which the gun

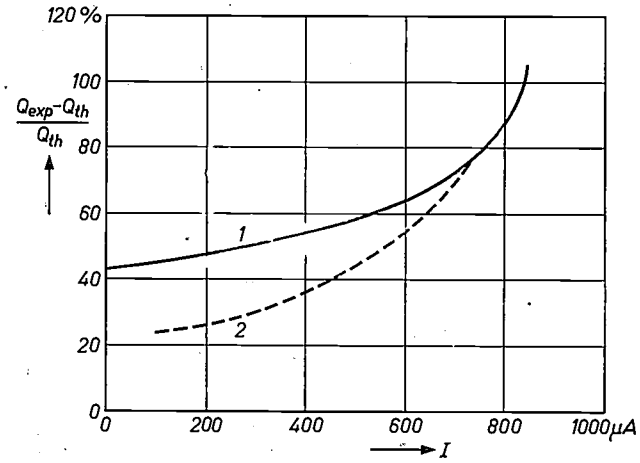


Fig. 6. Percentage deviation of the experimental figure of merit Q_{exp} from the theoretical value Q_{th}, for a certain type of electron gun to be used in a television picture tube. For curve *1* the beam current I was varied by means of the cathode voltage, for curve *2* by means of the cathode temperature. (Taken from [4]).)

was intended. The grid was positioned perpendicular to the beam axis. Between this grid and the screen there was an additional focusing coil (not shown in fig. 4, and not energized for the measurement of D). After the measurement of D, *without changing* the setting of the focusing coil for the *spot* (*1* in fig. 4), the *additional* focusing coil was energized to such an extent as to produce on the screen an image of the *grid*. The diameter $2r_i$ was then determined by counting the number of imaged grid wires.

C. WEBER *).

[4]) J. Hasker and H. Groendijk, Measurement and calculation of the figure of merit of a cathode-ray tube, Philips Res. Repts. 17, 401-418, 1962 (No. 5).

*) Philips Research Laboratories, Eindhoven.

THE BEHAVIOUR OF TRANSISTORS WITH INCREASING FREQUENCY

by M. BEUN *) and L. J. TUMMERS *).

537.311.33:621.382.333

As the frequency increases, certain effects gradually become apparent in a transistor which are bound up with the periodic changes in the distribution of the hole and electron concentrations. These effects set a limit to the frequency at which the transistor can deliver gain. An analysis of these effects reveals the factors that limit the frequency range, and provides a basis for finding ways and means of widening this range. This has already been the subject of intensive research in laboratories all over the world and considerable progress has been achieved: whereas the first junction transistors could be used at frequencies no higher than a few kc/s, transistors capable of operating at frequencies of several hundreds of Mc/s are now commonplace, and they can even be made for frequencies up to 1000 Mc/s.

This article does not report on new developments, but attempts to throw light on a familiar subject in a way which may be helpful particularly to newcomers to transistor theory.

The complications in the behaviour of transistors with increasing frequency are bound up with the periodic changes that occur in the concentration pattern of the charge carriers (electrons and holes) in a transistor subjected to alternating voltages. The relation between the concentration pattern and the voltages and currents occurring in a transistor has been dealt with at some length in this journal in an article which explained the principles of transistor action [1]). This explanation will be briefly recapitulated here. Two concepts will then be introduced which have an important bearing on the behaviour of a transistor with increasing frequency, namely: diffusion capacitance and barrier capacitance. An analysis of the effects of frequency on the behaviour of the transistor reveals the factors that limit the frequency at which the transistor can still amplify a signal. We shall be concerned throughout with *P-N-P* transistors having homogeneous *P* and *N* regions, in which the concentrations of minority charge carriers (briefly, minority concentrations) in the emitter and collector are very small compared with the minority concentration in the base. This is the case in conventional alloyed junction transistors [2]). It is this type, mass-produced, that has made the transistor so popular. We shall furthermore assume that the various quantities vary only in the direction perpendicular to the *P-N* junctions (the x direction). When, on the basis of our analysis, we consider what steps can be taken to widen the frequency range of transistors, we find that alloyed-type transistors are not suitable for high-frequency applications. For this reason, transistors specially designed for operation at frequencies up to 100 Mc/s and higher are not made by the alloying process. These modern high-frequency transistors will not be discussed in this article in any detail, but mention will be made of the principal differences compared with alloyed transistors.

A good starting point for recapitulating the principles of transistor action can be found by examining an important aspect of the difference between semiconductors and metals, thereby making clear that semiconductors offer possibilities which metals do not. Crystal diodes and especially transistors are striking examples of the ingenious exploitation of these possibilities.

Semiconductors and metals

Electrical conduction in metals is due to the presence of mobile negative charge carriers, i.e. electrons. In this connection the term "electron gas" is often used. An electric current is produced by the action of an electric field, which gives the electrons a certain drift velocity. When a flow takes place in a normal gas, however, a concentration gradient is usually involved. It is obvious to ask whether there might also be a concentration gradient in the electron gas of a metal. The answer is that, owing to the mutual repulsion of the electrons, it is virtually impossible to cause the concentration in the electron gas of a homogeneous metal to differ perceptibly from the mean value. If the electron concentration in a homogeneous metal were not everywhere the same, a negative space charge would occur at places of increased concentration. The repulsive force between electrons would then prevail over the attractive force exerted on them by positive metal ions

*) Philips Research Laboratories, Eindhoven.

1) F.H. Stieltjes and L.J. Tummers, Simple theory of the junction transistor, Philips tech. Rev. **17**, 233-246, 1955/56.

2) P.J.W. Jochems, The alloy-diffusion technique for manufacturing high-frequency transistors, Philips tech. Rev. **24**, 231-239, 1962/63 (No.8). The alloying method is also dealt with briefly in this article.

which are bound to fixed places. At places of reduced electron concentration there would be a positive space charge, which would attract electrons towards it. It is this striving towards electrical neutrality — which is so strong that it is generally referred to as a neutrality "condition" — which rules out appreciable differences in concentration in the electron gas of homogeneous metals.

In semiconductors the situation is completely different. In addition to negative mobile charge carriers, semiconductors also contain positive mobile charge carriers, called holes [3]). In homogeneous semiconductors the neutrality condition still applies, i.e. that the total charge must be zero in each volume element, but this does not conflict with a distribution of the concentration of electrons (or holes) that differs from a uniform distribution. For any given distribution of the concentration of charge carriers of the one polarity, the charge carriers of the opposite polarity need only be distributed in such a way that no space charge arises (see fig. 2*b* and *c*, where two examples are given). *In semiconductors, then, apart from currents caused by an electric field, currents can also flow as a result of a concentration gradient, called diffusion currents* [4]). *P-N* junctions provide a means of locally controlling the concentrations of charge carriers, and hence of producing diffusion currents, and this is done in crystal diodes and transistors.

The condition that there should be no space charge by no means implies that the hole concentration should everywhere be equal to the electron concentration, for the ions incorporated in the crystal lattice — the donors and acceptors — also contribute to the space charge. A difference must in fact exist between the concentrations of electrons and holes equal to the difference between the concentrations of ionized donors and acceptors.

We shall now recapitulate the principles underlying the operation of crystal diodes and transistors.

Crystal diode and transistor action at low frequencies

Two fundamental theorems

We base the explanation of crystal diode and transistor action on two theorems concerning the behaviour of the minorities [5]). In order not to be led away from the main theme the proof of these theorems, and the exact formulation of the conditions under which they apply, are given in an Appendix.

The first theorem concerns the behaviour of the minorities in homogeneous *P-N* regions, but *outside* a barrier. By a barrier is meant the (very narrow) region on either side of a *P-N* junction where large space-charge densities prevail; outside the boundary planes of such a barrier there is electrical neutrality (see *fig. 1*; inside the barrier, then, the neutrality condition does not apply, but here a marked inhomogeneity exists, namely the *P-N* junction). In general, the transport of charge carriers will be

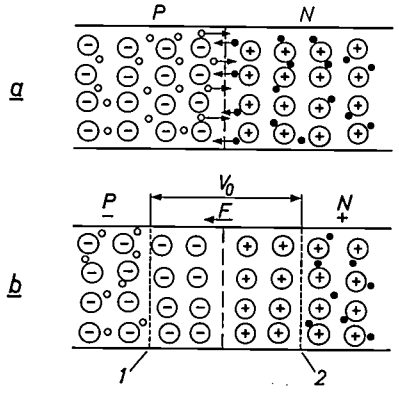


Fig. 1. Where a *P* region borders on an *N* region (*P-N* junction), holes diffuse from the *P* to the *N* region and electrons in the opposite direction (*a*). The charge carriers crossing the junction enter a region where they are minority charge carriers; they spread over this region and each ultimately combines with one of the numerous majority charge carriers. At each side of the boundary plane a layer forms in which the charge of the ions permanently incorporated in the crystal lattice (negative acceptor ions in the *P* region and positive donor ions in the *N* region) is no longer compensated by mobile charge carriers. These two layers form an electrical double layer — here called the barrier — which grows until the electric field *F* prevailing in the barrier prevents further diffusion (*b*). The transitions from the barrier to the neutral regions are assumed to be so abrupt that we can speak of "boundary planes" *1* and *2* of the barrier (see also figs. 8 and 24). The field is of course directed from the positive to the negative charge, i.e. from the *N* to the *P* region. The potential of the *N* region is therefore higher than that of the *P* region by a certain amount V_0 — the contact potential difference or diffusion voltage.

effected both by an electric field and by diffusion. We shall confine our considerations to *P* and *N* regions in which the equilibrium value of the majority concentration is many times greater than that of the minority concentration. In practice the difference is usually a factor of 10^3 to 10^{10}. In such a case the first theorem holds, which is: *in determining the minority current in a homogeneous P or N region outside a barrier, only diffusion need be taken into account.*

[3]) See e.g. J.C. van Vessem, The theory and construction of germanium diodes, Philips tech. Rev. **16**, 213-224, 1954/55. Here and in the article mentioned in footnote [1]) an explanation is given of such terms as *P* region, *N* region, majority and minority concentrations, generation and recombination.

[4]) In a gas in which the same temperature prevails everywhere — unless it is a highly rarefied gas (Knudsen gas) — the gas flows because a molecule undergoes on an average more collisions in the direction from high to low concentration than in the opposite direction. This is a different mechanism from that of diffusion. However, we will not go into such details here.

[5]) For brevity we shall use the terms minorities and majorities instead of minority and majority charge carriers.

This means that, as far as the minorities are concerned, the presence of an electric field can be disregarded. The behaviour of *majorities*, on the other hand, is governed both by the electric field and by diffusion. For that reason the behaviour of majorities is not so straightforward to describe as that of minorities. In order to explain the characteristics of crystal diodes and transistors, attention is therefore devoted in the first place to the minorities.

The second fundamental theorem concerns the concentrations of the minorities in the boundary planes of a barrier, i.e. in the planes *1* and *2* in fig. 1*b*. (The mobile charge carriers represented in this figure are majorities.) Between these boundary planes there exists a spontaneous potential difference: the contact potential difference or diffusion voltage V_0 (see fig. 1 and Appendix). If we connect the P and N regions to the poles of a battery, the voltage between the boundary planes of the barrier will change by a certain amount V_u, which we call the *external* voltage across the barrier. The second fundamental theorem is: *the minority concentration in the boundary plane of a barrier is proportional to* $\exp(qV_u/kT)$. Here q is the absolute value of the electron charge, k is Boltzmann's constant and T the absolute temperature. If the P region is connected to the positive pole of the battery, V_u must be taken as positive (to be remembered from the p of positive and P region).

Where $V_u = 0$ (no voltage applied across the crystal), then outside the barrier the equilibrium concentrations prevail: in the P region p_{P0} and n_{P0}, in the N region p_{N0} and n_{N0} [6]). (Equilibrium concentrations are the concentrations at which, per unit time and per unit volume, as many holes and electrons are generated as disappear by recombination [3]).) If a voltage V_u is across the barrier, then according to the second theorem the minority concentrations in the boundary planes *1* and *2* are respectively:

$$n(1) = n_{P0}\, e^{qV_u/kT}\ ,\quad p(2) = p_{N0}\, e^{qV_u/kT}\ . \qquad \ldots \quad (1a \text{ and } b)$$

It should be added that both postulates are applicable only as long as the minority concentration is small compared with the equilibrium concentration of the majorities (see Appendix). This implies that formulae 1a and 1b no longer hold at unlimitedly high positive values of V_u. In the following we shall assume that the condition referred to is fulfilled [7]).

We can now turn to the explanation of the rectification at a P-N junction, i.e. to the operation of a crystal diode.

Rectification at a P-N junction; crystal diode

Fig. 2 shows the concentration pattern of electrons and holes at a P-N junction for the cases where V_u is zero, positive or negative. The case $V_u = 0$ (fig. 2*a*) has already been mentioned. If V_u is positive, then according to formulae (1a) and (1b) the minority concentrations in the boundary planes *1* and *2* are pulled upwards (fig. 2*b*). The processes of generation and recombination endeavour, however, to maintain the equilibrium concentrations. Consequently, as the distance to the barrier increases, the concentrations gradually approach the equilibrium values. The deviations from the equilibrium values decrease exponentially with the distances to the boundary planes of the barrier [8]). The distance over which such a deviation diminishes by a factor e is called the diffusion length in the region concerned. This length depends among other things on the regularity of the crystal lattice: the more lattice defects the shorter the length. In germanium a normal value is e.g. 100 μm. In fig. 2*b* and *c* the diffusion lengths in the N and P regions are indicated respectively by L_N and L_P. Recalling our first theorem, we can at once conclude from the gradients of the concentration lines in fig.2*b* that starting from the barrier minority currents of holes and electrons will flow respectively into the N and P region. From the decrease of the gradients as the distance to the barrier increases we see that the minority currents decrease, the transport of charge being gradually taken over by majority carriers (fig. 2*d*). Because of the neutrality conditions the majority concentrations must, at their higher level, change in the same way as the minority concentrations. We shall return presently to the behaviour of the majority charge carriers.

If V_u is negative the minority concentrations in the barrier boundary planes are forced downwards. The minority currents are then directed towards the barrier (fig. 2*c* and *e*).

It is already evident where this is leading to. With rising *positive* V_u the minority concentrations in the boundary planes increase rapidly with V_u and the same will apply to the minority currents. With

[6]) As regards the use of the letters p, n, P and N, we shall keep to the following convention: the lower-case letters p and n refer to charge carriers, the upper-case letters P and N to regions. For example, p_P is a hole concentration in a P region (i.e. a majority concentration) and p_N a hole concentration in an N region (i.e. a minority concentration). An additional suffix 0 denotes an equilibrium concentration.

[7]) The case where the minority concentration is no longer small compared with the majority concentration is dealt with by F.H. Stieltjes and L.J. Tummers, in: Behaviour of the transistor at high current densities, Philips tech. Rev. **18**, 61-68, 1956/57.

[8]) For a mathematical treatment see e.g. the article mentioned in footnote [3]), page 221.

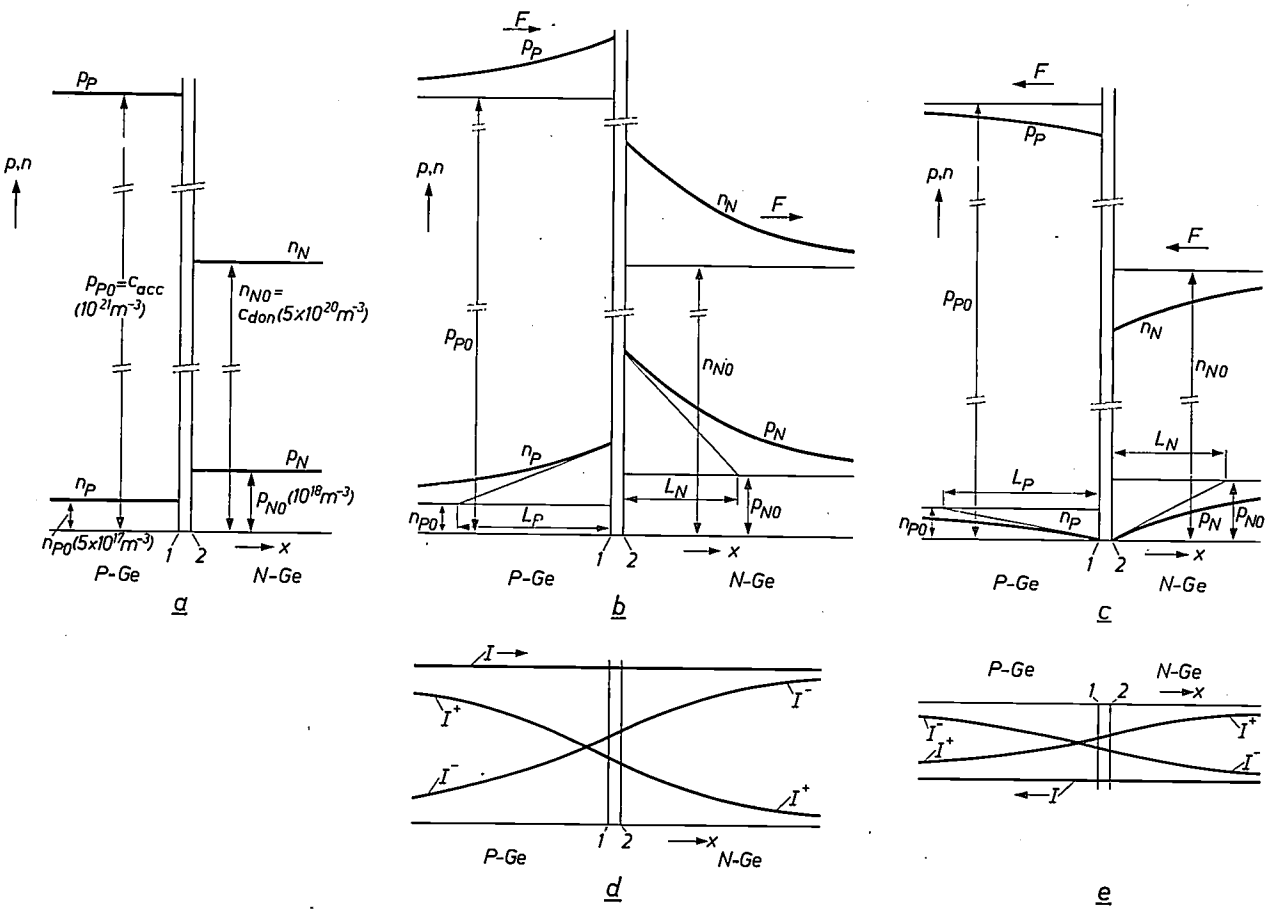


Fig. 2. Concentration pattern of electrons and holes at a P-N junction, *a*) without external voltage, *b*) with external voltage in the forward direction, and *c*) with external voltage in the reverse direction. The figures added in (*a*) indicate the order of magnitude of the concentrations involved. As the concentration scale is linear, two large breaks had to be made in it. The forward current flows to the right (*d*) and gradually changes from a hole current I^+ into an electron current I^-; the reverse current flows to the left (*e*) and gradually changes from an electron current into a hole current. In both cases the total current I is of course independent of position.

increasing *negative* V_u the minority concentrations in the boundary planes decrease, but, since they cannot drop below zero, the minority currents now approach a saturation value. Positive values of V_u (positive pole connected to the P region) will therefore correspond to the *forward direction*, and negative values to the *reverse direction*. This conclusion is indeed correct, as we shall now explain in more detail.

