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.

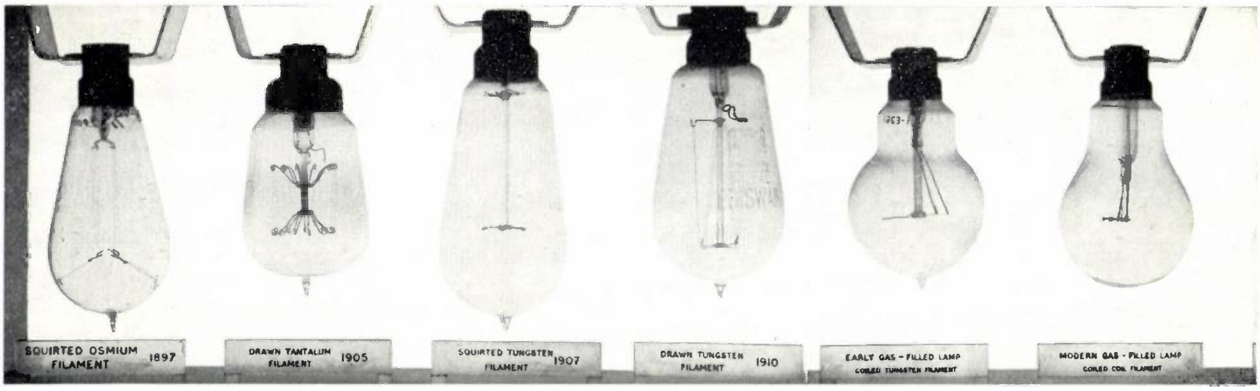

SQUIRTED OSMIUM FILAMENT 1897
DRAWN TANTALUM FILAMENT 1905
SQUIRTED TUNGSTEN FILAMENT 1907
DRAWN TUNGSTEN FILAMENT 1910
EARLY GAS - FILLED LAMP COILED TUNGSTEN FILAMENT
MODERN GAS - FILLED LAMP COILED COIL FILAMENT

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

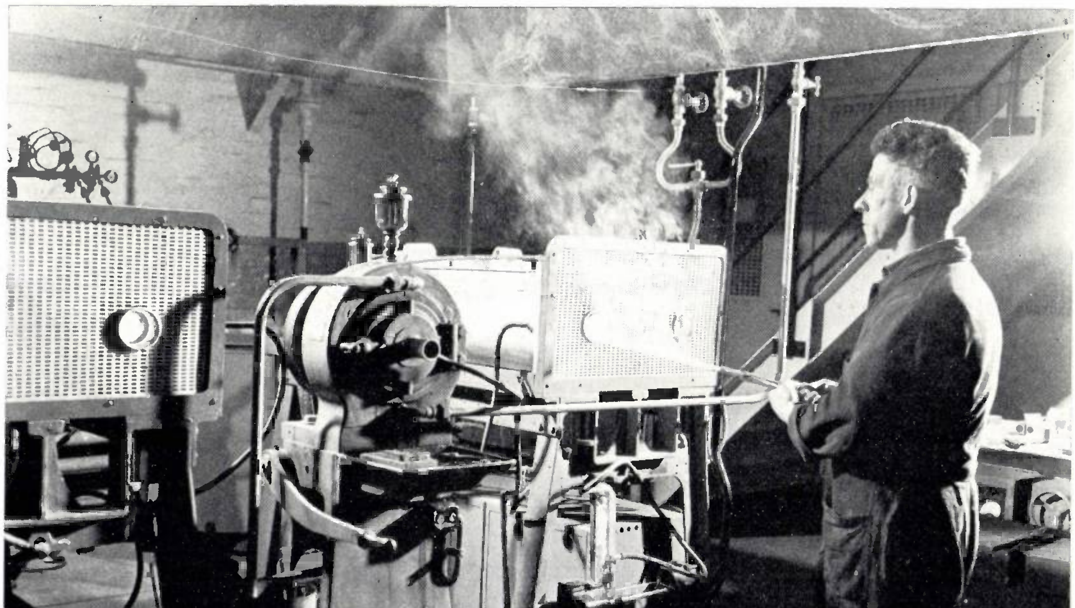
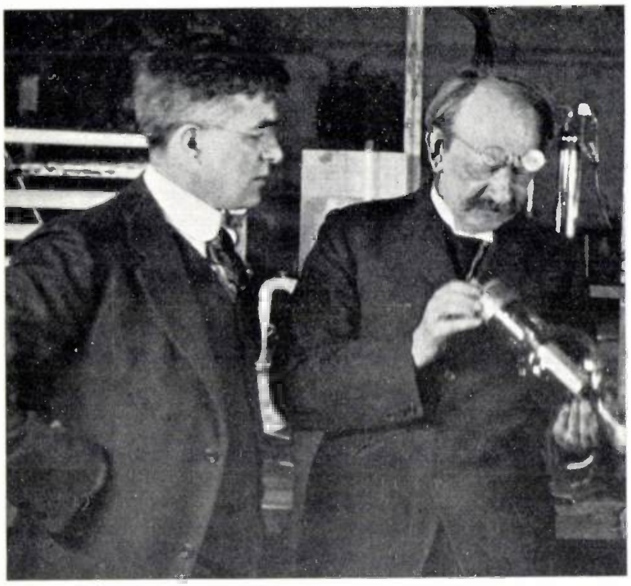
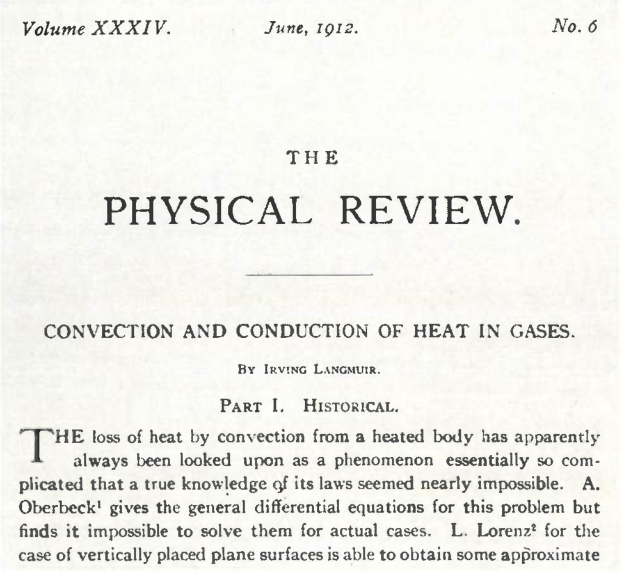


Fig. 2. Ductile tungsten wire is made by machine-hammering (swaging) sintered tungsten bars to increase their density after which they are drawn to the required thickness through hard-metal or diamond dies. The photograph shows a swaging machine, with the tungsten bar being introduced manually after 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 tempcrature 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:

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

where

$$\varphi_\mathrm{i} = 4.19 \int_0^{T_i} k \, \mathrm{d}T \,. \quad \dots \dots \quad (2)$$

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

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

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

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

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 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:

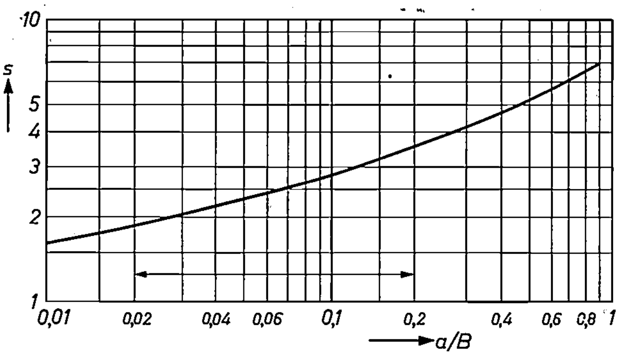


Fig. 4. Graphic representation of equation (4). In practice a/B varies only within the limits indicated along the abscissa.

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

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:

$$W_\mathrm{g} = C\, l \left(\frac{a}{B}\right)^{0.3} (\varphi_2 - \varphi_1) \,. \quad \dots \quad (6)$$

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:

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

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.

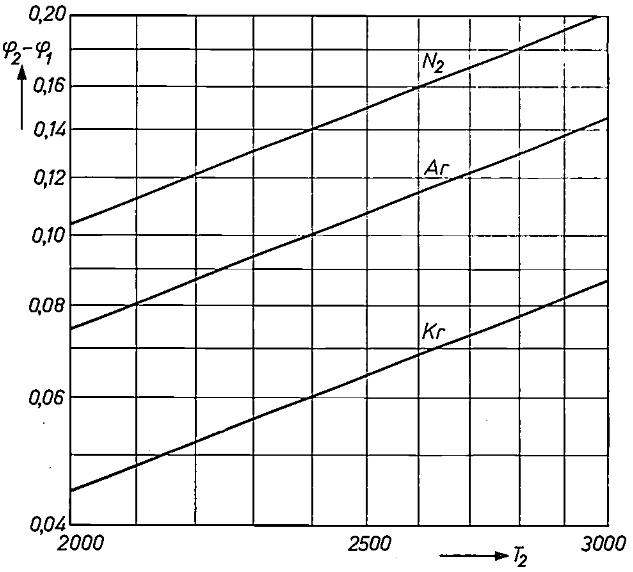


Fig. 5. Plot of the quantity φ_2—φ_1 in Langmuir's heat-loss equation (eq. 1) versus the filament temperature T_2 in °K, representing the theoretical values for different gases (in the case of N_2 no acccount is taken of dissociation).

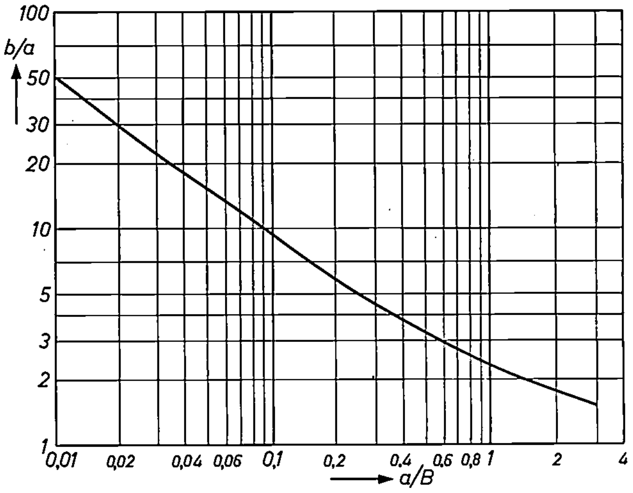


Fig. 6. Ratio of the diameter b of the stationary gas layer to the diameter a of the (cylindrical) incandescent filament, as a function of a/B, calculated using Langmuir's equation (8).

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

[5] Langmuir really gives the relation:

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

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

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-

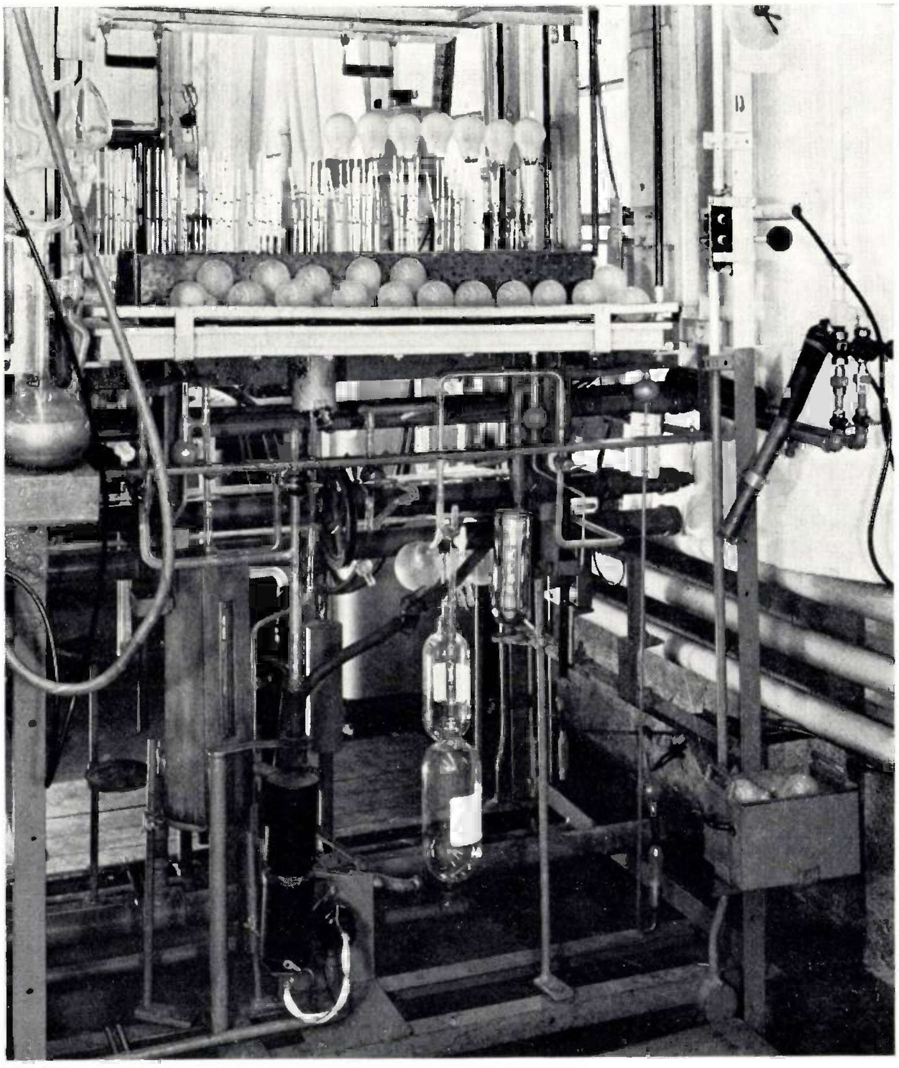
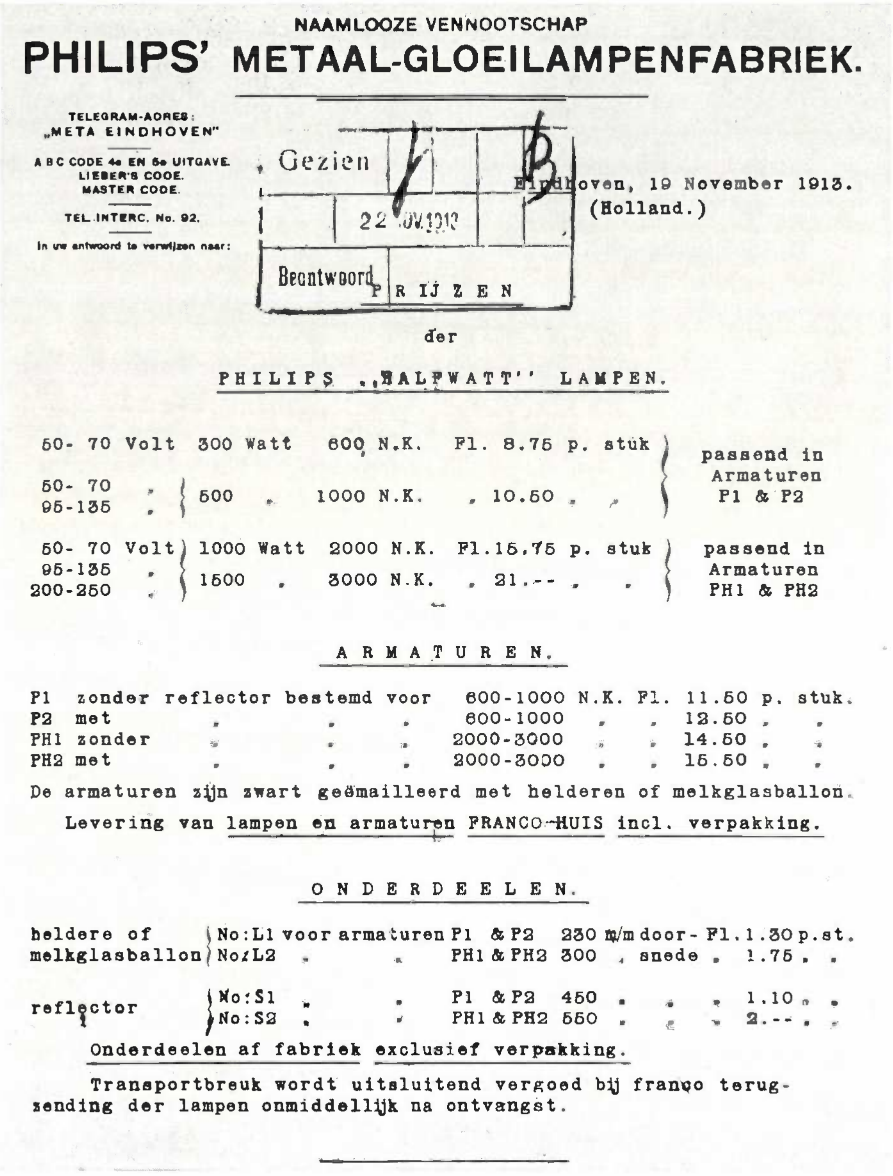


Fig. 7. When evacuating the lamp bulbs on the pumps the last traces of water vapour are removed by a liquid-air trap in the pump line (see the Dewar flask roughly in the middle of the photograph).

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]).

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*). 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]).



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.

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

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

[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.

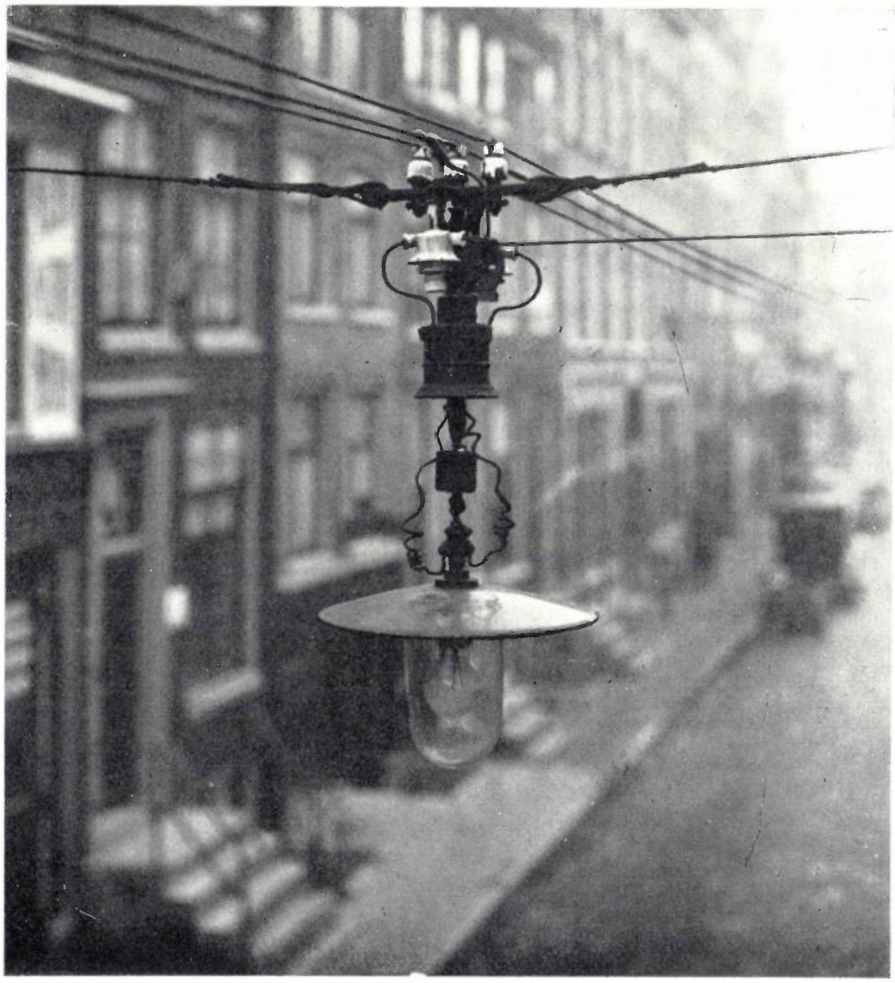
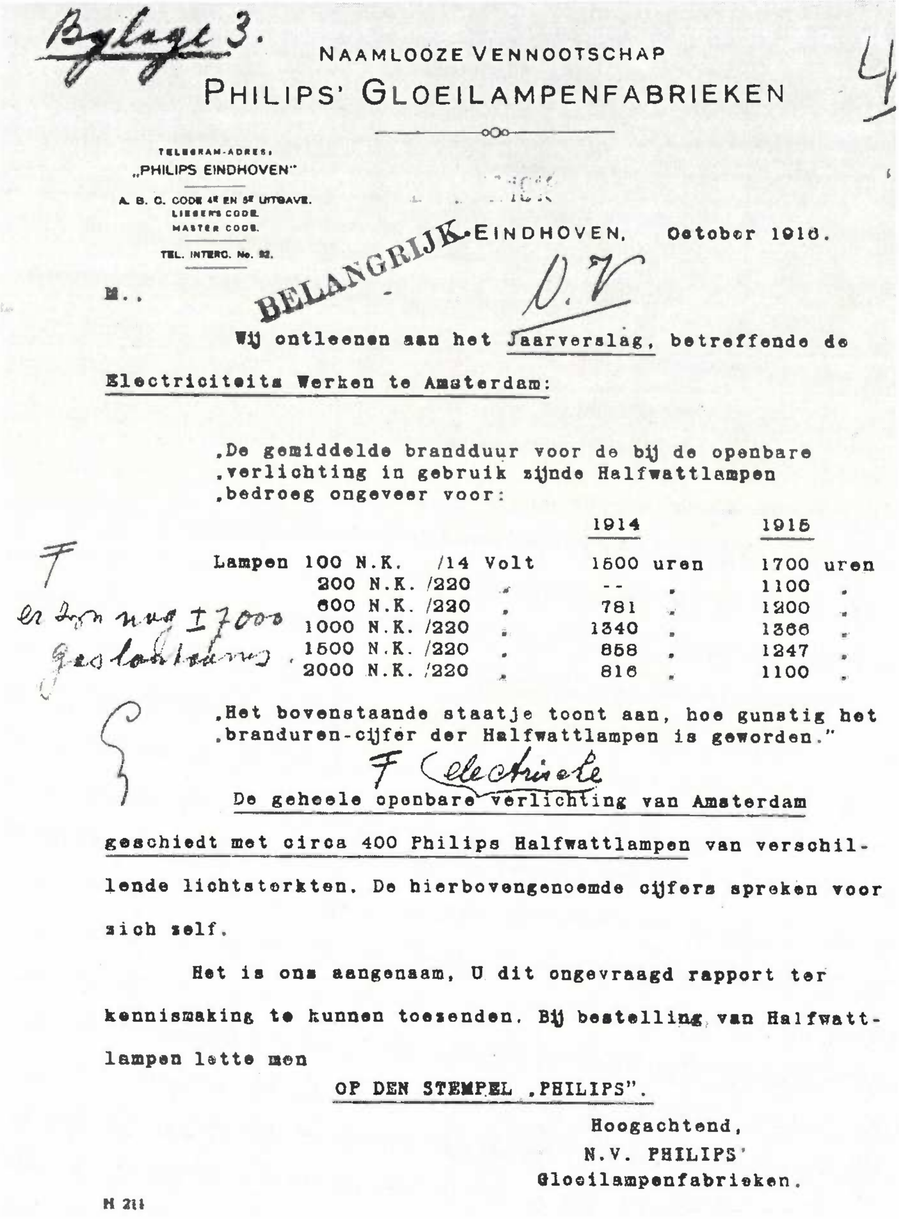


Fig. 9. Fitting containing a 100-candle gas-filled lamp, part of the lighting installation in an area of the Kerkstraat in Amsterdam in 1914. Each lamp required its own step-down transformer for 220/14 V. (Photograph by courtesy of the editor of "De Koppeling"; see also that journal, vol. **8**, 149, 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 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 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 even more unfavourably with argon than was predicted in the theory (see the values of a in Table I). The use of argon therefore considerably improved the heat balance.



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. 82.

BELANGRIJK

EINDHOVEN, October 1916.

O.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:

F

er zijn nog ± 7000 Gaslantaarns.

			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	„

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

F Electrische

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".

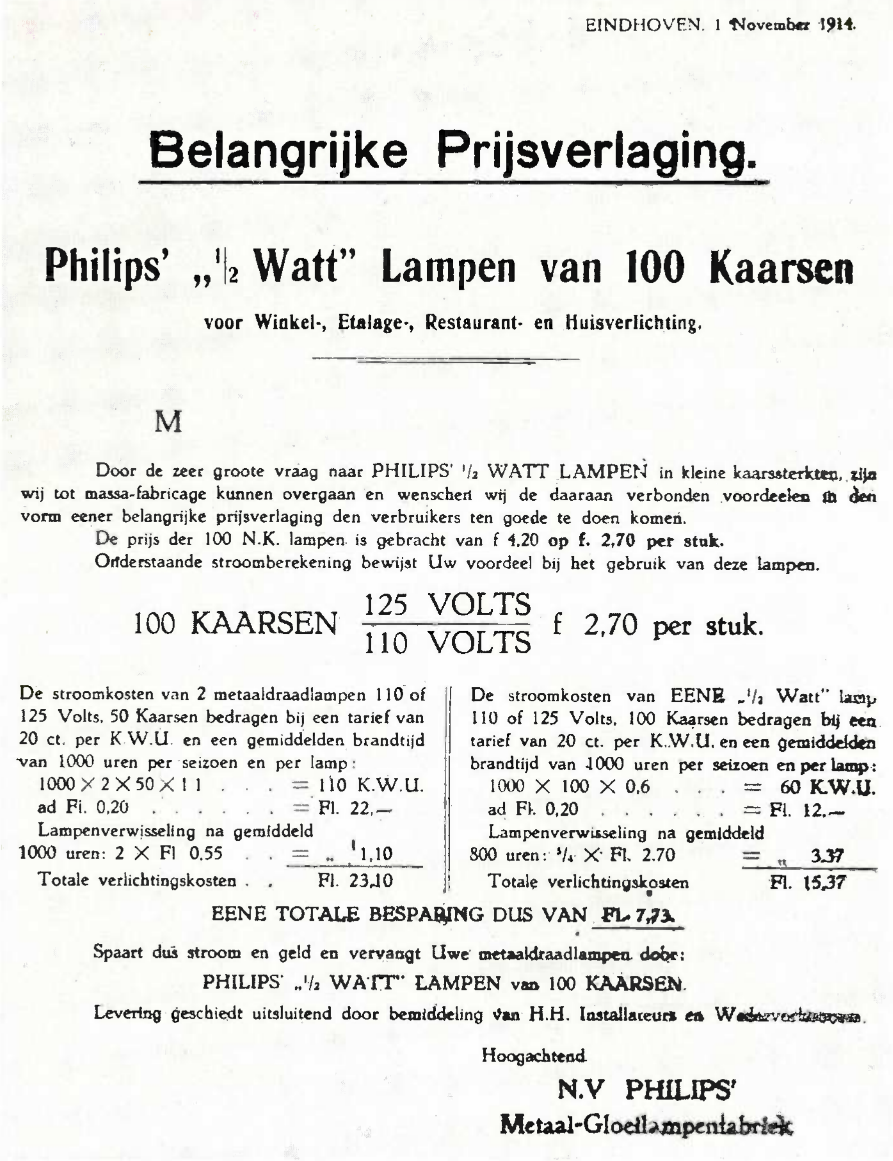
Hoogachtend,
N.V. PHILIPS'
Gloeilampenfabrieken.

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.

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' „½ Watt" Lampen van 100 Kaarsen

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

M

Door de zeer groote vraag naar PHILIPS' ½ 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 $\frac{\text{125 VOLTS}}{\text{110 VOLTS}}$ 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 „½ 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: 5/4 × 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' „½ 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

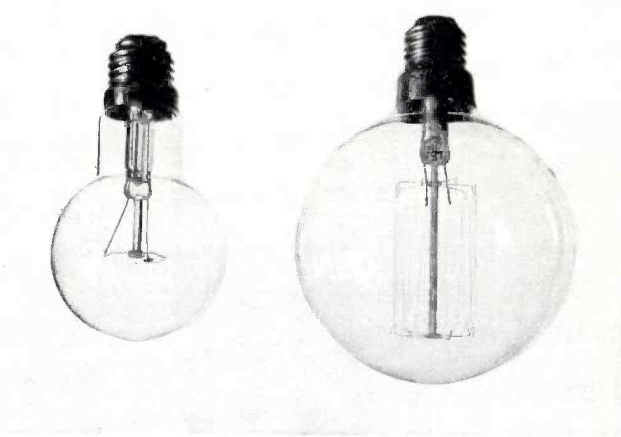


Fig. 12. The compact form of the filament and the favourable distribution of the gradually forming black tungsten deposits made it possible to produce gas-filled lamps (left) with a much smaller bulb than vacuum lamps of the same power.

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 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.

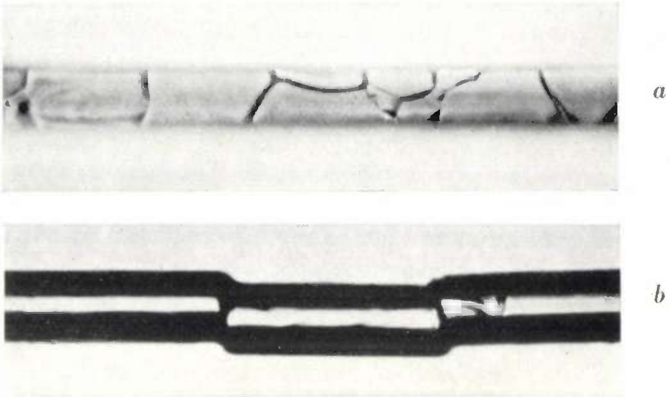


Fig. 13. *a*) Structure of a wire of pure tungsten after recrystallization.
b) "Offsetting" in a tungsten filament, occurring when a lamp having a filament with the structure shown in (*a*) had burnt for some time.

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

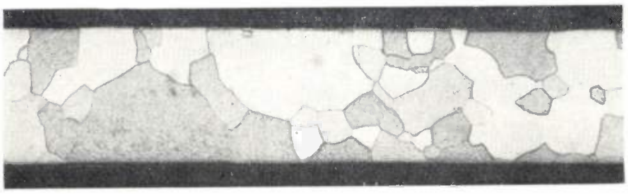


Fig. 14. Recrystallization texture of a tungsten wire containing impurities resulting from the Battersea process.

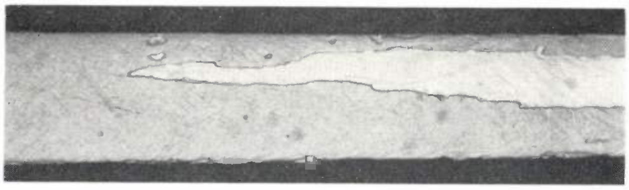


Fig. 15. Recrystallization texture of "doped tungsten", i.e. tungsten containing controlled amounts of additives. The elongated, wedge-shaped crystals give the wire considerable strength and prevent sagging when the wire is coiled.

Many theories have been put forward to explain the action of the additives ("dope") — N.B. added *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 improvement of luminous efficiency particularly at lower wattages [13]). The coiled-coil lamp for low watt-

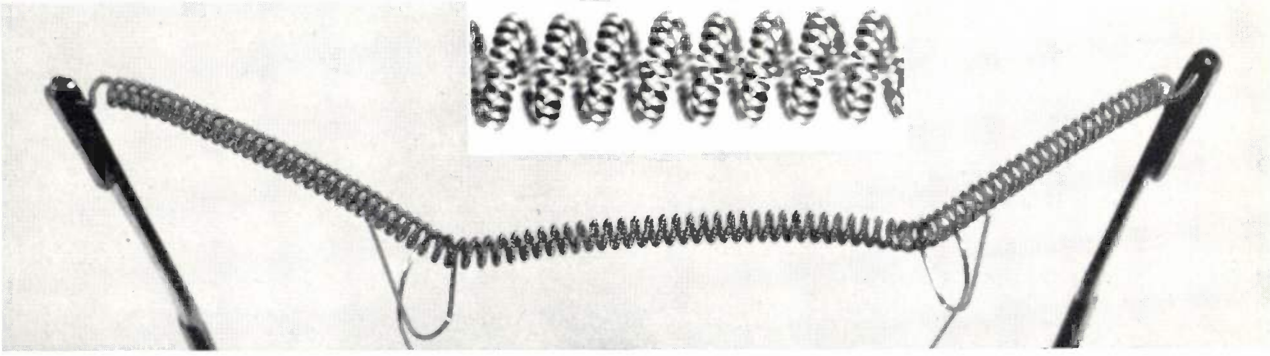


Fig. 16. Coiled-coil filament for a 100 W lamp. The inset shows a piece of the coiled-coil at higher magnification.

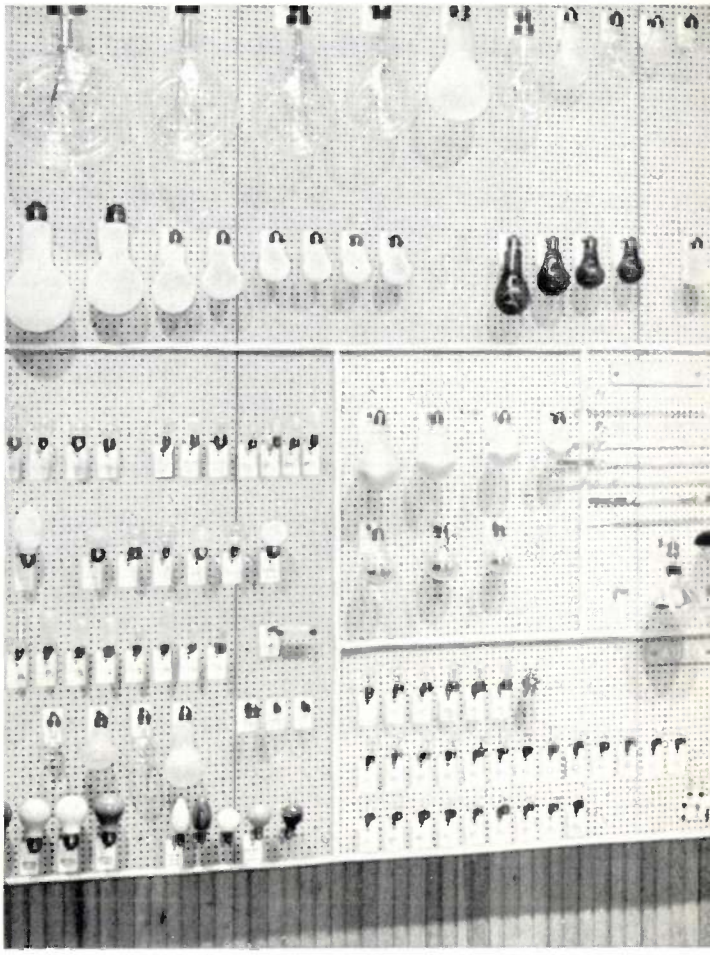
[11]) U.S.A. Patent 1410499, filed Feb. 1917, granted March 1922 in the name of A. Pacz.

[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.



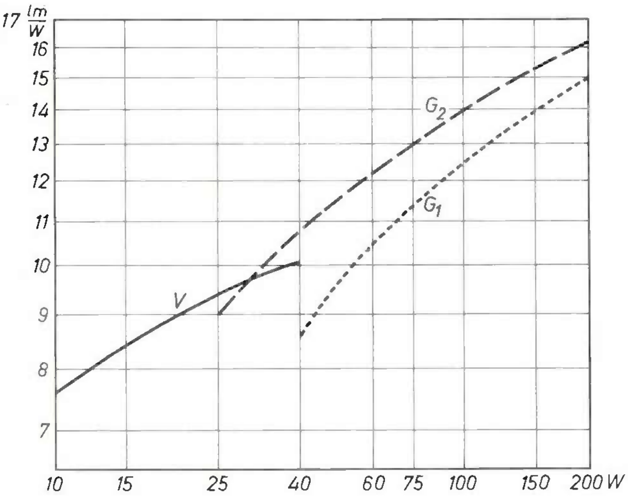


Fig. 17. Luminous efficiency of present-day vacuum lamps (V), of gas-filled single-coil lamps (G_1) and gas-filled coiled-coil lamps (G_2), as a function of wattage.

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.

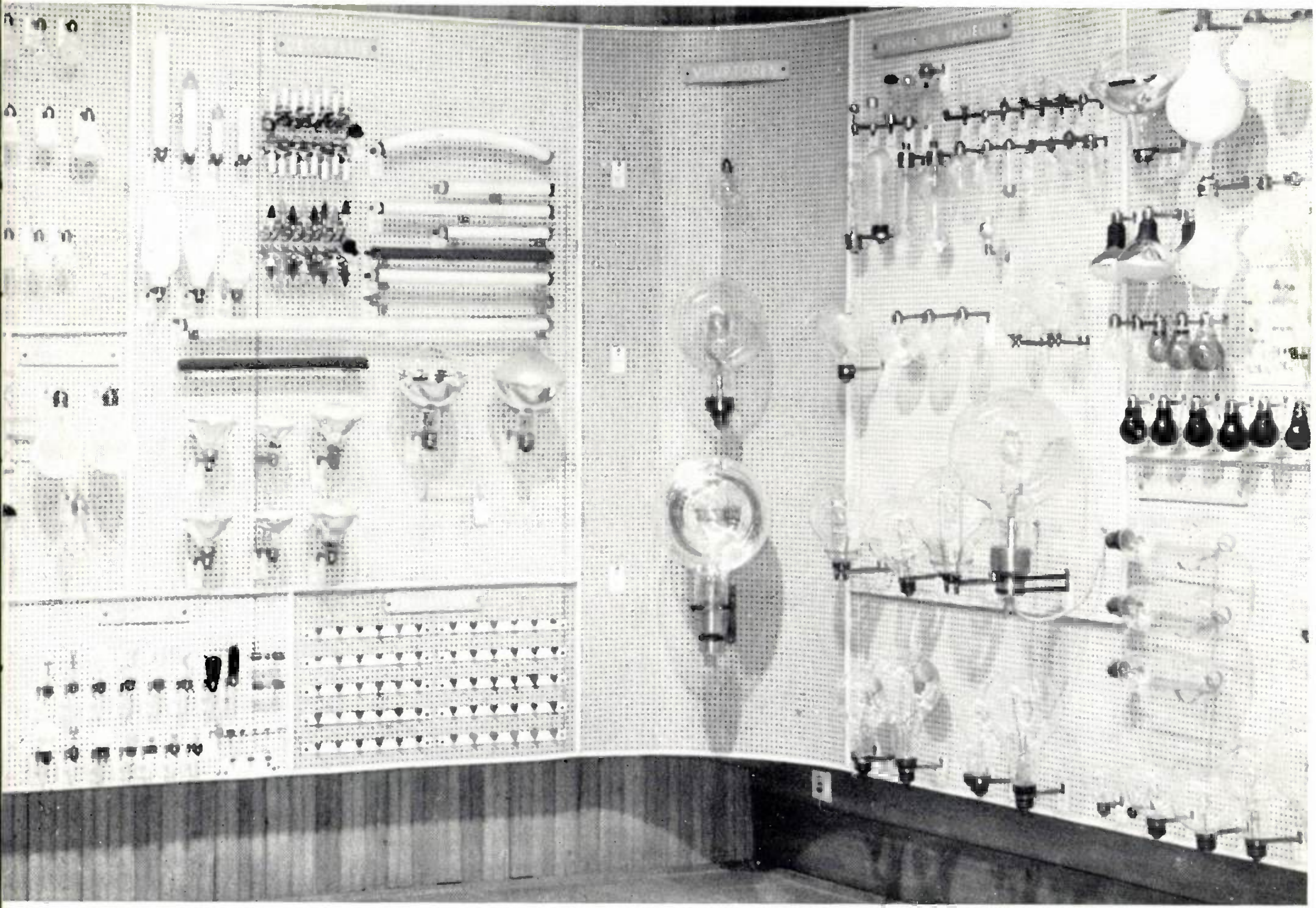


Fig. 18. A small selection from the tens of thousands of types of incandescent lamps nowadays produced by Philips. Most types come into the many categories of "special lamps" (projection lamps, car bulbs, window-display lamps, airfield, sports-field and lighthouse lamps, studio lamps, infrared-heating 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 $\frac{1}{2}$ 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/m^2 (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/m^2. 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/m^2 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).

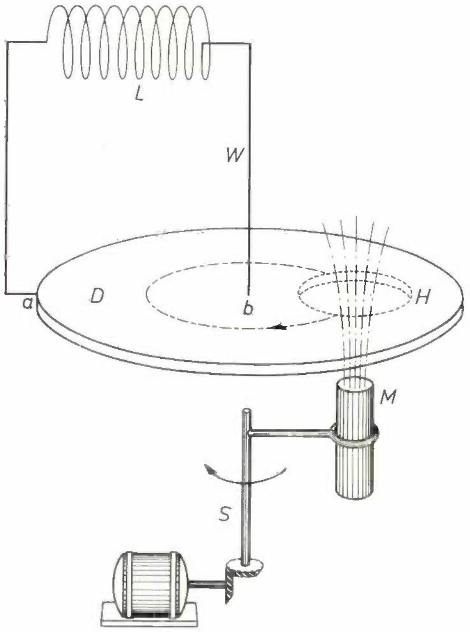


Fig. 1. Schematic diagram of the dynamo for generating a persistent current in a superconducting circuit. *D* lead disc. *W* circuit of "hard" superconducting material, connected to *D* at the rim (*a*) and in the centre (*b*). The shaft *S* mounted below the centre of *D* carries on an arm the bar magnet *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

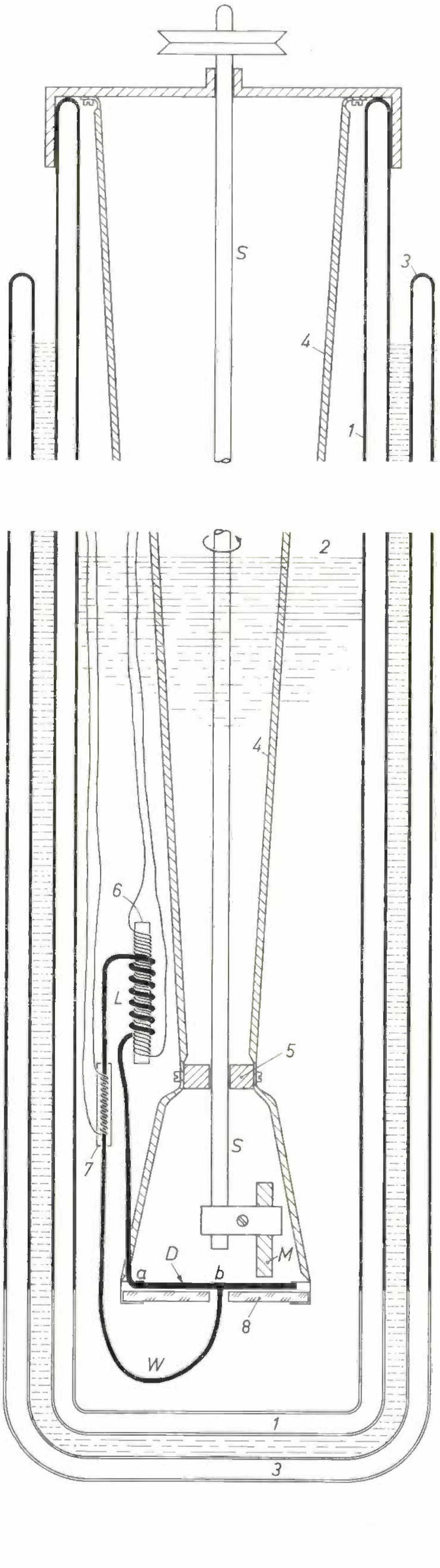


Fig. 2. Sketch showing the construction of the new dynamo and its positioning in a cryostat. The letters have the same meaning as in fig. 1. Other symbols are: *1* Dewar vessel of inner cryostat. *2* surface of the liquid helium. *3* Dewar vessel of outer cryostat; the space between *1* and *3* is filled 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).

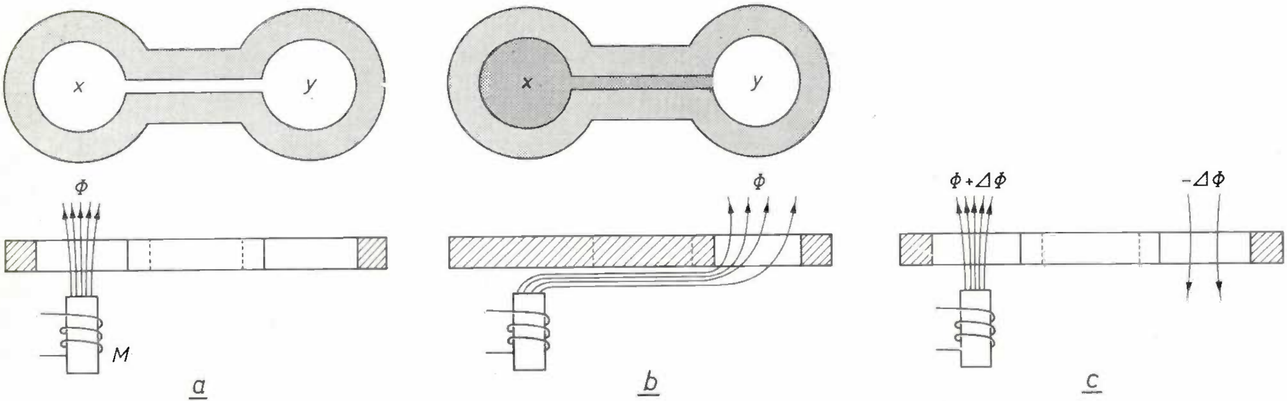


Fig. 3. Illustrating the properties of a superconducting ring circuit of complicated shape, in one part of which a flux is enclosed.

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.

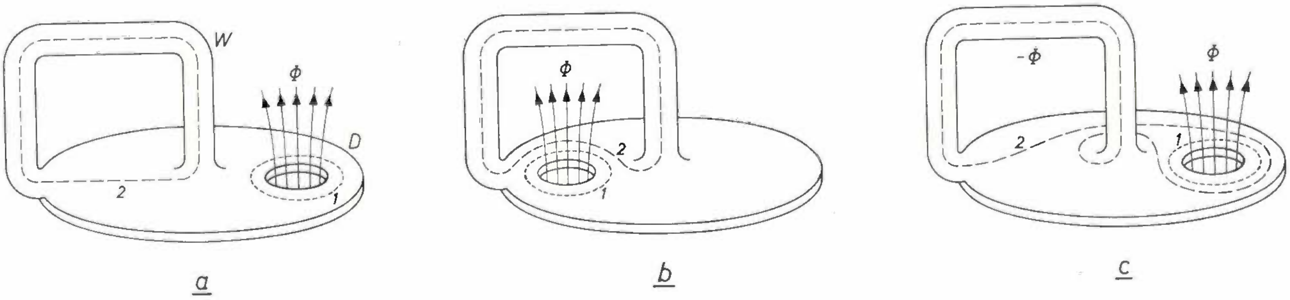


Fig. 4. Illustrating the operation of the new dynamo.

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.

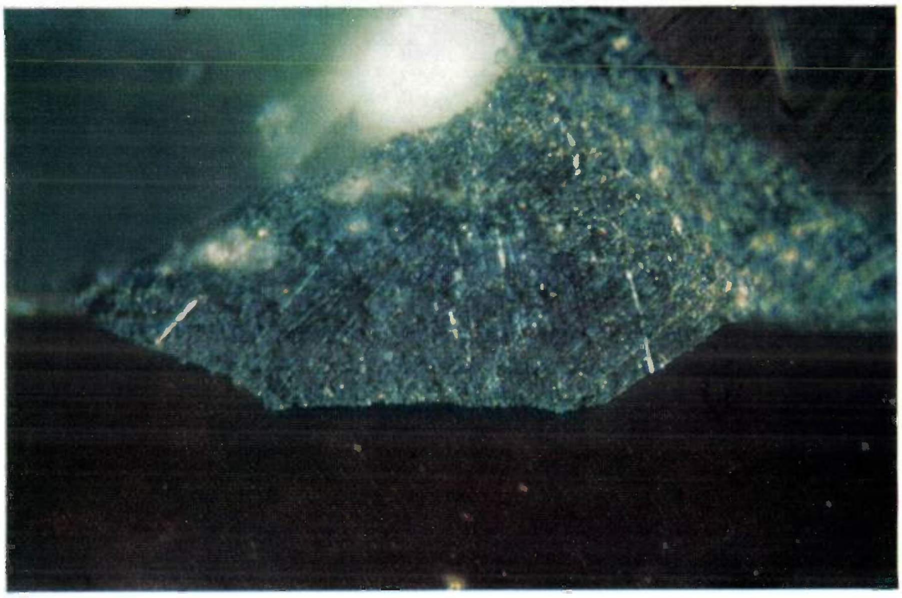


Fig. 1. Cross-section of a *P-N* junction in GaP with Au contact, photographed in polarized light on Kodachrome film (artificial-light reversal film), exposure 30 seconds. The *P-N* junction is roughly 0.5 mm long.

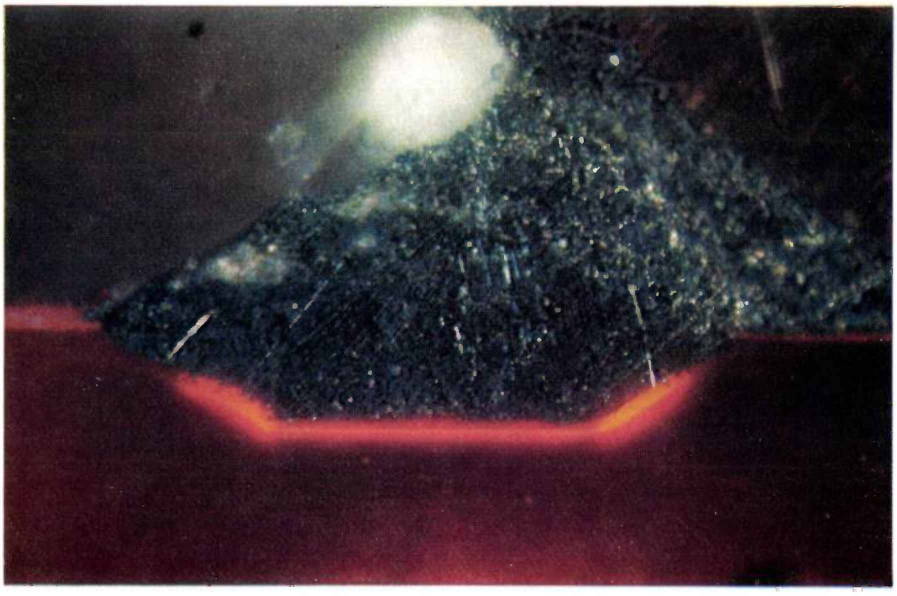


Fig. 2. The same *P-N* junction with forward current flowing (10 mA). Exposure 30 seconds.

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_0	= 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 . . .	$\Delta U_0/U_0$	= 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.

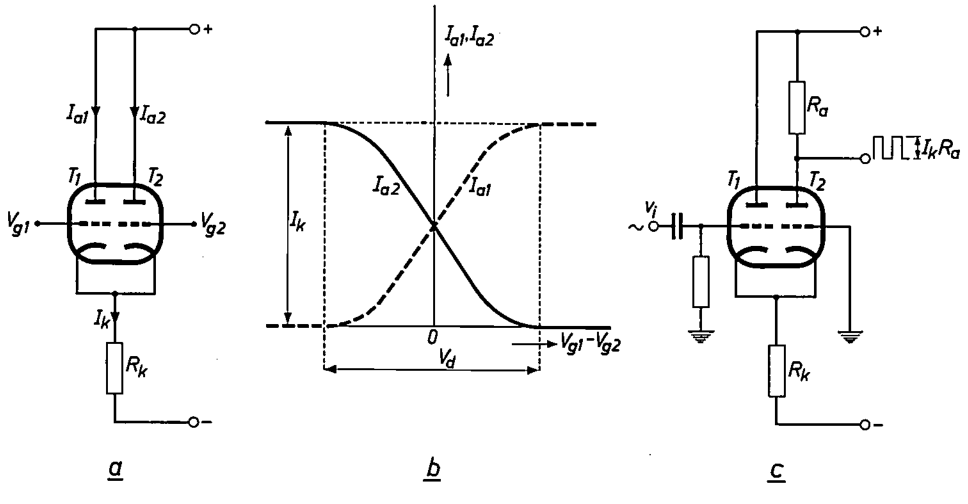


Fig. 1. *a*) Balanced stage with high resistance — here an ordinary resistance R_k — in the common cathode lead ("long-tailed pair").
b) The anode currents I_{a1} and I_{a2} of the triodes T_1 and T_2 in (*a*) are plotted versus the difference $V_{g1} - V_{g2}$ of the grid voltages. The difference V_d 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.

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_kR_a, and is thus independent of the amplitude of v_i. By stabilizing the supply voltages and using 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

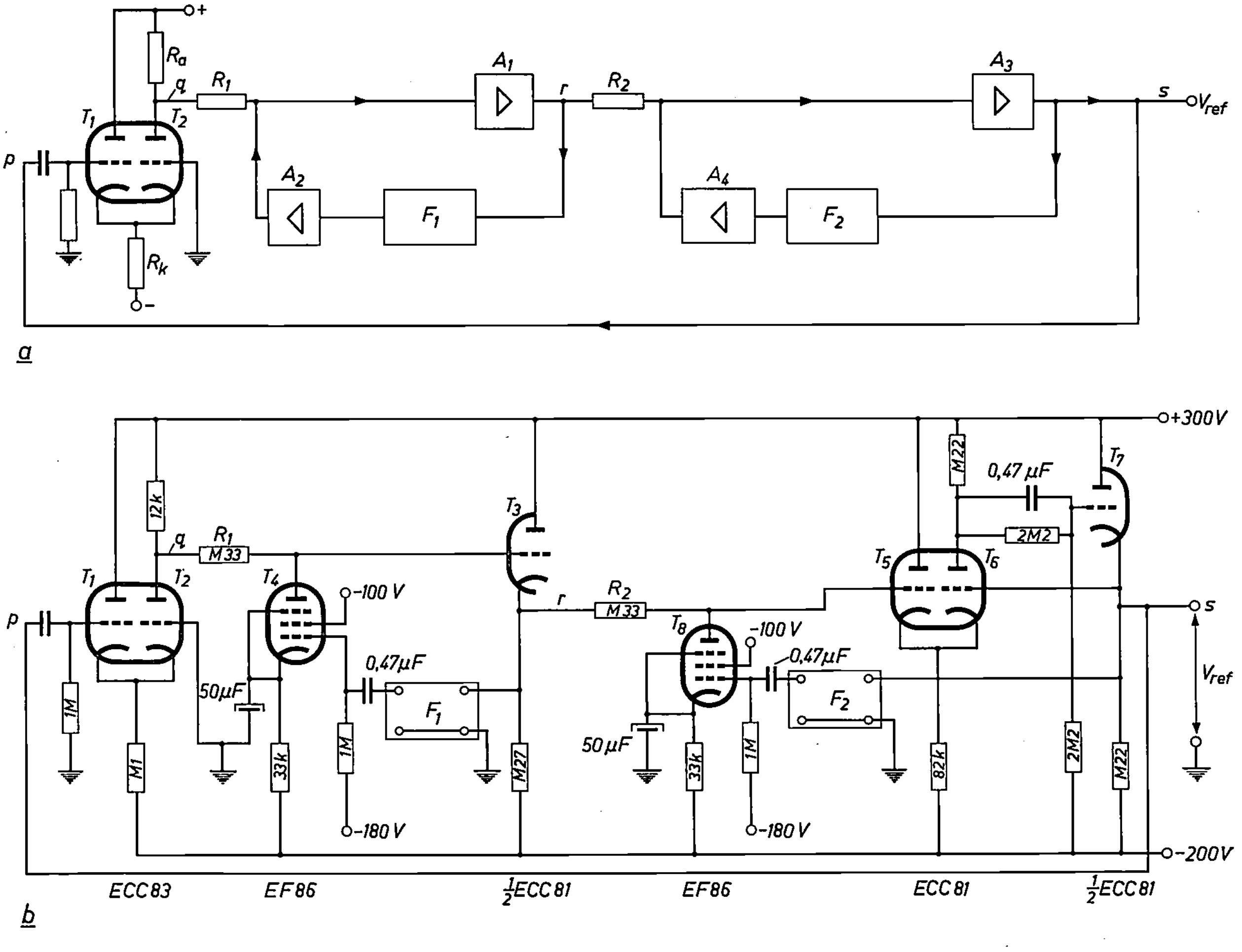


Fig. 2. *a*) Block diagram of the reference oscillator. T_1-T_2, R_k and R_a correspond to fig. 1c. At point *q* there is a constant square-wave voltage (frequency 80 c/s), the fundamental component of which passes through the amplifiers (cathode followers) A_1 and A_3. For the higher harmonics in the 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:

$$i_a = \alpha v + \beta v^2 + \gamma v^3 + \dots, \quad . \quad . \quad . \quad . \quad (1)$$

[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.

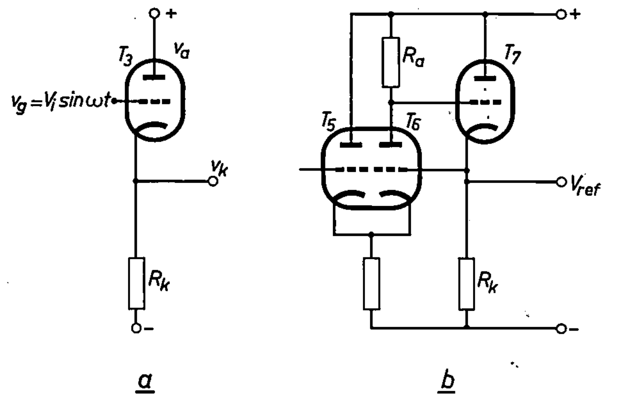


Fig. 3. *a*) Single cathode follower (like T_3 in fig. 2*b*): triode T_3 with cathode resistance R_k.
b) complex cathode follower (like T_5-T_6-T_7 in fig. 2*b*). The inherent distortion here is much lower than in the single type.

where

$$v = v_\mathrm{g} + \frac{v_\mathrm{a}}{\mu} - \left(1 + \frac{1}{\mu}\right) v_\mathrm{k} \, . \quad . \; . \; . \; . \; . \quad (2)$$

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:

$$v = a\, i_\mathrm{a} + b\, i_\mathrm{a}^2 + c\, i_\mathrm{a}^3 + \ldots, \quad . \; . \; . \; . \; . \quad (3)$$

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*), $v_\mathrm{a} = 0$ and $v_\mathrm{k} = i_\mathrm{a} R_\mathrm{k}$. Assuming $(1 + \mu^{-1})R_\mathrm{k} = r$, we find from (2) and (3) for the single cathode follower:

$$v_\mathrm{g} = (a + r) i_\mathrm{a} + b\, i_\mathrm{a}^2 + c\, i_\mathrm{a}^3 + \ldots . \quad . \; . \; . \; . \quad (5)$$

If we expand i_a in a power series to v_g:

$$i_\mathrm{a} = k_1 v_\mathrm{g} + k_2 v_\mathrm{g}^2 + k_3 v_\mathrm{g}^3 + \ldots, \quad . \; . \; . \; . \quad (6)$$

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, $v_\mathrm{g} = V_0 \cos \omega t$, this current i_a is given by the following series:

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

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 $v_\mathrm{k} = i_\mathrm{a} R_\mathrm{k}$):

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

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

$$\left(\frac{1}{\mu'} + \frac{2}{SR_\mathrm{a}}\right)^{-1} \gg 1 \, . \quad . \; . \; . \; . \; . \; . \; . \quad (9)$$

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

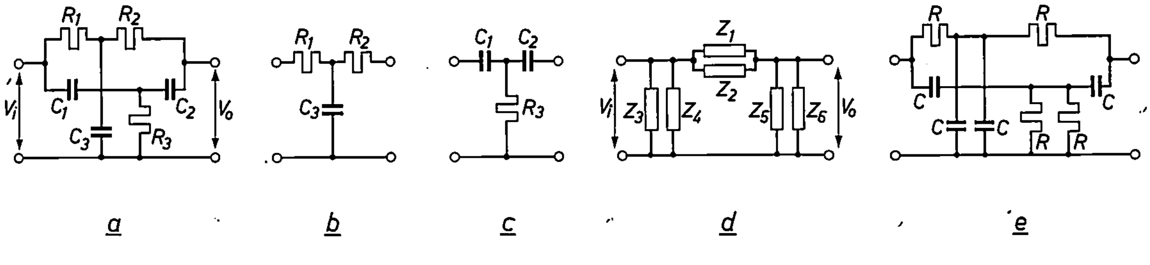


Fig. 4. *a*) Twin-T filter, consisting of resistors and capacitors. *b*) and *c*) The two T sections from which (*a*) is formed. *d*) The two T sections of (*a*) are transformed into *Π* sections. *e*) In the case of symmetry the network (*a*) can be formed from four identical resistors *R* and four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

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

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 $Z_1 \dots Z_6$ 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 $1/A_\mathrm{F}$ of the input to the output voltage of the filter we read from fig. 4*d*:

$$\frac{1}{A_\mathrm{F}} = \frac{V_\mathrm{i}}{V_\mathrm{o}} = 1 + \frac{Z_1Z_2(Z_5+Z_6)}{(Z_1+Z_2)Z_5Z_6}\,.$$

From this formula it is seen that the only case in which the filter blocks the signal completely ($A_\mathrm{F} = 0$, $V_\mathrm{i}/V_\mathrm{o} = \infty$) 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:

$$A_\mathrm{F} = \frac{V_\mathrm{o}}{V_\mathrm{i}} = \frac{\mathrm{j}Q\beta}{1+\mathrm{j}Q\beta}\,, \quad . \; . \; . \; . \quad (13)$$

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):

$$A_\mathrm{act} = \frac{V_\mathrm{r}}{V_\mathrm{q}} = B\frac{1+\mathrm{j}Q\beta}{1+\mathrm{j}(A_2+1)Q\beta}\,. \quad . \quad (14)$$

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 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):

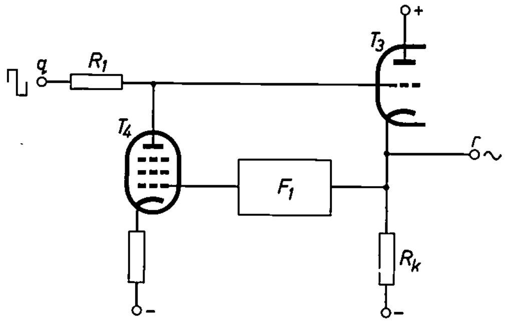


Fig. 5. The "passive" filter F_1 and the amplifiers A_1 and A_2 (fig. 2*a*) together form an active filter, with input terminal *q* and output terminal *r*. The fundamental component of the square-wave voltage at *q* is passed. The higher harmonics are considerably attenuated owing to the high gain of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

$$\begin{array}{llllll}
f = f_0 & 2f_0 & 3f_0 & 4f_0 & \dots & \infty, \\
\beta = 0 & 1\frac{1}{2} & 2\frac{2}{3} & 3\frac{3}{4} & \dots & \infty, \\
\left|\dfrac{A_\mathrm{act}}{B}\right| = 1 & \dfrac{2.85}{A_2} & \dfrac{1.80}{A_2} & \dfrac{1.46}{A_2} & \dots & \dfrac{1}{A_2}\,.
\end{array}$$

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	$= 2.5 \times 10^{-5}$,
third harmonic	d_3	$= 1.5 \times 10^{-5}$,
remaining harmonics	d_{rest}	$< 0.4 \times 10^{-5}$.

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 $kU_0 - V_{\text{ref}}$ 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 $kU_0 - V_{\text{ref}}$. 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.

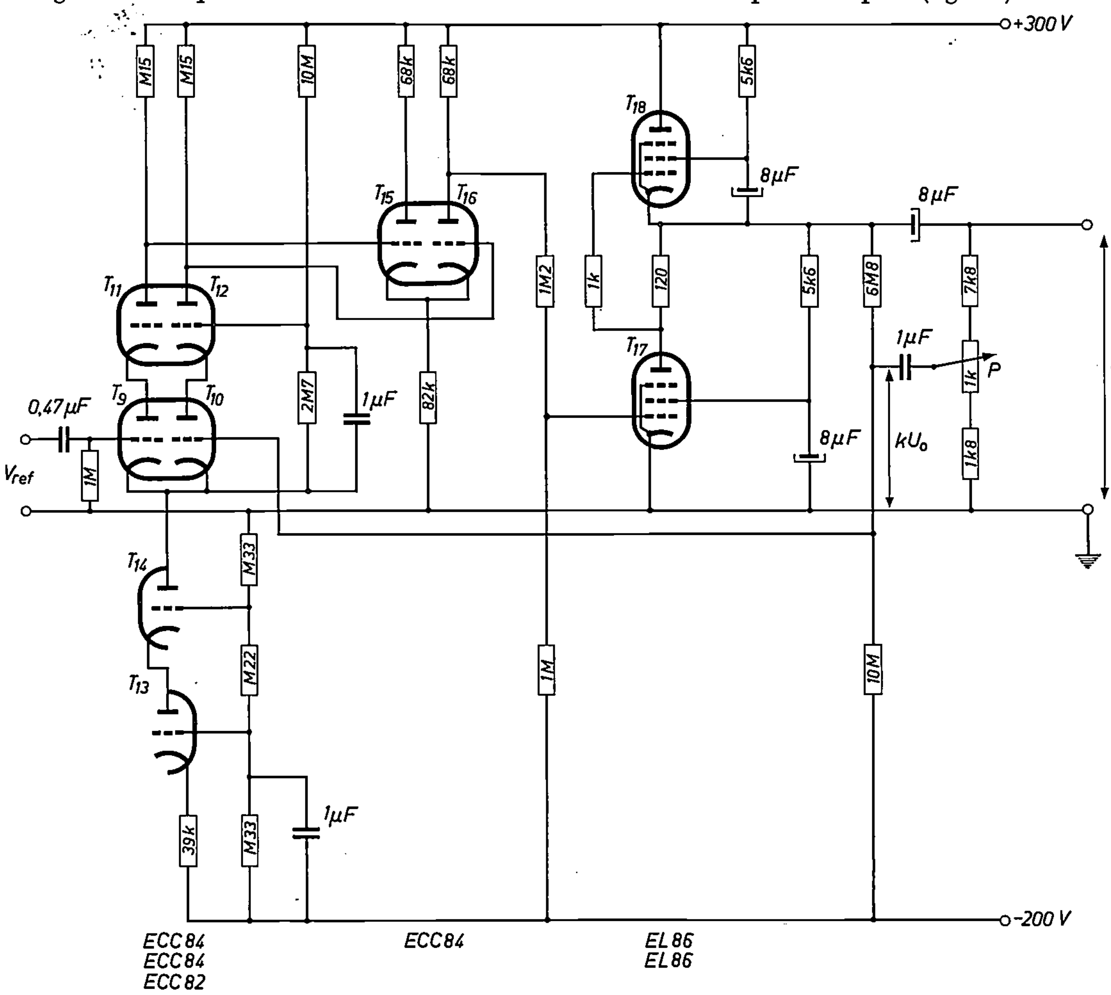


Fig. 6. Diagram of the output stage with two-stage driving circuit.
Driving circuit. First stage: difference amplifier with high rejection and high discrimination factors, consisting of cascodes T_9-T_{11} and T_{10}-T_{12} in push-pull with the cascode T_{13}-T_{14} in the common cathode lead. Second 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 $V_{\rm ref}$ 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 . .	$0.56 \times 10^{-5} \times U_0$.

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 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 $R_{\rm i}$ 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 measurements 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%.

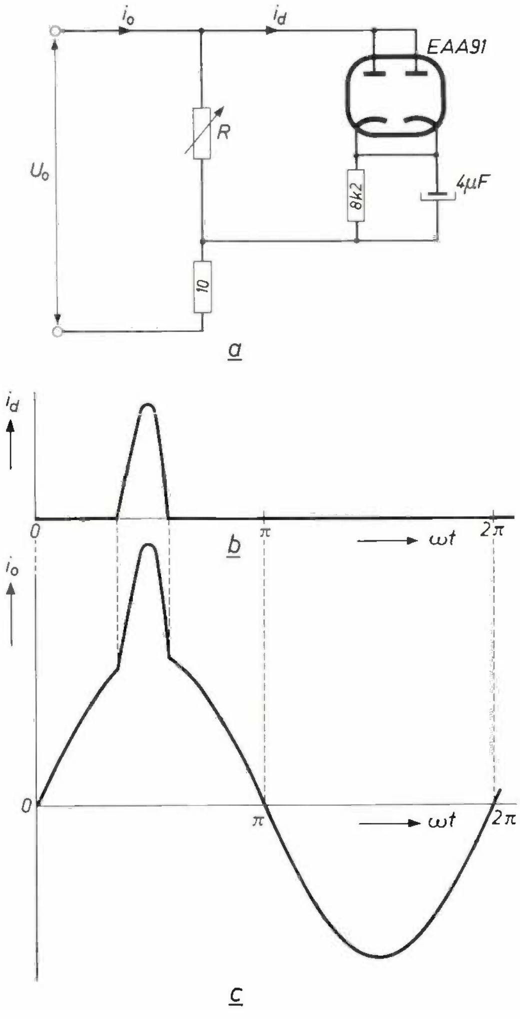


Fig. 7. *a*) Circuit used as non-linear load for distortion measurements. The load current i_o consists of a linear component, which flows through the variable resistor R, and a non-linear component i_d, which, during a small part of each period, charges up a capacitor of 4 µF through the diode EAA 91. 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.

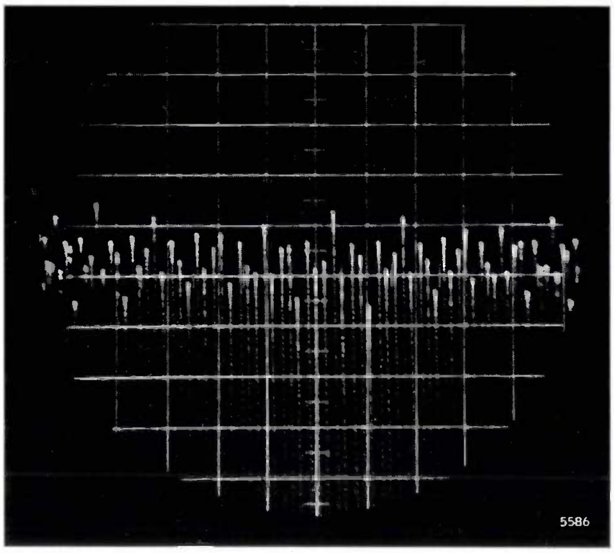


Fig. 8. Oscillogram obtained with a "voltage microscope" [5]), showing the peaks of the output voltage (amplitude approx. 70 V). The height of each square corresponds to about 14 mV (the zero line of the sine wave should therefore be imagined at about 35 metres below the peaks!). The maximum fluctuations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

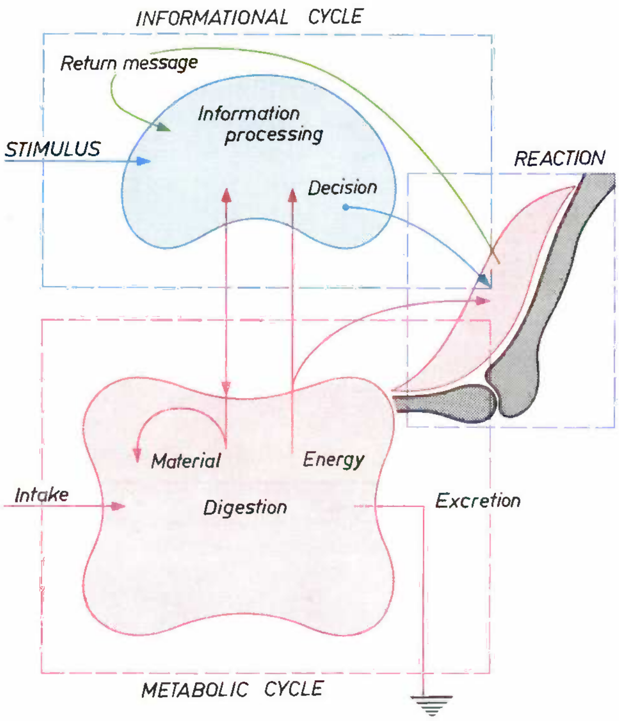


Fig. 1. Illustrating the reaction of an organism (human or animal) in connection with the informational cycle and the metabolic cycle.

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.).

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 coöperation 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 frequencv 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

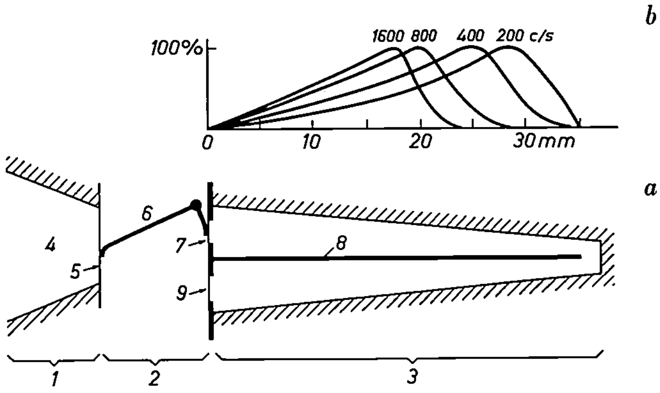


Fig. 1. *a*) Schematic representation of the human ear. From left to right the external ear (*1*), the middle ear (*2*) and the inner ear (*3*). The air vibrations pass down the ear canal *4* to the tympanic membrane *5*. The movement of the latter is transmitted via the middle ear ossicles *6* — represented 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. 3*a*) 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.