The total current is of course equal to the sum of the hole and electron currents in any given cross-section. To calculate the total current we must therefore consider a cross-section where both the hole current and the electron current can be determined. This is the case in the boundary planes of a barrier. The minority current is proportional to the deviation of the minority concentration from its equilibrium value. From our second theorem, we know that in a boundary plane this deviation is proportional to $\exp(qV_u/kT) - 1$. The minority current is therefore also proportional to this expression. In boundary plane *2* this therefore applies to the hole current, which is here the minority current. But the electron current, which is the majority current in boundary plane *2*, is almost equal to the electron current in boundary plane *1*, where it is the minority current. This is because any marked difference between these electron currents would of course mean that in the barrier substantially more electrons would disappear than are generated, or vice versa. The barrier, however, is only a few microns thick, and is thus thin compared with the diffusion length (e.g. 100 μm), which is a measure of the distance over which the hole current and electron current perceptibly merge with one another. In boundary plane *2*, therefore, not only is the hole current proportional to $\exp(qV_u/kT)-1$, but also the electron current, and the same conse-

quently applies to the total current I. The current-voltage characteristic of a P-N junction is therefore given by

$$I = I_s\,(e^{qV_u/kT} - 1), \quad \ldots\ldots \quad (2)$$

which function is represented in *fig. 3*. The rectifying action of a P-N junction is immediately apparent

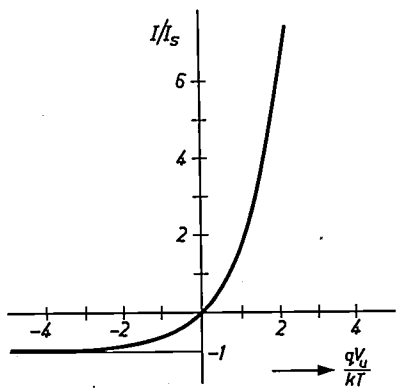


Fig. 3. Theoretical current-voltage characteristic of a P-N junction.

from this [9]). By putting $V_u = -\infty$ in formula (2) we see that the proportionality constant I_s represents the saturation current in the reverse direction.

A few more remarks are needed to complete the picture. In the P and N regions a majority concentration pattern is formed in such a way that in both regions a (weak) electric field exists which is exactly sufficient to conduct towards or away from the barrier the majorities necessary to maintain the minority currents at the other side of the barrier. The pattern required for this purpose, however, differs so little from that needed for electrical neutrality that in order to derive the concentration pattern of the majorities from that of the minorities, we need only make use of the neutrality condition. From fig. 2*b* and *c* it is further seen that the majorities are driven by the electric field F counter to the concentration gradient. (It should be noted that the behaviour of the majorities, which cannot directly be determined, is derived here from the behaviour of the minorities.)

Outside the barrier the electric field is apparently not exactly zero. This field causes a voltage drop; the external voltage V_u across the barrier, which appears in expression (2) for the rectification characteristic, is therefore not entirely identical with the voltage between the connection terminals.

Transistor action

After the foregoing, the explanation of transistor action is quite straightforward. In *fig. 4a* the concentrations p and n of holes and electrons are represented as a function of position in a P-N-P transistor, where the minority concentrations in emitter and collector are equal to one another and much smaller than the minority concentration in the base. In the boundary plane *2* between emitter barrier and base the minority concentration (holes) is given a large value $p(2)$ by applying a voltage in the forward direction across the emitter barrier. In the boundary plane *3* between base and collector barrier, the hole concentration is fixed at zero by an appreciable voltage in the reverse direction. This gives rise in the base to a steep concentration gradient and thus to a considerable "transit" or through current I_t of holes from emitter to collector. The holes are conducted towards the barrier by a (weak) electric field existing in the emitter, cross the emitter barrier, diffuse towards the collector barrier, where after crossing this barrier a (weak) electric field carries them away. The minority concentrations (electrons) in emitter and collector are in reality much smaller than appears in fig. 4*a*. The electron currents through the two P-N junctions may therefore in the first instance be neglected. The current I_t thus flows into the transistor at the emitter and emerges at the col-

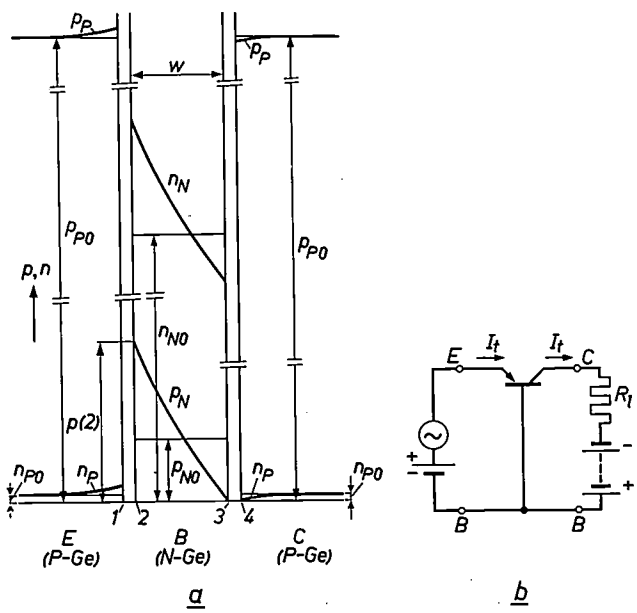


Fig. 4. *a*) The concentration pattern of holes and electrons in a P-N-P transistor, from which transistor action can be explained.
b) Common-base configuration in which the transistor can give voltage gain.

[9]) The absence of a perceptible generation or recombination surplus in the barrier is evidently necessary to the proper rectifying operation of a P-N junction. This is one of the reasons why a good crystal diode cannot be made by simply pressing together two pieces of germanium — one P type and the other N type. At the surface all kinds of disturbing effects occur, which result among other things in very fast generation and recombination. One can no longer assume, therefore, that the electron or hole current has the same value at both sides of the contact face.

lector virtually unchanged. If, between the emitter terminal E and the base terminal B, we superimpose a small alternating voltage on the direct voltage, then this alternating voltage appears across the emitter barrier and causes the hole concentration $p(2)$ in boundary plane *2* to vary. The concentration gradient in the base varies likewise, and so too therefore does the current I_t through the transistor. This current flows also through the load resistance R_l, incorporated in the path between the collector terminal C and the base terminal (fig. 4*b*). If R_l is made large enough, the alternating voltage between emitter and base terminals appears across this resistor in an amplified form. The circuit shown in fig. 4*b*, where the base terminal forms part both of the input and output pairs of terminals (for which reason it is called "common-base" connection) the transistor thus provides voltage gain. Assuming that the output current is equal to the input current (both I_t) the transistor also provides power gain of the same magnitude as the voltage gain.

When the voltage across R_l changes by a certain amount, the voltage across the collector barrier changes by an equal but opposite amount (fig. 4*b*). In the explanation of transistor action given above it is assumed that the end point at the collector side of the hole-concentration line in the base (fig. 4*a*) is fixed, so that this point does not shift when the voltage across the collector barrier changes. This assumption is not, however, entirely correct. For according to our second fundamental theorem, the relation between the hole concentration $p(3)$ in boundary plane *3* and the voltage across the collector barrier is given by the same exponential curve applicable to the relation between $p(2)$ and the voltage across the emitter barrier (*fig. 5.*) Taking this into account, we find that voltage amplification is possible if this curve is steeper at the operating point relating to the emitter barrier than at the operating point relating to the collector barrier. The ratio of the slopes at these operating points is the maximum voltage gain that can be obtained.

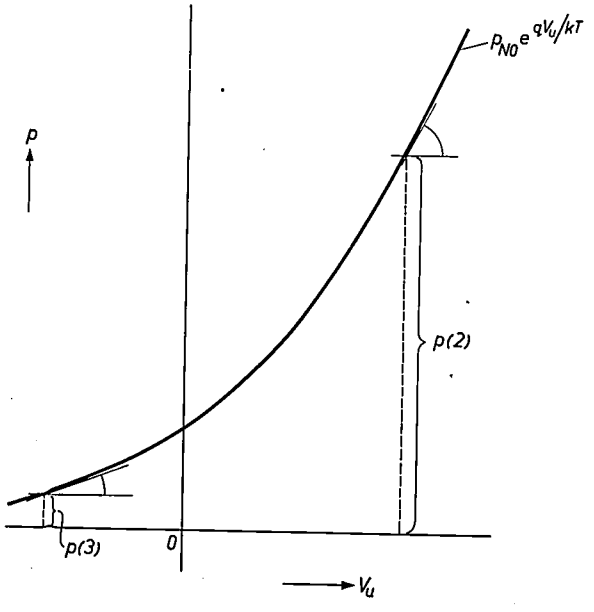


Fig. 5. The exponential relation between the hole concentration p and the external barrier voltage V_u, applicable to the boundary planes *2* and *3* in fig. 4*a*.

There is another reason why the said end point of the hole concentration line does not remain absolutely fixed, and that is the fact, presently to be discussed, that the barrier thickness increases with the reverse voltage. This causes the end point to shift horizontally towards the left as the reverse voltage across the collector barrier rises. This shift, too, sets a limit to the maximum available voltage gain.

We assume that the equilibrium value of the minority concentration in the emitter (electrons) is negligible compared with that in the base (holes). It follows from fig. 4*a* that in this case the current across the emitter barrier consists solely of holes. We further assume that the influence of recombination and generation in the base is negligible. The hole current entering the base across the emitter barrier will then leave, unchanged, across the collector barrier. For the holes the concentration line in the base is then straight, from which it follows that the same holds for the electrons. This amounts to assuming emitter and base efficiencies of 100% [10]).

In the simple explanation given of transistor action we are concerned with the voltages across the barriers. As pointed out on page 160, these voltages are not identical with those between the terminals. An important quantity in this connection is the internal base resistance, but for the present we shall neglect this. In order to keep it in mind we shall denote the voltage across the emitter and collector barriers, for a base without resistance, by $V_{EB'}$ and $V_{CB'}$ respectively, that is with an accent on the B denoting the base. It should also be noted that the sequence of the suffixes for the voltages implies a sign convention: a positive value means that the terminal (or region) to which the first suffix refers is positive with respect to that to which the second suffix refers. We can thus, if we wish, read: $V_{EB'} = V_E - V_{B'}$.

Changes of concentration pattern when the voltages change

Diffusion capacitance

We have seen that the passage of current through a crystal diode or a transistor is associated with certain concentration patterns of holes and electrons

[10]) Emitter and base efficiency are dealt with in detail in the article mentioned under footnote [1]), page 239.

in the P and N regions. If the current changes, then the concentration pattern must also change. Instead of first considering a separate P-N junction we shall turn straight away to the transistor.

To get some idea of the effects to be expected, we begin with the case where the voltage across the emmitter barrier is suddenly increased by a certain amount. In the boundary plane between emitter barrier and base the minority concentration then increases suddenly too, after which the new pattern gradually takes shape (*fig. 6*). The electron concentration will of course slavishly follow that of the holes. The electrical charges represented by the holes

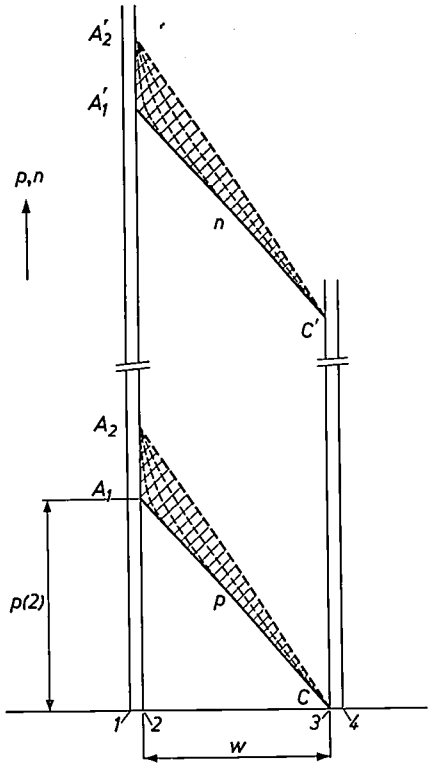


Fig. 6. After a sudden increase in the concentrations of holes and electrons in boundary plane *2* between emitter barrier and base, the concentration lines A_1C and A'_1C' gradually change to A_2C and A'_2C'. The hole and electron contents of the base thereby increase by equal amounts which are represented by hatched triangles. w is the thickness of the base. The thin dashed lines represent some intermediate phases.

and the electrons are proportional to the areas under the relevant concentration lines. It can be seen in fig. 6 that to form the new pattern, as much positive as negative charge is necessary, just as in the charging of a capacitor. At the same time, however, it is seen that, unlike the case in an ordinary capacitor, the charges do not follow the voltage without delay. The situation rather resembles that of a distributed capacitance, as encountered in transmission lines. Considerable correspondence in this respect indeed exists between the transistor and a transmission line [11]). The manner in which the new concentration pattern comes about is governed by the diffusion of the minorities, hence the term "diffusion capacitance".

We consider the voltage across the emitter barrier as the input signal, and the voltage over a resistance R_l in the collector circuit (fig. 4*b*) as the output signal. The output signal is proportional to the current in the collector circuit, that is with the *slope* of the hole concentration line in barrier boundary plane *3* (fig. 6). From this figure it can be seen that it takes some time before a perceptible signal appears at the output side, and moreover that the signal rises only gradually to full strength (*fig. 7a* and *b*). If the input voltage is abruptly returned to the initial value, the output signal again follows it gradually. The output signal is therefore delayed and distorted with respect to the input signal. If the input signal is a short pulse, the output signal does not even reach its full strengtb (fig. 7*c* and *d*). When the input pulse is made shorter and shorter, the output signal diminishes in strength. Extremely short pulses, which — as we know from Fourier analysis — contain a large proportion of high frequencies, are evidently no longer amplified by the transistor.

As our second case we assume that the voltage $V_{EB'}$ across the emitter barrier changes so slowly that the concentration pattern remains constantly

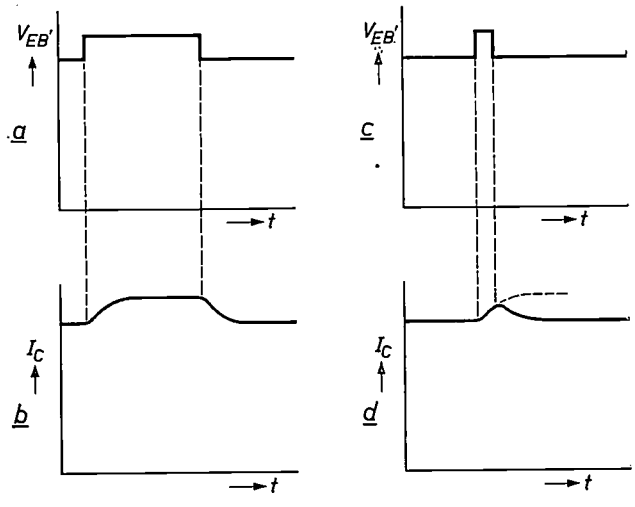


Fig. 7. Sudden changes in the voltage $V_{EB'}$ across the emitter barrier (*a*) cause — as follows from fig. 6 — gradual changes in the current I_C across the collector barrier (*b*). A short voltage pulse (*c*) has relatively little influence on I_C (*d*).