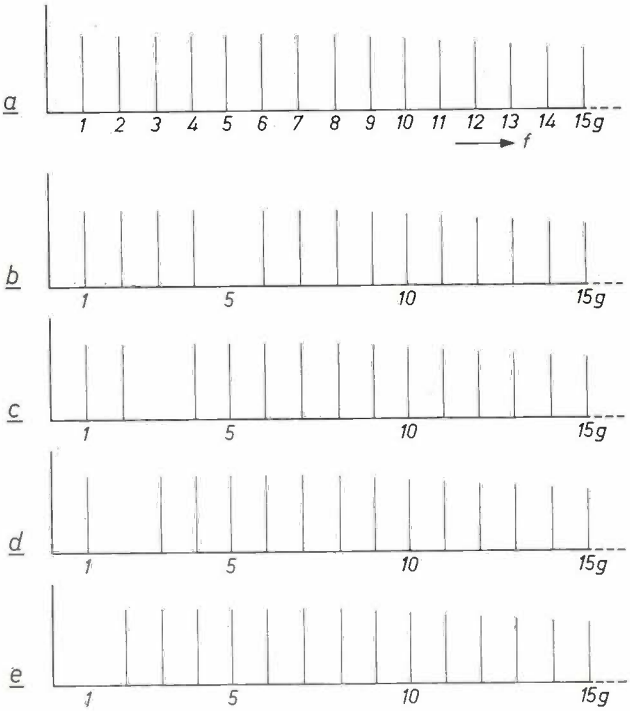


Fig. 2. *a*) Spectrum of a tone pulse, i.e. a tone produced by a periodic pulse of very short duration. The intensity of the spectral lines is plotted versus their frequency (*f*) expressed as a multiple of the fundamental frequency *g*. The spectrum consists of numerous harmonics of roughly equal strength.
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.

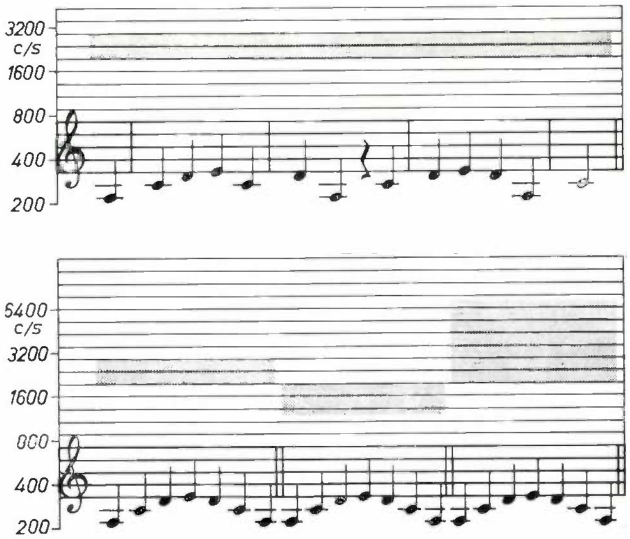


Fig. 3. *a*) When a melody is made of tone pulses (repetition frequency 200-300 c/s) of which only the harmonics between 2000 and 3000 c/s are passed, the pitch still corresponds to that of the scored notes. The range of harmonics passed is represented by the shaded strip in the lines above the staff. (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.

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

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 $V_\mathrm{i}(t)$ and the output signal $V_\mathrm{u}(t)$ may in general be formulated as:

$$V_\mathrm{u}(t) = C\,[V_\mathrm{i}(t) + \delta_1 V_\mathrm{i}^2(t) + \dots] \quad . \; . \; . \; . \; . \; . \qquad (1)$$

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

$$V_\mathrm{i}(t) = A\,(\sin \omega_1 t + \sin \omega_2 t), \quad . \; . \; . \; . \; . \; . \; . \qquad (2)$$

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

$$V_\mathrm{u}(t) = V_\mathrm{i}(t) + \delta_1 A^2\,[1 - \tfrac{1}{2}\cos 2\omega_1 t - \tfrac{1}{2}\cos 2\omega_2 t + \\ + \cos(\omega_1 + \omega_2)t + \cos(\omega_1 - \omega_2)t] + \dots\;.$$

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.* IV
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.

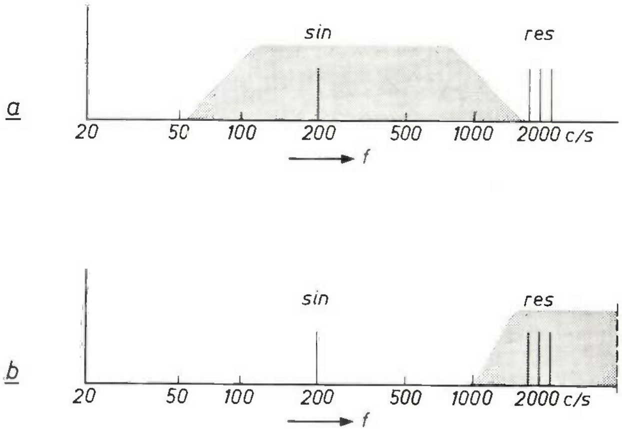


Fig. 4. Masking by noise shows that the tonal residue does not arise as a difference tone in the ear. If the noise spectrum is chosen to correspond roughly with the shaded part in (*a*), a pure tone of 200 c/s (*sin*) disappears, but not the tonal residue, which is heard to have the same pitch and originates 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 ($f_\mathrm{p} = 200$ c/s). Next, we raise all frequencies by the same amount, e.g. 10 c/s, and repeat this a few t mes 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. V

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.

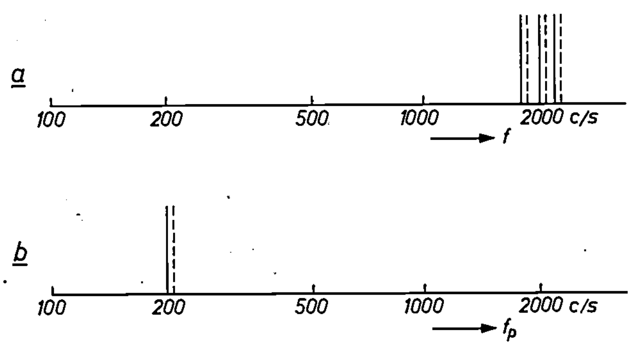


Fig. 5. When the components of a complex sound (*a*) are equidistantly shifted, the pitch of the residue (which is the frequency f_p of a pure tone of equal pitch, shown in *b*), does not remain constant but changes proportionally with the frequency f of the middle component.

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 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.

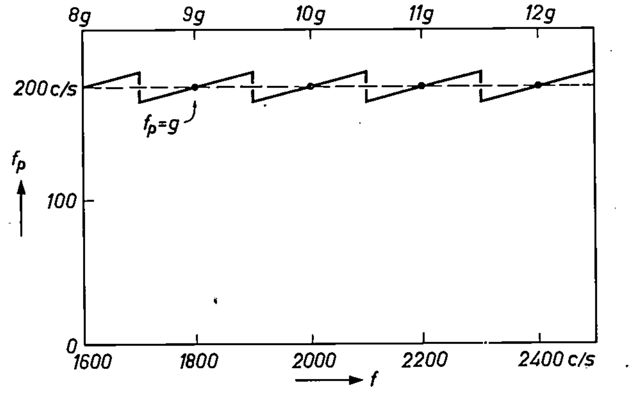


Fig. 6. When the three components of a complex sound (middle frequency f, frequency difference g) are equidistantly shifted, f_p does not rise continuously with f. When f has covered roughly half the distance to the next multiple of g, f_p makes a jump so that, when that multiple is reached, f_p is again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

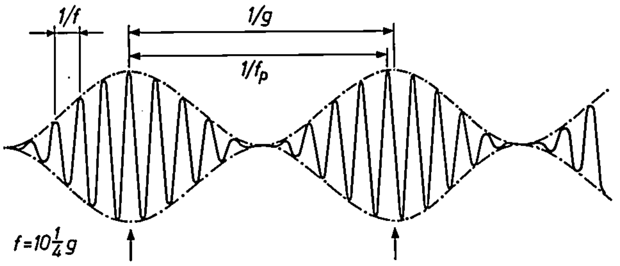


Fig. 7. Explanation of the effect that, when a complex of three components is equidistantly shifted, the pitch of the residue f_p changes proportionally with the middle frequency f. The sound signal is represented by the solid line (sinusoidal vibration of frequency f, modulated 100% by the frequency 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.

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.

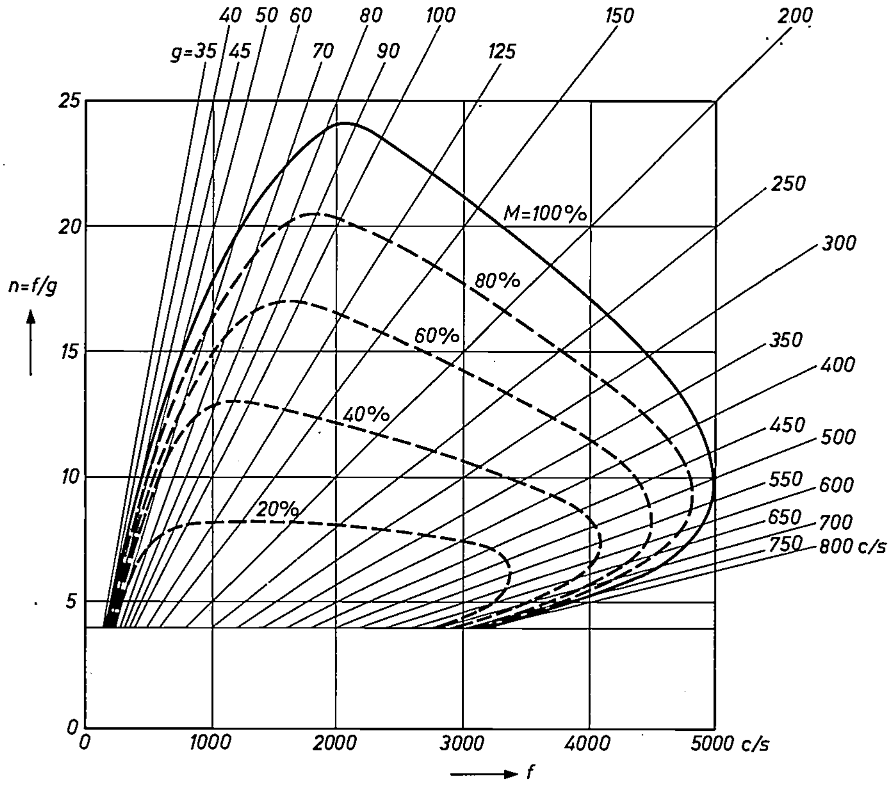


Fig. 8. Existence region of the tonal residue of a complex sound consisting of three neighbouring harmonics (middle frequency f, fundamental frequency g): the tonal residue is heard in the region bounded by the solid contour. If the two outer components are attenuated (experimentally by modulating f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

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-

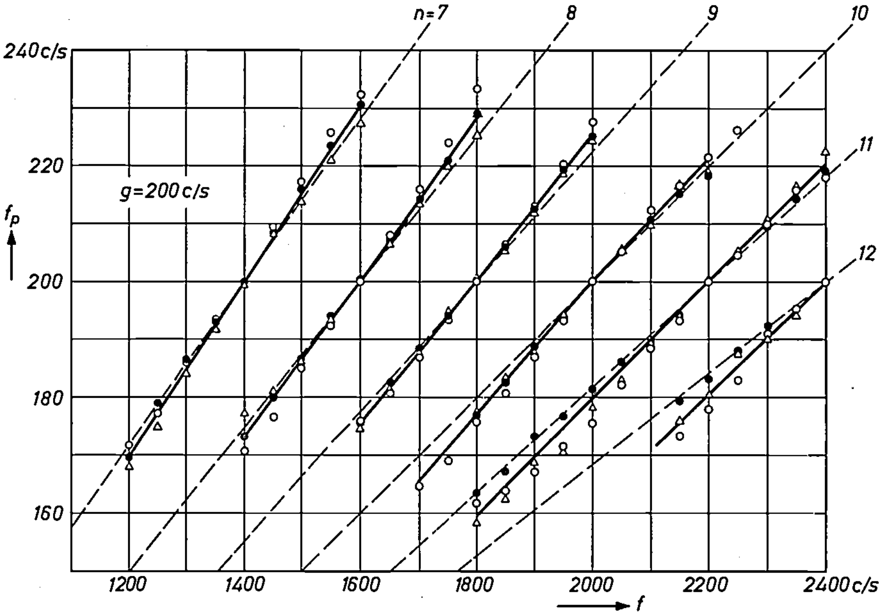


Fig. 9. The change in the pitch f_p of a complex sound of three components when the middle frequency f is varied can, if the listener concentrates on it, be followed for a short time beyond the frequency at which an unsuspecting listener hears a jump in f_p (cf. fig. 6). This means that the pitch of 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,

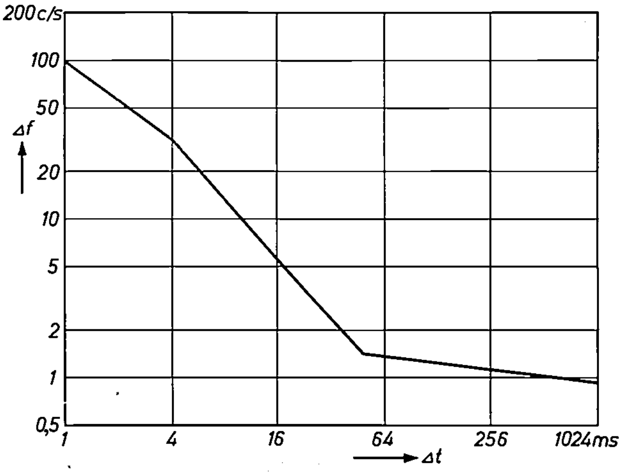


Fig. 10. The distinctness with which a short sinusoidal burst is heard as a tone depends on the duration Δt of the burst. As Δt is decreased the listener becomes more and more uncertain of the pitch. The graph shows the result of experiments in which the subject listened to pairs of tone bursts (1000 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 $\Delta f \Delta t =$ 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 $\Delta f \Delta t =$ 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]).

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

[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.

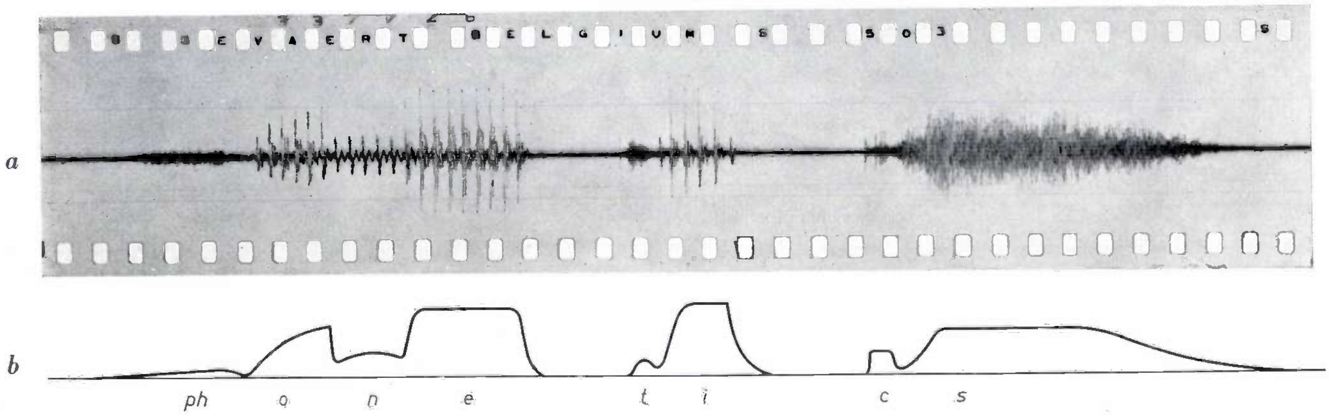


Fig. 1. *a*) Oscillogram of the word *phonetics*. Recorded on a continuously moving film while the oscilloscope time-base generator was switched off.
b) The amplitude envelope of the above pattern (giving the amplitude variation).

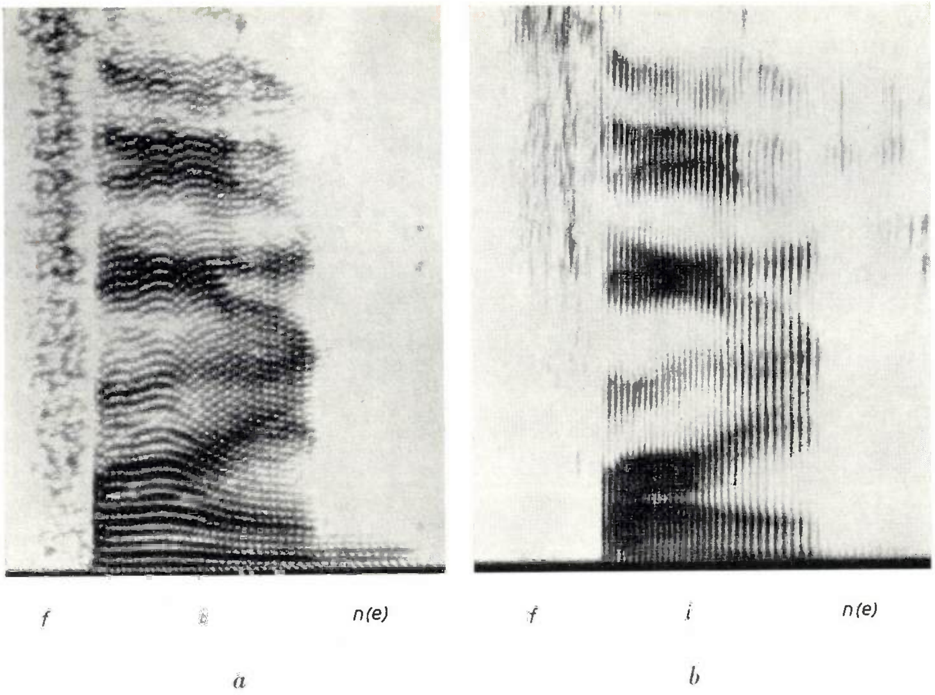


Fig. 2. Spectrograms of the word *fine*, *a*) recorded at high resolving power (approx. 50 c/s), *b*) at low resolving power (approx. 300 c/s). In (*a*) it is clear that *f* has a noise spectrum and the other sounds a line spectrum. The greater the density, the stronger is the spectral line. From (*b*) 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-

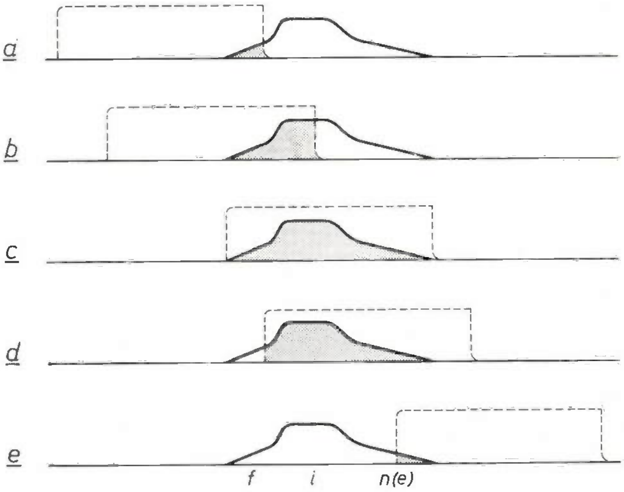


Fig. 3. Illustrating a phonetic analysis of the word *fine*, using a gating circuit. The word is represented by its envelope; the dashed line represents the response curve of the reproduction apparatus. The time interval during which the "gate" is open is shifted in steps of about 10 ms in relation to 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.

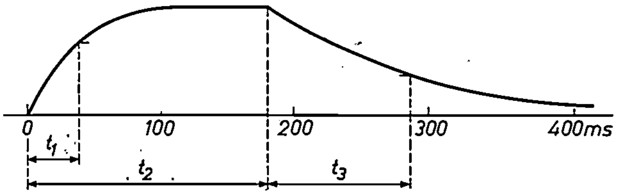


Fig. 4. In the I.P.O.'s method of speech synthesis the amplitude envelope of each sound segment is separately controlled by means of a *variable function gating circuit.* This is a special type of monostable multivibrator (univibrator) which makes it possible to adjust separately the length (t_2) of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

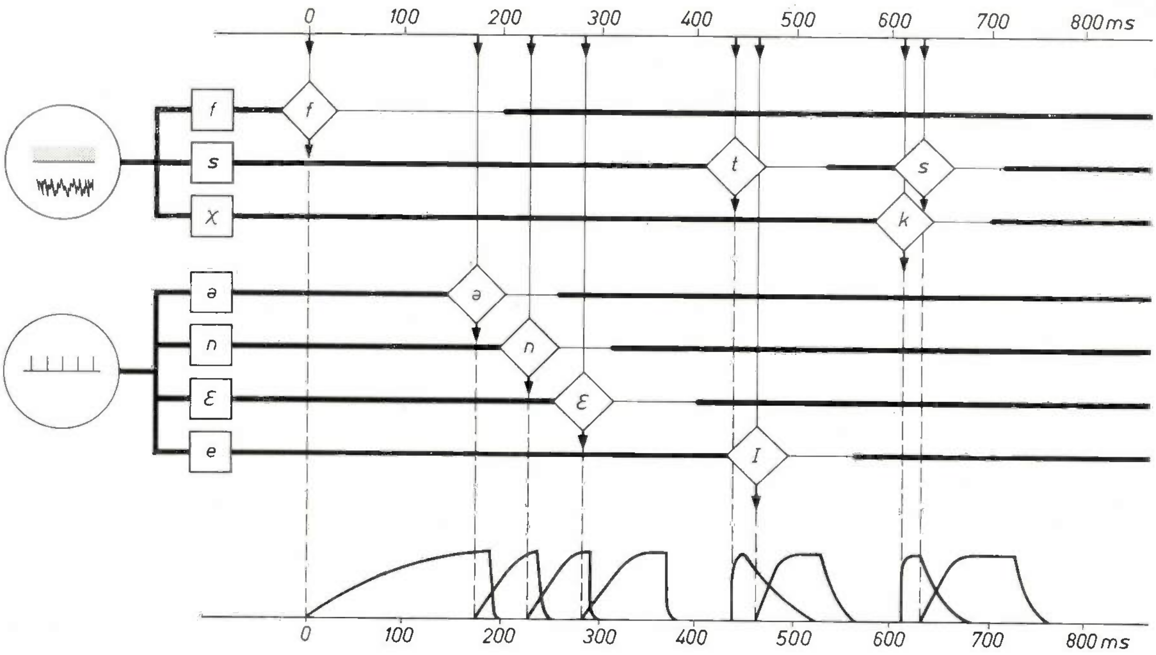


Fig. 5. Schematic illustration of how the word *phonetics* is synthesized. The consonants *f*, *s*, *t* and *k* are obtained by filtering the spectrum of a single noise source; the other sounds are produced by means of a generator of periodic very narrow pulses, of whose spectrum *two* suitably chosen 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]).

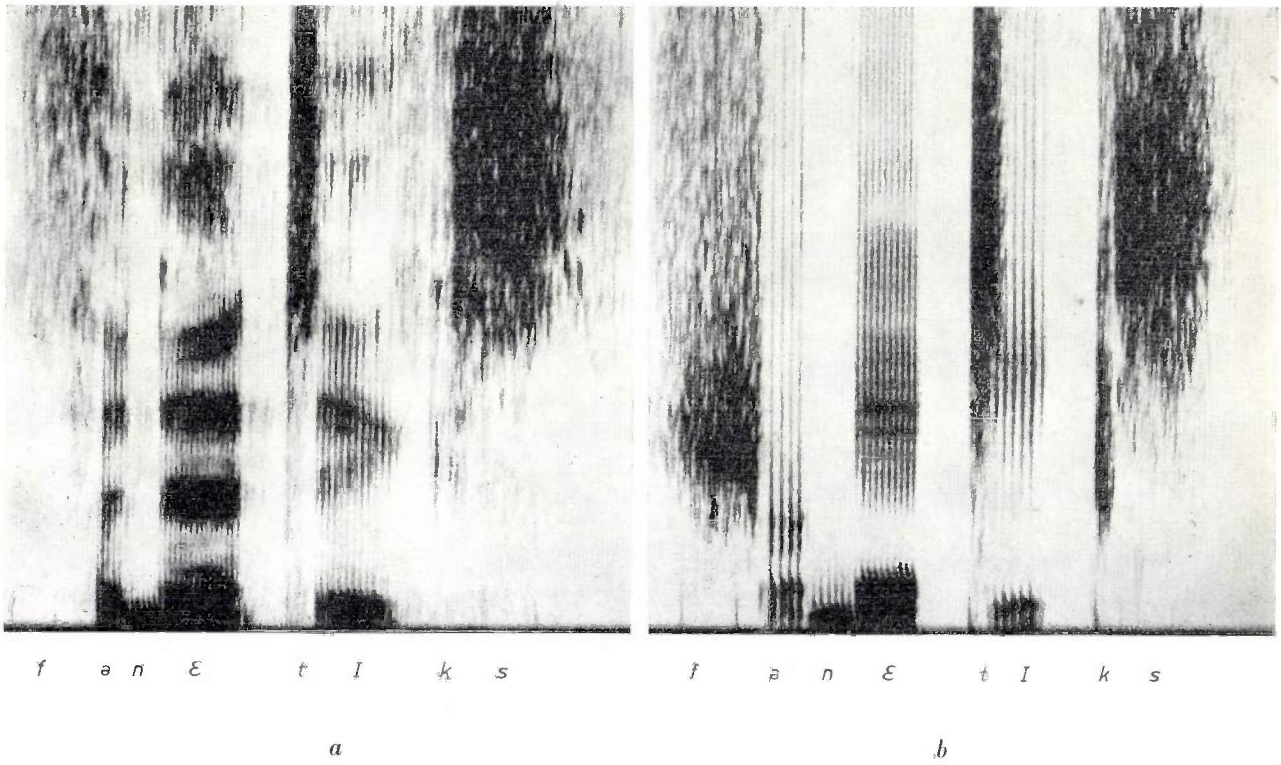


Fig. 6. Spectrograms of the word *phonetics*, *a*) spoken, *b*) synthetic. In spite of the simplification of the spectra in the synthesis (cf. fig. 5) and the absence of continuous transitions, as e.g. during the vowel sound *I*, the synthetic word is completely intelligible.

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.

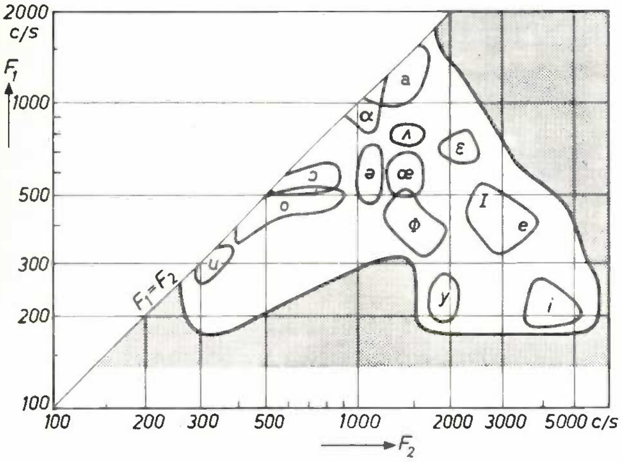


Fig. 7. In order to produce, as in the IPOVOX, vowels and vowel-like consonants with the aid of only two formants (frequency regions F_1 and F_2) it is necessary, as found by listening experiments, that in a plot of F_1 versus F_2 each of the sounds should come within the appropriate contour.

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

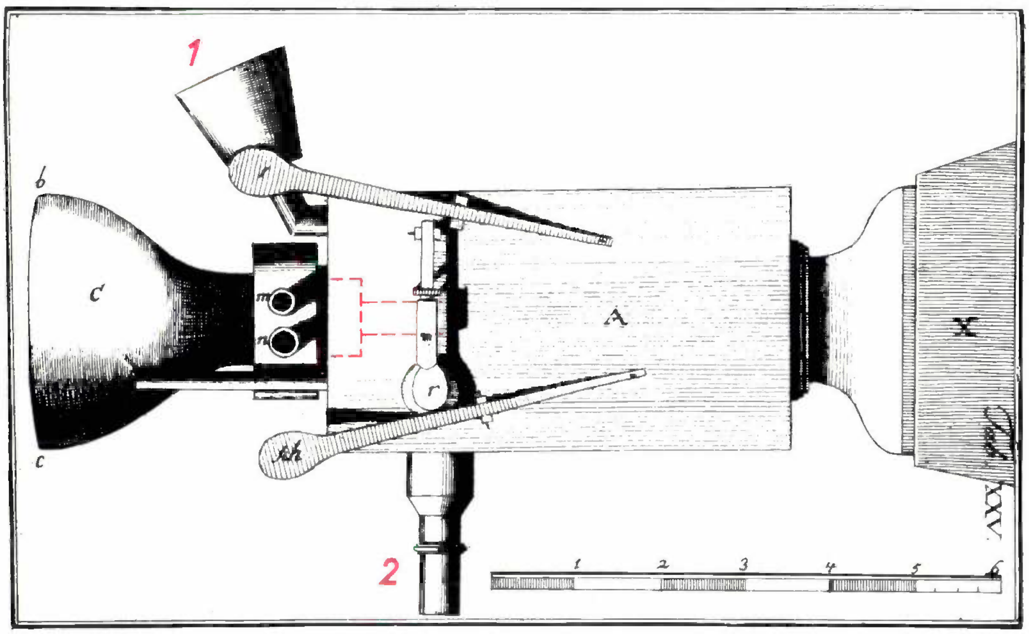


Fig. 1. Von Kempelen's speaking machine (top view), reproduced from his book [1]). The features printed in red are additions to the original drawing.

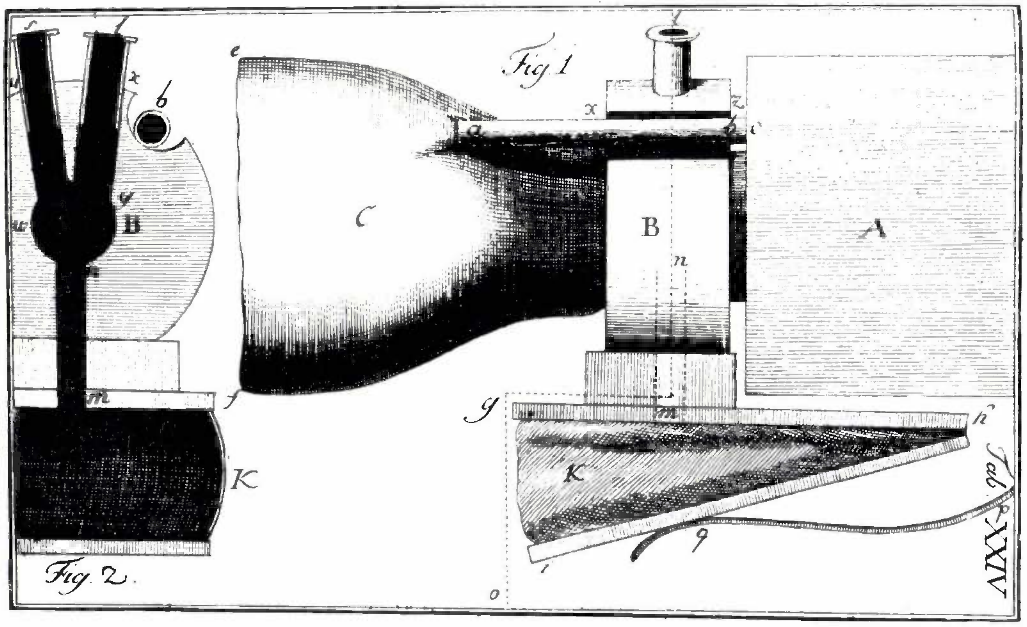


Fig. 2. The machine seen from one side, with a cross-section through the "nostrils".

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.

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.

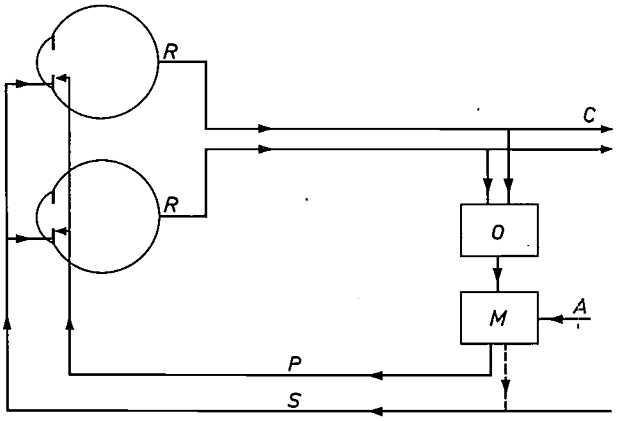


Fig. 1. Highly simplified diagram of the pupillary reflex mechanism. *R* the retinal receptors which absorb the light entering the pupil. *O* the "receptive centre" in the brain stem where the signals from both eyes are combined. *M* the motor centre, also in the brain stem, from which signals are transmitted 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*.

*) Institute for Perception Research, Eindhoven.

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.

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.

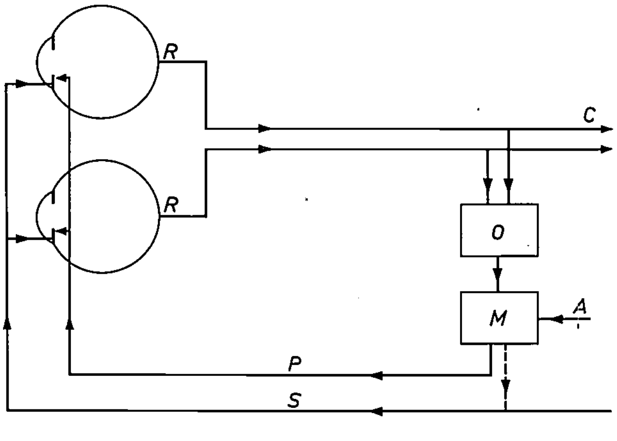


Fig. 1. Highly simplified diagram of the pupillary reflex mechanism. *R* the retinal receptors which absorb the light entering the pupil. *O* the "receptive centre" in the brain stem where the signals from both eyes are combined. *M* the motor centre, also in the brain stem, from which signals are transmitted 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*.

*) Institute for Perception Research, Eindhoven.

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.

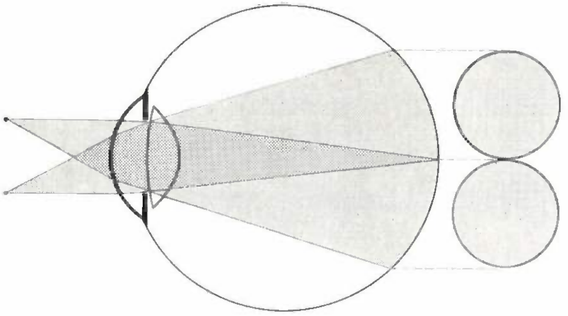


Fig. 2. Two point sources of light, immediately in front of the eye, project two beams of light on to the retina. The beams will overlap or be seen separately depending on the spacing between the light sources, the pupil diameter and the accommodation of the eye. When the eye is *not* accommodated, parallel 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

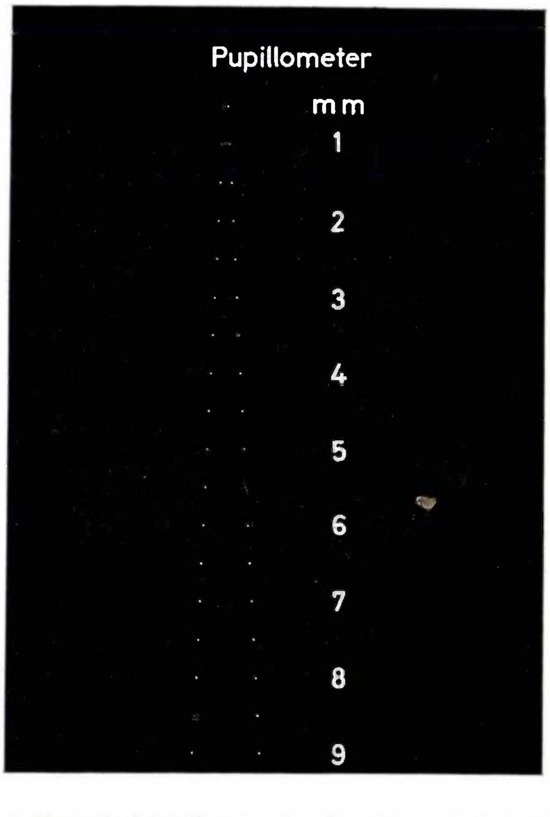


Fig. 3. Entoptical pupillometer, based on the principle sketched in fig. 2. A blackened photographic negative contains pairs of light-transmitting dots. The spacing of the dots is marked beside every other pair [2]).