[11]) J. te Winkel, Transmission-line analogue of a drift transistor, Philips Res. Repts. 14, 52-64, 1959. The transistor we are concerned with can be treated as a special case, i.e. as a drift transistor with a zero drift field.

in step with it. The charge then follows the voltage without any perceptible delay. We are now, therefore, no longer concerned with a distributed but with a lumped diffusion capacitance C_d. This diffusion capacitance is given by the increase of the positive charge Q per volt increase of $V_{EB'}$, so that $C_d = dQ/dV_{EB'}$. Since the displacement of points A and A' (fig. 6) is not a linear but an exponential function of $V_{EB'}$, it follows that Q too is an exponential function of $V_{EB'}$. We must therefore write C_d as a differential quotient: it is a "differential" capacitance dependent on the voltage.

An expression for C_d can easily be derived. This will presently stand us in good stead when we examine quantitatively the influence of frequency on the behaviour of the transistor. The area of the right-angled triangle under the concentration lines for holes (fig.6) represents the number of holes in the base per unit cross-section. The rectangular sides of this triangle are the base thickness w and the hole concentration $p(2)$ at the emitter, so that the area is $\frac{1}{2}p(2)w$. The base of a transistor of transverse cross-section O thus contains $\frac{1}{2}p(2)wO$ holes, and these represent a charge $Q = \frac{1}{2}p(2)wOq$. Differentiation with respect to $V_{EB'}$ where $p(2)$ as a function of $V_{EB'}$ follows from formula (1b) with $V_u = V_{EB'}$, yields:

$$C_d = \tfrac{1}{2}Ow\frac{q^2}{kT}p(2)\,.$$

The hole concentration $p(2)$ is closely related to the minority diffusion current I_t through the base, this current being proportional to the concentration gradient $p(2)/w$ in the base, so that:

$$I_t = qD_p\frac{p(2)}{w}O\,. \quad \ldots\ldots \quad (3)$$

The proportionality constant D_p is the diffusion constant for the holes (minorities) in the base. Using this relation to eliminate $p(2)$ from the expression for C_d, we find:

$$C_d = \frac{qw^2}{2kTD_p}I_t\,. \quad \ldots\ldots \quad (4)$$

In a good transistor I_t is easily measured because it is almost equal to the current entering the transistor at the emitter, and also to the current emerging at the collector (this follows from fig. 4*a*). According to formula (4) the diffusion capacitance is proportional to the "biasing current" I_t.

The insertion of conventional values in formula (4) gives some idea of the magnitude of the diffusion capacitance. Given $w = 50\,\mu m$, $D_p = 4.4\times10^{-3}\,m^2s^{-1}$, $q/kT = 40\,V^{-1}$ and $I_t = 1$ mA, we obtain $C_d \approx 10^{-2}\,\mu F$.

Barrier capacitance

In a barrier region, too, the concentration pattern must adapt itself to the voltage, and this again involves a capacitance, which we call the barrier capacitance. We consider an abrupt P-N junction, i.e. one in which the P region changes abruptly to the N region as shown in fig. 1. To make the barrier capacitance easy to calculate, we assume that the hole concentration p at the barrier boundary plane *1* drops abruptly from the value c_{acc} (the concentration of acceptors in the P region) to zero, and that the electron concentration n at the barrier boundary plane *2* drops abruptly from the value c_{don} (the donor concentration in the N region) to zero (*fig. 8a*). This schematic concentration pattern is of course only a rough approximation of the real situation, but the calculations based on it nevertheless yield useful results [12]). As explained in the caption to fig. 1,

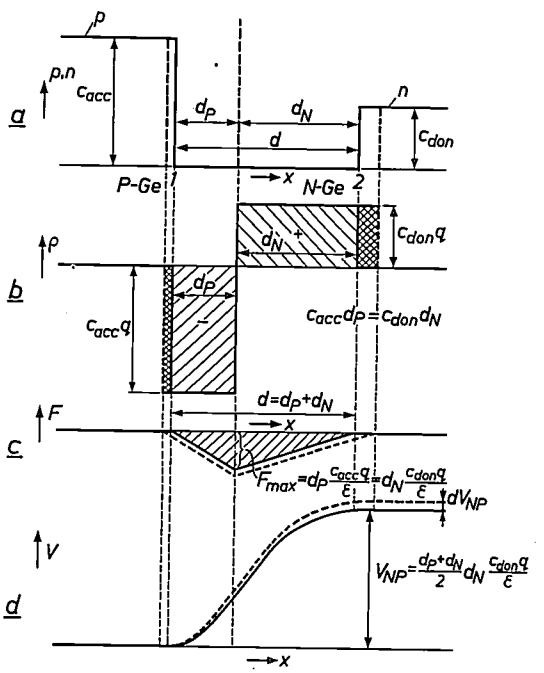


Fig. 8. Illustrating the barrier capacitance. The area of each of the hatched rectangles in (*b*), after division by the dielectric constant ε, yields the value F_{max} in (*c*); the area of the hatched triangle in (*c*) yields the potential difference V_{NP} in (*d*).

the region between the boundary planes constitutes an electrical double layer. The space-charge density ϱ in this layer is represented in fig. 8*b*. Since the total negative charge must be equal to the total positive charge, we can write

$$c_{acc}\,d_P = c_{don}\,d_N\,. \quad \ldots\ldots \quad (5)$$

(See fig. 8 for the meaning of d_P and d_N.)

[12]) For an exact calculation of the concentration pattern in a barrier see: W. Shockley, The theory of P-N junctions in semi-conductors and P-N junction transistors, Bell Syst. tech. J. 28, 435-489, 1949.

Since the concentrations of acceptors and donors are virtually equal to the majority concentrations in the P and N regions respectively, we can already draw from (5) the important conclusion that the barrier extends farthest into the region with the smaller majority concentration, i.e. the region with the lower conductivity.

From the elementary theory of electricity we know that, for the situation represented in fig. 8 (one-dimensional problem), there exists between the field F and the space-charge density ϱ the relation $F = \int (\varrho/\varepsilon)\mathrm{d}x$ (where ε is the dielectric constant), and between the potential V and the field strength F the relation $V = -\int F\,\mathrm{d}x$. Thus, the variation of the field strength in the barrier follows the straight lines drawn in fig. 8*c*, and the area of the hatched triangle in this figure is a measure of the potential V_{NP} of the N region relative to the P region, i.e. of the voltage across the barrier (fig. 8*d*). This leads to the formulae written in fig. 8*c* and *d* for the maximum field strength $F_{\max}$ in the barrier and for V_{NP}. It also follows from fig. 8*c* that the voltage V_{NP} over the space-charge layers in the P and N regions is divided into parts in the ratio $d_P : d_N$; these parts are thus inversely proportional to c_{acc} and c_{don}. If, then, there is a great difference between c_{acc} and c_{don}, that is a great difference in conductivity between P and N regions, the voltage is almost entirely across the space-charge layer in the region with the lower conductivity.

If the voltage across the barrier increases by $\mathrm{d}V_{NP}$ the charges will increase by $-\mathrm{d}Q$ and $+\mathrm{d}Q$ (the double-hatched strips in fig. 8*b*). As far as these additional charges are concerned, the situation corresponds to that in a capacitor whose plates are separated by a distance d and a medium with dielectric constant ε. Just as in such a capacitor, the proportionality factor between the increases of charge and voltage is given by the "capacitance", here termed "barrier capacitance":

$$C_{\mathrm{b}} = \frac{O\varepsilon}{d} \qquad \ldots\ldots\ldots \quad (6)$$

(where O is the cross-sectional area of the crystal), and the following expression holds for $\mathrm{d}Q$:

$$\mathrm{d}Q = C_{\mathrm{b}}\,\mathrm{d}V.$$

Since, as opposed to an ordinary capacitor, the distance d is not constant but increases with the charge (see fig. 8*b*), the barrier capacitance — as in the case of the change in the charges in the base — is a "differential" capacitance.

The dependence of d, and hence of C_{b}, on the barrier voltage V_{NP} can be derived from the formula for V_{NP} in fig. 8*d*. Eliminating d_P from this formula with the aid of (5), and solving for d_N, we find:

$$d_N = \sqrt{\frac{c_{\mathrm{acc}}}{c_{\mathrm{don}}\,(c_{\mathrm{don}} + c_{\mathrm{acc}})}\,\frac{2\varepsilon}{q}}\,\sqrt{V_{NP}}. \qquad (7)$$

According to (5) we can find d_P if we multiply (7) throughout by $c_{\mathrm{don}}/c_{\mathrm{acc}}$. Evidently d_N and d_P are both proportional to $\sqrt{V_{NP}}$, which therefore also applies to their sum d.

The barrier capacitance C_{b} is inversely proportional to d (see (6)), and therefore also inversely proportional to $\sqrt{V_{NP}}$. For C_{b} we find the expression:

$$C_{\mathrm{b}} = O\,\sqrt{\frac{c_{\mathrm{acc}}\,c_{\mathrm{don}}}{c_{\mathrm{acc}} + c_{\mathrm{don}}}\,\frac{\varepsilon q}{2}}\,\sqrt{V_{NP}}. \qquad (8)$$

Here V_{NP} is the *total* voltage between the N and P regions, and thus includes the spontaneous diffusion voltage V_0 (see fig. 1; the calculation of V_0 will be found in the Appendix, formula 19). To gain some idea of the magnitude of C_{b}, we insert in (8) conventional values for P-N-P alloyed transistors, e.g. $O = 3\times10^{-7}\,\mathrm{m}^2$, $c_{\mathrm{don}} = 10^{20}\,\mathrm{m}^{-3}$ and $c_{\mathrm{acc}} = 10^{24}\,\mathrm{m}^{-3}$. At these values of c_{don} and c_{acc} we find $V_0 = 0.3$ volt; with a reverse voltage of 2 volts we therefore have $V_{NP} = 2.3$ volts. Since $\varepsilon = 16\varepsilon_0 = 16\times8.85\times10^{-12}$ farad/m and $q = 1.6\times10^{-19}$ coulomb, we find that C_{b} is roughly 14 pF. The barrier capacitances of alloyed transistors are thus of the order of 10 pF.

Behaviour of transistors with increasing frequency

In our investigation of the behaviour of transistors with increasing frequencies, we make a distinction between two complementary cases where there exists across one of the two barriers a DC voltage which is kept rigorously constant, while a small alternating voltage is superimposed on the DC voltage across the other barrier. The fact that the DC voltage across the barrier is kept constant obviously implies that this barrier is short-circuited to alternating current.

Collector barrier short-circuited to alternating current

First we take the case in which the reverse voltage across the collector barrier is kept constant, so that this barrier is short-circuited to alternating current (*fig. 9a*). This means that the concentrations of holes and electrons in the boundary planes of the collector barrier are fixed. We assume that the reverse voltage is high enough for the hole concentration in boundary plane *3* to be practically zero (fig. 9*b*). Points C and C' are then the fixed end points of the concentration lines in the base. Across the emitter barrier there is a small alternating voltage $v_{EB'}$, which is superimposed on the DC voltage in the forward direction across this barrier. This alternating voltage causes

the points A and A' in fig. 9*b* to oscillate vertically, e.g. between A_1 and A_2, or A'_1 and A'_2 respectively. To a first approximation the oscillations are sinusoidal and in phase with the voltage. Strictly speaking, A and A' also oscillate in the horizontal direction, because the barrier thickness depends on the voltage (see fig. 8). The horizontal oscillation is negligible, however, compared with the vertical.

We first look at the holes — that is to say the minorities — and we assume that the frequency is very low. Even then the oscillating hole concentration line cannot always be exactly straight. If it were, then at any given moment as many holes would leave the base at the collector as enter at the emitter, and there would be no holes available to change the hole content of the base. An extreme case (corresponding to an infinitely large amplitude at zero frequency) is that where the point A has a uniform velocity v. If A moves upwards, the concentration line will be concave (*fig. 10a*), for this means that more holes enter from the emitter than leave towards the collector. If point A moves downwards, the concentration line will be convex (fig.10*b*). The concentration pattern — the solid curve AC — can be regarded as the superposition of a straight line AC, point A of which moves at a uniform velocity, and a curved line which cuts the zero line at emitter and collector. The straight line corresponds to a through current I_t of holes, entering the base at the emitter and leaving it unchanged at the collector. This current uniformly increases or decreases, depending on whether A moves uniformly upwards or downwards. The curved line EC corresponds to currents that change the hole pattern in the base — the hole storage currents. It can be seen that a hole storage current flows not only over the emitter barrier but also over the collector barrier. A simple calculation shows that a state is possible where the curved line EC does not change during the movement of the straight line, and that the slopes of the curved line at emitter and collector are in the ratio of 2:1. Two-thirds of the hole storage current of the diffusion capacitance is therefore over the emitter barrier and one-third over the collector barrier.

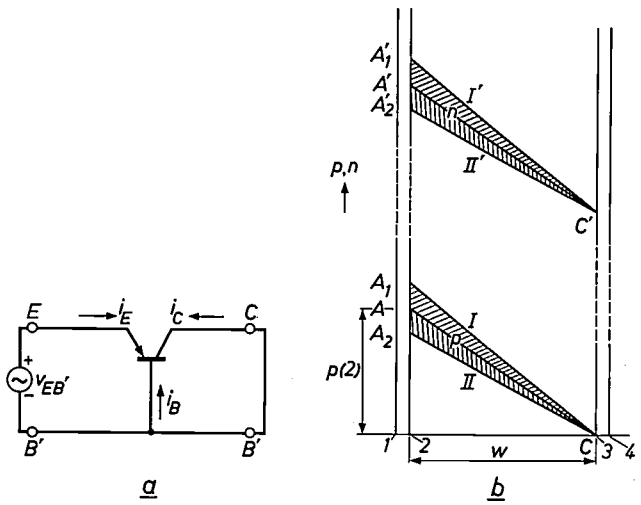


Fig. 9. *a*) The first fundamental case in an analysis of the behaviour of a transistor with increasing frequency: the collector barrier is short-circuited to alternating current; the biasing voltage across the emitter barrier has an alternating voltage $v_{EB'}$ superimposed on it. Biasing voltages are not indicated. The arrows establish the sign convention for currents, which are counted positive when directed towards the transistor.
b) The corresponding concentration pattern in the base: the concentration lines for holes (p) and electrons (n) oscillate around the fixed points C and C' between the respective positions I and II and I' and II'.

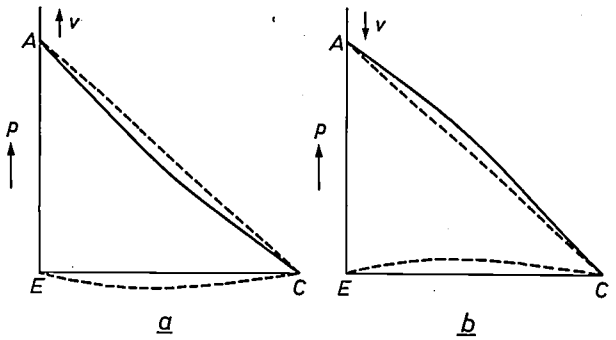


Fig. 10. When point A of the hole concentration line moves upwards at a velocity v, then AC must be slightly concave (*a*); when A moves downwards, then AC is convex (*b*). In both cases the concentration pattern can be regarded as the superposition of a moving straight line AC and a stationary curve EC (dashed lines). Calculation shows that the slope of the curve at E is twice as steep as at C.

We derive the above-mentioned ratio of 2 : 1 by introducing the location coordinate x, with $x = 0$ at the collector C, and by considering a slice of thickness dx and cross-section equal to unit area (*fig. 11*). The hatched strip then represents the hole

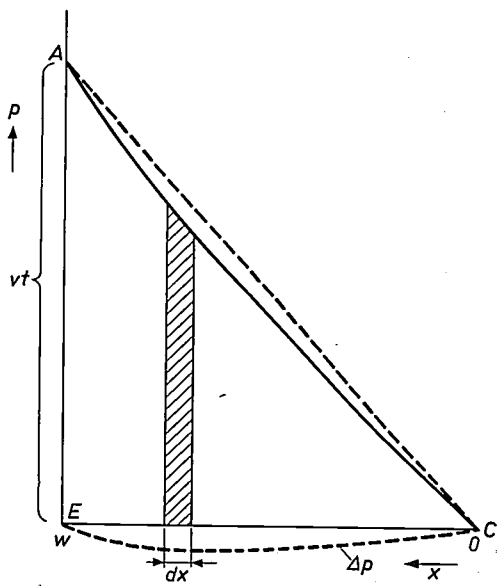


Fig. 11. To illustrate the derivation of the 2:1 ratio of the slopes at emitter and collector of the curve Δp as a function of position x.

content of the slice. From the condition that the amount of holes entering this slice in excess of the amount flowing out represents the increase in the hole concentration p, we obtain the partial differential equation:

$$D_p \frac{\partial^2 p}{\partial x^2} = \frac{\partial p}{\partial t}.$$

In a situation as assumed in fig. 10, the concentration at the emitter can be represented by vt, while at any given location x it can be represented by $p = \Delta p + (x/w)\,vt$, where Δp is the deviation from the linear distribution. (In fig. 11, Δp is negative.) After substituting for p in the differential equation we find for Δp the differential equation:

$$D_p \frac{\partial^2 \Delta p}{\partial x^2} = \frac{v}{w} x + \frac{\partial \Delta p}{\partial t}.$$

To find the solution for an unchanging form of the curve EC, we put $\partial \Delta p/\partial t = 0$. Integrating twice and filling in the boundary conditions ($\Delta p = 0$ for $x = 0$ and for $x = w$) then yields the required solution:

$$\Delta p = \frac{vw^2}{6D_p}\left[\left(\frac{x}{w}\right)^3 - \frac{x}{w}\right].$$

By differentiating with respect to x we find the slope of the curve as a function of x. At the emitter ($x = w$) this slope is found to be $2vw/6D_p$, and at the collector ($x = 0$) it is $-vw/6D_p$. Apart from the sign, then, the slope at the emitter is in fact twice as steep as at the collector.