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

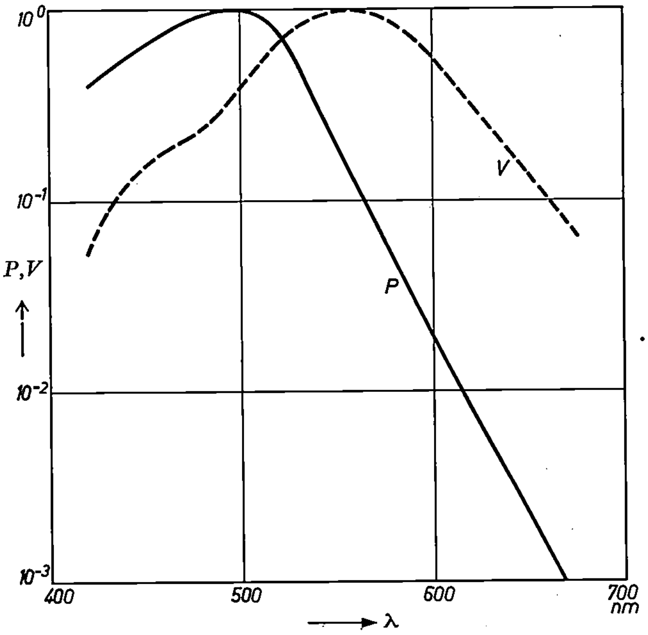


Fig. 4. The relative sensitivity P of the pupil and the relative luminous efficiency V (for photopic vision) as a function of wavelength λ. In all cases a static state was measured. Tbe field of view for both measurements was 18°.

[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 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.

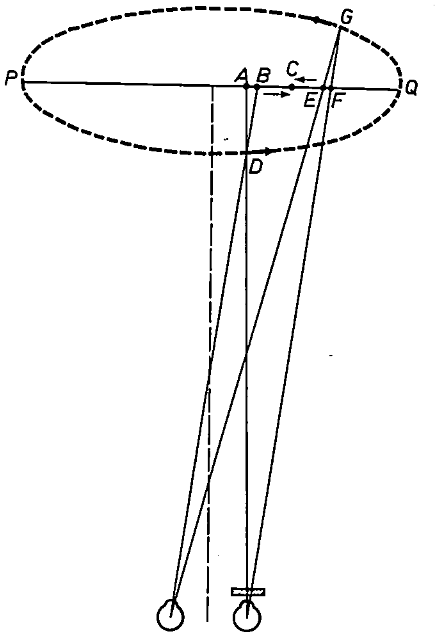


Fig. 5. The Pulfrich effect. An imaginary pendulum suspended above the drawing swings in the plane perpendicular to the paper through the horizontal line *PQ*. There is a delay in observing the pendulum — the perception delay.

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.

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 determinc 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_l and the light intensity I:

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

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 perccption 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.

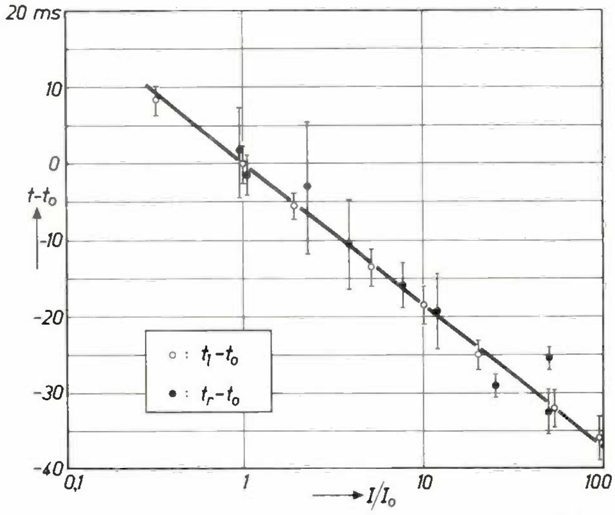


Fig. 6. In a certain range of intensities the perception delay t_l in the observation of a flash of light decreases linearly with the logarithm of the intensity I of the flash (open circles). I_0 is the intensity of the second flash, which is kept constant, and t_0 is the perception delay for $I = I_0$.

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

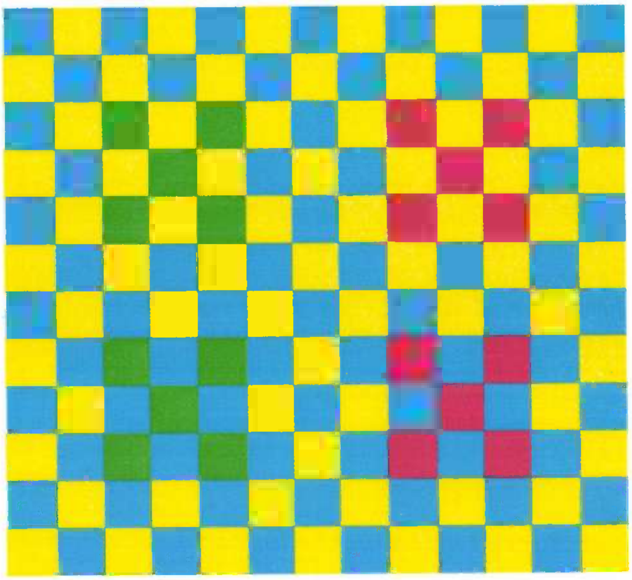


Fig. 7. The red (green) squares in the top and bottom halves of the figure are identical. The fact that they do not seem to be so is due to their different surrounds. The effect of the surround on the perception of a colour is most marked if the figure is viewed from a distance or from a small angle.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

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.

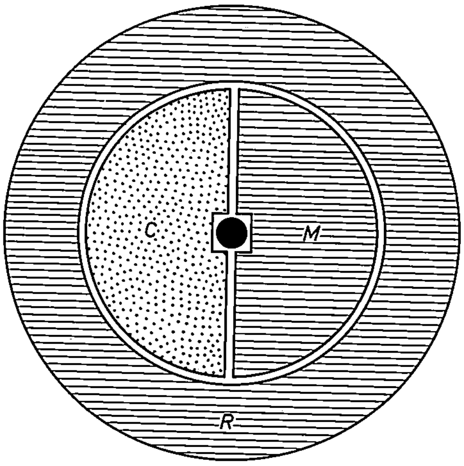


Fig. 8. The picture seen by a subject in experiments concerning the influence of the surround on the observed brightness of a given object, using the compensating method. The left eye sees only the dotted calibrating field C, and the right eye the hatched fields, i.e. the ring field R and the test field 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, $L_{c1} \dots L_{c4}$. 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 ôf 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.

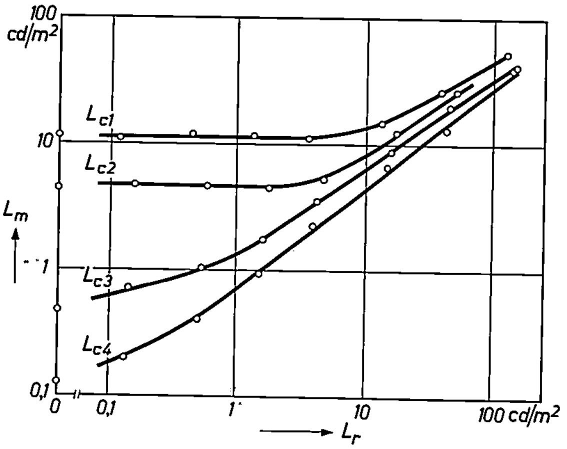


Fig. 9. The luminance L_m of the test field as a function of the luminance L_r of the ring, at four luminances L_c of the calibrating field (L_m being adjusted by the subject to produce equal brightness in left and right eyes). The upward trend of the curves indicates that the sensitivity of the part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

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.

[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.

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.

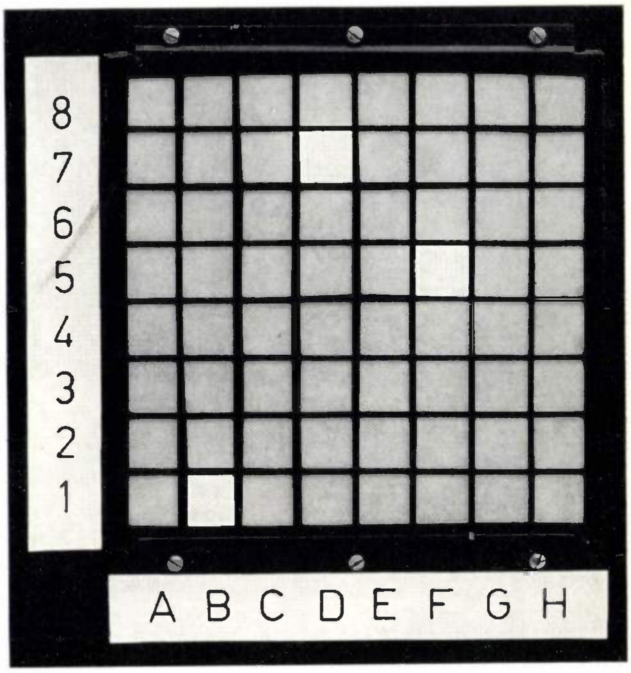


Fig. 1. The optical square. The subject must give the coordinates of the smaller squares which are illuminated for a short time (e.g. 0.1 s) by the experimenter. In this situation, the identification of one square corresponds to the intake of an amount of information equal to $\log_2 64 = 6$ bits; two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

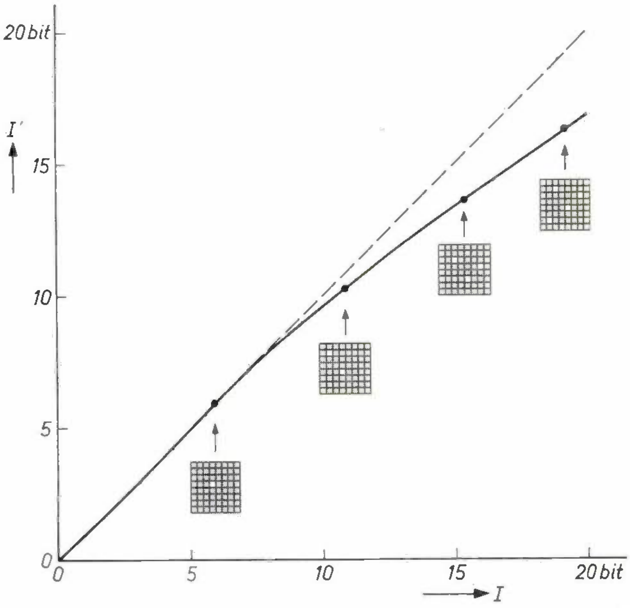


Fig. 2. The amount of information I presented in the optical square and the amount of information I' taken in by the subject are set out along the horizontal and vertical axes respectively. When more than about 9 bits are presented in 0.1 s, a steadily increasing proportion of the information is lost.

[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.

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 abovementioned 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.

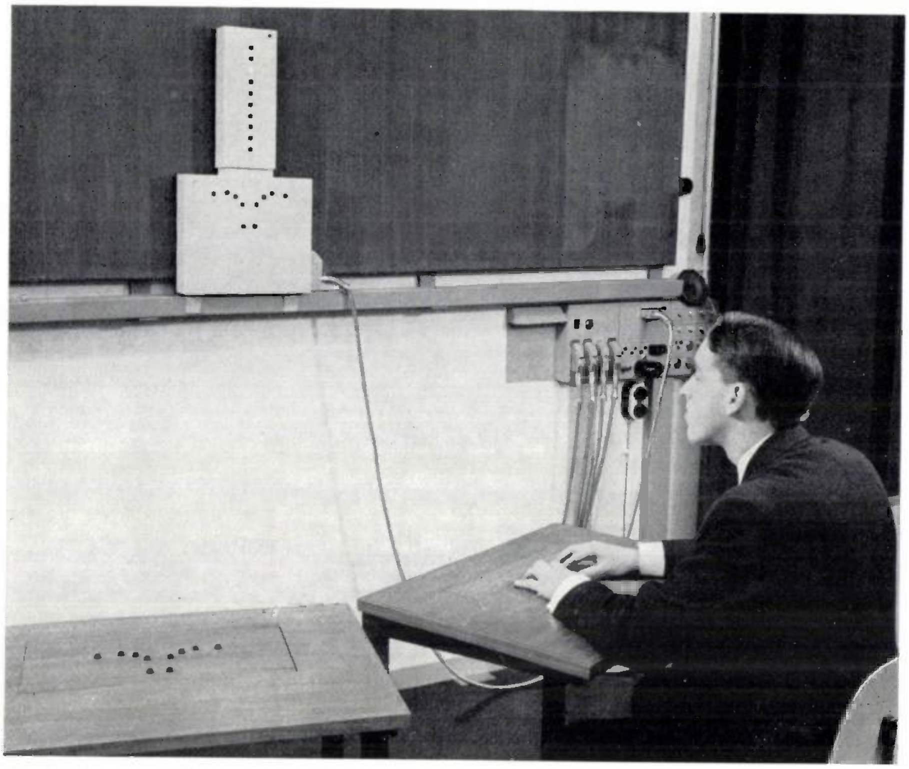


Fig. 3. Set-up which can be used to investigate the influence of the compatibility (degree of similarity between stimulus and response) on the reaction time. The subject must react as quickly as possible to the flashing of one of the lights on the stimulus panel (top left) by depressing the appropriate 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

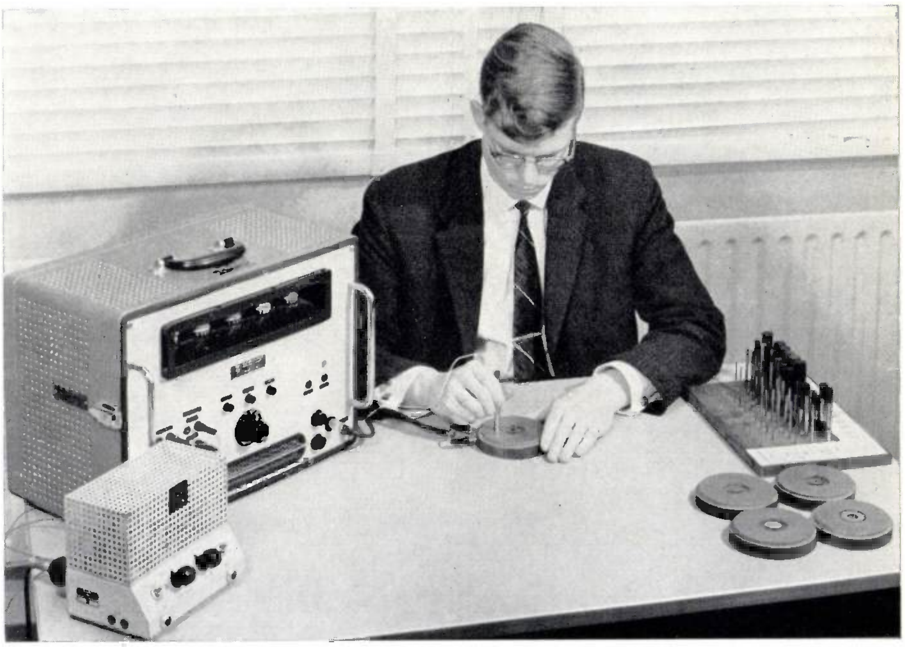


Fig. 4. Set-up for measuring the performance time for pin mounting (sticking a peg in a hole). The time taken for the operation is measured by the electronic counter on the left of the subject. This counter is set into action as soon as the pin leaves the starting circle (thus breaking an electric circuit), 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 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.

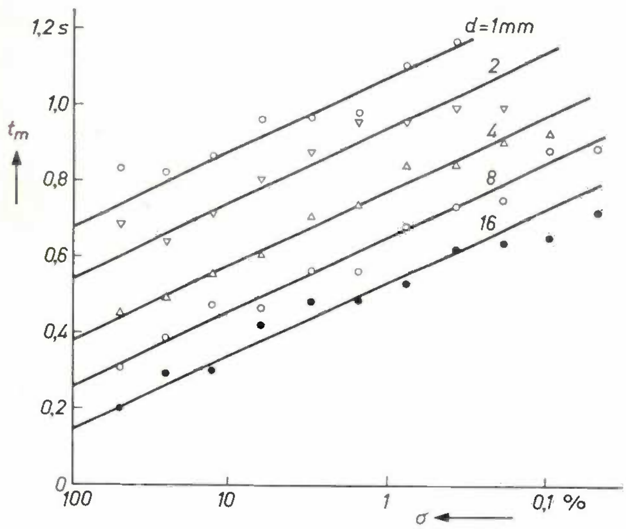


Fig. 5. Relationship between the assembly time t_m and the relative tolerance σ of the pins for various values of the diameter d of the hole. It may be seen that a logarithmic relationship exists both between t_m and σ and between t_m and d.

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

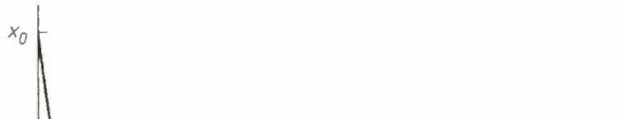


Fig. 6. In explanation of the logarithmic relationship between the assembly time t_m and the tolerance σ. If the error x (i.e. the distance between the axis of the hole and that of the pin) always decreases by the same factor in the same length of time, the pin (diameter d_p) moves to the hole (diameter 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.

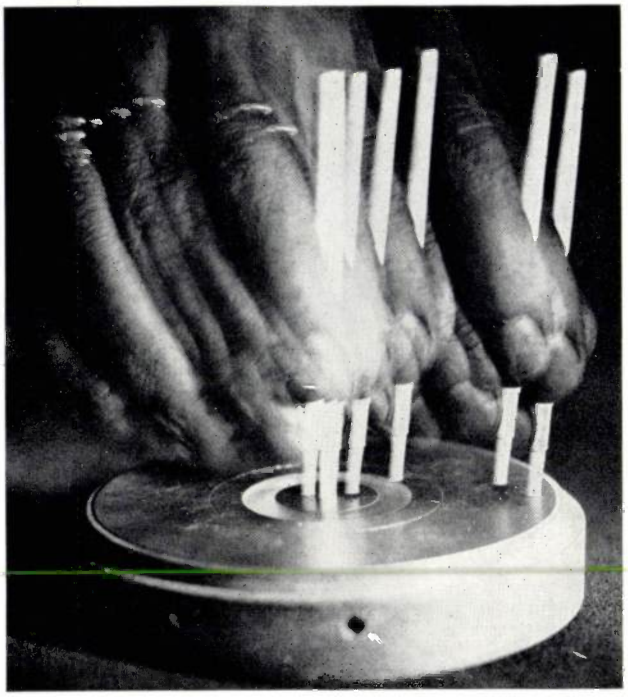


Fig. 7. Stroboscopic photo during the mounting of a pin in a hole of diameter 4 mm and a relative tolerance of 0.1%. The light was switched on at intervals of 0.1 s. It may be seen from the positions of the wedding ring that it took 0.3 s to move the pin from the top of the hole to the bottom.

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 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.

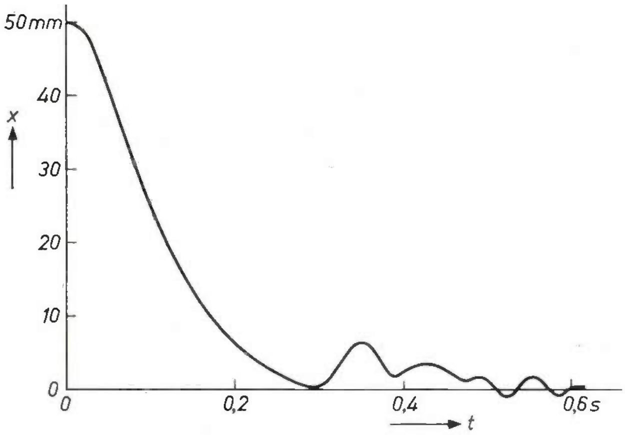


Fig. 8. The distance-time diagram for the assembly operation of fig. 7.

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 (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

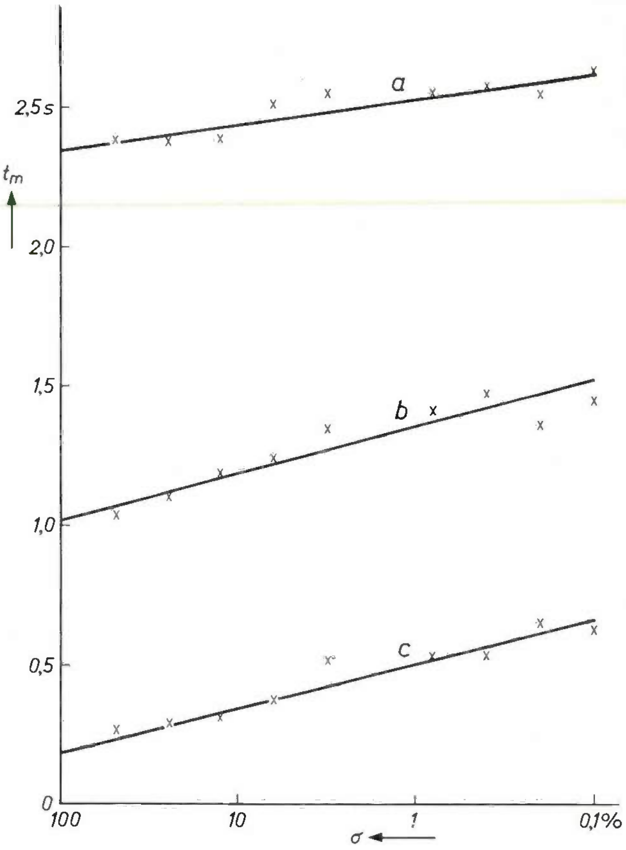


Fig. 9. Relationship between the assembly time t_m and the tolerance σ when the pin mounting is carried out *a*) slowly, *b*) at a medium rate, *c*) as fast as possible. The intercepts made by the various lines on the vertical axis represent the times needed to bring the pin from the starting circle 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.

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.

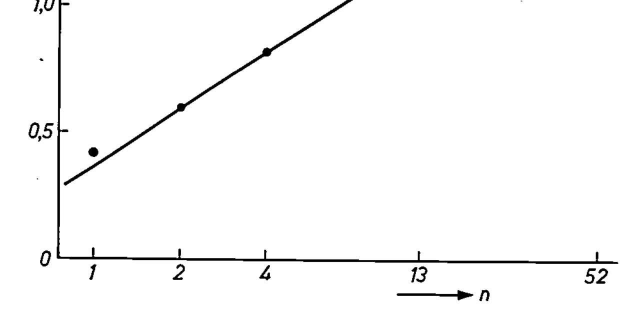


Fig. 10. Average dealing times t_s as a function of the number n of piles into which a pack of cards must be dealt. A linear relationship is found between the mean dealing time and the logarithm of the number of piles.

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.

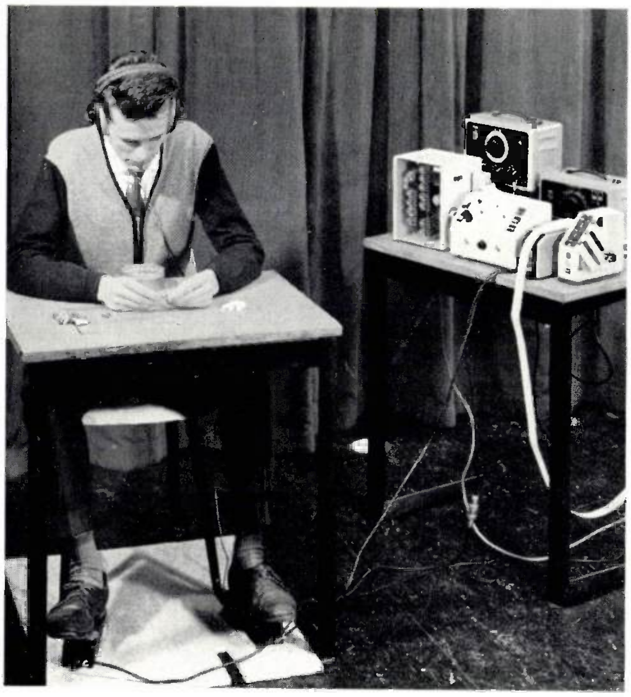


Fig. 11. The capacity meter developed in the Institute for Perception Research. In order to measure the difficulty of a perceptual task (here the assembly of a washer and a nut on a bolt), we investigate how the performance in this task deteriorates as the standard task is simultaneously carried out 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.

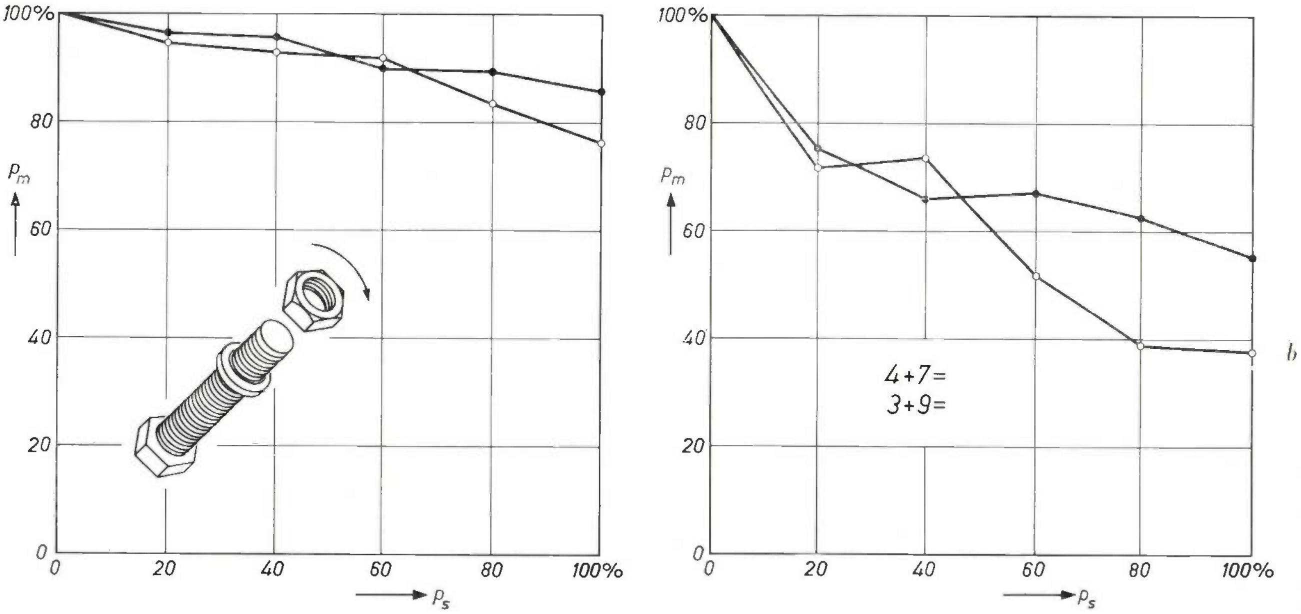


Fig. 12. Performance p_m in two perceptual tasks, *a*) assembly and *b*) simple arithmetic, as a function of the performance p_s in the standard task, in each case for two subjects. The performance is measured as the number of operations which the subject can carry out per unit time under the given conditions, 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.

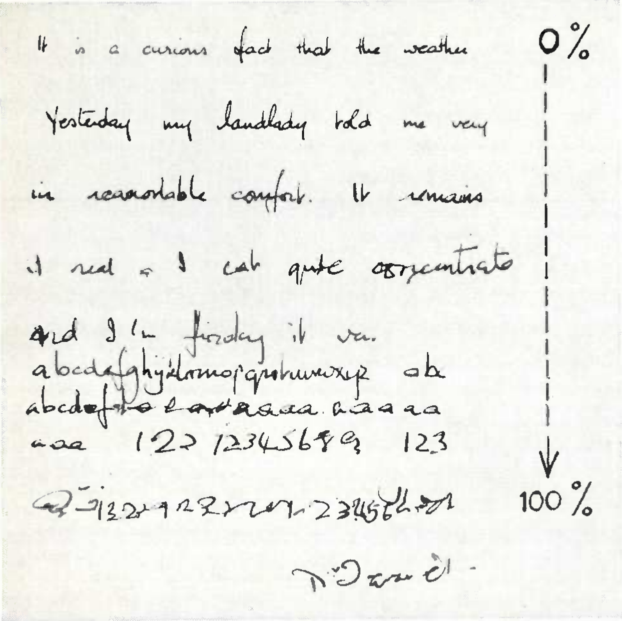

It is a curious fact that the weather
Yesterday my landlady told me very
in reasonable comfort. It remains
it read a I cat qute ozzycentrato
and I'm finding it ver.
abcdefghijklmnopqrstuvwxyz ab
abcdef aaaaa aaaaa
aaa 122 12345679 123
0%
100%

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.

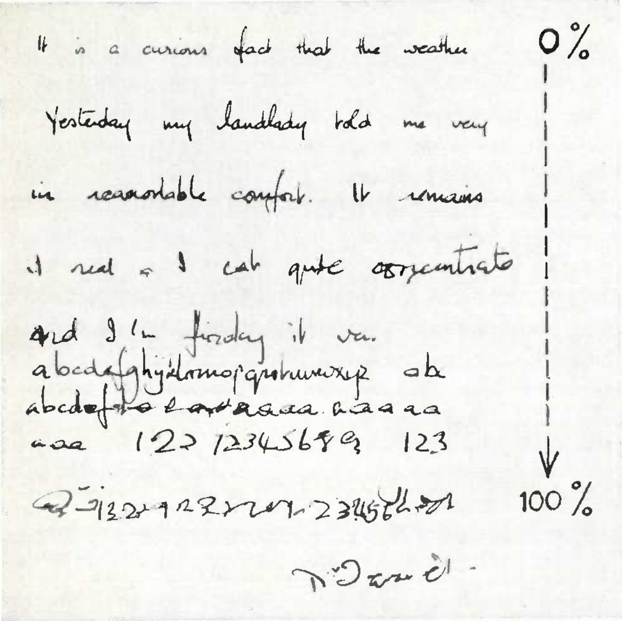

It is a curious fact that the weather
Yesterday my landlady told me very
in reasonable comfort. It remains
I read & I cant quite concentrate
and I'm finding it ver.
abcdefghijklmnopqrstuvwxyz ab
123 123456789 123
0%
100%

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.

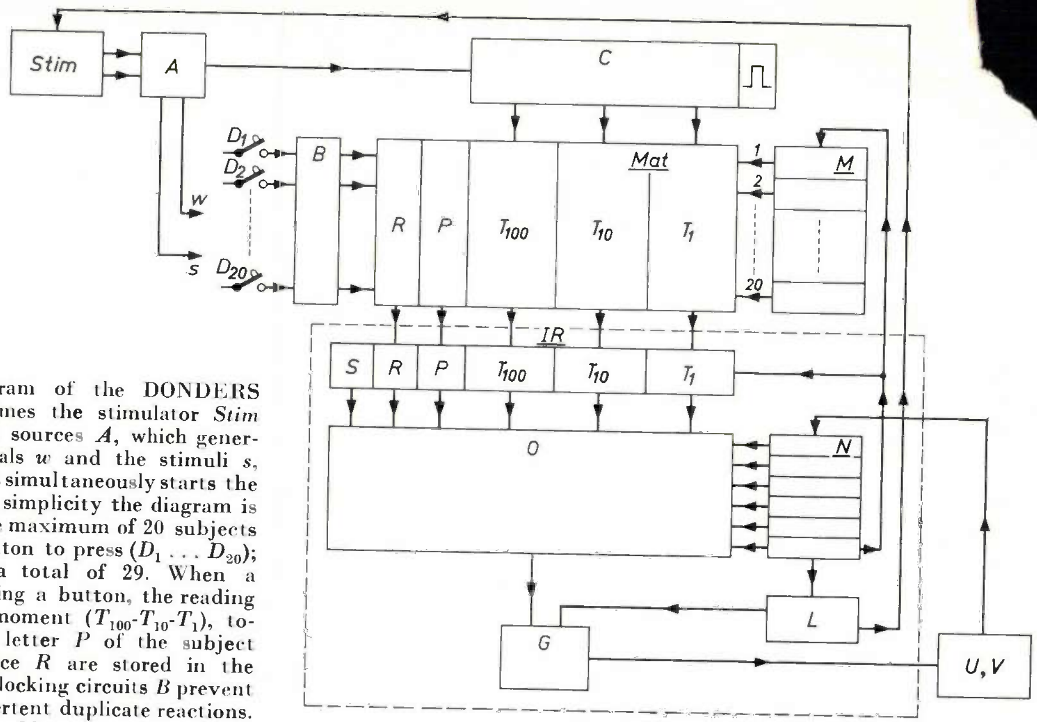


Fig. 1. Block diagram of the DONDERS (simplified). At set times the stimulator *Stim* activates the physical sources *A*, which generate the warning signals *w* and the stimuli *s*, and with each stimulus simultaneously starts the electronic clock *C*. For simplicity the diagram is drawn as if each of 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.

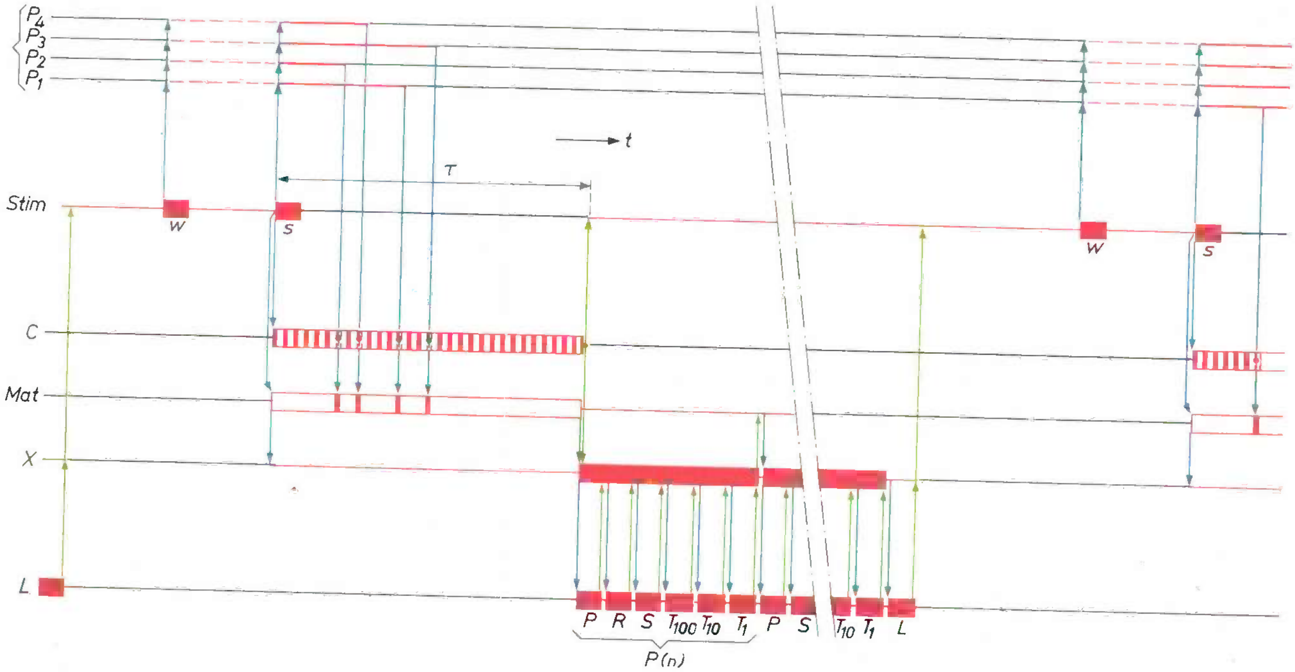


Fig. 2. Time-sequence diagram of the DONDERS. For the sake of clarity the number of subjects is limited here to four ($P_1 \ldots P_4$). Black lines denote the *resting state*, broken red lines the *alert phase*, solid red lines the *active phase*, blue arrows the *transmission of information*, and green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

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 20×6 =120 symbols. This storage of figures consti-

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

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:

$$
\begin{aligned}
0 &= 10000, & 5 &= 00101,\\
1 &= 10001, & 6 &= 00110,\\
2 &= 10010, & 7 &= 00111,\\
3 &= 10011, & 8 &= 01000,\\
4 &= 10100, & 9 &= 01001.
\end{aligned}
$$

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

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 complete 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

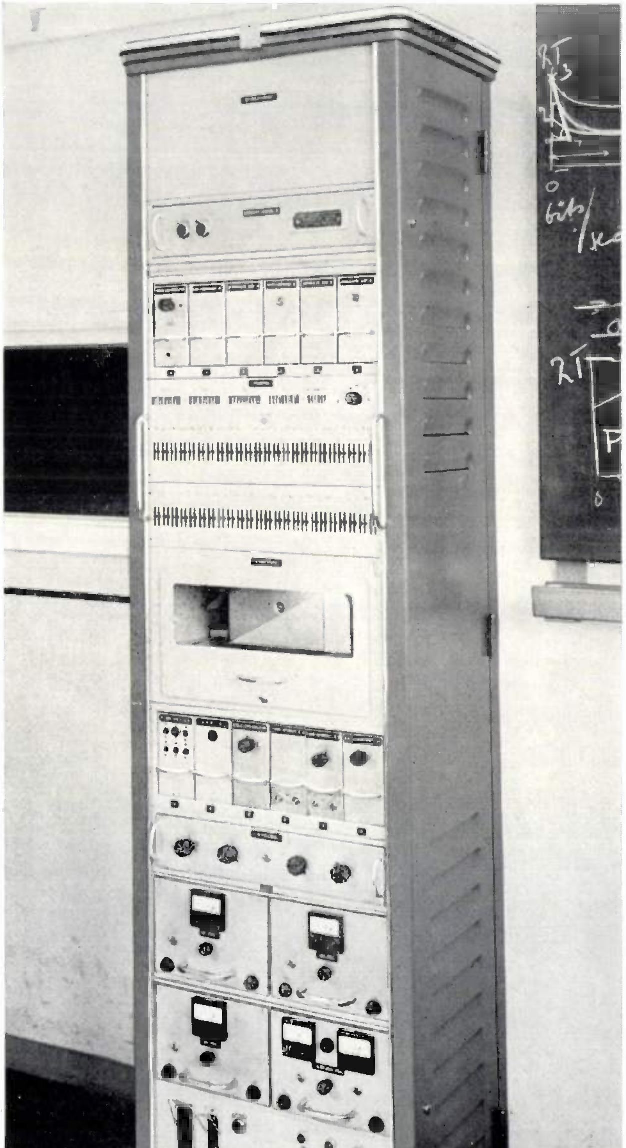


Fig. 3. The stimulator, which controls the progress of an experiment.

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

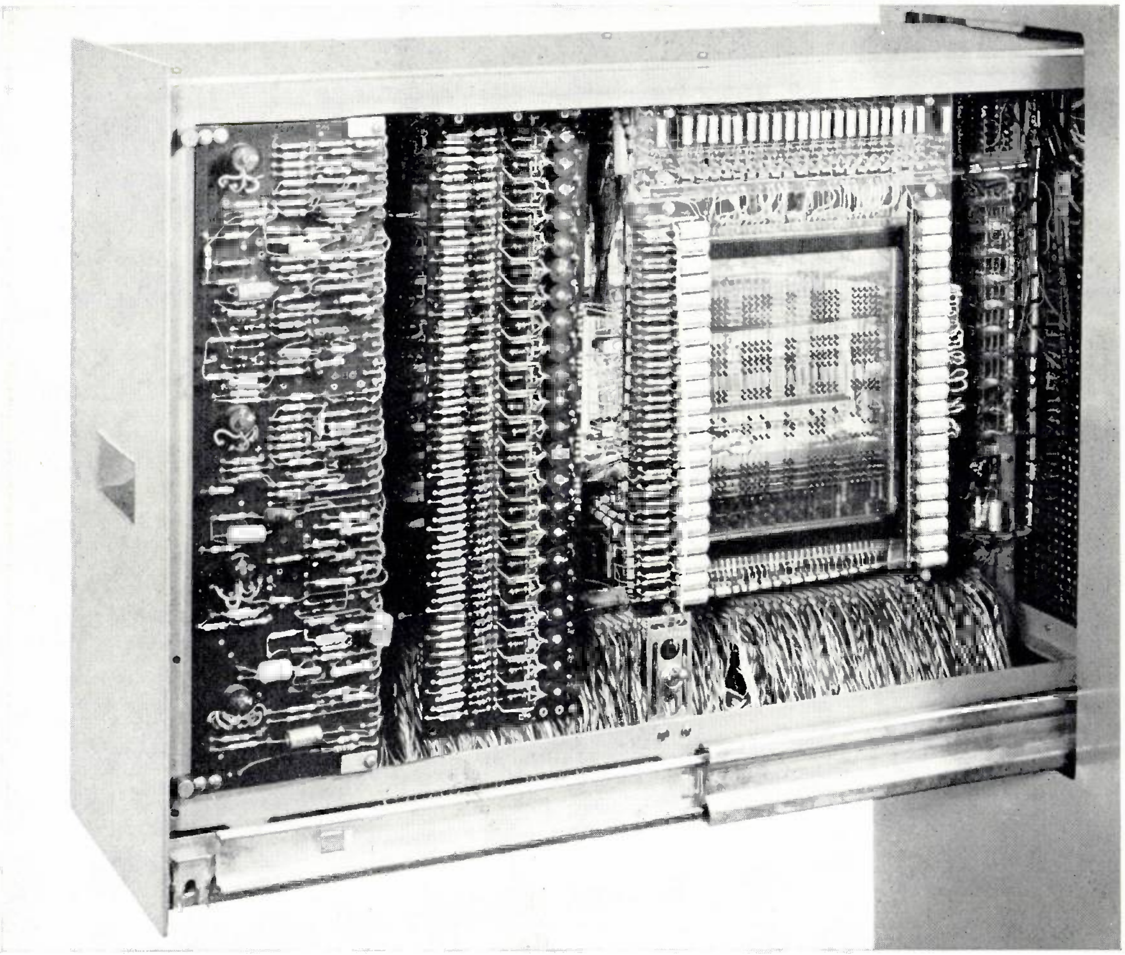


Fig. 4. The circuit wiring behind the operator's presetting panel (see fig. 7) is contained in a drawer, which is here shown open. Through the transparent panel right of centre can be seen part of the cores of the matrix memory.

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.

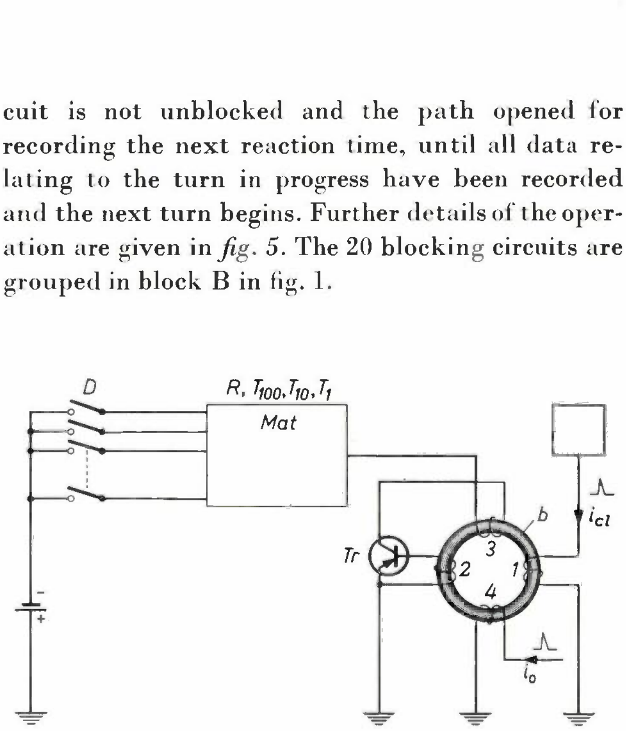


Fig. 5. Circuit which prevents the recording of inadvertent reactions and at the same time blocks the matrix memory during the movements of the electronic clock.

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 thc 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}, . . . , 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.

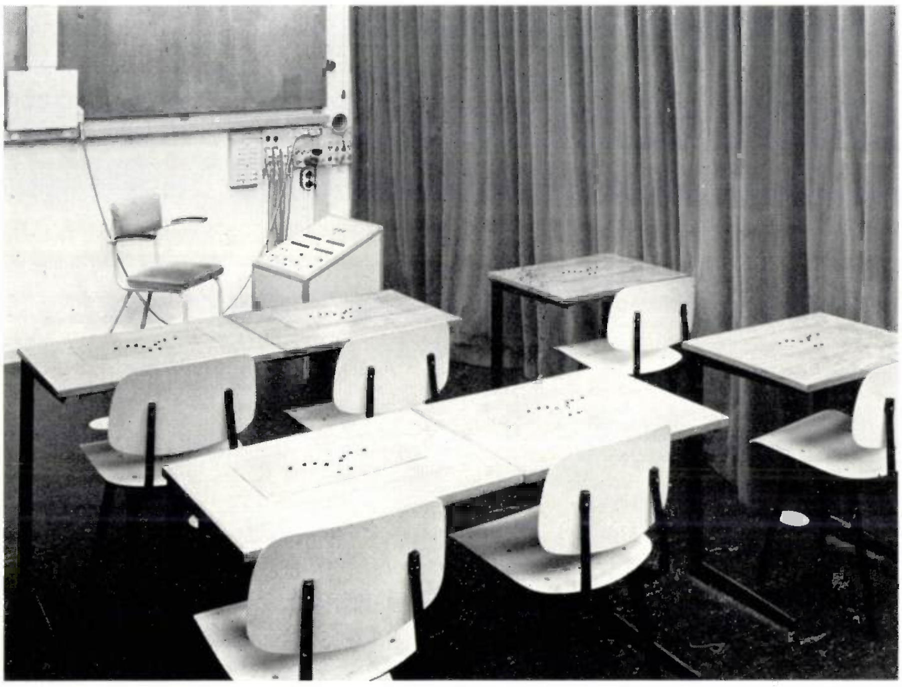


Fig. 6. View of the test room, showing six of the twenty desks at which the subjects sit. In the background can be seen the experimenter's desk. Top left, two optical stimulus sources, consisting of lamps in various configurations (more clearly seen in the photograph on page 58 of this number).

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.

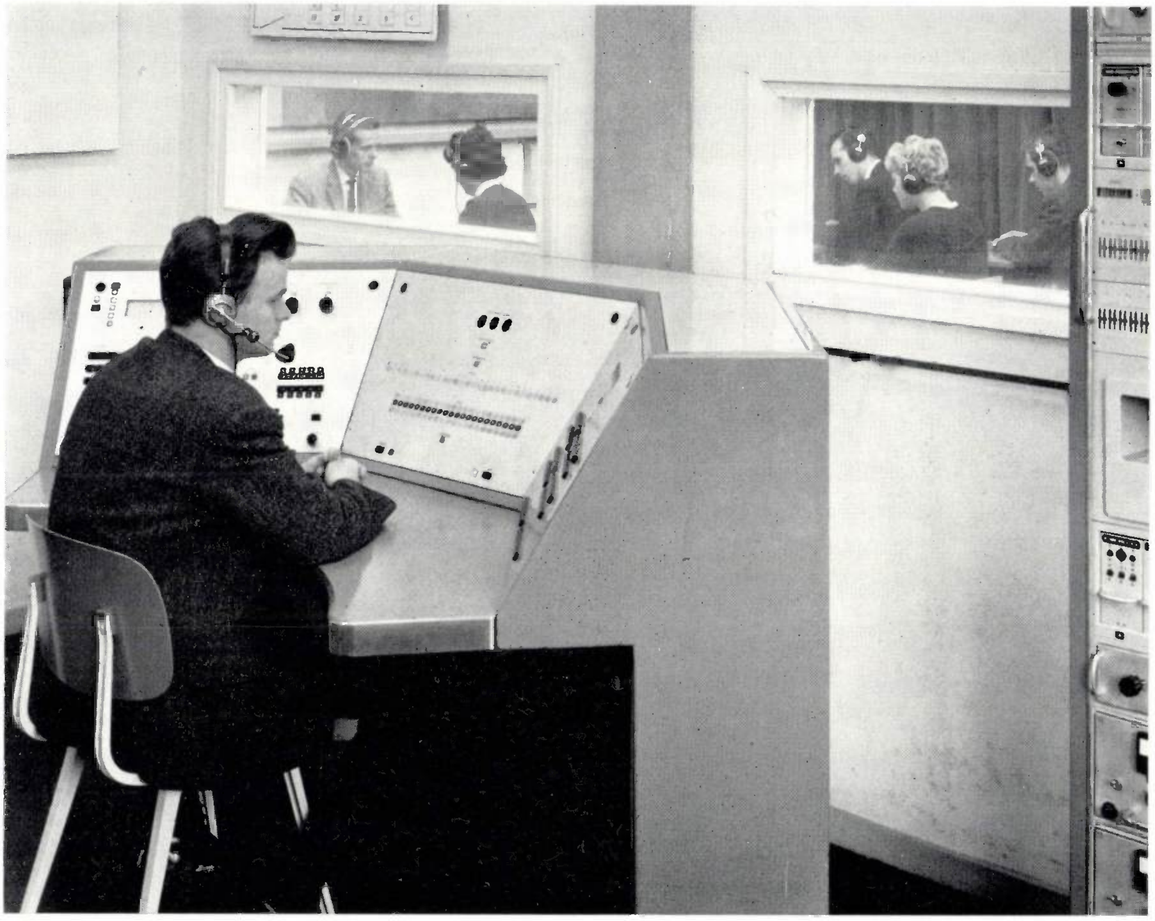


Fig. 7. The operator's console, in which the DONDERS is mounted. On his left the operator has the presetting panel, facing him is the monitoring panel, and on his right is the control panel. A duplicate of the monitoring panel is on the experimenter's desk (fig. 6). On the extreme right, a part of the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

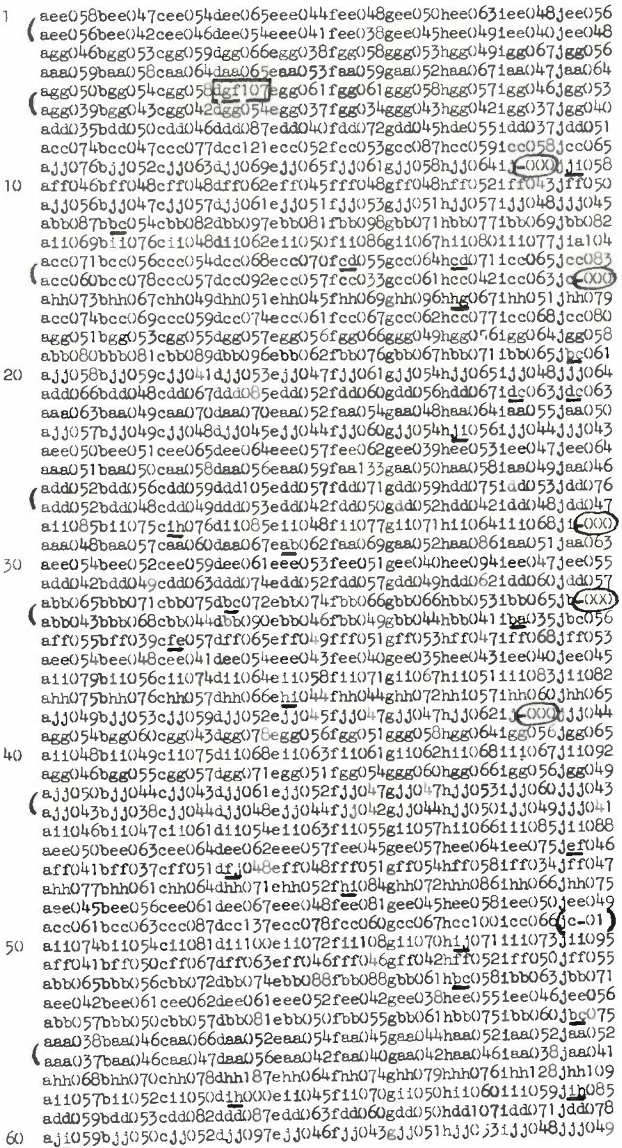
Results

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-*



```
1   aee058bee047cee054dee065eee044fee048gee050hee063iee048jee056
    aee056bee042cee046dee054eee041fee038gee045hee049iee040jee048
    agg046bgg053cgg059dgg066egg038fgg058ggg053hgg049igg067jgg056
    aaa059baa058caa064daa065eaa053faa059gaa052haa067iaa047jaa064
    agg050bgg054cgg058dgf107egg061fgg061ggg058hgg057igg046jgg053
    agg039bgg043cgg042dgg054egg037fgg034ggg043hgg042igg037jgg040
    add035bdd050cdd046ddd087edd040fdd072gdd045hde055idd037jdd051
    acc074bcc047ccc077dcc121ecc052fcc053gcc087hcc059icc058jcc065
    ajj076bjj052cjj063djj069ejj065fjj061gjj058hjj064ij-000jji058
10  aff046bff048cff048dff062eff045fff048gff048hff052iff043jff050
    ajj056bjj047cjj057djj061ejj051fjj053gjj051hjj057ijj048jjj045
    abb087bbc054cbb082dbb097ebb081fbb098gbb071hbb077ibb069jbb082
    aii069bii076cii048dii062eii050fii086gii067hii080iii077jia104
    acc071bcc056ccc054dcc068ecc070fcd055gcc064hcd071icc065jcc083
    acc060bcc078ccc057dcc092ecc057fcc033gcc061hcc042icc063jc-000
    ahh073bhh067chh049dhh051ehh045fhh069ghh096hhg067ihh051jhh079
    acc074bcc069ccc059dcc074ecc061fcc067gcc062hcc077icc068jcc080
    agg051bgg053cgg055dgg057egg056fgg066ggg049hgg056igg064jgg058
    abb080bbb081cbb089dbb096ebb062fbb076gbb067hbb071ibb065jbc061
20  ajj058bjj059cjj041djj053ejj047fjj061gjj054hjj065ijj048jjj064
    add066bdd048cdd067ddd085edd052fdd060gdd056hdd067idc063jdc063
    aaa063baa049caa070daa070eaa052faa054gaa048haa064iaa055jaa050
    ajj057bjj049cjj048djj045ejj044fjj060gjj054hji056ijj044jjj043
    aee050bee051cee065dee064eee057fee062gee039hee053iee047jee064
    aaa051baa050caa058daa056eaa059faa133gaa050haa058iaa049jaa046
    add052bdd056cdd059ddd105edd057fdd071gdd059hdd075idd053jdd076
    add052bdd048cdd049ddd053edd042fdd050gdd052hdd042idd048jdd047
    aii085bii075cih076dii085eii048fii077gii071hii064iii068ji-000
    aaa048baa057caa060daa067eab062faa069gaa052haa086iaa051jaa063
30  aee054bee052cee059dee061eee053fee051gee040hee094iee047jee055
    add042bdd049cdd063ddd074edd052fdd057gdd049hdd062idd060jdd057
    abb065bbb071cbb075dbc072ebb074fbb066gbb066hbb053ibb065jb-000
    abb043bbb068cbb044dbb090ebb046fbb049gbb044hbb041iba035jbc056
    aff055bff039cfe057dff065eff049fff051gff053hff047iff068jff053
    aee054bee048cee041dee054eee043fee040gee035hee043iee040jee045
    aii079bii056cii074dii064eii058fii071gii067hii051iii083jii082
    ahh075bhh076chh057dhh066ehi044fhh044ghh072hhi057ihh060jhh065
    ajj049bjj053cjj059djj052ejj045fjj047gjj047hjj062ij-000jjj044
    agg054bgg060cgg043dgg078egg056fgg051ggg058hgg064igg056jgg065
40  aii048bii049cii075dii068eii063fii061gii062hii068iii067jii092
    agg046bgg055cgg057dgg071egg051fgg054ggg060hgg066igg056jgg049
    ajj050bjj044cjj043djj061ejj052fjj047gjj047hjj053ijj060jjj043
    ajj043bjj038cjj044djj048ejj044fjj042gjj044hjj050ijj049jjj041
    aii046bii047cii061dii054eii063fii055gii057hii066iii085jii088
    aee050bee063cee064dee062eee057fee045gee057hee064iee075jef046
    aff041bff037cff051dfj048eff048fff051gff054hff058iff034jff047
    ahh077bhh061chh064dhh071ehh052fhi084ghh072hhh086ihh066jhh075
    aee045bee056cee061dee067eee048fee081gee045hee058iee050jee049
    acc061bcc063ccc087dcc137ecc078fcc060gcc067hcc100icc066(jc-01)
50  aii074bii054cii081dii100eii072fii108gii070hij071iii073jii095
    aff041bff050cff067dff063eff046fff046gff042hff052iff050jff055
    abb065bbb056cbb072dbb074ebb088fbb088gbb061hbc058ibb063jbb071
    aee042bee061cee062dee061eee052fee042gee038hee055iee046jee056
    abb057bbb050cbb057dbb081ebb050fbb055gbb061hbb075ibb060jbc075
    aaa038baa046caa066daa052eaa054faa045gaa044haa052iaa052jaa052
    aaa037baa046caa047daa056eaa042faa040gaa042haa046iaa038jaa041
    ahh068bhh070chh078dhh187ehh064fhh074ghh079hhh076ihh128jhh109
    aii057bii052cii050dih000eii045fii070gii050hii060iii059jih085
    add059bdd053cdd082ddd087edd063fdd060gdd050hdd107idd071jdd078
60  aji059bjj050cjj052djj097ejj046fjj043gjj051hjj053ijj048jjj049
```

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

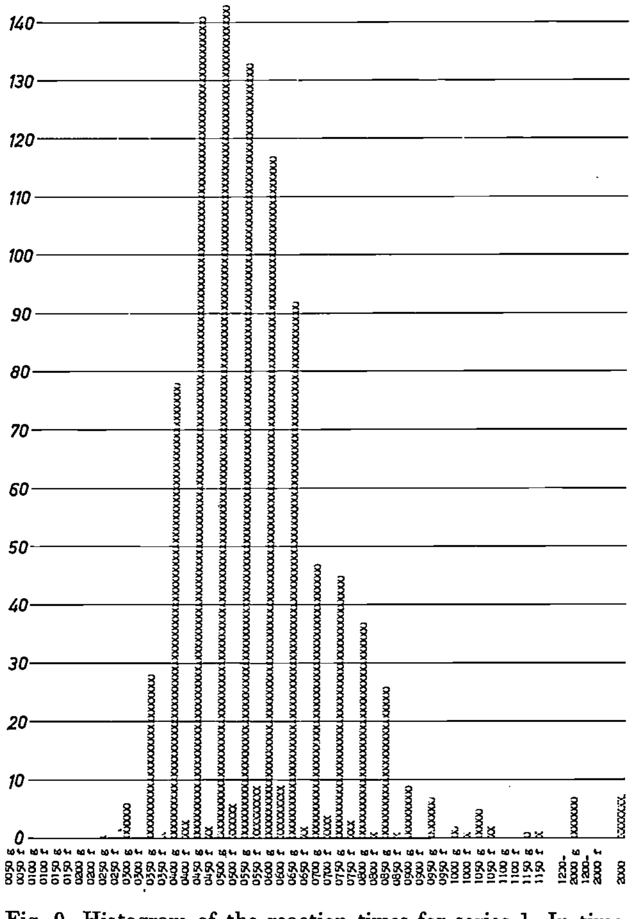


Fig. 9. Histogram of the reaction times for series 1. In time classes of 50 milliseconds the correct results (g) are first typed out with the symbol 0, then the wrong reactions (f) with the symbol X. The distribution curve is the result of 100 measurements on ten subjects.

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.

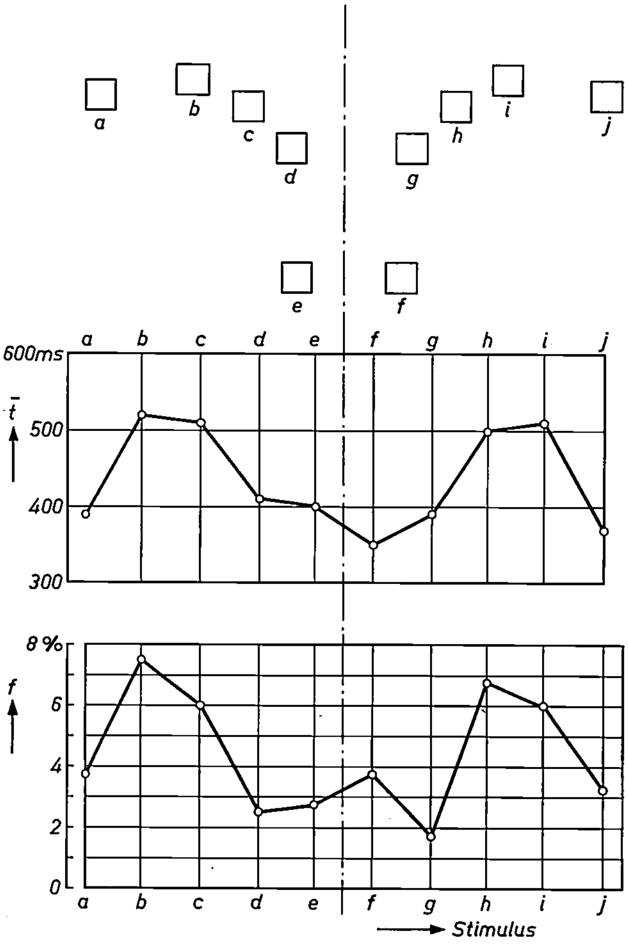
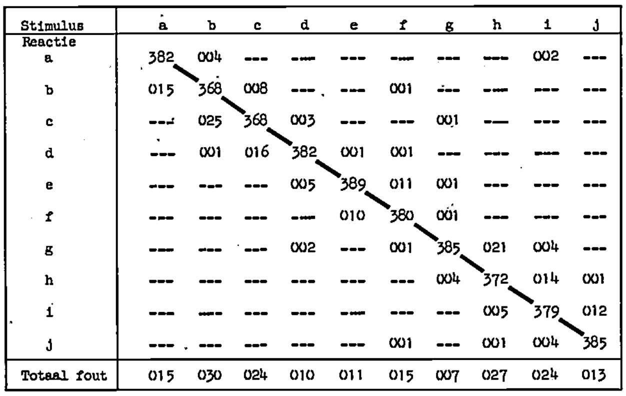


Fig. 11. The average reaction time $\bar{t}$ and the error percentage f pertaining to the stimuli *a-j* for the total experiment. Above: the configuration of the stimulus lamps (likewise of the push-buttons).

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	a	b	c	d	e	f	g	h	i	j
Reactie										
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]).

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

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.)

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

[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.

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".

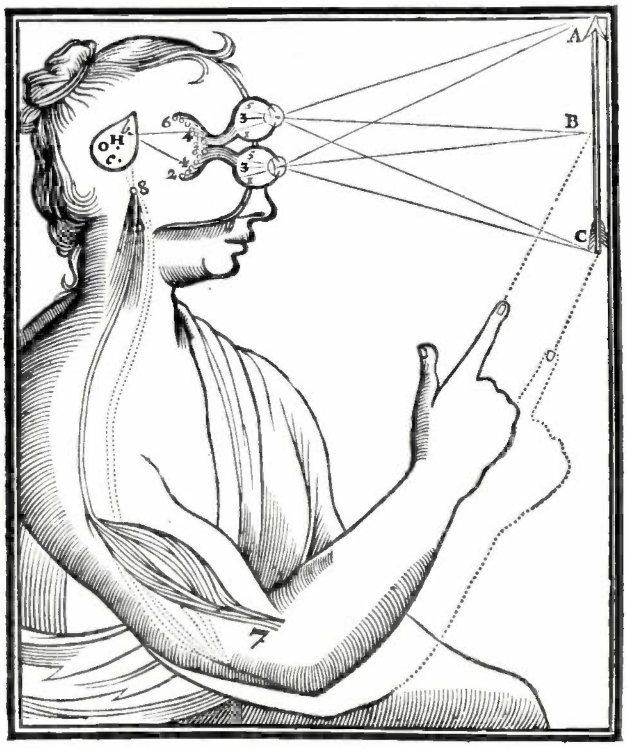


Fig. 2. Diagram from Descartes' Traité de l'homme. Paris 1662, representing the arising of a sensory impression and an active response it evokes. According to Descartes the connection between the two processes was to occur in the *epiphysis cerebri* or pineal gland, an unduplicated part of the brain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

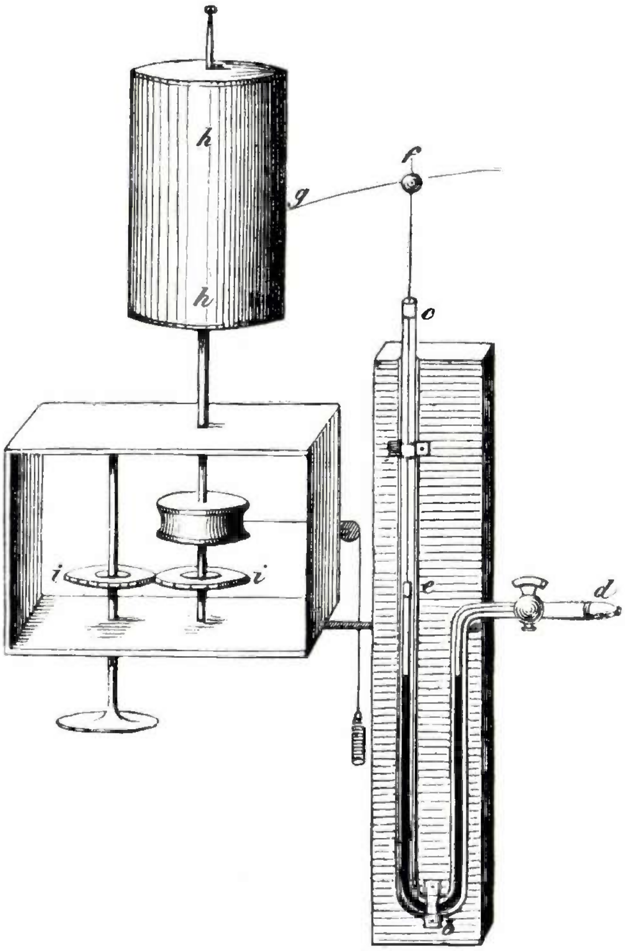


Fig. 3. Ludwig's kymograph, which he designed in 1847 for recording variations in blood pressure. Reproduced from K. Ludwig, Lehrbuch der Physiologie des Menschen, Winter, Heidelberg 1852-1856.

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. Genees- en Natuurkunde 3, 105-109, 1868.

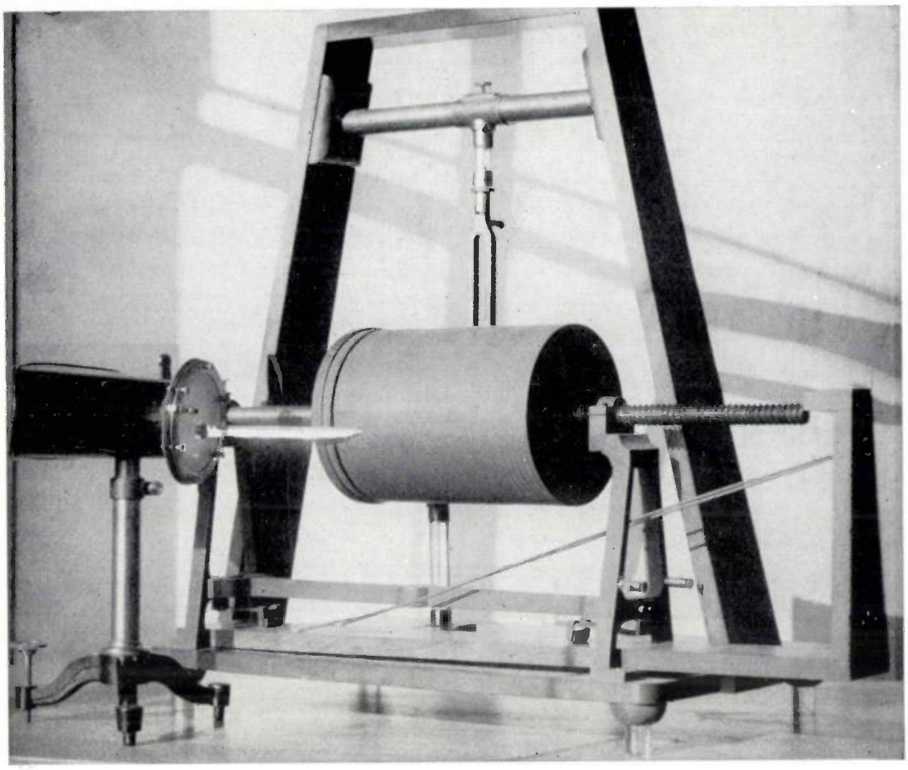


Fig. 4. Donders' noematachograph, which is still preserved in the Physiological Laboratory at Utrecht (photograph reproduced by permission of Prof. Jongbloed, director of that laboratory). In Donders' publication [8]) the apparatus is described in the following words.

"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 seamwelding 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*.

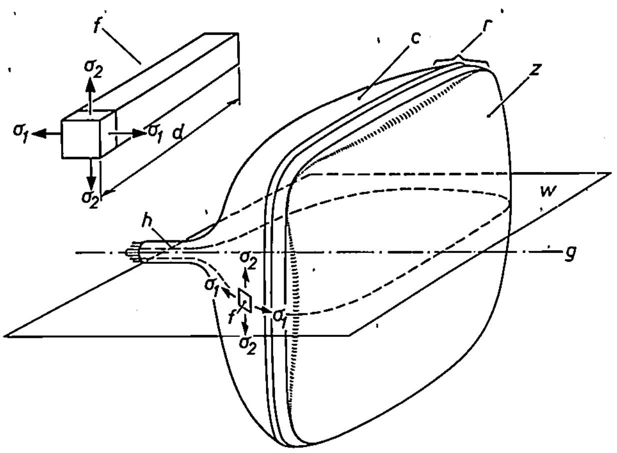


Fig. 1. The strength of an evacuated bulb exposed to atmospheric pressure is chiefly determined by the places where the tangential normal stresses at the outside surface are positive, and thus exercise a tensile force. σ_1 is here the component of the tangential normal stresses in the plane through a 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

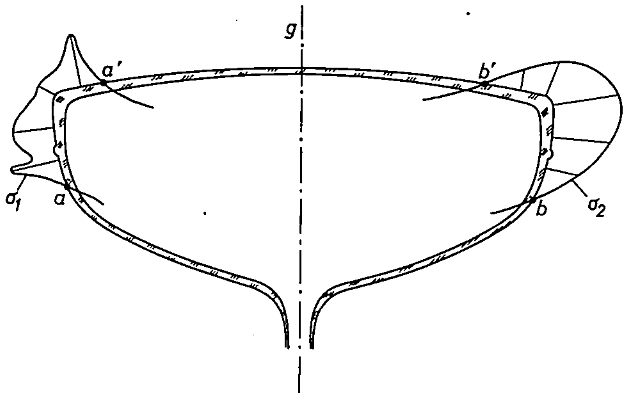


Fig. 2. Bulb cross-section in the plane of symmetry w showing the stresses σ_1 and σ_2 acting in the outer surface, which are represented graphically with the periphery as abscissa. Positive values, plotted outwards, indicate tensile stresses, and negative values, plotted inwards, are compressive stresses. 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.

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:

$$\Delta l = \frac{Kl}{ES}.$$

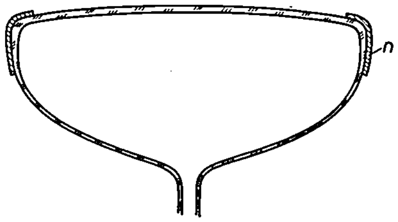


Fig. 3. Prevention of implosion by a reinforcement band n around the tensile stress region of the bulb.

[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.

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

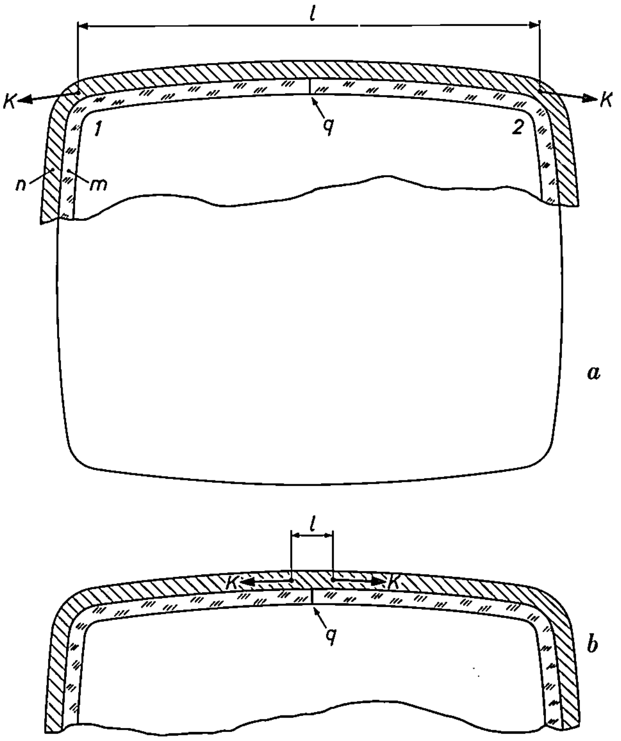


Fig. 4. The peripheral area of a bulb with reinforcement band, in a cross-section perpendicular to the tube axis, illustrating the effect of adhesion on the extent to which the band n limits the widening of a crack q in the glass m. K is the total force acting on the band when the crack is forced open 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*).