The partial differential equation for p can also be solved in the usual way for the case where the hole concentration p at the emitter does not change uniformly, but consists of a constant value on which a sinusoidal fluctuation is superimposed, while p at the collector is fixed at zero. This solution leads to the frequency scales shown as dashed lines in fig. 13 and to the dashed curve for α_i in fig. 15.

We return to the case where point A oscillates as a result of a small alternating voltage $v_{EB'} = \hat{v}_{EB'} \cos \omega t$ superimposed on the emitter DC voltage $V_{EB'}$. Disregarding the frequency entirely amounts to assuming that the hole concentration line at any moment corresponds to a stationary state (giving always a straight line). This was the assumption underlying the transistor theory presented in the article mentioned in footnote [1]). A better approximation, which reveals some important aspects of the frequency effect, can be obtained by assuming that the hole pattern at any instant is adapted to the instantaneous velocity of point A. The concentration line is then straight when A is in one of the extreme positions A_1 or A_2 (*fig. 12*), for then the velocity of A is zero. The hole concentration pattern can be regarded as superimposition of a straight line, which oscillates around the point C between the two extremes I and II, and a curve which cuts the zero line at emitter and collector, and which oscillates between its two extremes in the manner of a standing wave. At moments when the straight line is stationary, i.e. takes up one of the extreme positions I and II, the hole content of the base does not change, and the curve must thus coincide with the zero line. We shall examine which alternating currents are bound up with the oscillations of the straight line and of the curve, and draw these currents in the form of a vector diagram (*fig. 13*), which relates to an "intrinsic" transistor. By intrinsic is meant that only currents connected with the concentration pattern *outside* the barriers are considered, while the voltages are the pure barrier voltages, i.e. those between the boundary planes of the barriers.

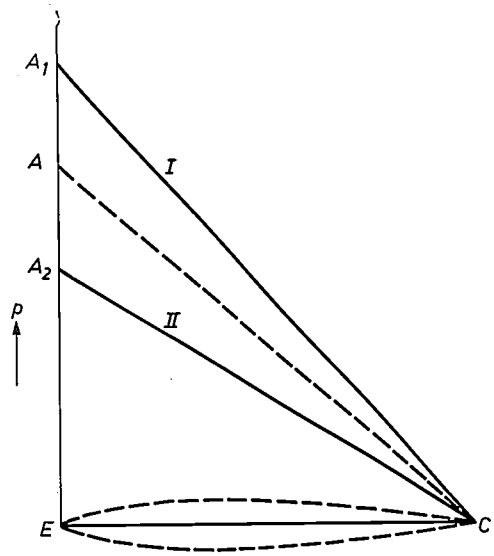


Fig. 12. If point A describes a slow oscillation, the hole concentration pattern can be regarded as the superposition of a straight line AC, oscillating around the fixed point C, and a curve EC oscillating in the manner of a standing wave. The extreme positions of EC correspond to the middle position of AC (dashed lines).

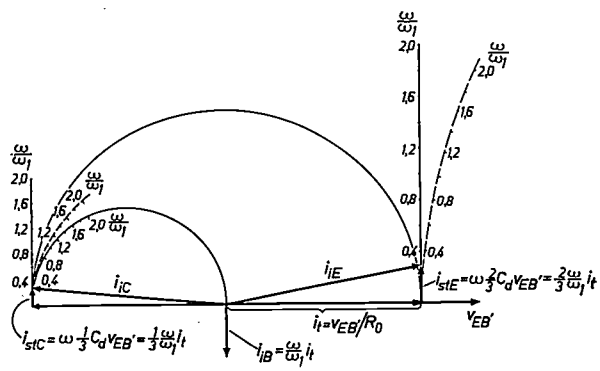


Fig. 13. Vector diagram of the "intrinsic" transistor when the reverse voltage across the collector barrier is kept constant, while the alternating voltage $v_{EB'}$ is superimposed on the forward voltage across the emitter barrier. According to the simplified representation in fig. 12, of the periodically alternating concentration pattern in the base, the ends of the vectors i_{iE} and i_{iC} for the emitter and collector currents shift upwards with increasing frequency along the vertical lines fitted with frequency scales ($\omega_1 = 1/R_0C_d$ is a characteristic frequency); rigorous mathematical treatment shows that these ends in fact move along the dashed curves. The scale along the smaller of the two semi-circles relates to an approximation in which the collector current is represented by the expression $i_{iC} = -v_{EB'}/R_0(1 + j\omega/3\omega_1)$.

The oscillation of the *straight* line AC corresponds to an alternating through current of holes, i_t. The amplitude of this current can be derived from formula (3), which represents the hole current I_t through the base in the case where the concentration line is straight. Inserting in (3) the expression (1b) for $p(2)$ — in which now $V_u = V_{EB'}$ — we find the bias current:

$$I_t = O\frac{qD_p}{w} p_{N0}\, e^{qV_{EB'}/kT}.$$

A small change $v_{EB'}$ in $V_{EB'}$ causes a current change $i_t = (dI_t/dV_{EB'})v_{EB'}$. Differentiating — disregarding the fact that w also depends slightly on $V_{EB'}$ because the barrier thickness does (fig.8) — yields:

$$\hat{i}_t = \hat{v}_{EB'}/R_0, \quad \ldots\ldots \quad (9)$$

where

$$R_0 = \frac{kT}{qI_t}. \quad \ldots\ldots \quad (10)$$

Since currents are counted positive if they are directed towards the transistor (fig. 9*a*), the contributions which i_t makes to the emitter and collector currents are respectively in phase and in antiphase with the alternating voltage $v_{EB'}$ across the emitter barrier (fig. 13).

The oscillation of the curve EC (fig. 12) corresponds to hole storage currents i_{stE} at the emitter and i_{stC} at the collector. These two storage currents reach their maximum values towards the transistor at moments when the hole content of the base increases fastest. This occurs when the straight line AC moves upwards at maximum speed, i.e. when $v_{EB'}$ goes from negative to positive through zero: i_{stE} and i_{stC} thus have a phase lead of 90° over $v_{EB'}$.

In order to plot i_{stE} and i_{stC} in fig. 13 we must not only know the phases but also the amplitudes. The latter follow from the condition that these two currents — the ratio between which is, as mentioned, 2 : 1 — must together deliver the alternating component ΔQ of the positive charge in the base. According to the definition of the diffusion capacitance C_d on page 163, $\Delta Q = C_d\hat{v}_{EB'}\cos\omega t$ (neglecting the fact that the concentration line is not always exactly straight). To deliver ΔQ, an alternating hole current $d\Delta Q/dt = \omega C_d\hat{v}_{EB'}\sin\omega t$ is needed. The currents from emitter and collector, which correspond to the charging of the diffusion capacitance, thus have the respective amplitudes $\frac{2}{3}\omega\, C_d\hat{v}_{EB'}$ and $\frac{1}{3}\omega\, C_d\hat{v}_{EB'}$.

The amplitudes of the currents that relate to the intrinsic transistor can be conveniently expressed in the amplitude $\hat{i}_t$ of the alternating through current. We then obtain:

$$\hat{i}_{stE} = \tfrac{2}{3}\frac{\omega}{\omega_1}\hat{i}_t \quad \text{and} \quad \hat{i}_{stC} = \tfrac{1}{3}\frac{\omega}{\omega_1}\hat{i}_t,$$

where (see also 9, 10 and 4):

$$\omega_1 = 1/R_0C_d = 2D_p/w^2. \quad \ldots\ldots \quad (11)$$

ω_1 is apparently a characteristic frequency of the intrinsic transistor.

Having dealt with the holes, we shall now turn our attention to the electrons. A glance at fig. 4*a* will show that no alternating current of electrons flows across the collector barrier. We have assumed the electron current across the emitter barrier to be negligibly small (page 161). The intrinsic emitter and collector alternating currents i_{iE} and i_{iC} are thus fully represented in the vector diagram; they consist entirely of holes. The electrons have to supply the base current i_{iB} of the intrinsic transistor. This current must be such that the vector sum of emitter, collector and base currents is zero at any instant. It follows from this that i_{iB} is equal but of opposite polarity to $i_{stE} + i_{stC}$. This completes the vector diagram (fig. 13).

To show in how far the simplifying assumption — that the concentration pattern is adapted at any instant to the instantaneous velocity at which the concentration at the emitter changes — in fact ties up with the exact theory, fig. 13 also indicates how i_{iE} and i_{iC} vary as a function of frequency in accordance with the rigorous mathematical treatment of the periodically varying hole pattern in the base. The merit of the simplified theory is that it makes it clear why only two-thirds of the positive charge for the diffusion capacitance is supplied across the emitter barrier and not the whole charge. As far as the amplitude of the collector current i_{iC} is concerned, however, the simplified theory is misleading: fig. 13 show sthat, according to this theory, this amplitude — the amplitude of $v_{EB'}$ assumed to be constant — increases with frequency, whereas according to the exact theory it actually decreases. The reason is bound up with the fact that the simplified theory takes no account of the time that elapses before the changes in concentration at the emitter make themselves felt at the collector (cf. fig. 7).

Common-base connection with short-circuited output

A circuit diagram is represented inside the dotted rectangle of *fig. 14*, the currents and voltages in which, for not too high frequencies, are described by the vector diagram in fig. 13. The circuit can thus be used as an equivalent circuit for the intrinsic transistor. The capacitor $C_{EB'}$ has been added as an

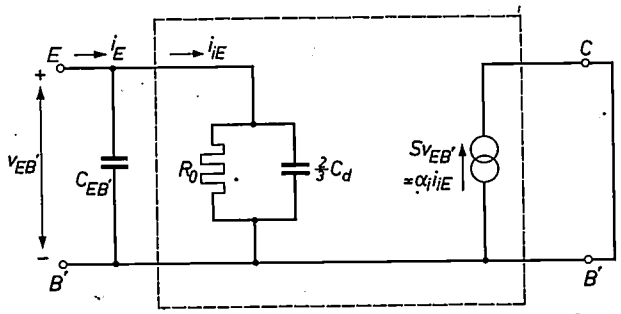


Fig. 14. Equivalent AC circuit for a transistor in common-base configuration with short-circuited output. The section inside the dashed lines relates to the intrinsic transistor and is based on the vector diagram of fig. 13.

external element to represent the capacitance of the emitter barrier. The complete diagram is an equivalent circuit for a transistor in the arrangement in fig. 9*a*, i.e. the common-base configuration with output short-circuited to alternating current. The influence of the frequency is expressed in the capacitors $\frac{2}{3}C_d$ en $C_{EB'}$ at the input side, and in a complex value for the strength of the current source at the output side. Since the holes for the diffusion capacitance are supplied only to the extent of two-thirds across the emitter barrier, only two-thirds of C_d is to be found at the input side; the remainder is accounted for in the complex strength of the current source. This strength can be expressed in terms of the input voltage $v_{EB'}$ by means of the complex transconductance S. A simple expression for S, which closely approximates to the exactly calculated behaviour of the transistor up to about the characteristic frequency ω_1, is:

$$S = \frac{1}{R_0\left(1 + j\,\frac{\omega}{3\omega_1}\right)}.$$

This expression corresponds to a vector i_{iC}, the end of which lies on the smaller of the semi-circles drawn in fig. 13.

The strength of the current source can of course also be expressed in terms of the input current i_{iE} of the transistor by means of the current amplification factor of the intrinsic transistor [13]). An examination of this current amplification factor: $\alpha_i = -i_{iC}/i_{iE}$, will make it clear that the properties of a transistor deteriorate with increasing frequency. Between the currents in question there exists a frequency-dependent phase difference (see fig. 13); α_i is therefore complex. In *fig. 15* it can be seen how α_i varies as a function of frequency. Here again the curves corresponding to our simplified theory and to the rigorous mathematical treatment are drawn. It can be seen that the simplified theory gives a reasonable presentation of the decreasing modulus, and especially of the argument of α_i, as long as ω is smaller than ω_1. The decrease of the modulus of α_i with increasing frequency is one of the reasons why the gain deteriorates towards higher frequencies. To counteract this decrease it is evidently necessary to raise the characteristic frequency $\omega_1 = 2D_p/w^2$ (see 11); in other words the diffusion constant must be large and the base thickness small. As far as this goes, *N-P-N* transistors are at an advantage compared with *P-N-P* transistors, for in germanium the diffusion constant for electrons is roughly twice as high as that for holes. Reducing the base thickness w is, however, the principal means of increasing ω_1.

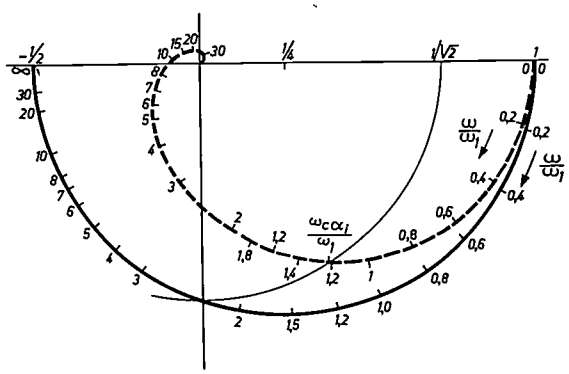


Fig. 15. Locus of the points which, in the complex plane, represent the current amplification factor α_i of the intrinsic transistor (really the transport factor β of the base [13])) when the frequency increases from zero to ∞. ω_1 ($= 2D_p/w^2$, see formula 11) is a characteristic frequency. The semi-circle corresponds to our simplified theory that the concentration pattern in the base is always adapted to the instantaneous velocity with which the concentration changes at the emitter. For not too high values of ω/ω_1 this curve is a reasonable approximation of the dashed spiral around the origin ($\alpha_i = \operatorname{sech} \sqrt{2j\omega/\omega_1}$), which corresponds to the exact solution of the differential equation for the diffusion of holes through the base. $\omega_{c\alpha i}$ is the cut-off angular frequency for α_i, that is the angular frequency at which α_i has dropped in magnitude by a factor $\sqrt{2}$.

Cut-off frequency of the intrinsic current amplification factor in common-base connection

The frequency behaviour of the current amplification factor is often characterized by what is termed the cut-off frequency. This is defined as the frequency at which the current amplification factor has dropped to $1/\sqrt{2}$ of the value at zero frequency. As may be seen from fig. 15, the cut-off (angular) frequency $\omega_{c\alpha i}$ of the intrinsic current amplification factor α_i is given by:

$$\omega_{c\alpha i} = 1.21\,\omega_1. \quad \ldots\ldots \quad (12)$$

[13]) It would be more correct here not to speak of the current amplification factor α_i but of the transport factor β of the base. At zero frequency, β is a real number and identical with what is termed "base efficiency" on page 239 of the article in footnote [1]). Assuming that the emitter current consists entirely of holes, then $\beta = \alpha_i$.

The cut-off frequency of α_i is equal, except for the factor 1.21, to the characteristic frequency ω_1, and therefore we can if we wish use $\omega_{c\alpha i}$ instead of ω_1 as the characteristic frequency.

The name cut-off frequency can be misleading, for it suggests that the transistor does not amplify above this frequency but does below it. This need not be the case, however, because the behaviour of a transistor is not governed purely by the current amplification factor; there are of course other transistor parameters that are important in this connection. Therefore $\omega_{c\alpha i}$ is the cut-off frequency of α_i but *not* of the transistor.

Emitter barrier short-circuited to alternating current

We shall now examine the complementary case where the forward voltage across the emitter barrier is kept constant — i.e. the barrier is short-circuited to alternating current, while the DC voltage across the collector barrier has superimposed on it a small alternating voltage $v_{CB'}$ (*fig. 16a*). As regards the concentration lines in the base this means that the end points A and A' are fixed (fig. 16*b*), while the end points C and C' oscillate horizontally, owing to the barrier thickness depending on the voltage (see fig. 8). Strictly speaking, C and C' also oscillate vertically, but when there is an appreciable reverse voltage across the collector barrier the vertical oscillation is negligible compared with the horizontal, which is exactly the converse of the oscillation of A and A' in fig. 9*b*.

In many respects this case resembles the previous one: here too, the concentration lines oscillate about fixed end points, while the hole and electron contents of the base show mutually identical periodic variations, represented by the hatched triangles in fig. 16*b*. There is again a through current (or transit current) in phase with the voltage $v_{CB'}$, and a diffusion capacitance gives rise to the flow of storage currents over the two barriers.

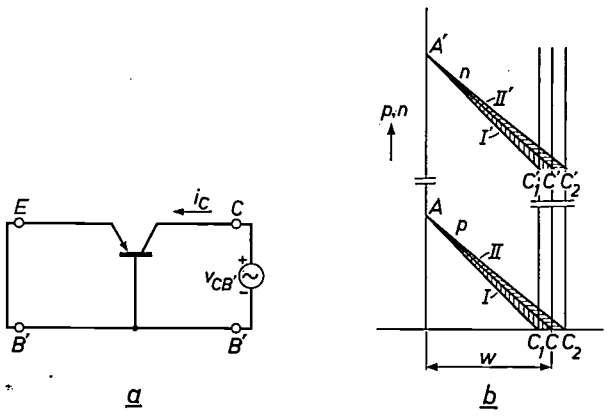


Fig. 16. *a*) The complementary case of fig. 9*a*; the emitter barrier is short-circuited to alternating current; an alternating voltage $v_{CB'}$ is superimposed on the biasing voltage across the collector barrier.
b) The corresponding concentration pattern in the base: the concentration lines for holes (p) and electrons (n) oscillate around the fixed points A and A' between the respective positions I and II and I' and II'.

Common-base connection with short-circuited input

Fig. 16*a* shows a transistor in common-base configuration with short-circuited input. It follows from the foregoing that a transistor in this arrangement can be represented by an equivalent circuit (*fig. 17*) constructed in the same way as that with short-

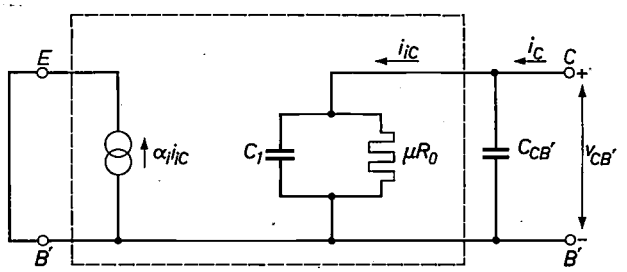


Fig. 17. Equivalent circuit for a transistor in common-base configuration with short-circuited input, drawn by analogy with the diagram in fig. 14 relating to short-circuited output.

circuited output (fig. 14). In the current source, on the left in fig. 17, is to be seen the same factor α_i as in fig. 14 on the right. In fig. 17 the intrinsic transistor contains on the right an impedance consisting of a resistance in parallel with a capacitance C_1, which account respectively for the through current and the storage current across the collector barrier. A capacitor $C_{CB'}$ is again added to represent the barrier capacitance.

An important difference compared with the previous case is that the resistance is now a certain factor μ higher, and the diffusion capacitance the same factor smaller. In alloyed transistors μ is of the order of magnitude of 10^3. The reason is that the slope of the concentration line is much less sensitive (by the factor μ; see formula 20 in Appendix) to voltage variations across the collector than across the emitter barrier. The result is that, of the capacitances C_1 and $C_{CB'}$, the latter now predominates, which is precisely the opposite of the situation in fig. 14.

Complete equivalent circuit for common-base connection; internal base resistance

In *fig. 18* the figures 14 and 17 are combined to form an equivalent circuit for a transistor in common-base connection, which is valid even if neither

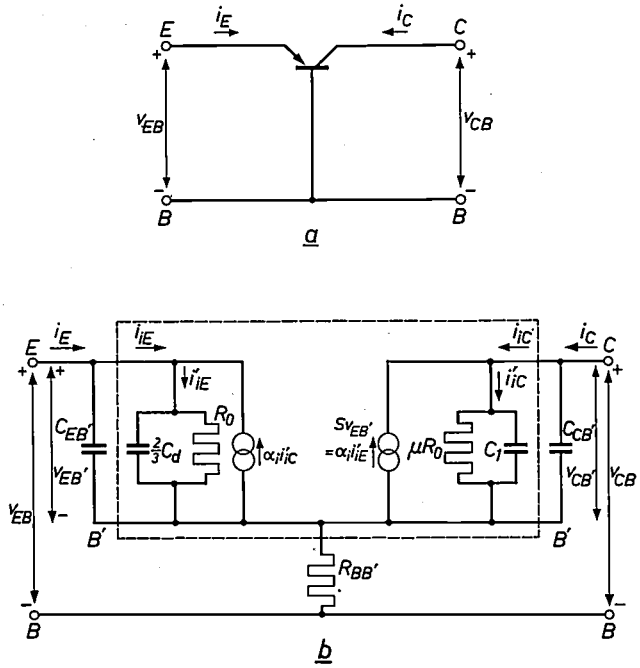


Fig. 18. *a*) A transistor in common-base configuration. *b*) Equivalent circuit for (*a*), obtained by combining figures 14 and 17, at the same time adding the internal base resistance $R_{BB'}$. The correctness of the diagram follows from the fact that, afterputting $R_{BB'}$ at zero, the diagram changes to that in fig. 14 or 17, depending on whether the output or input is short-circuited. Between the dashed lines is the intrinsic transistor.

the input nor the output are short-circuited. The inclusion of the internal base resistance $R_{BB'}$ accounts for the resistance met by the base current in its transverse passage through the crystal. It can be seen in this diagram that the voltages v_{EB} and v_{CB} between the transistor terminals are not identical with the voltages $v_{EB'}$ and $v_{CB'}$ that relate to the intrinsic transistor.

At low frequencies no or at least very little current goes through $R_{BB'}$, because the base current is then very small (fig. 13). As the frequency increases the base current increases, and $R_{BB'}$ then increasingly causes feedback from the output to the input. As this feedback is unwanted, the aim therefore is to keep the value of $R_{BB'}$ small.

As the frequency rises the capacitance $C_{CB'}$ of the collector barrier increasingly short-circuits the output current source. In order to expand the useful frequency range of a transistor it will therefore be necessary to make $C_{CB'}$ small.

Complete equivalent circuit for common-emitter connection

Considerations similar to those just discussed for a transistor in common-base connection may also be applied to a transistor in common-emitter connection. A complication, however, is that one must take into account the base loss at zero frequency, a loss which is of negligible importance in the common-base configuration. *Fig. 19* shows the common-emitter arrangement together with its equivalent circuit constructed on the basis of the considerations just referred to. In the latter circuit the diffusion capacitance C_d is now no longer partly but wholly between the input terminals. This is because the base current is the input current, and all electrons needed for changing the concentration pattern in the base have to pass the base contact (see page 167). The electrons thus follow only one path, whereas the holes take two paths, namely over both barriers. Furthermore, $R_{BB'}$ is now in series with the input capacitance, so that across $R_{BB'}$ a part of the input voltage appears that is lost to the transistor action. The higher the frequency the larger is this part, because the influence of C_d (and of $C_{EB'}$ if this is not negligible compared with C_d) then increases. The capacitance $C_{CB'}$ now gives rise to internal feedback and has thus, as far as this is concerned, assumed the role of $R_{BB'}$ in the common-base arrangement. In the common-emitter connection the frequency range is evidently limited by the same factors as in the common-base connection, but the adverse influences of $R_{BB'}$ and that of $C_{CB'}$ appear in different ways in these two configurations.

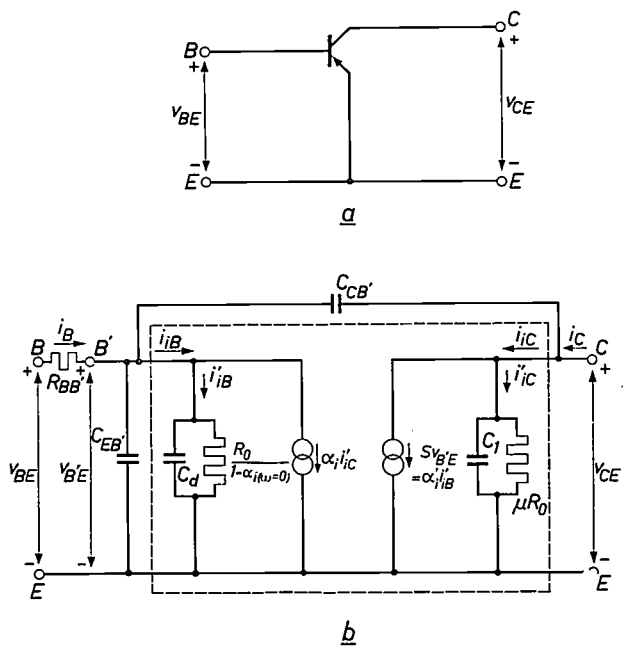


Fig. 19. *a*) The transistor in common-emitter configuration. *b*) Equivalent circuit for (*a*).

Common-emitter connection with short-circuited output

When the output of a transistor in common-emitter arrangement is short-circuited to alternating current, and an alternating voltage is applied to the input terminals, this voltage appears both across the emitter barrier and across the collector barrier (*fig. 20a*). This means that point *C* of the hole concen-

tration line in the base (fig. 9*b*) is not rigorously fixed. As mentioned on page 169, however, the effect of an alternating voltage across the collector barrier is negligible compared with the effect of an alternating voltage of equal magnitude across the emitter barrier (see also the Appendix). For that reason one can therefore still assume that in this case too the points C and C' of the concentration lines are fixed. The vector diagram for the intrinsic transistor in fig. 13 is therefore equally applicable to the common-emitter arrangement with short-circuited output. Fig. 20*b* shows the vector diagram from fig. 13 adapted to the common-emitter connection: instead of $v_{EB'}$ the input voltage is now $v_{B'E}$ ($= -v_{EB'}$). Consequently, as far as the current vectors are concerned, the diagram is a mirror image of fig. 13, so that right and left are interchanged.

We shall first correct the vector diagram to allow for the base efficiency [10]) before constructing from it an equivalent circuit for a transistor in common-emitter connection with short-circuited output. In the common-base arrangement, neglecting the base loss (base efficiency 100%) has scarcely any influence, but in common-emitter arrangement it would lead to infinitely high values of input impedance and current amplification factor at zero frequency. To avoid this, fig. 20*c* takes into account that — owing to the base loss — the emitter current at $\omega = 0$ is somewhat higher than $i_t = v_{B'E}/R_0$, whereas the collector current is somewhat lower. Consequently, at zero frequency there flows a certain base current $i_{iB(\omega=0)} \approx (1 - \alpha_{i(\omega=0)})i_t$ in phase with the input voltage $v_{B'E}$. (Here $\alpha_{i(\omega=0)}$ is the current amplification factor for the common-base circuit at $\omega = 0$, a factor which, due to the base loss, is somewhat lower than 1.) The movement of the ends of the vectors i_{iE} and i_{iC} as the frequency increases is hardly affected by the correction. The current $i_{iB(\omega=0)}$ is of interest only at low frequencies ($\omega \ll \omega_1$).

From fig. 20*c* we derive *fig. 21* as the equivalent circuit for a transistor in common-emitter connection with short-circuited output. At the input side there is now a resistance $R_0/(1-\alpha_{i(\omega=0)})$ with which no longer a part, but all of the diffusion capacitance C_d is in parallel. The strength of the current source can be expressed in the input voltage $v_{B'E}$ using the transconductance S. The latter has the same value as in the diagram in fig. 14 for the common-base arrangement, because — apart from signs — the same output current and the same input voltage are involved.

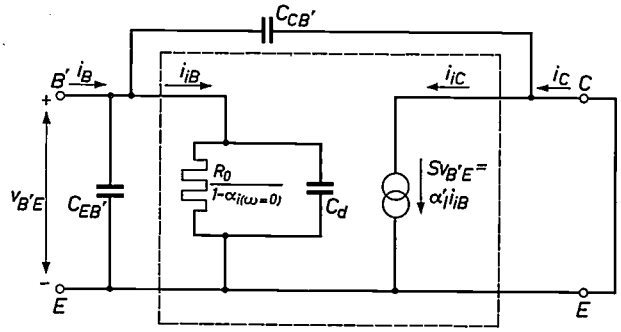


Fig. 21. Equivalent circuit for a transistor in common-emitter configuration with short-circuited output. The part between the dashed lines again relates to the intrinsic transistor and is based on the vector diagram in fig. 20*c*. Between the (complex) current amplification factors α_i' and α_i of the intrinsic transistor in common-emitter and common-base configuration, respectively, there exists the relation $\alpha_i' = \alpha_i/(1-\alpha_i)$.

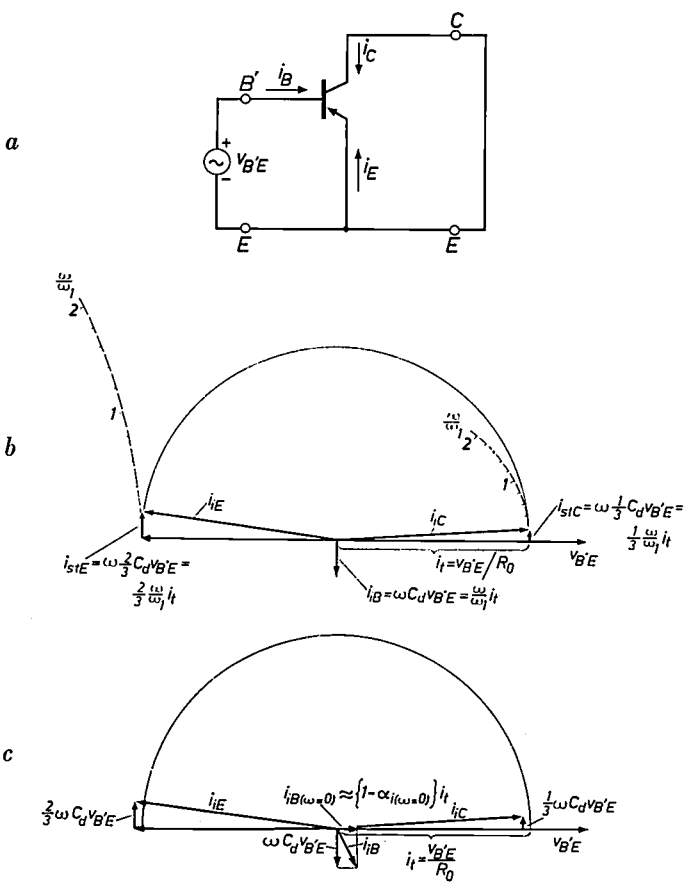


Fig. 20. *a*) A transistor in common-emitter configuration with short-circuited output.
b) Vector diagram for the intrinsic transistor, corresponding to (*a*).
c) The vector diagram (*b*) corrected for the base loss.

The effect of the barrier capacitances is taken into account by $C_{EB'}$ and $C_{CB'}$ respectively between emitter and base terminals and between collector and base terminals.

The strength of the current source can of course also be expressed in terms of the input current i_{iB} of the intrinsic transistor, i.e. as $\alpha_i' i_{iB}$. The current amplification factor α_i' is provided with an accent to distinguish it from the current amplification factor α_i of the common-base circuit. From the condition that the vectorial sum of base, emitter and collector currents should be zero it follows that $i_{iB} = -i_{iE} - i_{iC}$. After dividing by i_{iC} we find $1/\alpha_i' = (1/\alpha_i) - 1$, that is $\alpha_i' = \alpha_i/(1-\alpha_i)$. This well-known formula, then, retains its validity even with complex numbers.

Cut-off frequency of the intrinsic current amplification factor in common-emitter connection

For the current amplification factor α'_i in the common-emitter configuration we can again define a cut-off frequency $\omega_{c\alpha'_i}$ as the frequency at which the modulus of this factor has decreased by a factor $\sqrt{2}$. As we shall show, there exists between $\omega_{c\alpha'_i}$ and $\omega_{c\alpha_i}$ the simple relation:

$$\omega_{c\alpha'_i} = \omega_{c\alpha_i}/1.21\alpha'_{i(\omega=0)} = \omega_1/\alpha'_{i(\omega=0)} . \quad . \quad . \quad (13)$$

Since $\alpha'_{i(\omega=0)}$ is much greater than unity in a good transistor, it follows from this that $\omega_{c\alpha'_i}$ is much smaller than $\omega_{c\alpha_i}$. This does not, however, justify the conclusion that a transistor in common-base connection can be used up to much higher frequencies than the same transistor in common-emitter connection. Before examining this in some detail, we shall first derive

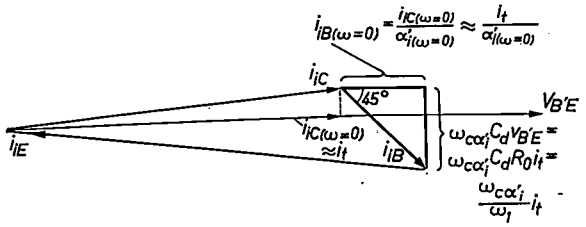


Fig. 22. Collector, base and emitter currents of the intrinsic transistor in common-emitter configuration with short-circuited output (cf. fig. 20*c*), drawn as a closed vectorial triangle. The situation shown, in which the vector i_{iB} makes an angle of 45° with the input voltage $v_{B'E}$, corresponds to the cut-off frequency $\omega_{c\alpha'_i}$ of the current amplification factor α'_i.

the above-mentioned relation with the aid of *fig.22*. In this figure, emitter, base and collector currents of the intrinsic transistor (cf. fig. 20*c*) are drawn as a closed vectorial triangle. The situation chosen is that in which the vector i_{iB} makes an angle of 45° with the input voltage $v_{B'E}$. This situation corresponds to a frequency at which i_{iB} is a factor of $\sqrt{2}$ *greater* than $i_{iB(\omega=0)}$, the value at $\omega = 0$. As regards amplitude the value of i_{iC} scarcely differs from that at $\omega = 0$, and therefore α'_i ($=i_{iC}/i_{iB}$) is about a factor $\sqrt{2}$ *smaller* than at $\omega = 0$. Fig.22 relates, then, to the situation at the cut-off frequency $\omega_{c\alpha'_i}$ of α'_i. Equating the sides containing the right angle of the thickly drawn triangle in fig. 22 gives: $\omega_{c\alpha'_i} = \omega_1/\alpha'_{i(\omega=0)}$, from which, using (12), we find the relation (13).