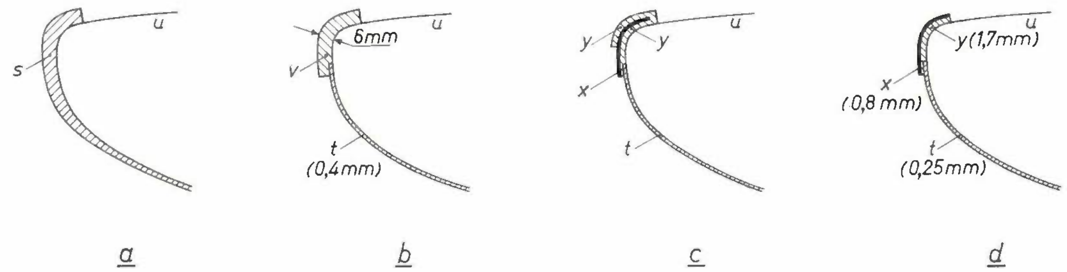


Fig. 5. Different versions of the reinforcement band, in a partial cross-section. The line *u* represents the periphery of the bulb.
a) Experimental version: polyester resin reinforced with glass fibre, *s*.
b) Preform version: the band is in two parts. The part *v* around the rim consists of resin reinforced 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.*

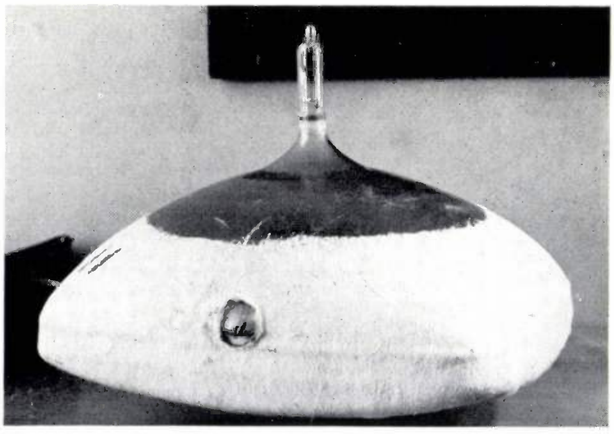


Fig. 6. Bulb with experimental band of polyester resin reinforced with glass fibre.

[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.

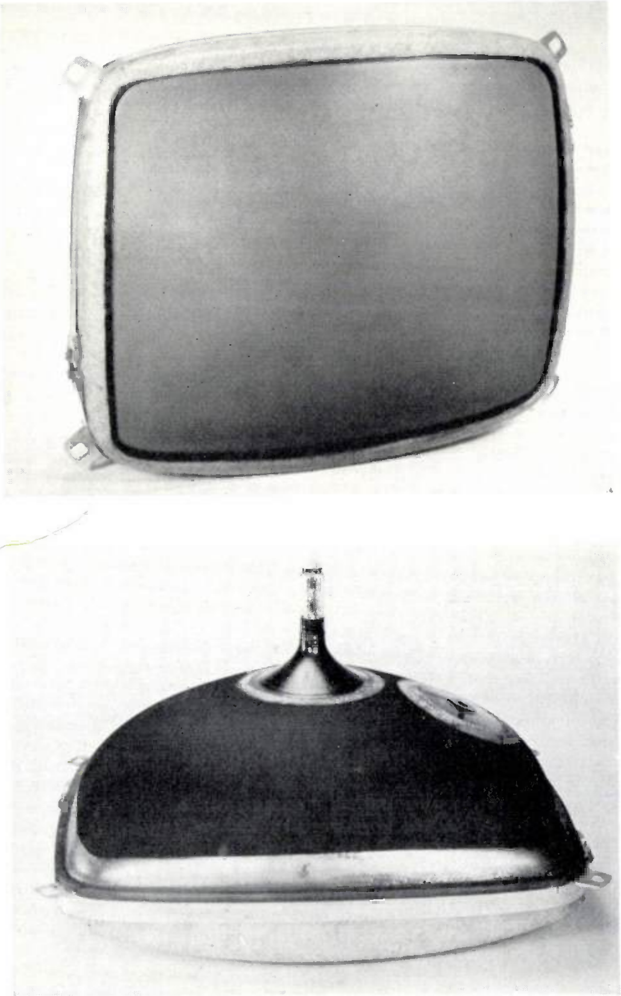


Fig. 7. Preform version.
a) Front view. The rim is surrounded by polyester resin reinforced by a preformed glass-fibre ring. The attachment lugs are fixed to the tube by the metal strip subsequently fitted.
b) Rear view. The band around the cone, resin reinforced with glass-fibre cloth, is largely concealed by the outside layer of graphite which, together with a conductive layer on the inside of the cone, forms the EHT smoothing capacitor.

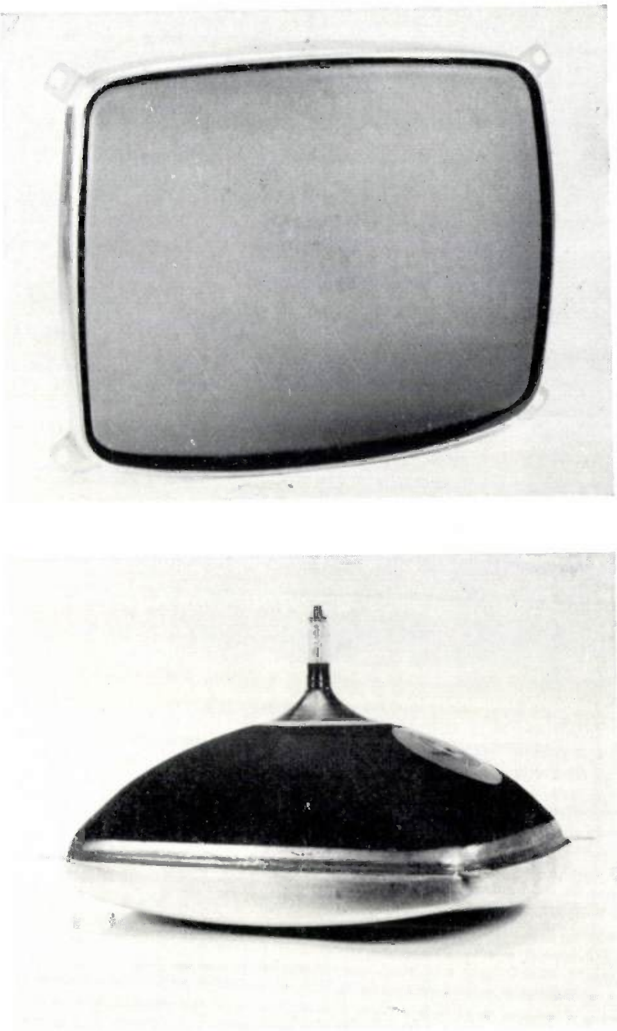


Fig. 8. The implosion-proof tube now in mass productione Around the rim is a metal band filled with resin; around ths- cone, under the graphite layer, is resin reinforced with glass-fibre cloth.

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

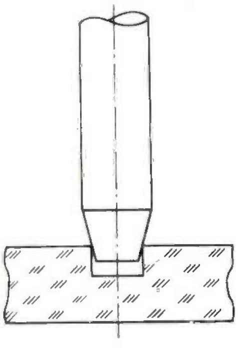


Fig. 9. Simulating the spontaneous fracture process by introducing small surface cracks in the tensile stress region. In the method shown here this is done by tapping lightly with a hammer on a tapered pin, placed in a blind hole drilled into the glass.

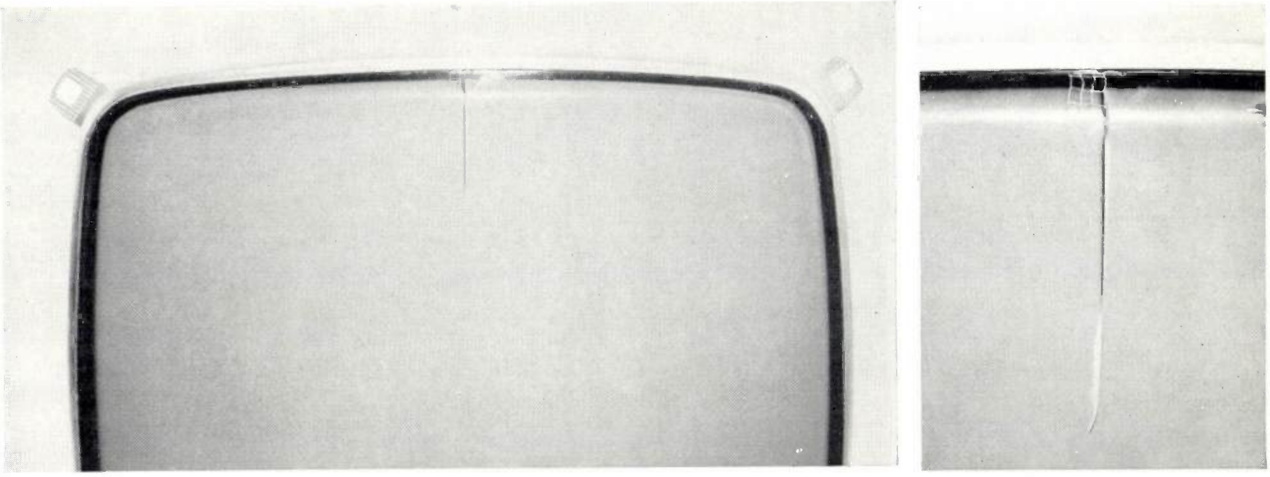


Fig. 10. The new tube with a "spontaneous" crack, simulated by producing surface abrasions near the rim (visible in the detail photograph on the right) and cooling this area rapidly with liquid nitrogen. The plane of fracture is usually slightly twisted. The fissure produced is so slight that it takes 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 setmaker. 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

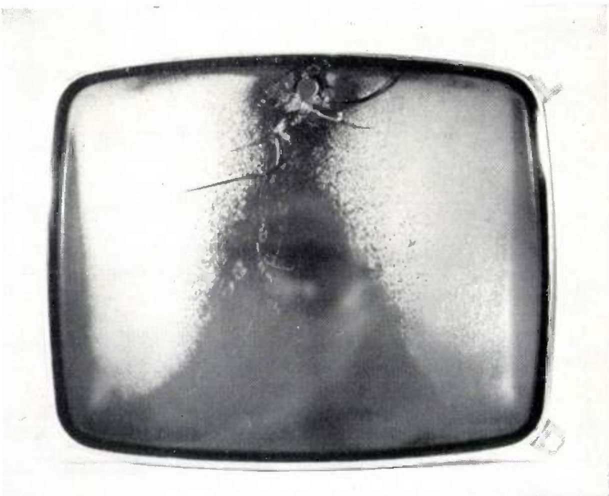


Fig. 11. Testing the new tube by the impact of a heavy steel ball on the tensile stress region of the bulb. The result is only a small hole and a few cracks in the face plate. Omission of the protective window is thus entirely warranted. The rapid ingress of air has blown away part of the fluorescent 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.

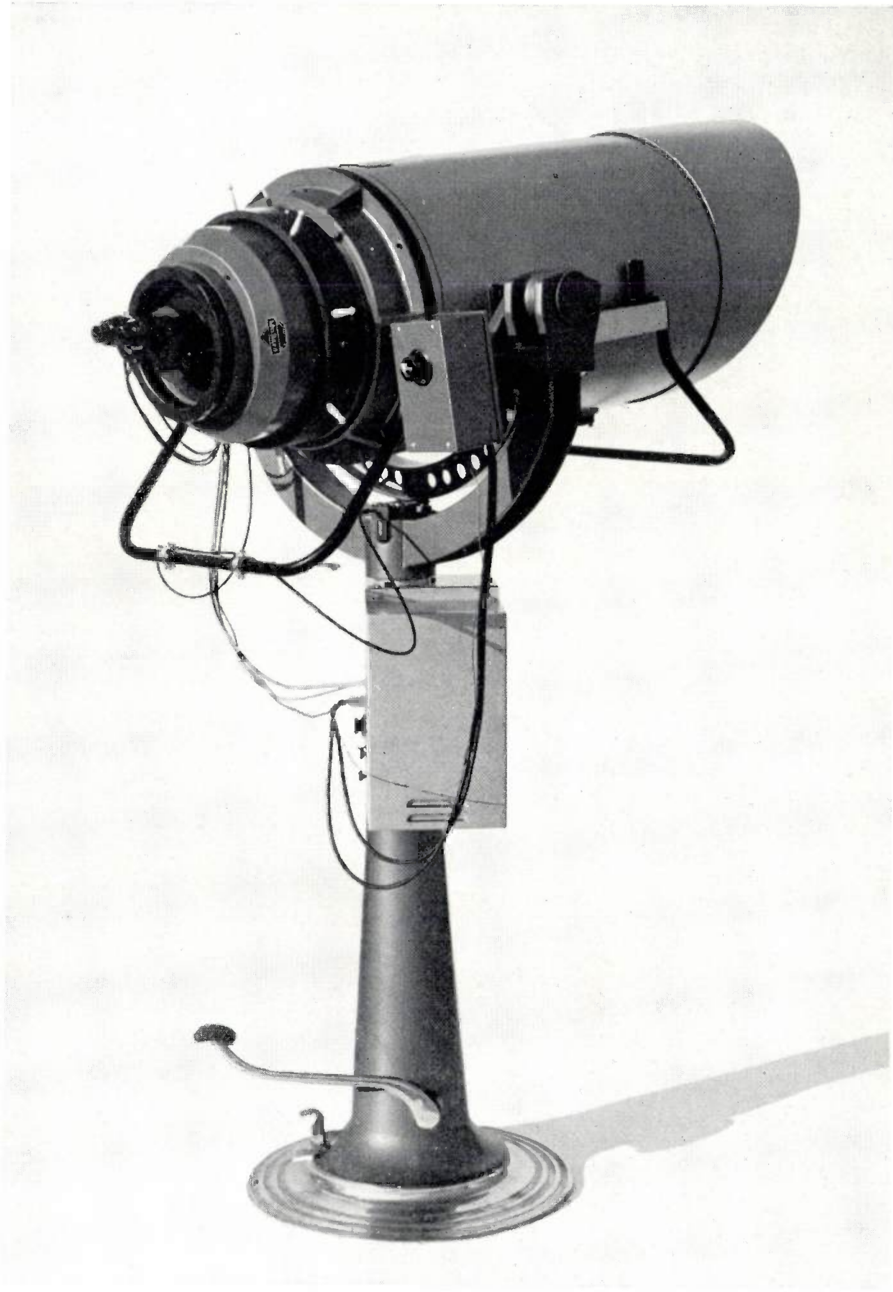


Fig. 1. Complete system for low light level vision developed at Mullard Research Laboratories. The system comprises an image intensifier tube with mirror type optics and a binocular microscope for observing the viewing screen. It is mounted in a clevis ring on a stand easily permitting adjustment of 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).

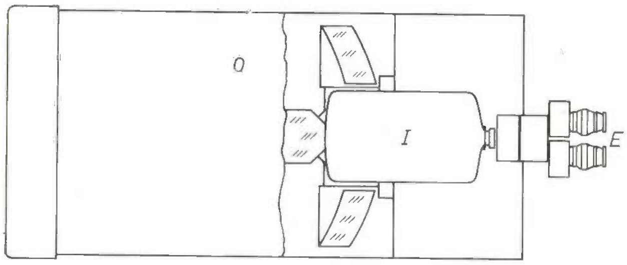


Fig. 2. Schematic diagram of the system. *O* mirror objective of 35 cm focal length and effective relative aperture $f: 1.0$. *I* image intensifier tube. *E* binocular viewer of magnification 13.5.

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]).

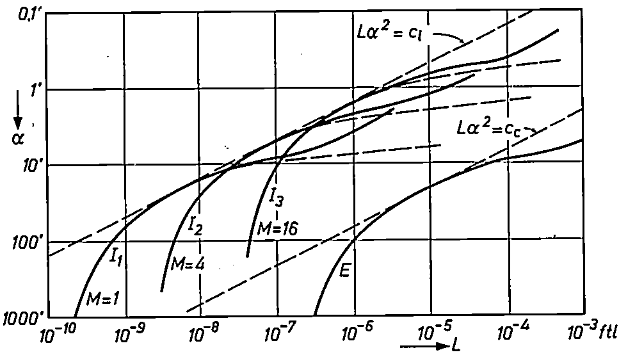


Fig. 3. Measured visual acuity curve E of the unaided eyes after complete adaptation [3]), compared with calculated curves for the eyes aided by an image intensifier system (of the type described but having unlimited resolving power) providing different overall magnifications: $M = 1$, $M = 4$ and $M = 16$. 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).

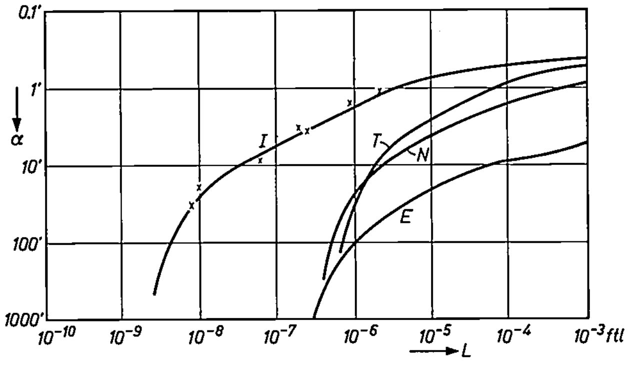


Fig. 4. Theoretical total performance curve I of the system for low light level vision, assuming optimum adjustment of magnification for each level. Points measured with our image intensifying system fit the curve very well. E acuity curve for the unaided eyes, T and N performance curve of the eyes aided 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.

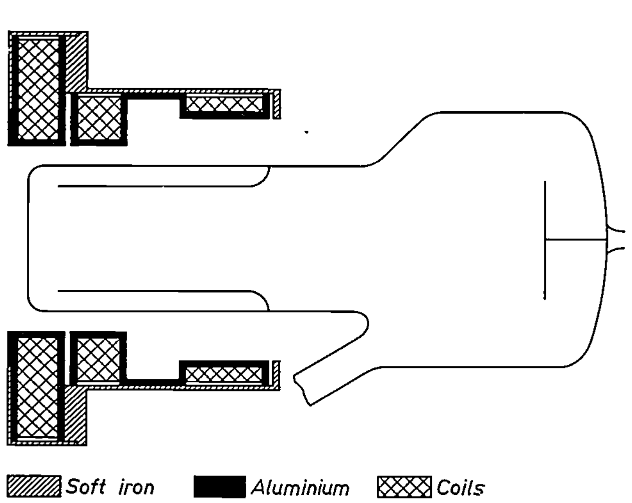


Fig. 5. Electron-optical system of the Philips image iconoscope, an early type of television camera tube, with magnetic focusing. By separately adjusting the currents through the three coil sections the overall magnification could be varied by a factor 2 while maintaining picture focus and picture rotation (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.

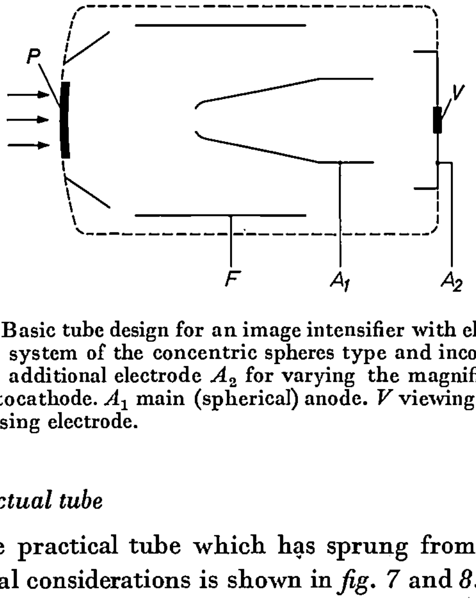


Fig. 6. Basic tube design for an image intensifier with electron-optical system of the concentric spheres type and incorporating an additional electrode A_2 for varying the magnification. P photocathode. A_1 main (spherical) anode. V viewing screen. F focusing electrode.

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.

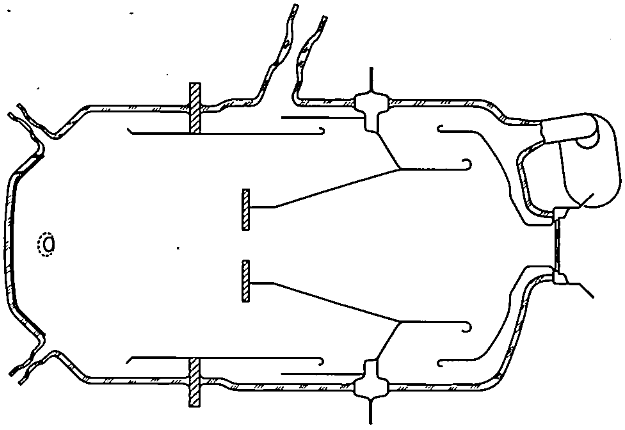


Fig. 7. Simplified cross-section of the tube in which the same elements as in fig. 6 can be found.

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.

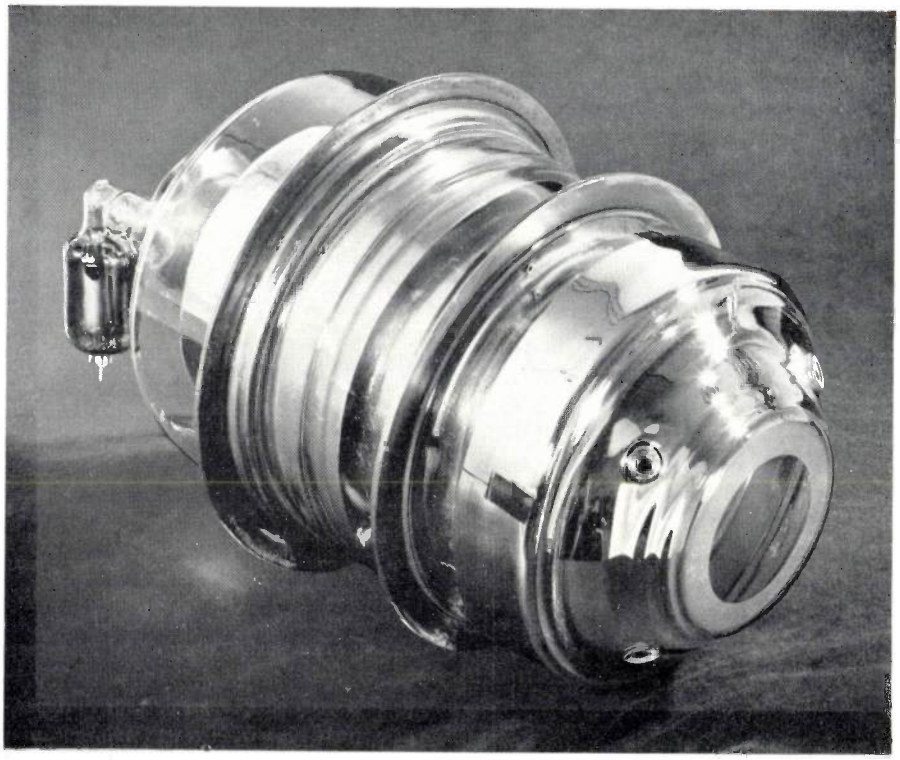


Fig. 8. Image intensifier tube for the low light level vision system developed at Mullard Research Laboratories. To the right the window with photocathode (6 cm diameter). The metal rings forming part of the tube envelope provide means for mounting the tube and at the same time for connecting the various 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.

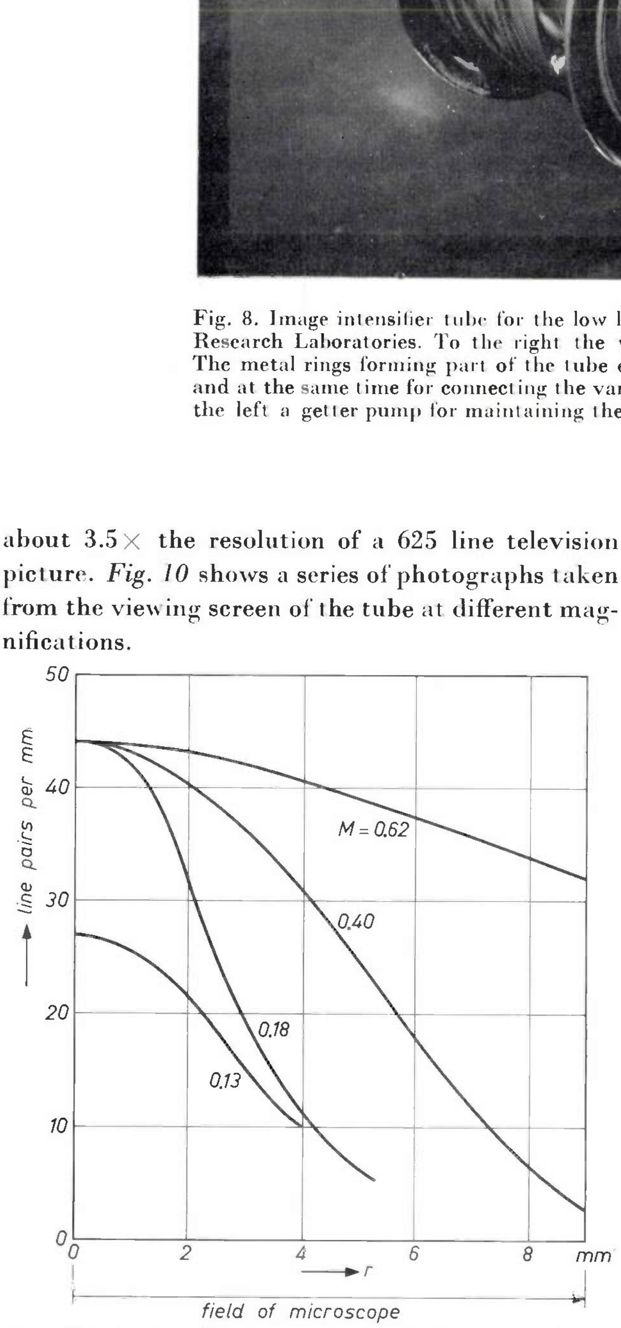


Fig. 9. Resolution of the image on the viewing screen plotted against the distance from the centre, for different magnifications.

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.

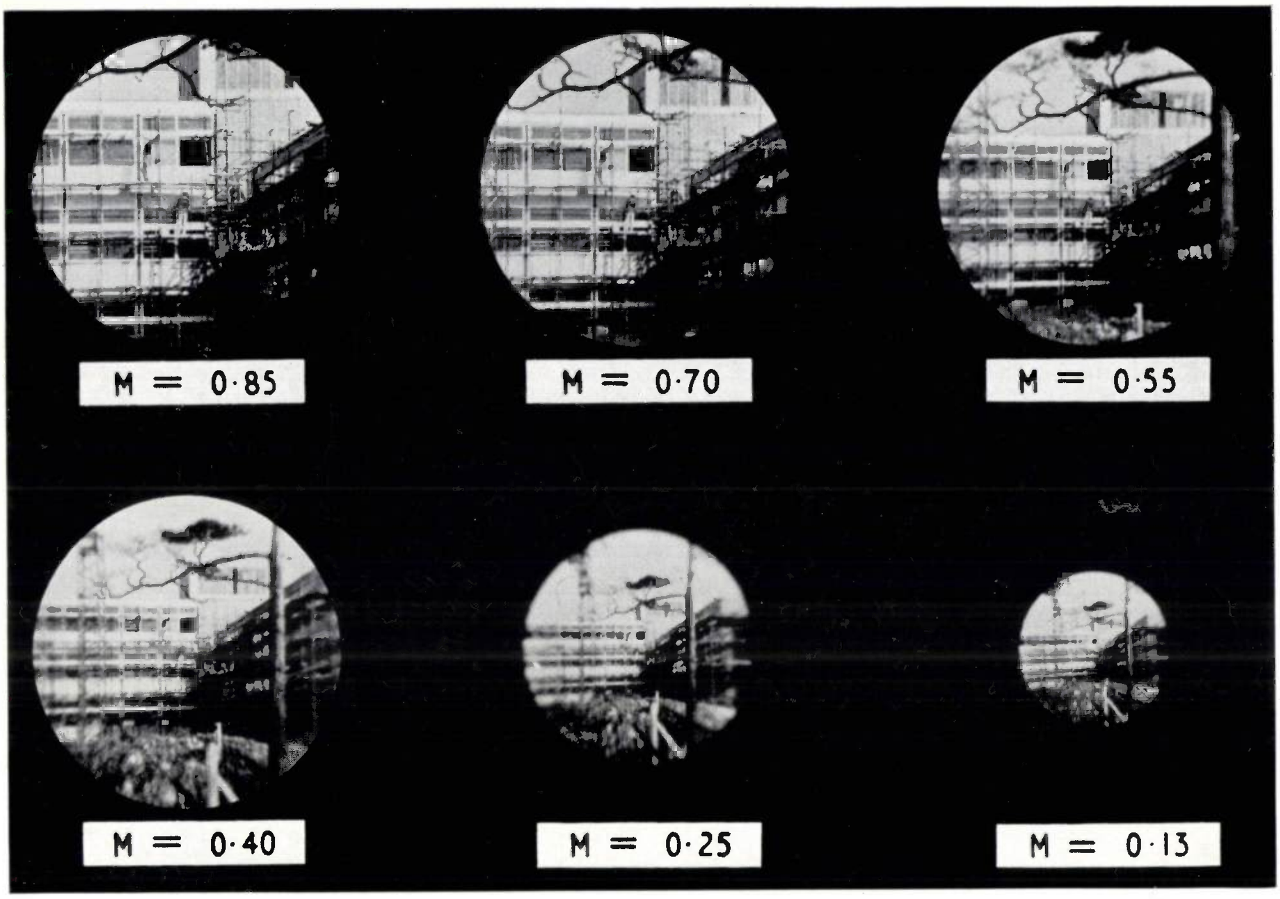


Fig. 10. Photographs taken from the viewing screen at different magnifications.

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.

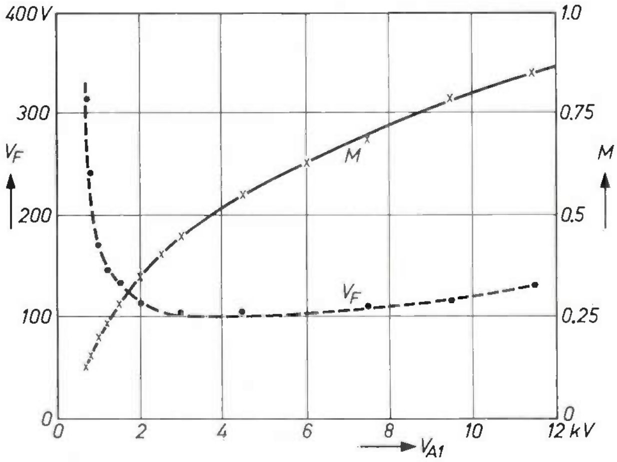


Fig. 11. Magnification M by the image intensifier tube, plotted against the voltage V_{A1}. The dotted curve represents the potential V_F required at the focus electrode in order to keep the image in focus.

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:

$$C_\mathrm{v} = \frac{C_{\mathrm{v}0}}{1+\hat{a}\cos pt}\,.$$

Here $C_{\mathrm{v}0}$ 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

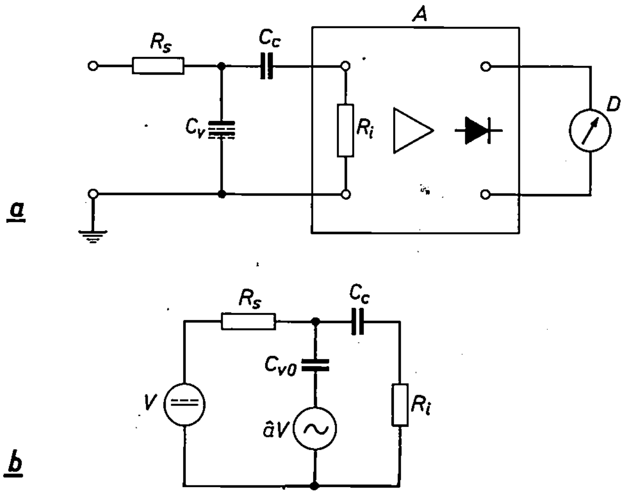


Fig. 1. *a*) Basic diagram of a vibrating-plate electrometer. C_v vibrating capacitor. C_c coupling capacitor. *A* AC voltage amplifier and rectifier. R_i input resistance of amplifier. *D* moving coil meter. R_s high series resistance.
b) Equivalent circuit, in which the vibrating capacitor is approximately represented by a fixed capacitor $C_{\mathrm{v}0}$ 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

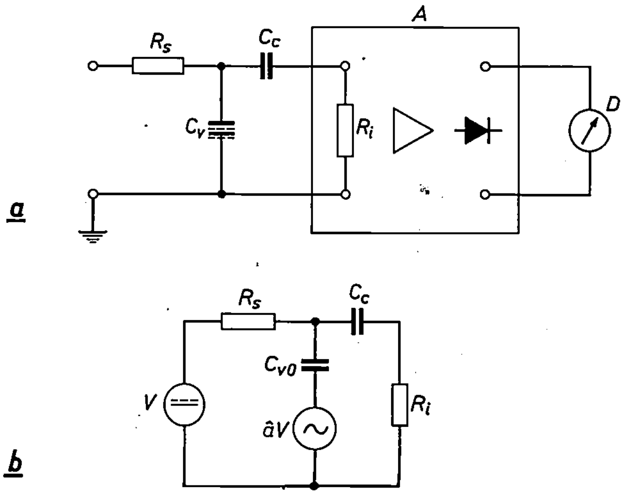


Fig. 1. *a*) Basic diagram of a vibrating-plate electrometer. C_V vibrating capacitor. C_c coupling capacitor. *A* AC voltage amplifier and rectifier. R_i input resistance of amplifier. *D* moving coil meter. R_s high series resistance.
b) Equivalent circuit, in which the vibrating capacitor is approximately 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:

$$v_V = \frac{q}{C_{V0}}(1 + \hat{a}\cos pt) \quad \ldots\ldots \quad (1)$$

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

$$pC_cR_s \gg 1, \quad \ldots\ldots \quad (2)$$

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.

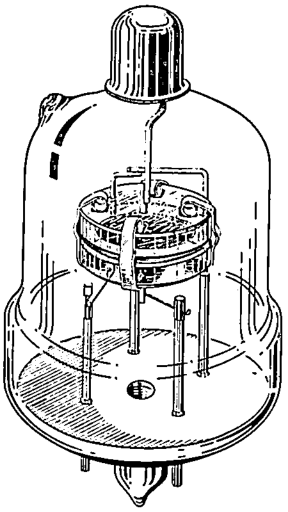


Fig. 2. The new vibrating capacitor, now produced by Philips' X-ray and Medical Apparatus Division. The vibrating capacitor is mounted in an evacuated bulb, which protects it from dust, moisture and other atmospheric influences. Roughly true size.

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).

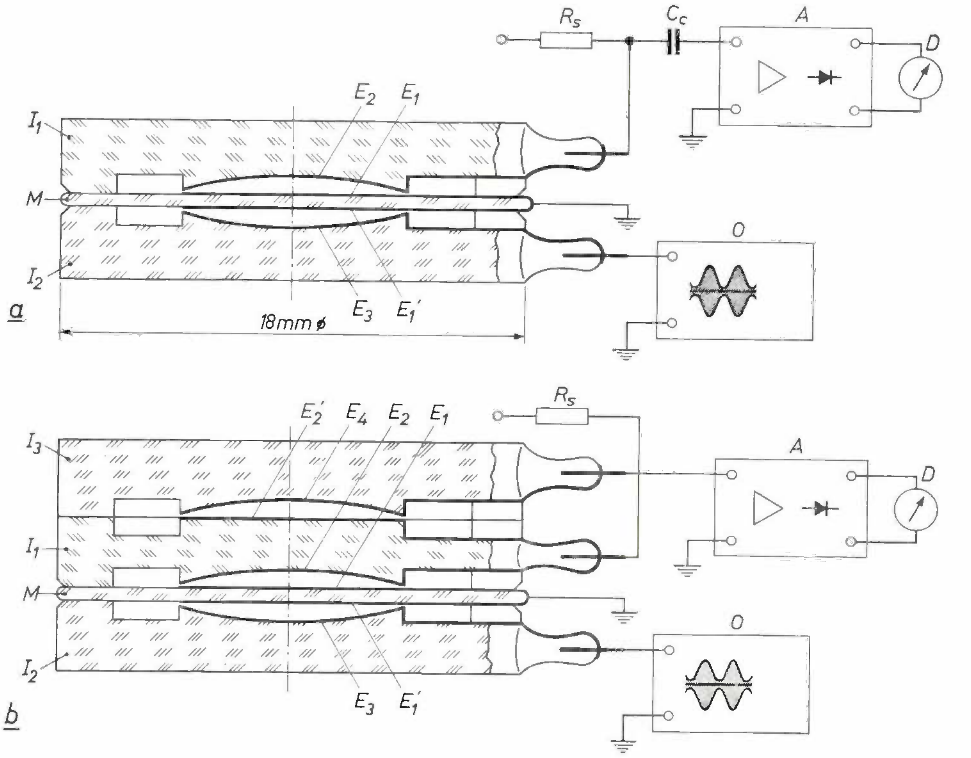


Fig. 3. *a*) Schematic cross-section of the vibrating capacitor. *M* round glass membrane (0.135 mm thick) the edge of which is clamped between glass insulators I_1 and I_2. The thick line indicates metal layers that serve as electrodes (E_1, E_1', E_2, E_3). Electrodes E_1 and E_2 form the vibrating 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.

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,

2) Low-frequency components with the angular frequencies *p* and 2*p*,

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 components. 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.