The value of $\omega_{c\alpha'_i}$ gives scarcely any indication of the merits at high frequencies of a transistor in common-emitter connection. This is demonstrated, for example, by the extreme case of zero base loss. In that case $\alpha'_i = \infty$, and $\omega_{c\alpha'_i} = 0$ according to (13). The fact that this by no means implies that this transistor in common-emitter arrangement can only be used at extremely low frequencies is apparent, for instance, from the vector diagram in fig. 20*b*, which relates to this case.

Common-emitter connection with short-circuited input

The common-emitter connection is identical with the common-base connection when the input in both cases is short-circuited. Fig. 17 is therefore equally valid for the common-emitter connection with short-circuited input. *Fig. 23* reproduces the diagram of fig. 17, now redrawn for common emitter. The equivalent circuit given in fig. 19*b* for a transistor in common-emitter connection, where neither the output nor the input is short-circuited, was obtained by combining figures 21 and 23 and adding the internal base resistance $R_{BB'}$.

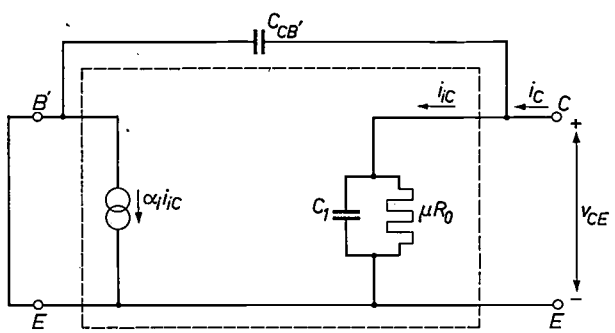


Fig. 23. Equivalent circuit of a transistor in common-emitter arrangement with short-circuited input. Since transistors in common-emitter and common-base configuration are identical when the input is short-circuited, this diagram — although drawn differently — is the same as that in fig. 17.

The three factors that limit the frequency range

Summarizing, we conclude that there are three principal factors that limit the frequency range of a transistor: the characteristic frequency ω_1, the capacitance $C_{CB'}$ of the collector barrier, and the internal base resistance $R_{BB'}$. In the case of transistors for operation at high frequencies, ω_1 must be large while $C_{CB'}$ and $R_{BB'}$ must be small. To increase the value of ω_1 it is necessary to make the base very thin. The higher we make ω_1 the lower becomes the diffusion capacitance C_d. If C_d becomes comparable with the capacitance $C_{EB'}$ of the emitter barrier, then steps should also be taken to reduce $C_{EB'}$. We shall show that these requirements are conflicting in the case of alloyed transistors, and that consequently this type of transistor is fundamentally unsuitable for high-frequency operation.

Fundamental unsuitability of alloyed transistors for high-frequency operation

In an alloyed transistor, owing to the nature of the manufacturing process, the equilibrium value of the majority concentration in the base is always small compared with that in the collector [2]). This leads to difficulties if one wishes to make the base very thin, which, as we have seen, is a necessary condition for a high-frequency transistor. The alloyed transistor does not therefore in principle lend itself for use at high frequencies.

To illustrate this, *fig. 24* shows two graphs similar to those in fig. 8 but corresponding to the situation in the collector barrier of a *P-N-P* alloyed transistor. In reality the donor concentration c_{don} in the base B compared with the acceptor concentration c_{acc} in the collector C is much smaller than it is represented to be in fig. 24. Practical values, fore xample, are $c_{don} = 10^{20}\ m^{-3}$ and $c_{acc} = 10^{24}\ m^{-3}$. It can be seen that the total thickness d of the barrier is almost equal to the distance d_N over which the barrier extends into the base material.

In formulae (7) and (8) for the barrier thickness d and the barrier capacitance C_d we can now neglect c_{don} with respect to c_{acc}. Taking into account moreover that the total voltage V_{NP} between the N and P regions consists of the sum of the diffusion voltage V_0 and the reverse voltage $V_{B'C}$, we find:

$$d \approx d_N = \sqrt{2\varepsilon/qc_{don}}\ \sqrt{V_0 + V_{B'C}} \qquad (14)$$

and

$$C_{CB'} = \frac{O\sqrt{\frac{1}{2}\varepsilon q c_{don}}}{\sqrt{V_0 + V_{B'C}}}. \qquad (15)$$

To obtain a wide useful frequency range, $C_{CB'}$ must be small. According to formula (15) it is favour-

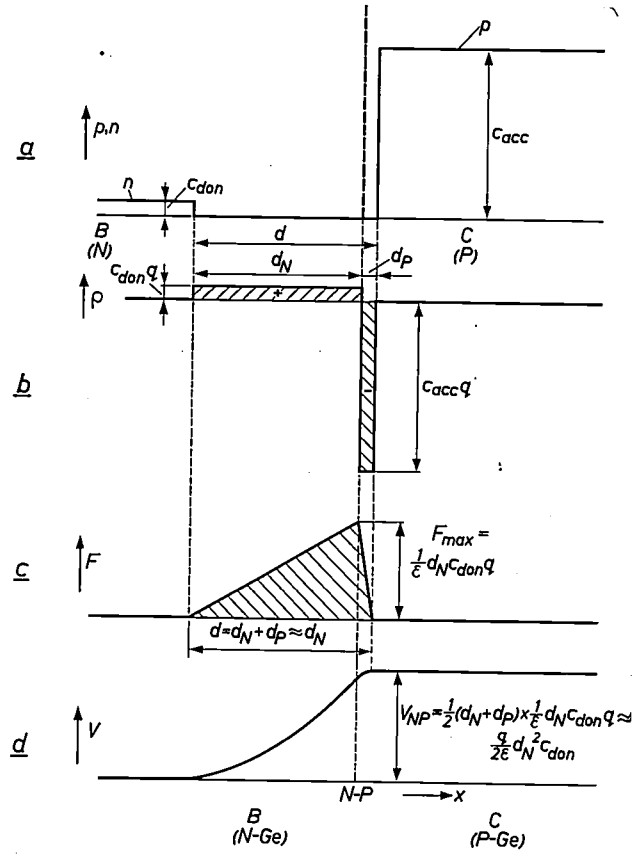


Fig. 24. The collector barrier in a *P-N-P* alloyed transistor; cf. fig.8.

able for this purpose to make both the barrier area O and the donor concentration c_{don} in the base small, but the reverse biasing voltage $V_{B'C}$ large. (V_0 increases and decreases slightly with c_{don}, but so little that V_0 may be regarded here as constant.)

A small value of O implies small transverse dimensions. Since the base thickness w must also be small (with a view to a high characteristic frequency ω_1), we see that high-frequency transistors must be as small as possible at least as far as the active transistor part is concerned.

The reverse biasing voltage $V_{B'C}$ cannot be raised ad libitum. Fig. 24*c* indicates that the maximum field strength F_{max} in the barrier has the value $qc_{don}\,d_N/\varepsilon$. Using formula (14) for d_N we find:

$$F_{max} = \sqrt{2q/\varepsilon}\ \sqrt{(V_0 + V_{B'C})c_{don}}. \quad . \quad (16)$$

Thus, F_{max} increases with increasing $V_{B'C}$. Roughly it can be said that if F_{max} exceeds a certain critical value F_{cr} the barrier will break down, i.e. the current will suddenly increase sharply with the voltage. The reverse voltage at which this effect occurs, called the breakdown voltage V_{br}, follows from (16):

$$V_{br} = \frac{\varepsilon F_{cr}^2}{2qc_{don}} - V_0. \quad . \quad . \quad . \quad . \quad (17)$$

We see from this that the breakdown voltage V_{br} is higher according as c_{don} is lower.

Since c_{don} is equal to the majority concentration in the base, and the majority concentration mainly determines the resistivity (a small majority concentration implies a high resistivity), we now have two reasons for choosing a base material with a high resistivity ϱ_B (e.g. about 1 ohm cm); these reasons are 1) to minimize the capacitance of the collector barrier, and 2) to have a high breakdown voltage for this barrier.

For the sake of completeness we mention a third reason for making ϱ_B high, which is to bring the emitter efficiency close to 100%. For this purpose the base resistivity ϱ_B must be high compared with the resistivity ϱ_E of the emitter material[10]). This condition can, however, also be met by making ϱ_E small, which usually proves to be possible in practice.

There are also three reasons why one should aim at a *low* value of ϱ_B (that is a high c_{don}). The first, obviously, is the wish to have a low internal base resistance $R_{BB'}$. In transistors for high frequencies the base thickness w must be made very small with a view to having a sufficiently high characteristic frequency ω_1. Since the base current is a majority current which flows transversely through the base, a thin base is not conducive to a low internal base resistance. To keep this within bounds, ϱ_B should not therefore be too high.

The second reason for keeping ϱ_B small is that the distance d_N over which the collector barrier, for a given reverse voltage, extends in the base material, is greater the higher is ϱ_B. In high-frequency transistors, with their extremely thin base, there is easily a danger of the collector barrier touching the emitter barrier — known as "punch-through" — whereby the base completely disappears. It is seen from formula (14) that d_N is inversely proportional to $\sqrt{c_{don}}$. The danger of punch-through can thus be reduced by making c_{don} large (i.e. ϱ_B small).

The third reason is the Early effect, i.e. the undesired feedback effect from the output to the input, an effect which also follows from the dependence of d_N on the reverse voltage. It is precisely when the base is very thin that periodic changes in d_N become significant (cf. fig. 16*b*).

The conflicting requirements in regard to the resistivity ϱ_B of the base material (summarized in the *table* below) make some compromise necessary, but even with the best compromise the useful frequency range goes no higher than about 10 Mc/s. The alloyed transistor is therefore fundamentally unsuitable for high frequencies. A practical drawback, moreover, is that the fabrication of the ex-

tremely thin base needed for high-frequency operation involves considerable technological problems [14]).

Summary of the conflicting requirements which the donor concentration c_{don} in the base material (or the resistivity ϱ_B of this material) should meet if an alloyed transistor is to be used at high frequencies.

	Wanted	Requirements in regard to	
		c_{don}	ϱ_B
1	Small collector-barrier capacitance $C_{CB'}$	small	large
2	High breakdown voltage of collector barrier	small	large
3	High emitter efficiency	small	large
4	Low base resistance	large	small
5	Little risk of punch-through	large	small
6	Minimum feedback due to Early effect	large	small

Transistors for high frequencies

A way out of this impasse can be found if the majority concentration in the base is made much greater than in the collector, which is exactly the opposite of the situation in alloyed transistors. The resistivity ϱ_C of the collector material is then much higher than ϱ_B, and the collector barrier extends almost completely into the collector material. Points 5 and 6 in the table are then ruled out, because no further trouble is caused by changes in the thickness of the collector barrier. The wishes under points 1 and 2 can now be fulfilled by making ϱ_C sufficiently high. A conflict remains only in regard to points 3 and 4. As we have already indicated, however, a favourable compromise is possible here: by using in the emitter an acceptor such as gallium or aluminium it is possible to make the majority concentration in the emitter high compared with that in the base, even though the majority concentration in the base is high enough to provide an acceptable internal base resistance $R_{BB'}$.

Modern *P-N-P* transistors for high frequencies are made in principle as follows. The starting material is *P*-type germanium in which the acceptor concentration (*fig. 25*) is sufficient for the collector region to have the required majority concentration. Donors are allowed to diffuse into this material under such conditions that an *N-P* junction forms at the appropriate depth. This is the collector junction. Next, by alloying, the acceptor concentration at the edge of the crystal over a certain depth is increased sufficiently to produce here a *P* region, which forms the emitter. At the *N-P* junction on the collector side the donor surplus of the base passes with no discontinuity into the acceptor surplus of the collector. In the immediate vicinity of this junction the donor and acceptor surpluses are therefore very small. This favours the formation of a thick barrier and hence a low barrier capacitance. The donor surplus, however, rises much faster and much higher than the acceptor surplus. Consequently, the collector

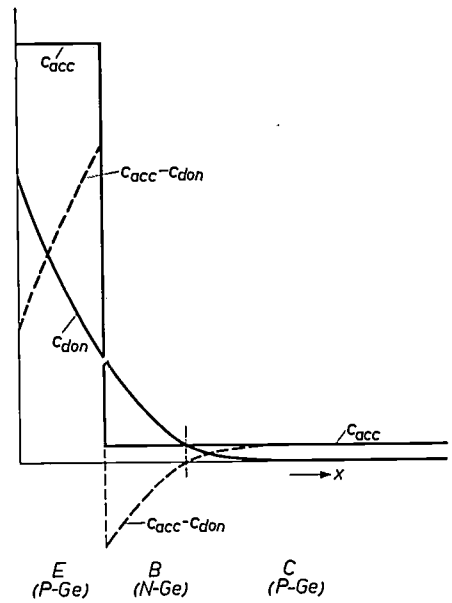


Fig. 25. The distributions of the concentrations of acceptors (c_{acc}) and donors (c_{don}) in a modern *P-N-P* high-frequency transistor (not to scale), where the donor concentration is obtained by diffusion. The dashed line represents the difference $c_{acc} - c_{don}$. Whether the region is *P*- or *N*-type depends on whether the acceptor or donor concentration is greater.

barrier still extends mainly into the collector material, and voltage changes across this barrier are mainly absorbed by changes in the thickness of the space-charge layer in this material. The resistivity of the base material is high at the collector but low at the emitter. The internal base resistance can therefore still be relatively low.

A favourable consequence of the inhomogeneous distribution of the donor surplus is the formation in the base of an electric field which is directed from emitter to collector and therefore aids the diffusion in driving the holes through the base. This field is known as the "drift field", hence transistors containing such a field are known as drift transistors [15]).

[14]) For a further discussion see the article in footnote [2]) on page 234.

[15]) H. Krömer, The drift transistor, Naturwissenschaften **40**, 578-579, 1953. The principle of the drift transistor is described in the British patent specification 769674, based on a priority application in the United States of America of 16th November 1951, in the name of W.G. Pfann.

The drift field increases the characteristic frequency ω_1.

Using a diffusion process to form the base layer has made it possible to reduce the base thickness from about 40 μm — which is normal in alloyed transistors — to a few μm and even to less than 1 μm. One of the methods of doing this is termed the alloy-diffusion technique, the principle of which was discussed in a recent article in this journal [2]). In this way transistors can be made that deliver a reasonable gain up to 1000 Mc/s.

Appendix

Proof of the first fundamental theorem

The hole current I^+ consists of a component I_F^+, supplied by the field F, and a component I_d^+, supplied by diffusion. The same applies to the electron current I^-, which consists of I_F^- and I_d^-. These component currents together deliver the total current I_{tot}:

$$I_{tot} = I_F^+ + I_F^- + I_d^+ + I_d^- .$$

Outside a barrier the concentration lines for majorities and minorities — because of the neutrality condition applicable there — have the same form, but they lie at entirely different "levels" (see fig. 2). The field currents are proportional to the concentrations themselves, but the diffusion currents are proportional to the concentration gradients, which are identical for both kinds of charge carriers. The minority field current is therefore very small compared with the majority field current, but the minority diffusion current is of the same order of magnitude as the majority diffusion current. This essentially underlies our first fundamental theorem, which states: *In determining the minority current in a homogeneous P or N region outside a barrier, only diffusion need be taken into account.* It is not so easy, however, to see at a glance all cases that might be encountered. For instance, one might think at first that, in the case of minorities, the field current would always be negligible compared with the diffusion current. But this is certainly not true at places where there is no concentration gradient, which will be the case at a considerable distance from a barrier (fig. 2). For there the diffusion current is zero and the minority current is 100% a field current. The entire minority current is then, however, negligible.

A formal proof that the minority field current is always negligible can be provided as follows:

$$I_F^+ = q\mu_p F p \qquad I^+ = -qD_p \frac{dp}{dx}$$

$$I_F^- = q\mu_n F n \qquad I_d^- = +qD_n \frac{dn}{dx}$$

In these formulae μ_p and μ_n denote respectively the mobility of holes and electrons. If, using these formulae, we determine I_F^+/I_d^+ and I_F/I_{tot}, taking into account that $dp/dx = dn/dx$ and, for germanium, $D_n/D_p = \mu_n/\mu_p \approx 2$, we find that there is a relation between I_F^+/I_d^+ and I_F^+/I_{tot} of the form represented by the curve in *fig. 26*. The curve is a rectangular hyperbola given by the equation:

$$\left(\frac{I_F^+}{I_d^+} - \frac{1}{\frac{2n}{p}+1}\right)\left(\frac{I_F^+}{I_{tot}} - \frac{1}{\frac{2n}{p}+1}\right) = \frac{1}{\left(\frac{2n}{p}+1\right)^2}.$$

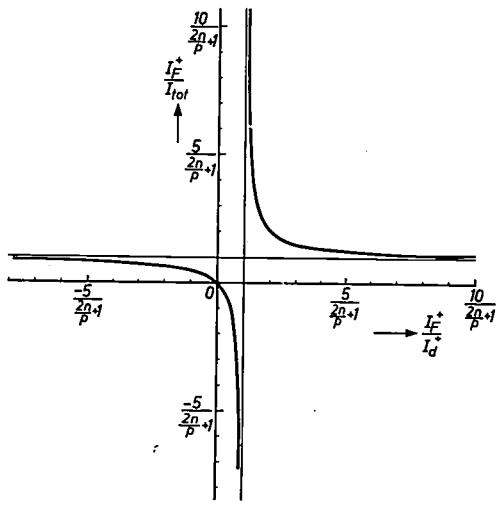


Fig. 26. Graphic illustration of the first fundamental theorem.