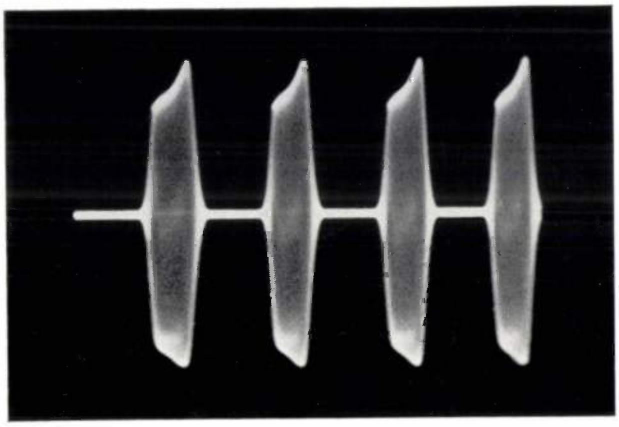


Fig. 4. Oscillogram of the voltage u_d of the oscillator *O* in fig. 3. The oscillator oscillates at certain intervals with a frequency of 1 Mc/s. The amplitude modulation is thus not sinusoidal but there is a periodic interruption. The repetition frequency is equal to the natural frequency of the membrane (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.

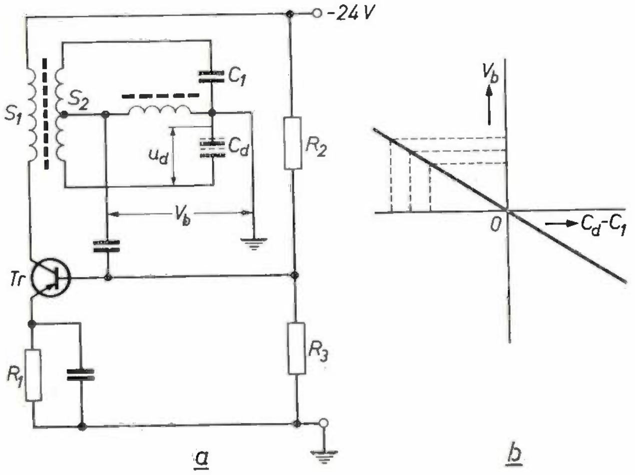


Fig. 5. *a*) Basic diagram of the oscillator. *Tr* transistor type OC 44. Of the coupled coils S_1 and S_2 the later is centre-tapped to form a bridge circuit with capacitors C_1 and C_d; C_d is formed by the electrodes E_1' and E_3 (fig. 3).
b) The diagonal voltage V_b in (*a*) as a function of the 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 $\alpha'_{\text{eff}} = 1/\beta$; 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

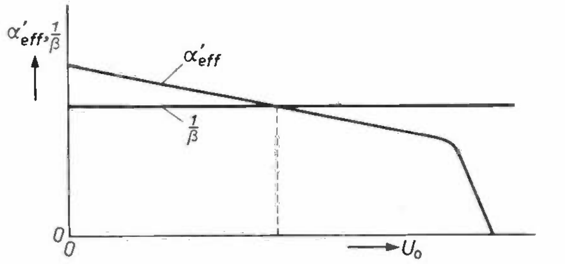


Fig. 6. Effective current amplification factor α'_{eff} of the transistor circuit in the oscillator and $1/\beta$ (β = feedback factor) as a function of the amplitude U_0 of the high-frequency voltage. The oscillation condition is satisfied at the amplitude at which $\alpha'_{\text{eff}} = 1/\beta$. 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 $\alpha'_{\mathrm{eff}} = 1/\beta$ 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.

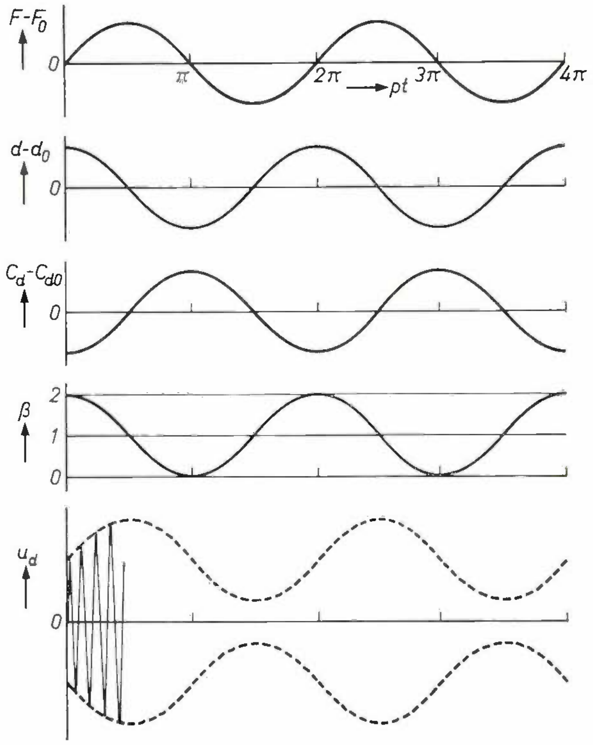


Fig. 7. The following are plotted as a function of pt:

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*).

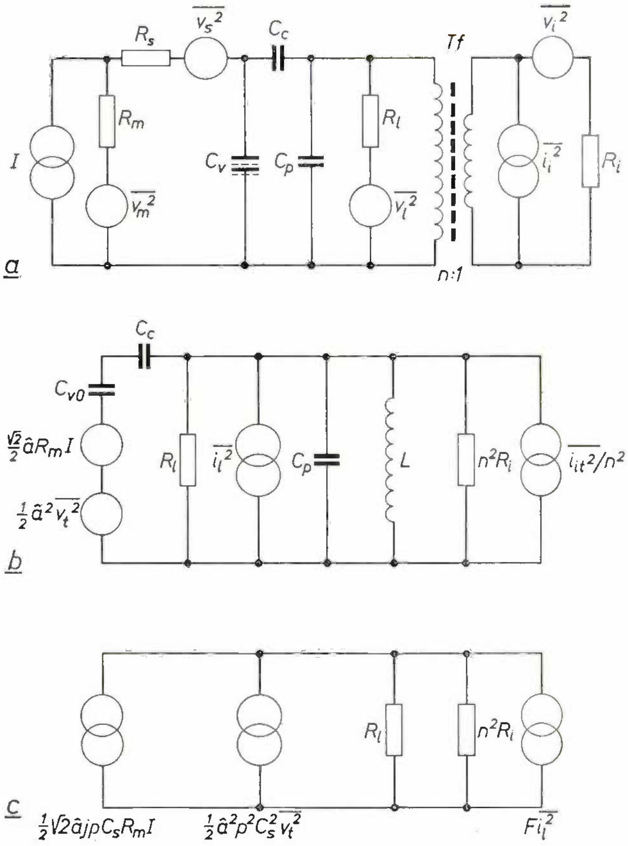


Fig. 8. *a*) First equivalent circuit of electrometer. The direct current I to be measured flows through the measuring resistance R_m. The amplifier has an input transformer *Tf* with a turns ratio $n:1$ and stray capacitance C_p: the resistance R_l represents the transformer losses. R_s, C_v, C_c and 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 $\overline{v_m^2}$, $\overline{v_s^2}$ and $\overline{v_l^2}$, the noise of the amplifier by the noise voltage source $\overline{v_i^2}$ and the noise current source $\overline{i_i^2}$.

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 $\overline{v_m^2}$ and $\overline{v_s^2}$ of (*a*) are combined to form a single noise-voltage source $\overline{v_t^2}$. The noise voltage source $\overline{v_l^2}$ is replaced by a noise current source $\overline{i_l^2}$. The resistance R_i on the secondary side is replaced by n^2R_i at the primary side, and the noise sources $\overline{i_i^2}$ and $\overline{v_i^2}$ at the secondary side by a single noise current source $\overline{i_{it}^2}/n^2$ 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 $\overline{i_l^2}$ and $\overline{i_{it}^2}/n^2$ in (*b*) are replaced by a single noise current source $F\,\overline{i_l^2}$, and the inductance L — by the addition of capacitance — is tuned to the frequency $p/2\pi$.

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

[3]) F. A. Muller, Het meten van stromen en spanningen met de trilplaatcondensator, Thesis Amsterdam 1951, p. 43 et seq.

source $\overline{v^2{}_\mathrm{m}}$. 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 $\overline{v_\mathrm{s}{}^2}$, $\overline{v_\mathrm{l}{}^2}$ and $\overline{v_\mathrm{i}{}^2}$ and the noise current source $\overline{i_\mathrm{i}{}^2}$ [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 $4kTBR_\mathrm{m}$ and a part $4kTBR_\mathrm{eq}$, 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:

$$R_\mathrm{eq} = R_\mathrm{s} + \frac{F}{\frac{1}{2}\hat{a}^2p^2C_\mathrm{s}{}^2R_\mathrm{l}} \quad . \; . \; . \; . \; . \quad (4)$$

Here C_s is the capacitance of C_v0 and C_c in series: $C_\mathrm{s} = C_\mathrm{v0}C_\mathrm{c}/(C_\mathrm{v0} + C_\mathrm{c})$, 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 $C_\mathrm{p} + C_\mathrm{s}$ already present (C_p being the stray capacitance) to give:

$$p^2L(C_\mathrm{p} + C_\mathrm{s} + C_\mathrm{e}) = 1 \, . \quad . \; . \; . \quad (5)$$

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 $v_\mathrm{l}{}^2$, $v_\mathrm{i}{}^2$ and $i_\mathrm{i}{}^2$ 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	$\overline{v_\mathrm{m}}^2$ and $\overline{v_\mathrm{s}{}^2}$.
From $\dfrac{p}{2\pi} - \frac{1}{2}$ to $\dfrac{p}{2\pi} + \frac{1}{2}$ c/s	$\overline{v_\mathrm{m}{}^2}$, $\overline{v_\mathrm{s}{}^2}$, $\overline{v_\mathrm{l}{}^2}$ (of these three $\overline{v_\mathrm{l}{}^2}$ is the most important, since $R_\mathrm{l} < R_\mathrm{m}+R_\mathrm{s}$), $\overline{v_\mathrm{i}{}^2}$ and $\overline{i_\mathrm{i}{}^2}$.

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 $\frac{1}{2}\sqrt{2}\,\hat{a}R_\mathrm{m}\,I$ 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 $\frac{1}{2}\hat{a}^2\,\overline{v_\mathrm{t}{}^2}$. The input impedance R_i transformed to the primary side is n^2R_i and $\overline{i_\mathrm{it}{}^2/n^2}$ appears in the p lace of the input noise sources $\overline{i_\mathrm{i}{}^2}$ and $\overline{v_\mathrm{i}{}^2}$.

According to the definition, the noise factor F is given by:

$$F = \frac{\overline{i_\mathrm{l}{}^2} + \overline{i_\mathrm{it}{}^2}/\mathrm{n}^2}{\overline{i_\mathrm{l}{}^2}} \, .$$

This yields:

$$\overline{i_\mathrm{l}{}^2} + \overline{i_\mathrm{it}}{}^2/n^2 = F\,\overline{i_\mathrm{l}{}^2} .$$

The second step consists among other things in replacing the parallel noise-current sources $\overline{i_\mathrm{l}{}^2}$ and $\overline{i_\mathrm{it}}{}^2/n^2$ by one single noise-current source $F\,\overline{\mathrm{i}_\mathrm{l}}{}^2$, 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 $\frac{1}{2}\sqrt{2}\hat{a}R_\mathrm{m}I$ and $\frac{1}{2}\hat{a}\overline{v_\mathrm{t}{}^2}$ in fig. 8*b* are replaced in fig. 8*c* by the equivalent current sources $\frac{1}{2}\sqrt{2}\,\hat{a}\mathrm{j}pC_\mathrm{s}R_\mathrm{m}I$ and $\frac{1}{2}\hat{a}^2p^2C_\mathrm{s}{}^2\overline{v_\mathrm{t}}{}^2$ 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*:

$$\frac{S}{N} = \frac{\frac{1}{2}\hat{a}^2p^2C_\mathrm{s}{}^2R_\mathrm{m}{}^2I^2}{\frac{1}{2}\hat{a}^2p^2C_\mathrm{s}{}^2\overline{v_\mathrm{t}{}^2} \;+\; F\overline{i_\mathrm{l}{}^2}} . \quad . \; . \; . \; . \; . \quad (6)$$

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:

$$\begin{aligned} N &= \tfrac{1}{2}\hat{a}^2p^2C_\mathrm{s}{}^2\overline{v_\mathrm{t}{}^2} \;+\; F\,\overline{i_\mathrm{l}{}^2} = \\ &= 4kTB \left\{ \tfrac{1}{2}\hat{a}^2p^2C_\mathrm{s}{}^2(R_\mathrm{m} + R_\mathrm{s}) + \frac{F}{R_\mathrm{l}} \right\} = \\ &= 2kTB\hat{a}^2p^2C_\mathrm{s}{}^2 \left(R_\mathrm{m} + R_\mathrm{s} + \frac{F}{\frac{1}{2}\hat{a}^2p^2C_\mathrm{s}{}^2R_\mathrm{l}} \right) \quad . \; . \; . \quad (7) \end{aligned}$$

This leads to:

$$R_\mathrm{eq} = R_\mathrm{s} + \frac{F}{\frac{1}{2}\hat{a}^2p^2C_\mathrm{s}{}^2R_\mathrm{l}} ,$$

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 $C_\mathrm{s}(\approx \frac{1}{2}C_\mathrm{v0} \approx \frac{1}{2}C_\mathrm{c})$ 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:

$$\frac{S}{N} = \frac{R_m}{R_m + R_{eq}} \frac{R_mI^2}{4kTB} . \quad . \quad . \quad (8)$$

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.

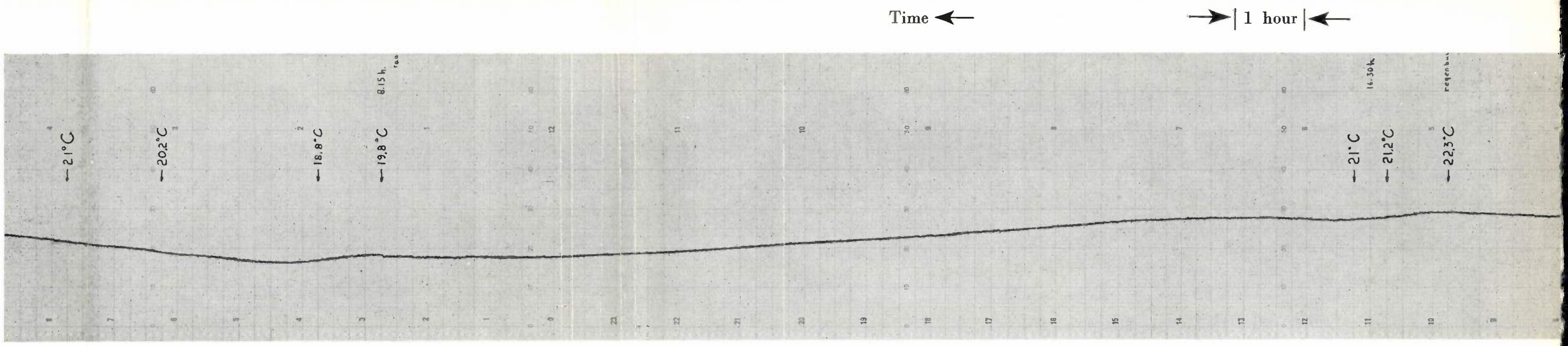


Fig. 9

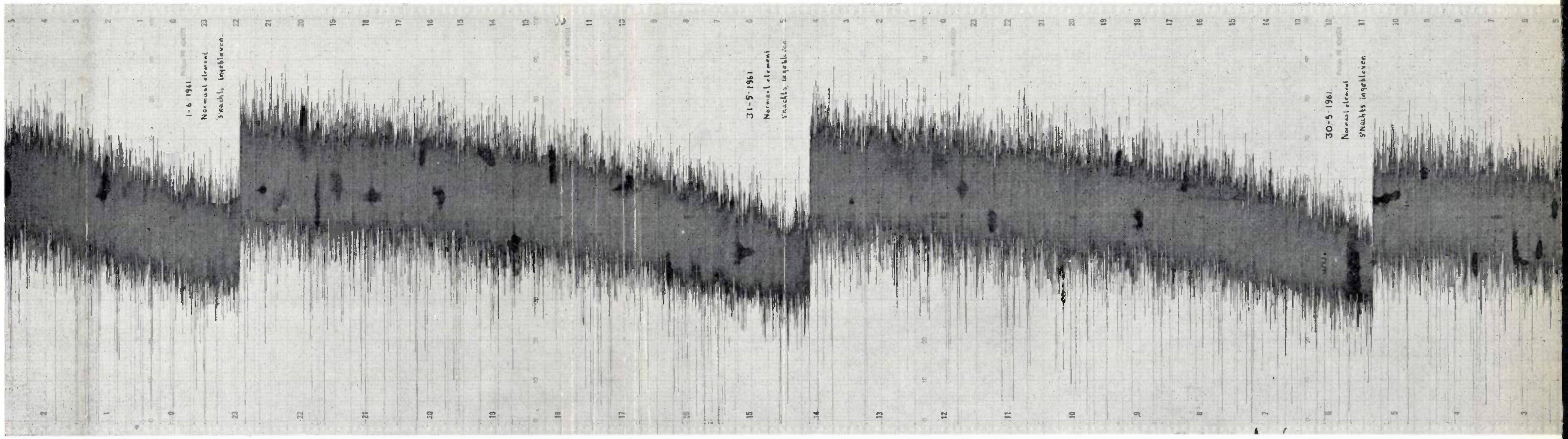


Fig. 11*a*

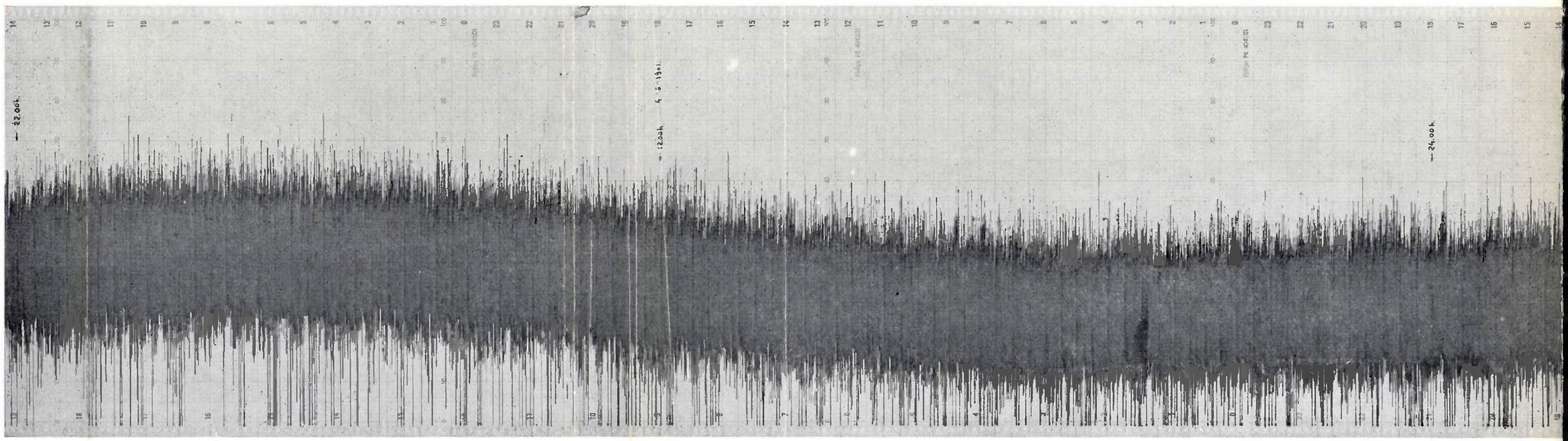


Fig. 11*b*

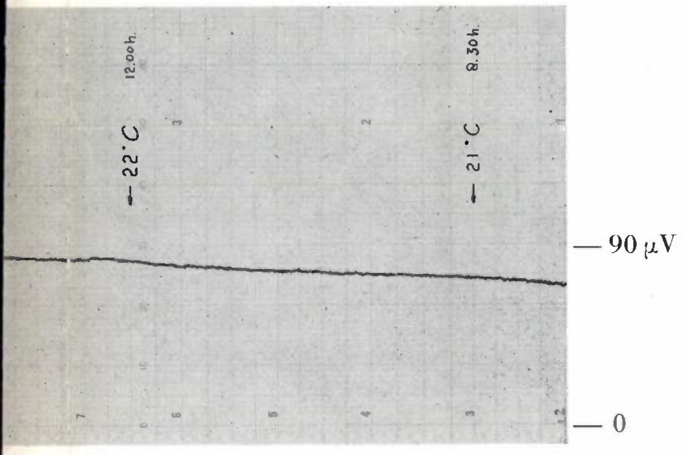


Fig. 9. Recording (during a period of more than 36 hours) of the output voltage with short-circuited input (R_s = 100 MΩ parallel with C_v). The noise is scarcely visible. The slight drift of the zero point is due to fluctuation of the ambient temperature and amounts to approximately 15 μV/°C.

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.

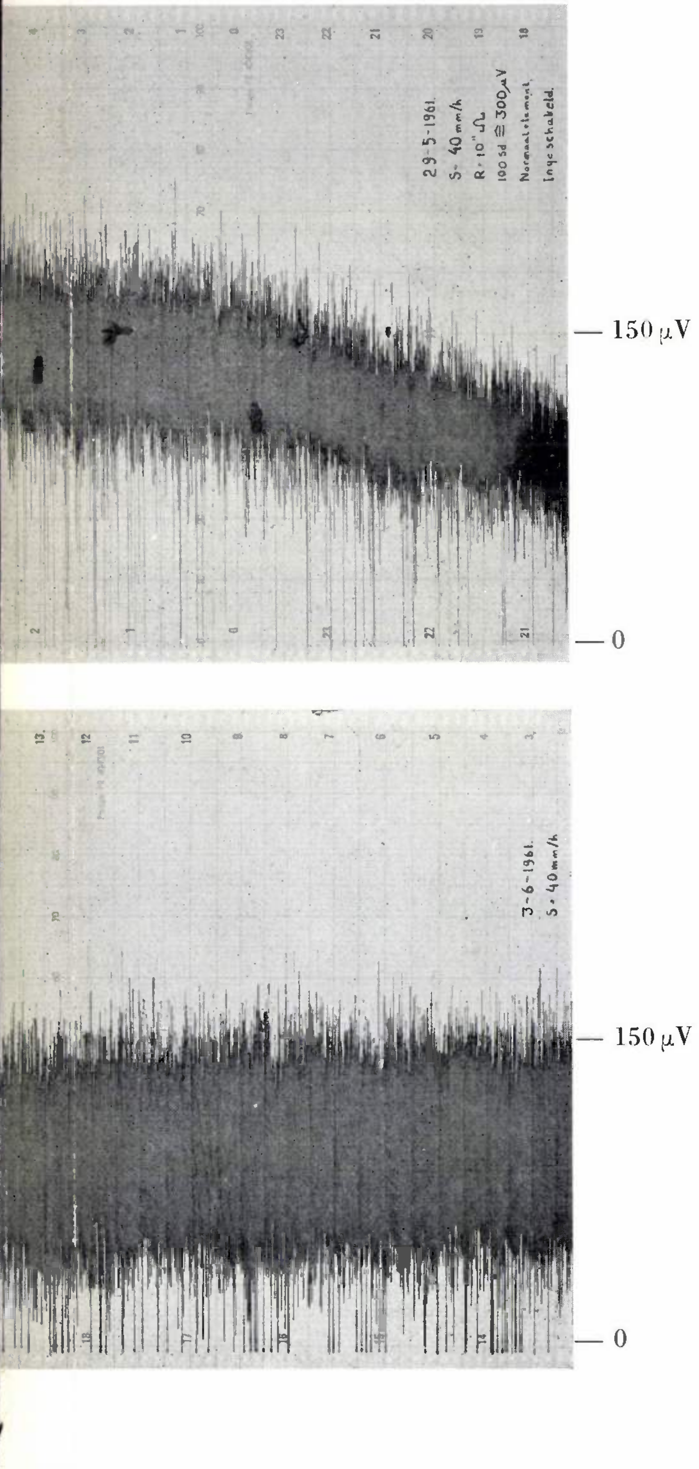


Fig. 11. Recording made while the input was connected to a measuring resistance R_m of 10^{11} ohms; this explains the much stronger noise than in fig. 9.

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.

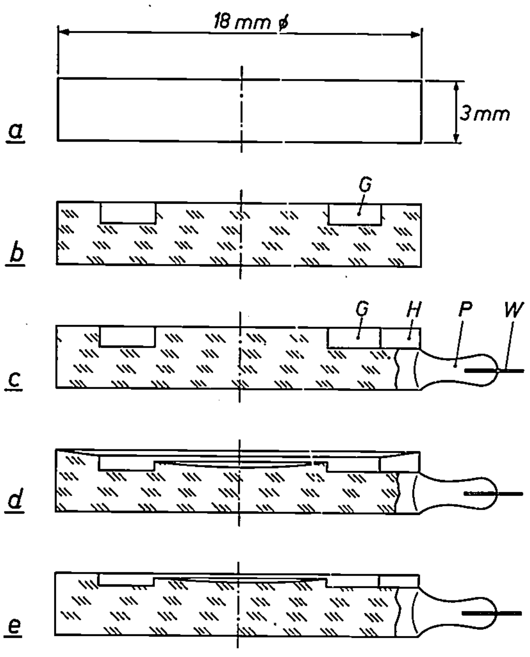


Fig. 10. Stages in the fabrication of one of the insulators (I_1 or I_2, fig. 3*a*).
a) Centreless ground hard-glass disc.
b) A concentric groove *G* is cut.
c) In the border a radial slot *H* is sawn, and on the rim a glass bead *P* with tungsten contact pin *W* is fused [8]).
d) One face is hollow- 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 $R_\mathrm{m} = 10^{11}$ 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 $R_\mathrm{m} = 10^{11}$ 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 $8\alpha,10\alpha$-progesterone and $8\alpha,10\alpha$-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.

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,

*) Philips' Research Laboratories, Eindhoven.

[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:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -a_1 y\,, \quad \frac{\mathrm{d}y}{\mathrm{d}t} = -a_2 x\,.$$

It is easy to eliminate $\mathrm{d}t$ from these equations. If, moreover, we suppose $a_1 = a_2$, then we get the simple equation:

$$x\,\mathrm{d}x = y\,\mathrm{d}y\,,$$

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 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.

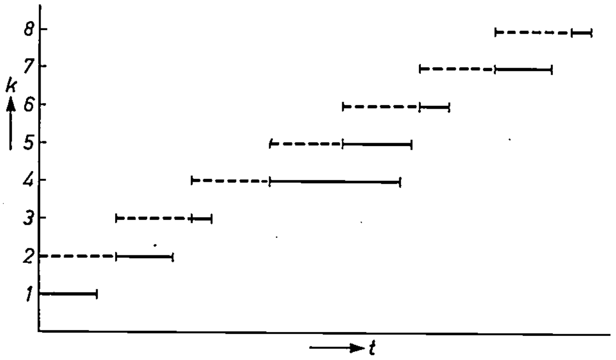


Fig. 1. Diagram of the arrival of orders ("jobs", with serial number k) at constant intervals (broken lines), but with various processing-times (continuous lines).

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:

$$p(t)\mathrm{d}t = m\,\mathrm{e}^{-mt}\,\mathrm{d}t\,. \quad \ldots\ldots \quad (1)$$

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

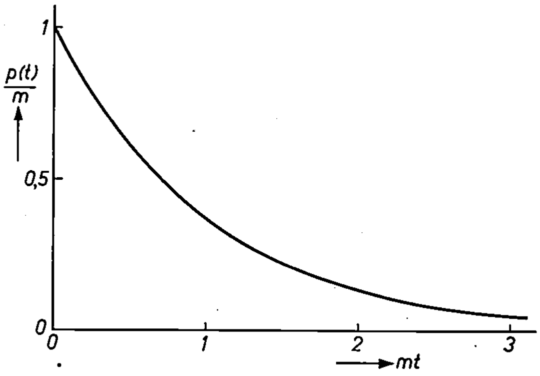


Fig. 2. For the occurrence of m random events per unit of time the probability $p(t)\mathrm{d}t$ that, starting from an arbitrary moment, the next event will take place in the interval between t and $t + \mathrm{d}t$ is given by $m\mathrm{e}^{-mt}\mathrm{d}t$. This distribution of probability is a special 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: $P(dt,1) = mdt$, 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 $\overline{w}$ is given by the formula of Pollaczek-Khintchine [6]):

$$\frac{\overline{w}}{\overline{t}} = \frac{\varrho}{2(1-\varrho)}\left(1 + \frac{\sigma^2}{\overline{t}^2}\right), \quad . \; . \; . \quad (2)$$

in which $\overline{t}$ is the average processing-time, ϱ the "utilization factor" of the station, i.e. the ratio between $\overline{t}$ and the average interval ($1/m$) between the arrivals, and σ the variance in the processing-times. This variance is defined by:

$$\sigma^2 = \int_0^\infty \varphi(t)\,(t-\overline{t})^2\,dt,$$

in which $\varphi(t)dt$ is the fraction of the processing-times lying between t and $t + dt$.

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 ($\varrho \equiv \overline{t}m = 1$) 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 $\overline{t} = 1/n$, 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 $\sigma = \overline{t}$. Hence formula (2) changes into the simpler form:

$$\frac{\overline{w}}{\overline{t}} = \frac{\varrho}{1-\varrho},$$

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.

[6]) See, for example, D. G. Kendall, Some problems in the theory of queues, J. Roy. Statist. Soc. **B 13**, 151-173, 1951.

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:

$$Q^* = \sqrt{\frac{2DF}{C_i}}. \quad \dots\dots \quad (3)$$

This formula was derived by Camp as long ago as 1922, but aroused little interest at the time [2]).

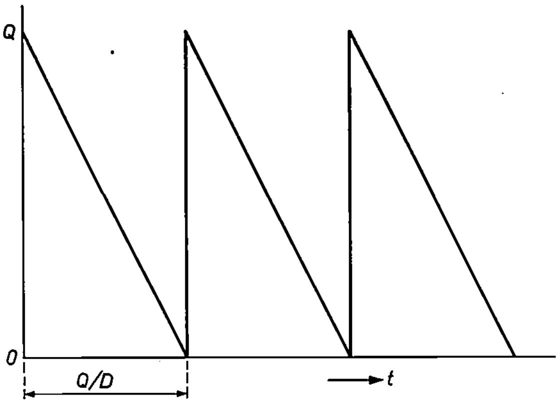


Fig. 3. The fluctuation of the stock of a product, as a function of the time, where a constant demand D is met from the stock and a series of a constant size Q is immediately manufactured every time the stock has been exhausted.

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:

$$\left.\begin{aligned} &0 \leqslant x \leqslant 4;\\ &0 \leqslant y \leqslant 3;\\ &1 \leqslant x + y \leqslant 4.\end{aligned}\right\} \quad . \; . \; . \; . \; . \; . \; . \; . \quad (4)$$

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:

$$\begin{aligned} E &= x(2b - a) + by + 4a + b \\ &= b\{(2 - a/b)\,x + y + 1 + 4a/b\}\,. \quad . \; . \quad (5)\end{aligned}$$

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.

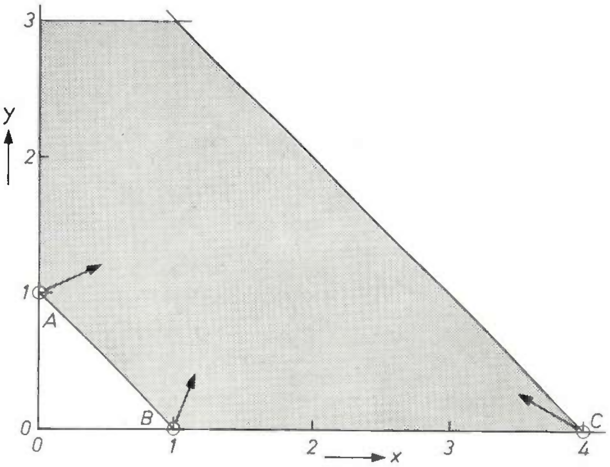


Fig. 4. Graph illustrating the process of linear programming. The aim is to distribute a certain production over three periods with various demands such that the sum of the additional cost of overtime (a per product) and that of maintaining a stock (b per product per period) is a minimum.

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 $(\mathrm{d}y/\mathrm{d}x)$ 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

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{2 - a/b}\,. \quad . \; . \; . \; . \; . \quad (6)$$

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]). 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 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*.

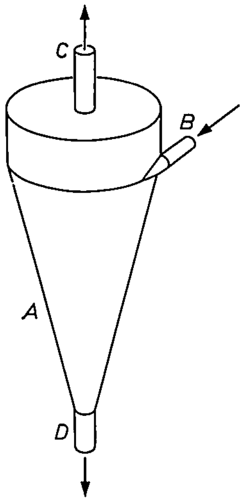


Fig. 1. Hydrocylone, used to remove the fine particles from the phosphor. The powder, mixed with a large amount of water, is pumped through the tube *B* into the vessel *A* (height e.g. 10 cm). The particles are separated into two fractions by a complicated interplay of forces in the body of water rotating 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*.

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

*) 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.

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.

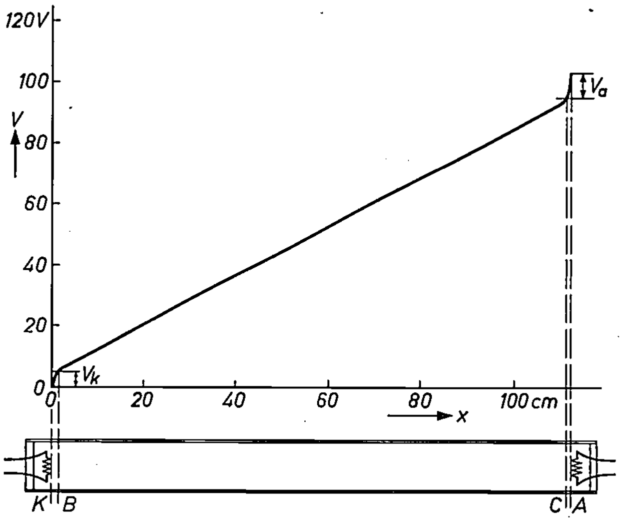


Fig. 2. The variation of the potential V in a 40 W "TL" lamp with the distance x to the cathode K. A is the anode. In the range BC (the positive column), where the desired mercury radiation is produced, the variation is linear so that the field strength is independent of x. Relatively large voltages 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 interrelated. 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.

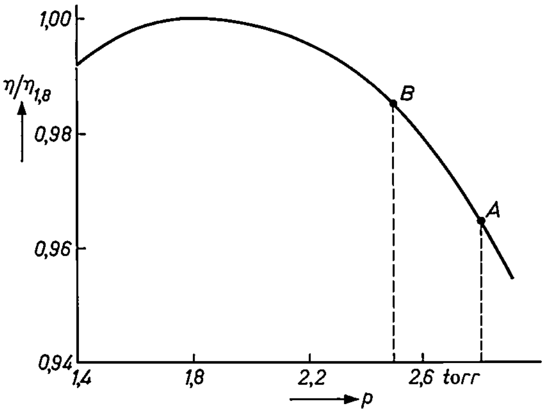


Fig. 3. The luminous efficiency η, expressed as a fraction of the maximum value $\eta_{1.8}$, as a function of the pressure p of the argon filling. By reducing the pressure from the normal value of 2.8 torr (point A) to 2.5 torr (B), the luminous efficiency has been increased by about 2%.

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

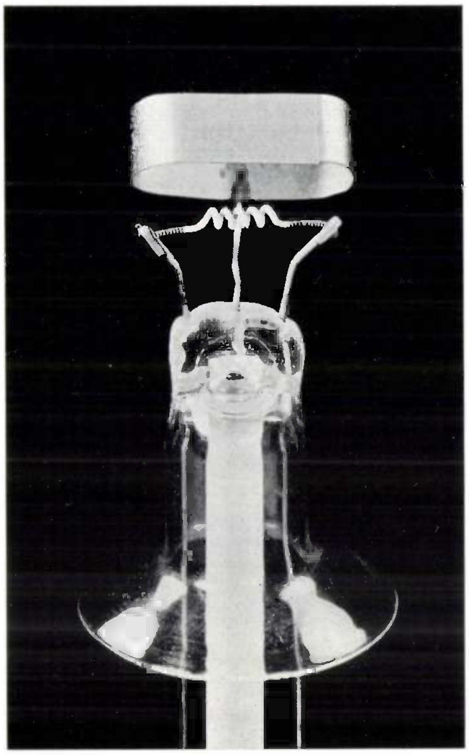


Fig. 4. If the electrode coil is surrounded by an insulated flat metal ring, the "electrode losses" are found to decrease. The luminous efficiency thus increases by about 2%. — In order to make the coil itself visible in the photo, the ring has been bent backwards. Magnification about 1.5 ×.