In an N region, where $n \gg p$, it can be seen from fig. 26 that always at least one of the quantities $|I_F^+/I_d^+|$ and $|I_F^+/I_{tot}|$ is extremely small. A small value of $|I_F^+/I_d^+|$ means that the influence of the field is negligible compared with that of the diffusion. But even if $|I_F^+/I^+|$ is not very small, the field effect is still negligible because then I_F^+/I_{tot} tot is certainly very small, so that I_F^+ in that case makes a negligible contribution to the *total* current [7]).

The diffusion voltage V_0

When a P-N crystal is in a state of equilibrium, then according to the Boltzmann distribution [16]) the hole and electron concentrations are related by exponential functions with the potential V:

$$p = p_0\, e^{-qV/kT} \text{ and } n = n_0\, e^{+qV/kT}. \quad \ldots \quad \text{(18a and b)}$$

The values p_0 and n_0 are the values at the place where we set $V = 0$. Multiplying together the expressions for p and n gives the familiar result that, in equilibrium, the product of p and n is independent of V, and hence of the location in the crystal.

From (18) an expression can be derived for the spontaneous potential difference between an N and a P region, i.e. for the diffusion voltage V_0. Outside the barrier, p has the equilibrium values p_{P0} and p_{N0} in the P and the N regions respectively. If we choose $V = 0$ in the P region, then V_0 is the potential of the N region. From (18a) it then follows that:

$$p_{N0} = p_{P0}\exp(-qV_0/kT),$$

so that

$$V_0 = \frac{kT}{q} \ln(p_{P0}/p_{N0}). \quad \ldots \quad \text{(19)}$$

If (18b) is used instead of (18a) we find $\ln(n_{N0}/n_{P0})$ instead of $\ln(p_{P0}/p_{N0})$, in agreement with the constancy of the product $p \times n$. Given $p_{P0} = 10^{24}$ m^{-3} and $p_{N0} = 5 \times 10^{18}$ m^{-3}, which correspond to the values of c_{acc} in the P region and c_{don} in the N region mentioned on page 172, we find for V_0 at $T = 300$ °K the approximate value of 0.3 V.

Proof of the second fundamental theorem

In treating the concentration pattern in a barrier, the starting assumption is that an equilibrium state may be assumed in

[16]) See G. Joos, Theoretical physics, 2nd edition, chapter 24, London 1951.

the barrier even though an external voltage V_u is applied to the crystal, so that current flows. The proof of our second fundamental theorem, which stated: *The minority concentration in the boundary plane of a barrier is proportional to* exp (qV_u/kT) (see page 158), is then perfectly straightforward. We have counted V_u as positive if it lowers the potential of the N region with respect to the P region (see page 158). $V_0 - V_u$ therefore represents the total voltage between the barrier boundary planes *2* and *1* (fig. 2). Formulae (18a and b) then give: $p(2) = p(1)$ exp $\{-q(V_0 - V_u)/kT\}$ and $n(2) = n(1)$ exp $\{q(V_0 - V_u)/kT\}$. These are only *two* equations for the four unknown concentrations in the boundary planes. The neutrality condition, applied in boundary plane *1* and in boundary plane *2*, yields two further equations: $p(1) - n(1) = p_{P0} - n_{P0}$ and $n(2) - p(2) = n_{N0} - p_{N0}$. The solution of the four equations with four unknowns is straightforward. For $n(1)$ and $p(2)$ expressions are found which, for small values of V_u, yield by approximation the formulae (1a) and (1b). The approximation is valid only provided V_u is so small that the minority concentrations are small compared with the equilibrium concentrations of the majorities; this, incidentally, is the same condition that applied to the validity of the first theorem [7]). In that case the last two equations can be written as $p(1) \approx p_{P0}$ and $n(2) \approx n_{N0}$, and we find the relations (1a) and (1b).

Calculation of the factor μ

On page 169 it was mentioned that the slope of the hole concentration line in the base is a factor μ less sensitive to a change in the voltage $V_{CB'}$ across the collector barrier than to an equal change of the voltage $V_{EB'}$ across the emitter barrier. To calculate μ we note that μ is equal to the ratio of the changes $\Delta V_{CB'}$ and $\Delta V_{EB'}$, which compensate each other's effect. Since the slope in question is given by $p(2)/w$ (see e.g. fig. 9*b*), we find — partially differentiating the expression $p(2)/w$ with respect to $V_{EB'}$ and $V_{CB'}$:

$$0 = \frac{1}{w}\frac{\mathrm{d}p(2)}{\mathrm{d}V_{EB'}}\,\Delta V_{EB'} - \frac{p(2)}{w^2}\frac{\mathrm{d}w}{\mathrm{d}V_{CB'}}\,\Delta V_{CB'},$$

so that:

$$\mu = \frac{\Delta V_{CB'}}{\Delta V_{EB'}} = \frac{w}{p(2)}\,\frac{\mathrm{d}p(2)/\mathrm{d}V_{EB'}}{\mathrm{d}w/\mathrm{d}V_{CB'}}.$$

From formula (1b) it follows that $\mathrm{d}p(2)/\mathrm{d}V_{EB'} = p(2)q/kT$. To determine $\mathrm{d}w/\mathrm{d}V_{CB'}$, we use formula (7). Obviously $\mathrm{d}w = -\mathrm{d}d_N$ and $\mathrm{d}V_{CB'} = -\mathrm{d}V_{NP}$, so that $\mathrm{d}w/\mathrm{d}V_{CB'} = \mathrm{d}d_N/\mathrm{d}V_{NP}$, from which with (7) it follows that $\mathrm{d}w/\mathrm{d}V_{CB'} = d_N/2V_{NP}$. With this we find:

$$\mu = 2\,\frac{w}{d_N}\,\frac{V_{NP}}{kT/q}. \quad \ldots\ldots\ldots \quad (20)$$

Since $w \gg d_N$ and $V_{NP} \gg kT/q$, μ is a large number. In an alloyed transistor a normal value is about 1000. An AC voltage across the collector-base barrier thus has approximately 1000 times less influence on the slope of the hole line than the same AC voltage across the emitter-base barrier.

Summary. The article is didactic in intention. Transistor action, both in principle and for the purpose of numerical calculations, can be explained in terms of the concentration pattern of electrons and holes in the base. It is not necessary to use the quantum-theoretical band model. Two fundamental theorems are formulated which elucidate the behaviour of the minority charge carriers. From the latter the behaviour of the majority charge carriers is derived.

With increasing frequency certain effects become operative which are bound up with the periodic changes in the concentration pattern. The concepts diffusion capacitance and barrier capacitances are introduced to account respectively for changes of pattern outside and inside the barriers. In the analysis two complementary cases are considered where one barrier has only the DC biasing voltage across it, while across the other barrier there is also a small AC voltage. The alternating currents flowing in these cases are examined in respect of amplitude and phase. On this basis equivalent circuits are constructed for transistors in common-base and common-emitter configurations. It is shown that the frequency at which a transistor still gives a reasonable gain is limited by three quantities: a characteristic frequency (closely connected with the diffusion capacitance), the collector-barrier capacitance, and the internal base-resistance. In a discussion of means of widening the useful frequency range it is explained why alloyed junction transistors are fundamentally not suitable for high-frequency operation and why transistors with a diffused base are.

Proof of the two fundamental theorems is given in an appendix.

THE AUDIBILITY OF PHASE ERRORS

621.391.832.22

In the transmission of a signal via an apparatus or installation the output signal should ideally be equal to the input signal, apart from a proportionality constant. As a rule it is not, the signal suffering distortion which may be either linear or non-linear. We shall consider here only linear distortion.

The linear distortion in a transmission can be characterized in various ways. In the transmission of video signals importance is attached to the signal as a function of time, and use is made of the (*step-function*) transient response [1]). In sound transmission more attention is paid to the signal as an aggregate of Fourier components and the *frequency response characteristic* is used. For a complete description of the latter, the transmission of a sinusoidal signal both in amplitude and in phase should be known as a function of frequency. The usual practice is to consider only the amplitude characteristic, it being assumed that if this is reasonably flat the transmission is sufficiently free of distortion, and the phase response characteristic is disregarded. There is no doubt that in straightforward cases this is quite acceptable and well matched to the mechanism of hearing. But it is equally evident that it can never be made a generally valid criterion, because it is possible to create circumstances where the output signal is distinctly heard to differ from the input signal though the amplitude characteristic is flat.

[1]) See e.g. J. Haantjes, Judging an amplifier by means of the transient characteristic, Philips tech. Rev. 6, 193-201, 1941. Also: A. van Weel, Phase linearity of television receivers, Philips tech. Rev. 18, 33-51, 1956/57.

the barrier even though an external voltage V_u is applied to the crystal, so that current flows. The proof of our second fundamental theorem, which stated: *The minority concentration in the boundary plane of a barrier is proportional to* exp (qV_u/kT) (see page 158), is then perfectly straightforward. We have counted V_u as positive if it lowers the potential of the N region with respect to the P region (see page 158). $V_0 - V_u$ therefore represents the total voltage between the barrier boundary planes *2* and *1* (fig. 2). Formulae (18a and b) then give: $p(2) = p(1)$ exp $\{-q(V_0 - V_u)/kT\}$ and $n(2) = n(1)$ exp $\{q(V_0 - V_u)/kT\}$. These are only *two* equations for the four unknown concentrations in the boundary planes. The neutrality condition, applied in boundary plane *1* and in boundary plane *2*, yields two further equations: $p(1) - n(1) = p_{P0} - n_{P0}$ and $n(2) - p(2) = n_{N0} - p_{N0}$. The solution of the four equations with four unknowns is straightforward. For $n(1)$ and $p(2)$ expressions are found which, for small values of V_u, yield by approximation the formulae (1a) and (1b). The approximation is valid only provided V_u is so small that the minority concentrations are small compared with the equilibrium concentrations of the majorities; this, incidentally, is the same condition that applied to the validity of the first theorem [7]). In that case the last two equations can be written as $p(1) \approx p_{P0}$ and $n(2) \approx n_{N0}$, and we find the relations (1a) and (1b).

Calculation of the factor μ

On page 169 it was mentioned that the slope of the hole concentration line in the base is a factor μ less sensitive to a change in the voltage $V_{CB'}$ across the collector barrier than to an equal change of the voltage $V_{EB'}$ across the emitter barrier. To calculate μ we note that μ is equal to the ratio of the changes $\Delta V_{CB'}$ and $\Delta V_{EB'}$, which compensate each other's effect. Since the slope in question is given by $p(2)/w$ (see e.g. fig. 9*b*), we find — partially differentiating the expression $p(2)/w$ with respect to $V_{EB'}$ and $V_{CB'}$:

$$0 = \frac{1}{w}\frac{\mathrm{d}p(2)}{\mathrm{d}V_{EB'}}\,\Delta V_{EB'} - \frac{p(2)}{w^2}\frac{\mathrm{d}w}{\mathrm{d}V_{CB'}}\,\Delta V_{CB'},$$

so that:

$$\mu = \frac{\Delta V_{CB'}}{\Delta V_{EB'}} = \frac{w}{p(2)}\,\frac{\mathrm{d}p(2)/\mathrm{d}V_{EB'}}{\mathrm{d}w/\mathrm{d}V_{CB'}}.$$

From formula (1b) it follows that $\mathrm{d}p(2)/\mathrm{d}V_{EB'} = p(2)q/kT$. To determine $\mathrm{d}w/\mathrm{d}V_{CB'}$, we use formula (7). Obviously $\mathrm{d}w = -\mathrm{d}d_N$ and $\mathrm{d}V_{CB'} = -\mathrm{d}V_{NP}$, so that $\mathrm{d}w/\mathrm{d}V_{CB'} = \mathrm{d}d_N/\mathrm{d}V_{NP}$, from which with (7) it follows that $\mathrm{d}w/\mathrm{d}V_{CB'} = d_N/2V_{NP}$. With this we find:

$$\mu = 2\,\frac{w}{d_N}\,\frac{V_{NP}}{kT/q}. \quad \ldots\ldots\ldots \quad (20)$$

Since $w \gg d_N$ and $V_{NP} \gg kT/q$, μ is a large number. In an alloyed transistor a normal value is about 1000. An AC voltage across the collector-base barrier thus has approximately 1000 times less influence on the slope of the hole line than the same AC voltage across the emitter-base barrier.

Summary. The article is didactic in intention. Transistor action, both in principle and for the purpose of numerical calculations, can be explained in terms of the concentration pattern of electrons and holes in the base. It is not necessary to use the quantum-theoretical band model. Two fundamental theorems are formulated which elucidate the behaviour of the minority charge carriers. From the latter the behaviour of the majority charge carriers is derived.

With increasing frequency certain effects become operative which are bound up with the periodic changes in the concentration pattern. The concepts diffusion capacitance and barrier capacitances are introduced to account respectively for changes of pattern outside and inside the barriers. In the analysis two complementary cases are considered where one barrier has only the DC biasing voltage across it, while across the other barrier there is also a small AC voltage. The alternating currents flowing in these cases are examined in respect of amplitude and phase. On this basis equivalent circuits are constructed for transistors in common-base and common-emitter configurations. It is shown that the frequency at which a transistor still gives a reasonable gain is limited by three quantities: a characteristic frequency (closely connected with the diffusion capacitance), the collector-barrier capacitance, and the internal base-resistance. In a discussion of means of widening the useful frequency range it is explained why alloyed junction transistors are fundamentally not suitable for high-frequency operation and why transistors with a diffused base are.

Proof of the two fundamental theorems is given in an appendix.

THE AUDIBILITY OF PHASE ERRORS

621.391.832.22

In the transmission of a signal via an apparatus or installation the output signal should ideally be equal to the input signal, apart from a proportionality constant. As a rule it is not, the signal suffering distortion which may be either linear or non-linear. We shall consider here only linear distortion.

The linear distortion in a transmission can be characterized in various ways. In the transmission of video signals importance is attached to the signal as a function of time, and use is made of the (*step-function*) transient response [1]). In sound transmission more attention is paid to the signal as an aggregate of Fourier components and the *frequency response characteristic* is used. For a complete description of the latter, the transmission of a sinusoidal signal both in amplitude and in phase should be known as a function of frequency. The usual practice is to consider only the amplitude characteristic, it being assumed that if this is reasonably flat the transmission is sufficiently free of distortion, and the phase response characteristic is disregarded. There is no doubt that in straightforward cases this is quite acceptable and well matched to the mechanism of hearing. But it is equally evident that it can never be made a generally valid criterion, because it is possible to create circumstances where the output signal is distinctly heard to differ from the input signal though the amplitude characteristic is flat.

[1]) See e.g. J. Haantjes, Judging an amplifier by means of the transient characteristic, Philips tech. Rev. 6, 193-201, 1941. Also: A. van Weel, Phase linearity of television receivers, Philips tech. Rev. 18, 33-51, 1956/57.

This can be illustrated with a signal to which artificial reverberation has been added. The reverberation is built up from a series of echoes in such a way that the amplitude characteristic is perfectly flat [2]; the audible difference compared with the input signal (which is without reverberation) must then, when described mathematically, be attributed to phase shifts. Another example will be mentioned presently.

It may now be asked in what situation a phase distortion is just perceptible and how it is perceived; and further, whether in electro-acoustical practice phase errors have any effect on sound quality, which nowadays has to meet such high standards.

The first question can be answered as far as a simple phase error is concerned. In listening experiments with a transmission having an entirely flat amplitude characteristic we introduced a more or less abrupt phase shift of 360° at a frequency somewhere in the audio range (see *fig. 1*). The circuit used for this purpose is shown in *fig. 2*. The phase "jump" has no audible effect on a steady test signal, but it has on test signals in which numerous or marked discontinuities occur. This leads, then, to an analysis on the basis of step functions and to a description which in fact amounts to using the transient response mentioned at the beginning. When a step-function signal is applied to the input, the phase behaviour referred to gives rise at the output to a damped oscillation,

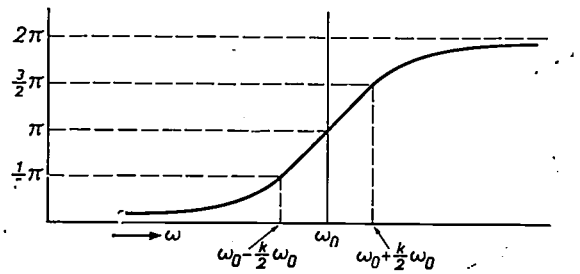


Fig. 1. Phase response characteristic used for audibility experiments. This curve which comprises a single phase "jump" of 360° near the angular frequency ω_0, is characteristic of the transfer function

$$\frac{1 - \frac{\omega^2}{\omega_0^2} - jk\,\frac{\omega}{\omega_0}}{1 - \frac{\omega^2}{\omega_0^2} + jk\,\frac{\omega}{\omega_0}}.$$

The parameter k governs the steepness of the phase shift. The amplitude characteristic of this function is flat (absolute value for all frequencies = 1). As can be demonstrated from network theory this transfer function is relatively easy to produce, by appropriately combining the input signal and the response of a simple LC circuit to that signal.

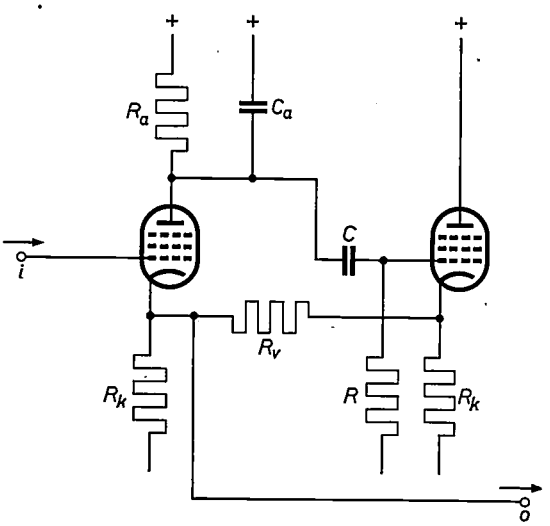


Fig. 2. Possible circuit for obtaining the transfer function mentioned in fig.1. Unlike the basically simple method indicated there, no LC circuit is used here, in fact not even a single inductance. For low values of ω_0 this is an advantage, not only because it is difficult to make the large inductances needed for this purpose, but also because parasitic capacitances or resistances of coils adversely influence the flat amplitude characteristic required. Dimensioning is given by the following relations, which hold for $SR_k \gg 1$:

$$R_a C_a = RC = \frac{1}{\omega_0}, \qquad \frac{1}{k} = SZ_k = \frac{1}{2} SR_a,$$

where Z_k is the cathode impedance, taking into account R_k, R_v and the coupling with the neighbouring valve. (Both valves are assumed to be identical.)

the amplitude and decrement of which depend closely on the steepness of the phase jump. The effect is perceptible if the decay time determined from the response (measured, like a reverberation time, as the time in which the level drops 60 dB) is longer than about 50 ms. This value corresponds to a phase-response slope of roughly 4° per cycle. The audibility tests are done in an anechoic room. In a room which is not anechoic, the effects can also be heard, but only for an even steeper phase-response slope.

This answers the question as far as this special case of a simple phase error is concerned, but in view of the numerous cases possible no general answer can be given for any arbitrary phase error.

Let us now consider the practical electro-acoustical situation. The main source of linear distortion that needs to be investigated in this repect is the loudspeaker, with its erratic amplitude and phase response characteristic. In order to get an idea of the influence of phase effects in this case, we used a method which makes it possible, with the same amplitude characteristic, to compare a phase-linear with a phase-distorted loudspeaker reproduction. This method and its results will now briefly be discussed.

The loudspeaker fed with the audio signal to be reproduced is set up in an anechoic room. The sound

[2]) M. R. Schroeder and B. F. Logan, "Colourless" artificial reverberation, J. Audio Engng. Soc. 9, 192-197, 1962 (No.3).

from the loudspeaker is recorded on magnetic tape via a microphone. The recording naturally shows amplitude and phase deviations with respect to the original which are governed not only by the loudspeaker response but also by those of the microphone and magnetic recording; the first, however, will give by far the greatest distortion.

The magnetic tape in its turn is played back through the same loudspeaker, and the sound is recorded on a second magnetic tape. Plainly, the latter recording shows double the amplitude error (in dB) and double the phase error (in degrees). However, if the tape with the first recording is played back in the reverse direction for the re-recording, the latter will exhibit *no* phase errors because, related to the original phase of the signal, the phase shifts of the first and second transmissions are of equal magnitude but of opposite sign. The double amplitude error of course remains. By listening to the re-recordings thus obtained it is possible, with the same amplitude characteristic, to make a comparison between signals with and without phase errors.

To avoid an audible difference being masked by the amplitude error, variable filters were used to flatten the amplitude characteristic of the whole system. Obviously only the coarse irregularities could be flattened in this way, and not the fine structure of the loudspeaker response characteristic.

Instead of two successive recordings, four, six or more can be used so as to magnify the phase deviation and make the difference better audible. In our experiments we worked with a maximum of six recordings, giving a sixfold magnification of the phase error. Both speech and music signals were investigated. The above-mentioned experiments with a single abrupt phase shift had already made it clear that a speech signal was much more likely to give rise to audible effects than a music signal. Our loudspeaker experiments confirmed this. With music signals no difference was audible even using six successive recordings, whereas with speech signals and the same number of recordings an effect was just recognizable.

The nature of the effect is difficult to describe. It is bound up with the onset of the signal, which it tends to blur. This may be understood as follows. From the theory of the step-function transient response it is known that, in phase-linear reproduction, the output transient as a function of time is symmetrical with respect to the instant of the "jump", in other words it is just as strong *before* as after the onset. This need not be a physical unreality provided the whole step-function response is sufficiently delayed in relation to the input signal (see *fig. 3*).

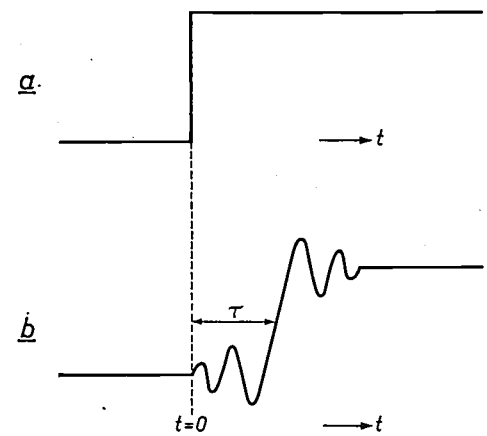


Fig. 3. Example of a phase-linear response (*b*) to a step-function signal (*a*) at the time $t = 0$. The response shows symmetry with respect to $t = \tau$. Physical reality of course demands that the transient should begin after $t = 0$, implying that there must necessarily be a delay τ.

From the manner in which we have obtained the phase-linear characteristic it can also be understood that the result will not only be followed but also preceded by a kind of "reverberation". This is the blurring effect referred to.

The effects mentioned were of course produced very artificially and have little relation to anything that might be encountered in practice, for they were only discovered from a sixfold recording. The final conclusion as regards the practical effect of phase errors is therefore a confirmation of the view that, as individual distortion in electro-acoustical apparatus, they have no influence on the sound quality, even where the highest standards are set.

K. TEER *).

*) Philips Research Laboratories, Eindhoven.

RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF THE PHILIPS LABORATORIES AND FACTORIES

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

3121: H. Martinides, K. Nienhuis and K. van Duuren: The building-up and the spread of the discharge in halogen counters (Proc. 5th int. Conf. on ionization phenomena in gases, Munich 1961, Vol. I, pp. 756-762, North-Holland Publ. Co., Amsterdam 1962).

3122: L. A. Ellenkamp: Instrument for recording coercive force as a function of temperature (Rev. sci. Instr. **33**, 383-384, 1962, No. 3).

3123: W. Albers, C. Haas, H. Ober, G. R. Schodder and J. D. Wasscher: Preparation and properties of mixed crystals $SnS_{(1-x)}Se_x$ (Phys. Chem. Solids **23**, 215-220, March 1962).

3124: K. J. de Vos, W. A. J. J. Velge, M. G. van der Steeg and H. Zijlstra: Permanent magnetic properties of iron-cobalt phosphides (J. appl. Phys. **33**, 1320-1322, 1962, suppl. to No. 3).

3125: J. C. Balder and C. Kramer: Video transmission by delta modulation using tunnel diodes (Proc. Inst. Radio Engrs. **50**, 428-431, 1962, No. 4).

3126: K. van Duuren: Electrical characteristics of halogen-filled Geiger counters (Le Vide **16**, 235-248, 1961, No. 95).

3127: J. J. van Loef: A note on activation analysis with neutron-induced threshold reactions (Nukleonik **4**, 151-152, 1962, No. 3).

3128: J. Verweel: Permeability and ferromagnetic resonance line width of some ferrites with garnet structure (Proc. Instn. Electr. Engrs. **109 B**, suppl. No. 21, 95-98, 1962).

3129: H. Bosma: On the principle of stripline circulation (Proc. Instn. Electr. Engrs. **109 B**, suppl. No. 21, 137-146, 1962).

3130*: H. Bremmer: The pulse solution connected with the Sommerfeld problem for a dipole in the interface between two dielectrics (Electromagnetic waves, editor R. E. Langer, pp. 39-64, Univ. Wisconsin Press, Madison 1962).

3131: G. J. M. Ahsmann: Some remarks on the anode fall in the Faraday dark space (as **3121**, pp. 306-314).

3132: C. J. M. Rooymans: A new compound in the Na_2O-Fe_2O_3 system (J. Phys. Soc. Japan **17**, 722-723, 1962, No. 4).

3133: C. Haas: Infrared absorption in heavily doped *N*-type germanium (Phys. Rev. **125**, 1965-1971, 1962, No. 6).

3134: R. Dijkstra, J. de Jonge and M. F. Lammers: The kinetics of the reaction of phenol and formaldehyde (Rec. Trav. chim. Pays-Bas **81**, 285-296, 1962, No. 3).

3135: B. van der Veen: Een meer-lokettenprobleem met overwerk (Statistica neerl. **16**, 195-204, 1962, No. 2). (A multiple-server problem with overtime; in Dutch.)

3136: W. J. Oosterkamp: Möglichkeiten zur Steigerung der Detailerkennbarkeit im Röntgenbild (Ärztl. Forschung **16**, I/122-I/129, 1962, No. 3). (Possibilities of improving detail-perceptibility in X-ray images; in German.)

3137: G. Klein: Some aspects of the measurement of small voltages (Mesucora 1961, Congr. int. Mesure - contrôle - régulation - automatisme, Paris, pp. 109-118).

3138: U. Enz: Permeability, crystalline anisotropy and magnetostriction of polycrystalline manganese-zinc-ferrous ferrite (Proc.Instn.Electr. Engrs. **109 B**, suppl. No. 21, 246-247, 1962).

3139: H. Mooijweer: Enige wiskundige aspecten van parametrische versterkers (Ingenieur **74**, O 37-O 43, 1962, No. 20). (Some mathematical aspects of parametric amplifiers; in Dutch.)

3140: M. J. Sparnaay: Corrections of the theory of the stability of hydrophobic colloids (the specific influence of ions) (Rec. Trav. chim. Pays-Bas **81**, 395-416, 1962, No. 5).

3141: P. F. Bongers: Thermodynamische en kristallografische eigenschappen van enkele verbindingen der overgangselementen (Chem. Weekblad **58**, 313-319, 1962, No. 26). (Thermodynamic and crystallographic properties of some compounds of the transition elements; in Dutch.)

3142: J. A. W. van Laar: Should the cold check test for furniture lacquers be substituted by more scientific test methods? (VI. FATIPEC-Kongress 1962, pp. 430-436, Verlag Chemie, Weinheim 1962).

3143: H. G. Grimmeiss, W. Kischio and H. Koelmans: *P-N*-junction photovoltaic effect in zinc-doped GaP (Solid-state electronics **5**, 155-159, May-June 1962).

3144: S. van Houten: Mechanical losses in Li-doped NiO semiconductors (Phys. Chem. Solids **23**, 1045-1048, Aug. 1962).

3145: F. K. Lotgering: Paramagnetic susceptibilities of Fe^{2+} and Ni^{2+} ions at tetrahedral or octahedral sites of oxides (Phys. Chem. Solids **23**, 1153-1167, Aug. 1962).

3146: D. J. van Ooijen: The electrical resistance of hydrogen-charged nickel wires (Phys. Chem. Solids **23**, 1173-1175, Aug. 1962).

3147: H. Koopman: Hydrolysis of 2-substituted 4,6-dichloro-1,3,5-triazines (Rec. Trav. chim. Pays-Bas **81**, 465-474, 1962, No. 5).

3148: P. A. van Zwieten and H. O. Huisman: Synthesis and physiological properties of some heterocyclic-aromatic sulphides and sulphones, II. Synthesis of some aryl-pyrimidyl, aryl-thiazolyl and aryl-thienyl sulphides (Rec. Trav. chim. Pays-Bas **81**, 554-564, 1962, No. 6).

3149: J. Cornelissen, H. W. J. H. Meyer, A. M. Kruithof and H. C. Hamaker: The mechanical behavior of pristine glass rods (Advances in glass technology, pp. 488-510, Plenum Press, New York 1962).

3150: J. H. Haanstra and C. Haas: Infrared spectrum of an acceptor in ZnTe (Physics Letters, Amsterdam, **2**, 21-22, 1962, No. 1).

3151: J. F. Marchand: A new type of parallel cryotron (Physics Letters, Amsterdam, **2**, 57-58, 1962, No. 2).

3152: T. J. Viersma: Designing load-compensated fast-response hydraulic servos (Control Engng. **9**, No. 5, 111-114, 1962).

3153: K. Compaan and H. Zijlstra: Effective-field approach to the problem of interacting polarized particles (Phys. Rev. **126**, 1722-1723, 1962, No. 5).

3154: L. A. Æ. Sluyterman: Photo-oxidation, sensitized by proflavine, of a number of protein constituents (Biochim. biophys. Acta **60**, 557-561, 1962, No. 3).

3155: C. A. van Sluis and J. H. Stuy: On the inactivation of transforming deoxyribonucleic acid by heat (Biochem. biophys. Res. Comm. **7**, 213-219, 1962, No. 3).

3156: J. F. Marchand and J. Volger: Radiation-induced transport of magnetic flux along a superconducting sheet (Physics Letters, Amsterdam, **2**, 118-119, 1962, No. 3).

3157: G. Blasse and E. W. Gorter: Some magnetic properties of spinels with compositions $NiFe_{2-x}V_xO_4$ (J. Phys. Soc. Japan **17**, suppl. B-I, 176-180, 1962).

3158: J. Smit, F. K. Lotgering and R. P. van Stapele: Anisotropy of ferrous ions in spinels (J. Phys. Soc. Japan **17**, suppl. B-I, 268-272, 1962).

3159: P. A. van Zwieten, M. Gerstenfeld and H. O. Huisman: Synthesis and physiological properties of some heterocyclic-aromatic sulphides and sulphones, III. Synthesis of some heterocyclic-aromatic sulphones (Rec. Trav. chim. Pays-Bas **81**, 604-615, 1962, No. 7).

3160: P. A. van Zwieten, J. Meltzer and H. O. Huisman: Synthesis and physiological properties of some heterocyclic-aromatic sulphides and sulphones, IV. Biological investigations (Rec. Trav. chim. Pays-Bas **81**, 616-623, 1962, No. 7).

3161: A. H. Gomes de Mesquita: On the structure of triphenylmethyl perchlorate. A crystallographic investigation (thesis G.U. Amsterdam, Sept. 1962).

3162: D. A. Schreuder: Aufgehellte Fahrbahndecken und lichttechnische Probleme (Asphalt- und Teerstrassen, No. 16, 144-153, Kirschbaum, Bad Godesberg 1962). (Illuminated road surfaces and associated lighting problems; in German.)

3163: A. T. Vink and C. Z. van Doorn: Complex fine-structure of GaP photoluminescence at 4.2 °K (Physics Letters, Amsterdam, **1**, 332-333, 1962, No. 8).

3164: J. Dieleman, R. S. Title and W. V. Smith: Paramagnetic resonance studies of Cr^+ in cubic and hexagonal ZnS (Physics Letters, Amsterdam, **1**, 334-335, 1962, No. 8).

3165: C. W. Berghout: Phase equilibria in superconducting niobium-zirconium alloys (Physics Letters, Amsterdam, **1**, 292-295, 1962, No. 7).

3166: S. van Houten: Preparation of single crystals of cadmium oxide (Nature **195**, 484-485, 1962, No. 4840).

3167: S. van Houten: Magnetic interaction in EuS, EuSe, and EuTe (Physics Letters, Amsterdam, **2**, 215-216, 1962, No. 5).

3168: J. Volger and P. S. Admiraal: A dynamo for generating a persistent current in a superconducting circuit (Physics Letters, Amsterdam, **2**, 257-259, 1962, No. 5).
See also Philips tech. Rev. **25**, 16-19, 1963/64 (No. 1).

3169*: J. D. Fast: Entropy. The significance of the concept of entropy and its applications in science and technology (Philips Technical Library, XII + 313 pp., 69 fig., Eindhoven 1962).

3170*: N. V. Franssen: Stereofonica (Philips Technical Library, VIII + 91 pp., 64 fig., Eindhoven 1962). (Stereophonics; in Dutch.)

3171: A. Schmitz: Nomograph for the preparation of germanium tunnel diodes (Solid-state electronics **5**, 354-357, Sept.-Oct. 1962).

3172: C. Kooy: Anisotropic exaggerated grain growth and sintering in $MnFe_2O_4$ and $Y_3Fe_5O_{12}$ (Science of ceramics, Proc. Conf. Oxford 1961, Vol. 1, pp. 21-34, Academic Press, London 1962).

3173: G. H. Jonker and W. Noorlander: Grain size of sintered barium titanate (as **3172**, pp. 255-264).

3174: O. Drexler and B. R. Schat: Development of ceramic materials with a dielectric constant of 10000 at room temperature (as **3172**, pp. 239-254).

3175: J. Hornstra: On the type of point defects formed after crossing of dislocations (Acta metallurgica **10**, 987-988, 1962, No. 10).