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

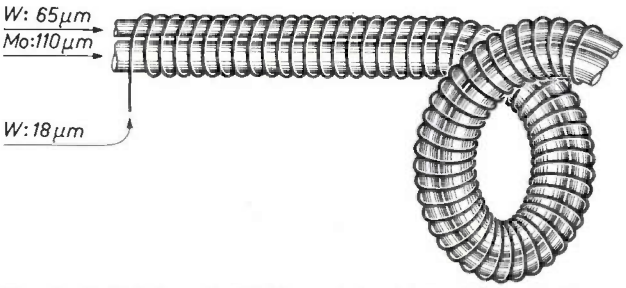


Fig. 5. A "triply coiled" filament is obtained by winding a compound "wire" into a coiled coil. The compound wire consists of a tungsten wire (diameter 65 μm) and a molybdenum wire (110 μm) placed side by side, and a very thin (18 μm) tungsten wire wound round both of them. The molybdenum wire is then 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.

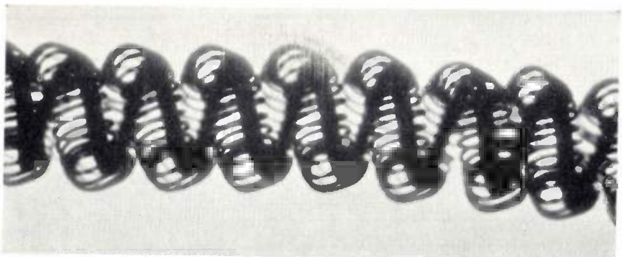


Fig. 6. Photo of part of a triple coil not yet coated with emitting material. Magnification 10×. (This picture shows a straight part of the coil.)

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.

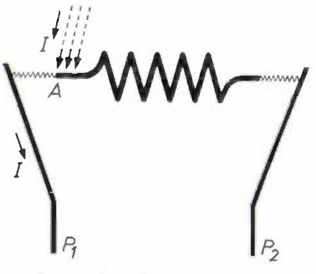


Fig. 7. In a new electrode, the discharge is concentrated at the junction *A* between the part of the coil coated with emitting material and the uncoated part (see also fig. 4) at the side of the pole P_1 through which the discharge current *I* leaves the lamp. This is because the current through a burning 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

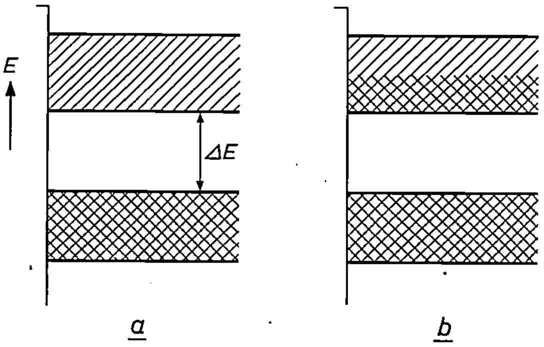


Fig. 1. Energy band scheme of an insulator (*a*) and of a metal (*b*). The electron energy E is set out vertically; the top of the ordinate axis gives the energy of an electron which is just outside the material. The horizontal coordinate has no physical meaning here, but may be used to represent a space 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*.

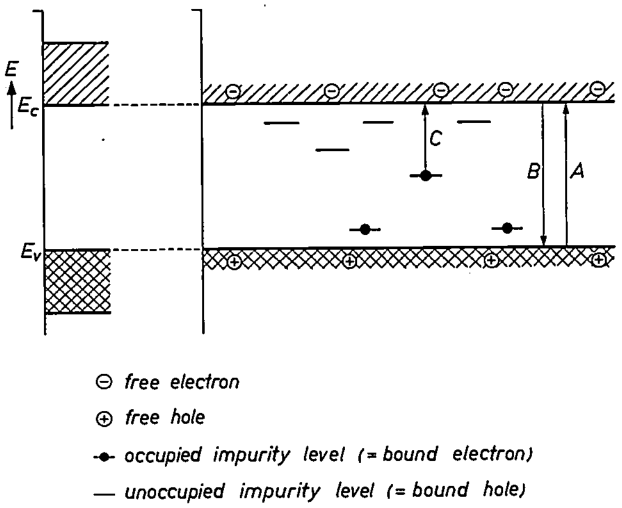

E
E_c
E_v
C
B
A
⊖ free electron
⊕ free hole
-●- occupied impurity level (= bound electron)
— unoccupied impurity level (= bound hole)

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.

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*).

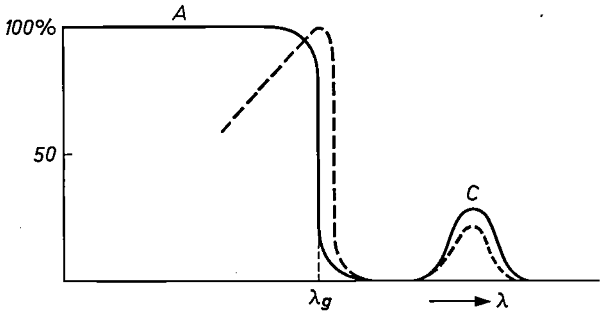


Fig. 3. Light absorption (solid line) and photocurrent (dotted line) as a function of the wavelength of the light incident on a photoconductor. The maximum value is put at 100% for both. The region *A* of the absorption spectrum corresponds to the excitation of electrons out of the valence band (transition *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).

[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.

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:

$$f(E) = \frac{1}{1 + \exp\,[(E - E_{\mathrm{F}})/kT]} . \qquad (\mathrm{I},1)$$

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 $E < E_{\mathrm{F}}$ 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 $E = E_{\mathrm{F}}$. The width of this region is greater the higher the value of T (*fig. 4a*). For $E = E_{\mathrm{F}}$ 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-

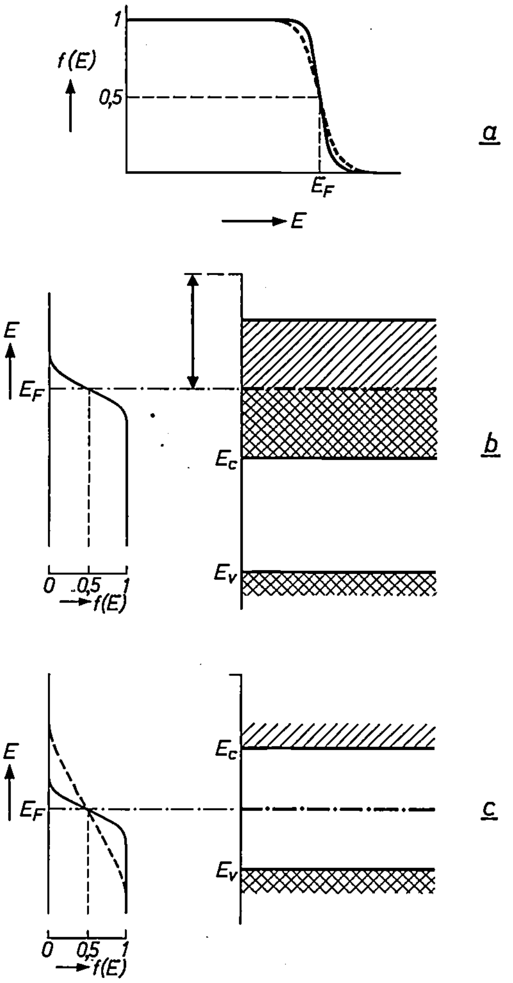


Fig. 4. *a*) The occupation $f(E)$ of the energy levels in accordance with the Fermi-Dirac distribution function. Except in a narrow transition region around $E = E_{\mathrm{F}}$, all levels with $E < E_{\mathrm{F}}$ are occupied by electrons and all levels with $E > E_{\mathrm{F}}$ are unoccupied. The 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 $E_{\mathrm{c}} - E_{\mathrm{v}}$ 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 cm^3 — 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:

$$n = N_c \exp\,[-(E_c-E_F)/kT], \quad . \; . \; . \quad (I, 2a)$$

and

$$p = N_v \exp\,[-(E_F-E_v)/kT]. \quad . \; . \; . \; . \quad (I,2b)$$

From this it follows that:

$$np \sim N_c N_v \exp\,[-(E_c-E_v)/kT]. \quad . \; . \; . \quad (I, 3)$$

The quantity N_c is given by the relation:

$$N_c = 2\left(\frac{2\pi\, m_n^*\, kT}{h^2}\right)^{3/2}. \quad . \; . \; . \; . \quad (I, 4)$$

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

$$G = \Delta n/\tau, \text{ or } \Delta n = G\tau. \quad . \; . \; . \quad (I, 5)$$

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:

$$v = -\mu F. \quad . \; . \; . \; . \; . \; . \quad (I, 6)$$

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 $-\Delta nev$, follows from formulae (I,5) and (I,6):

$$j_f = -\Delta nev = \mu\tau eFG \quad . \; . \; . \; . \; . \quad (I, 7)$$

(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:

$$D = \frac{kT}{e}\,\mu\,. \qquad \text{(I, 8)}$$

If we consider an electron current flowing in one particular direction (x), the complete current equation is thus:

$$\Delta j_{\mathrm{n,x}} = \Delta n\, e\, \mu_{\mathrm{n}} F_{\mathrm{x}} + e D_{\mathrm{n}} \frac{\mathrm{d}(\Delta n)}{\mathrm{d}x}\,. \qquad \text{(I, 9)}$$

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-

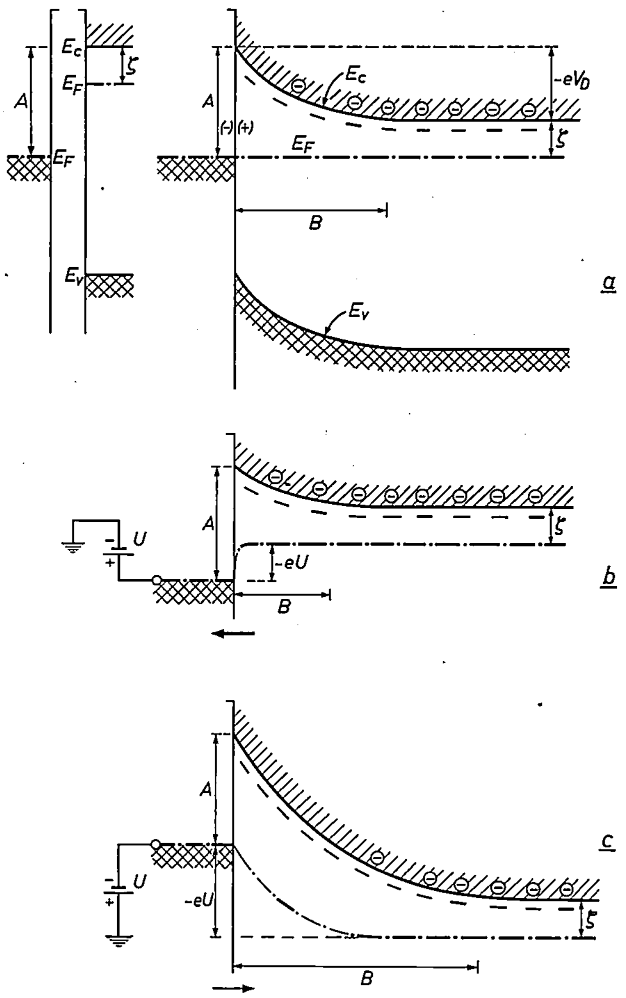


Fig. 5. *a*) Band scheme of the contact between a metal and a semiconductor for the case where the thermionic work function of the semiconductor is smaller than that of the metal, i.e. where the Fermi level in the semiconductor is higher before the contact is made (see sketch on the left). When the contact 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:

$$\frac{d^2V}{dx^2} = -\frac{\varrho}{\varepsilon} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (I,10)$$

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 barrier 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:

$$j_n = \text{const.}\ \exp[eU/kT - 1] . \quad \text{(I, 11)}$$

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 have no blocking action, although of course its resistance is not completely independent of the direction of the current.

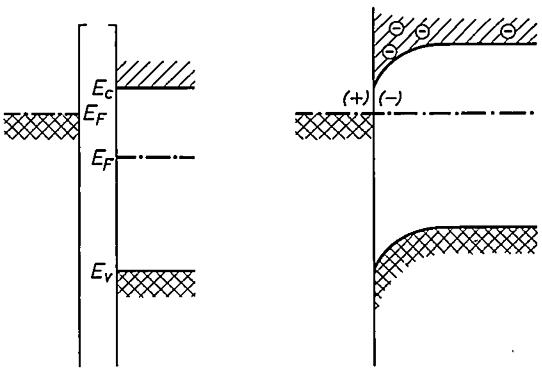


Fig. 6. Band scheme of the contact of a metal and an N-type semiconductor which has a *higher* thermionic work function (lower Fermi level) than the metal. The sketch on the left shows the situation before the contact is made. In the semiconductor a negative space charge is formed, which is concentrated in a thin layer. There is no potential barrier.

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 described 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^+)

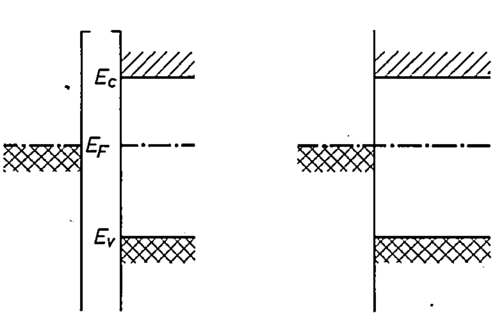


Fig. 7. Band scheme of a metal and a semiconductor with *identical* thermionic work functions (neutral contact). The establishment of the contact causes no change either in the metal or in the semiconductor.

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.

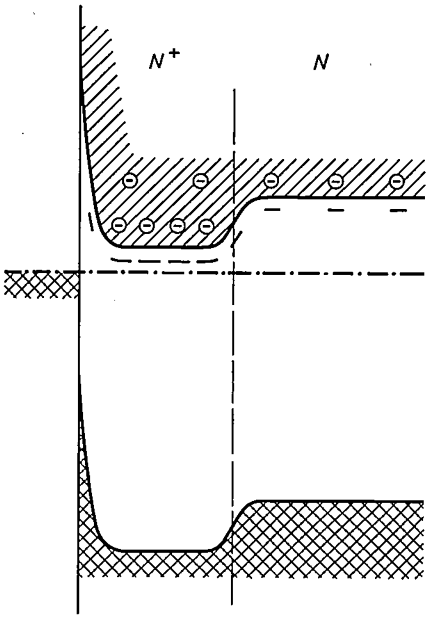


Fig. 8. Band scheme of an injecting contact, produced from a blocking contact by causing atoms to diffuse from the metal into the semiconductor. If these atoms can act as donors, a large concentration of free electrons is obtained in the boundary layer. At the position of the vertical dashed line the 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:

$$\eta = \frac{\tau_\mathrm{n}}{T_\mathrm{n}} + \frac{\tau_\mathrm{p}}{T_\mathrm{p}}, \quad . \quad . \quad . \quad . \quad (\mathrm{I}, 12)$$

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.

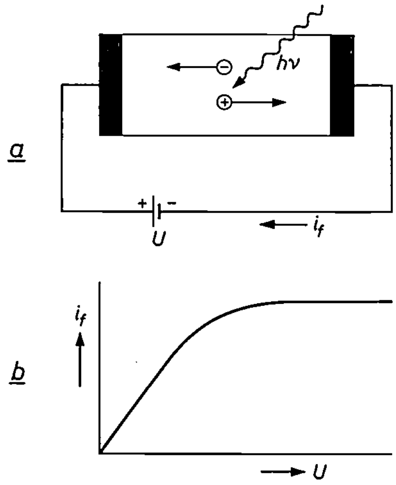


Fig. 9. *a*) Schematic representation of the flow of the photocurrent i_f in a semiconductor in which both kinds of charge carriers have adequate mobility and which is provided with completely blocking contacts.
b) The photocurrent i_f, under constant illumination, as a function of the applied voltage 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:

$$\eta = \frac{j}{eGd} = \frac{S_n}{d} + \frac{S_p}{d}\,. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (\mathrm{I},13)$$

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, 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.

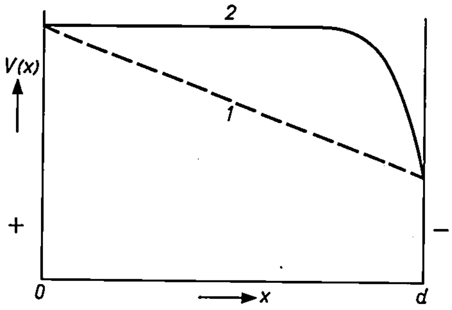


Fig. 10. Potential distribution in a non-illuminated (curve *1*) and in an illuminated (curve *2*) semiconductor with blocking contacts (distance between contacts *d*) where the holes have low mobility. The curvature of curve *2* is due to the occurrence of a positive space charge at the cathode: there the field strength is very high, at the anode it is zero.

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 researoh 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. 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 $j_{\rm ms}$ of the partial electron current from metal to semiconductor may therefore be written (omitting the suffix n):

$$j_{\rm ms} = j_0 \exp\,[-A/kT],$$

and the density in the opposite direction:

$$j_{\rm sm} = j_0 \exp\,[(-A+eU)/kT].$$

The density j of the net current is therefore:

$$j = j_0 \exp \frac{-A}{kT}\left(\exp\frac{eU}{kT}-1\right). \quad . \; . \; . \quad (\mathrm{I},\,14)$$

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:

$$j_{\rm block} \approx j_0 \exp\,(-A/kT).$$

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:

$$F = -\frac{Ne}{\varepsilon}(B-x)\,. \quad . \; . \; . \; . \; . \quad (\mathrm{I},\,15)$$

The variation of V with x can be found by a further integration. Suitably choosing the zero point of V, this gives:

$$V = -\frac{Ne}{2\varepsilon}(B-x)^2\,. \quad . \; . \; . \; . \; . \quad (\mathrm{I},\,16)$$

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.

[10]) See the article by Gudden and Pohl in reference [9]).

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:

$$B = \sqrt{\frac{2\varepsilon V_D}{eN}}. \quad . \; . \; . \; . \; . \quad (\mathrm{I,\ 17a})$$

When an external voltage U is applied (and provided still that $n \ll N$) the expression becomes:

$$B = \sqrt{\frac{2\varepsilon\,(V_D - U)}{eN}}. \quad . \; . \; . \quad (\mathrm{I,\ 17b})$$

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:

$$j(U) \approx -n(0)\, e\mu \sqrt{\frac{2eN(V_D - U)}{\varepsilon}}. \quad . \; . \; . \quad (\mathrm{I,18})$$

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:

$$j_{\mathrm{diff}} = eD\left(\frac{\mathrm{d}n(x)}{\mathrm{d}x}\right)_{x=0} =$$

$$= -n(\infty)\frac{e}{kT}\frac{\mathrm{d}V(0)}{\mathrm{d}x}\,eD\,\exp\,[-e\{V(0) - U\}/kT],$$

which, with the aid of (I,8), can be reduced to:

$$j_{\mathrm{diff}} = n(0)\mu F(0)\, e\, \exp(eU/kT).$$

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. Plooš 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 gammaverdeeld 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.

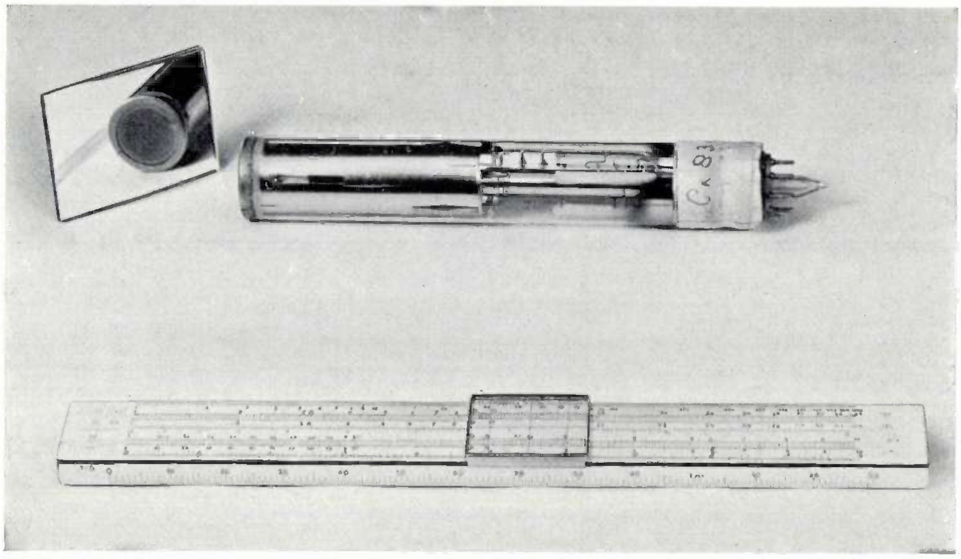


Fig. 1. A pick-up tube of the "Plumbicon" type. The tube is in the form of a cylinder having a diameter of 3 cm and a length of 19 cm. The target diameter is 2 cm.

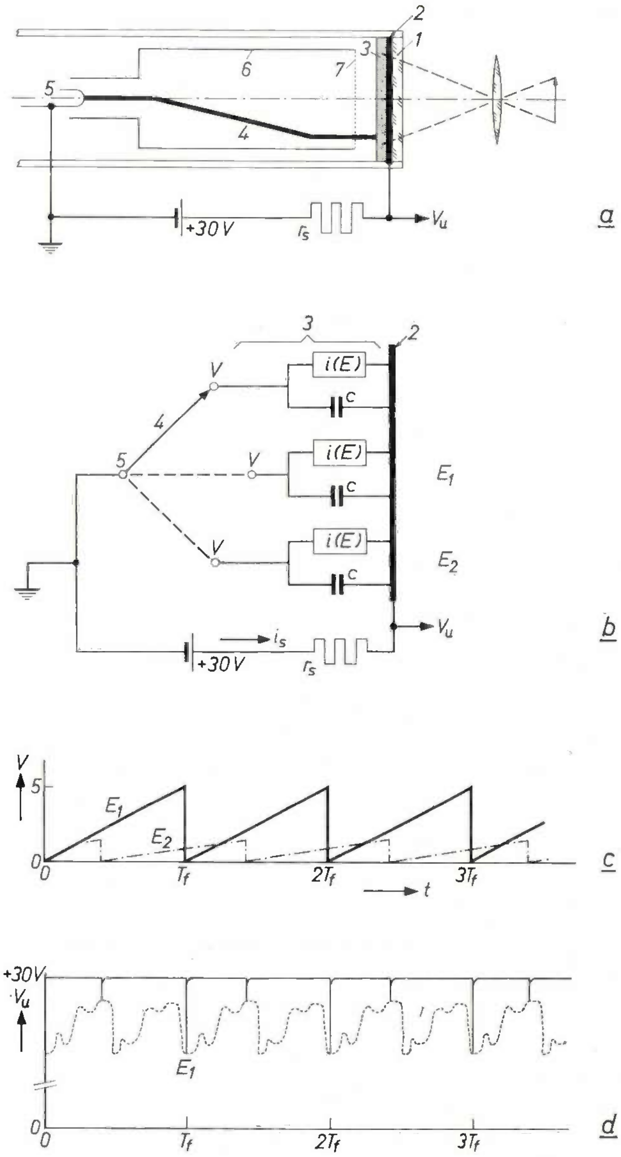


Fig. 2. *a*) Electrode layout of the "Plumbicon" (schematic). In the front of the tube (on the right) is a glass window *1*, on the inside of which have been applied, in that order, a transparent, conductive SnO_2 layer *2* and a photoconducting layer of PbO constituting the target *3*. An image of the 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_r, 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 avery 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

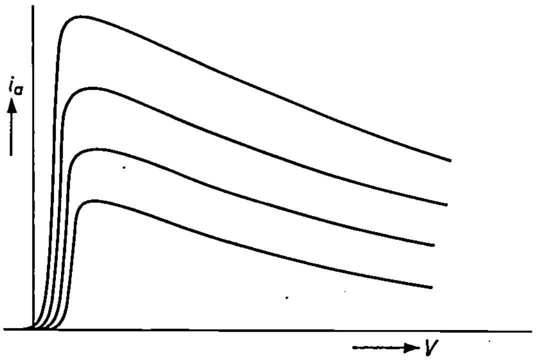


Fig.3. The net flow of current i_a to a surface element of a PbO target under continuous electron bombardment as a function of the potential V of the target surface (the cathode of the electron gun is assumed to have zero potential). As V increases, i_a also initially increases but subsequently falls 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 target ($U = V_u - V$; see fig. 2). The plotted quantity i_b is again the value of the current per surface element.

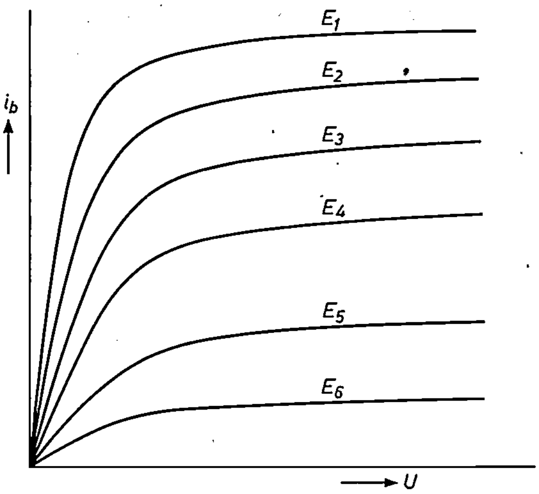


Fig.4. The variation of i_b, the photo-current that removes (negative) charge from an elementary area of the free target surface, with the potential difference U across the target. The curves, which are schematic, hold for different values E_i of the illumination. The stronger the illumination, the heavicr is the current i_b.

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

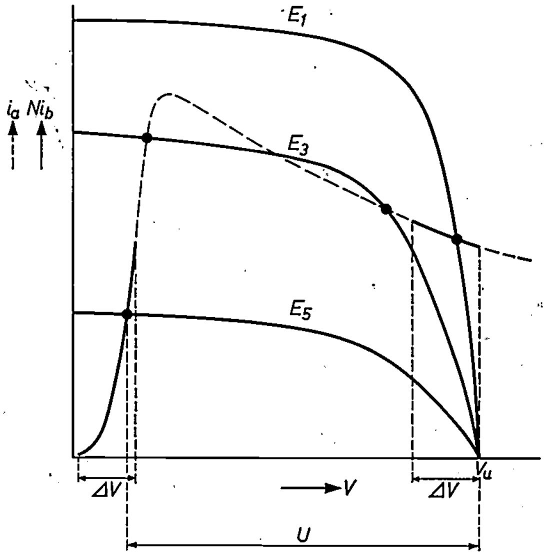


Fig. 5. To explain the fact that the interval ΔV can fluctuate in the neighbourhood of the cathode potential of the electron gun ($V \approx 0$), as well as in the neighbourhood of the potential of the signal plate ($V \approx V_u$).

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:

$$\mathrm{d}Q = (i_a - i_b)\mathrm{d}t \approx i_a\mathrm{d}t = c\mathrm{d}V.$$

Here $\mathrm{d}Q$ is the charge supplied in time $\mathrm{d}t$ and c is the capacitance of the layer per surface element. It follows from this that $\mathrm{d}V/i_a = \mathrm{d}t/c$. In the time T_p, V decreases from V_2 to V_1 and so:

$$\int_{V_2}^{V_1} \frac{\mathrm{d}V}{i_a} = \frac{T_p}{c}.$$

For the potential increase following from the flow of the discharge (photo-) current, we have by analogy:

$$\int_{V_1}^{V_2} -\frac{\mathrm{d}V}{i_b} = \frac{T_f}{c},$$

or, as $T_f = NT_p$:

$$\int_{V_2}^{V_1} \frac{\mathrm{d}V}{Ni_b} = \frac{T_p}{c}.$$

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.

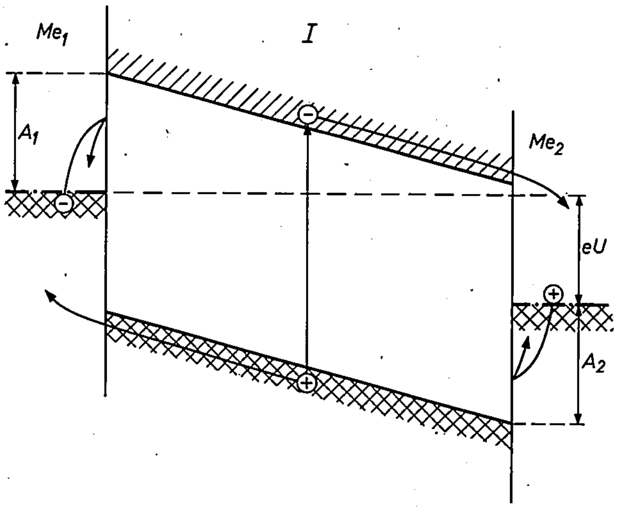


Fig. 6. Energy-band diagram (greatly simplified) of an intrinsic semiconductor (I) fitted with two metal contacts (Me_1 and Me_2). Me_1 has such a high work function (A_1) that electrons contained in it have very little chance of crossing the interface and entering the conduction band of the semiconductor. 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.*7a* shows the details and fig.*7b* 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.*7c*. Fig.*7d* 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,

[6]) See F.A. Kröger, G. Diemer and H.A. Klasens, Phys. Rev. **103**, 279, 1956.

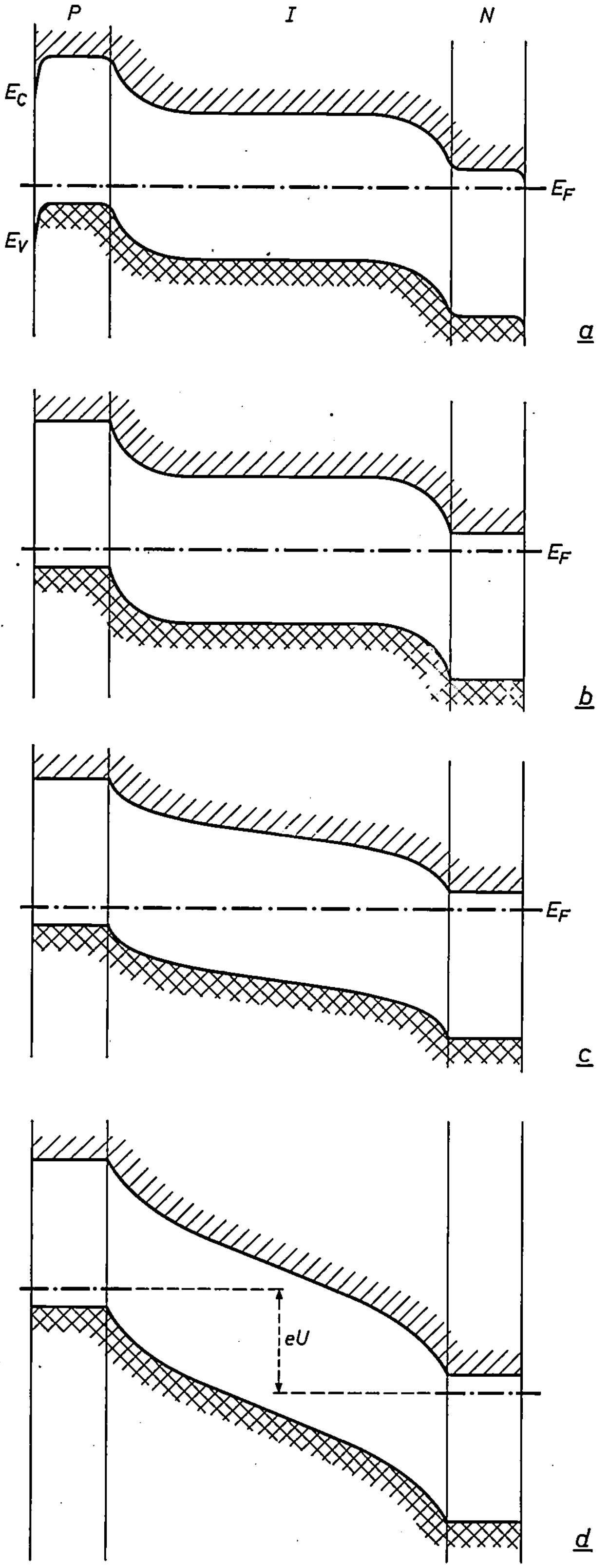


Fig. 7. *a*) Energy-band diagram of *P-I-N* diode. The curvature shown by the bands in the *I* layer is due to the fact that in practice this layer can never be completely free of impurities. In the case illustrated here, tbe concentration of impurities is relatively high. Diagram (*b*) is a simplified version of (*a*), the marked curvature exhibited in the two junctions having been converted into an abrupt changc 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.

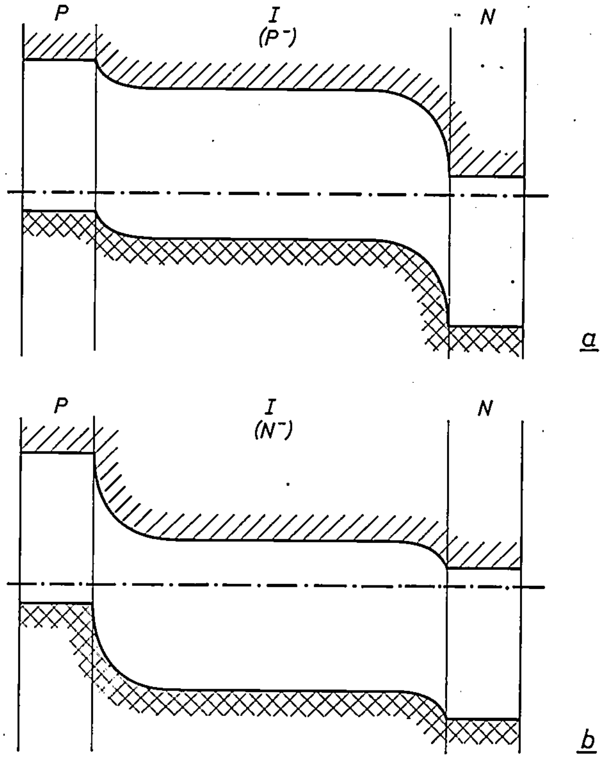


Fig. 8. Band diagrams simplified in the same way as fig. 7*b* and relating to *P-I-N* diodes whose *I* layers are slightly *P*-type (*a*) and slightly *N*-type (*b*). In these cases there is a fairly steep fall-off of potential close to one contact and a region with a gradual potential fall-off (and consequently a very low field-strength) near the other.

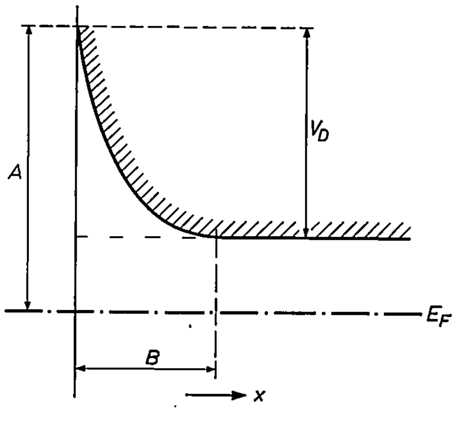


Fig.9. Near the interface between two regions of different conduction type a potential barrier arises which can be described in terms of the work function *A*, the diffusion potential V_D, and the width *B* of the barrier [7]).

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

$$B = \sqrt{\frac{2\varepsilon(V_D + U)}{eN_D}}, \quad . \; . \; . \; . \quad (1)$$

and the capacitance c' per unit area by

$$c' = \frac{\varepsilon}{B} = \sqrt{\frac{\varepsilon\, eN_D}{2\,(V_D + U)}}. \quad . \; . \; . \quad (2)$$

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,

$$F = \frac{eN_D(x - B)}{\varepsilon}, \quad . \; . \; . \; . \; . \; . \quad (3)$$

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

$$j_n(U) = n(0)\, e\mu_n\, F(0) = -n(0)\, e\mu_n \sqrt{\frac{2eN_D(V_D + U)}{\varepsilon}}, \quad . \; . \; . \quad (4)$$

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

$$n(0) = N_c\, e^{-A/kT}. \quad . \; . \; . \; . \; . \; . \quad (5)$$

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 cm^2, to a current density of $j_n \approx 3 \times 10^{-9}$ A/cm^2). 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 (cm^2/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 cm^2. The value of V_D has been put equal to zero.)

N_D μ_n	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.

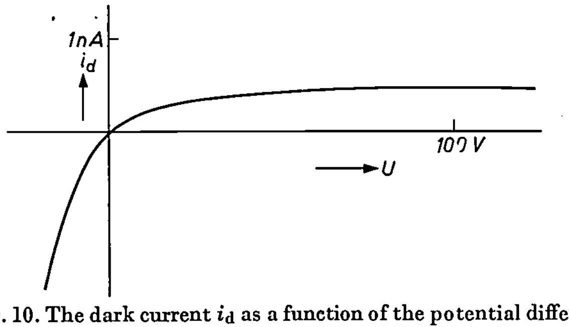


Fig. 10. The dark current i_d as a function of the potential difference U across the target. At no value of U likely to occur in practice does the dark current exceed 0.5×10^{-9} A. As might be expected, the curve resembles a diode characteristic. (In fact, the signal-plate potential V_u has been plotted 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* 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 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 tha- the zones just referred to extend pretty well throught 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.

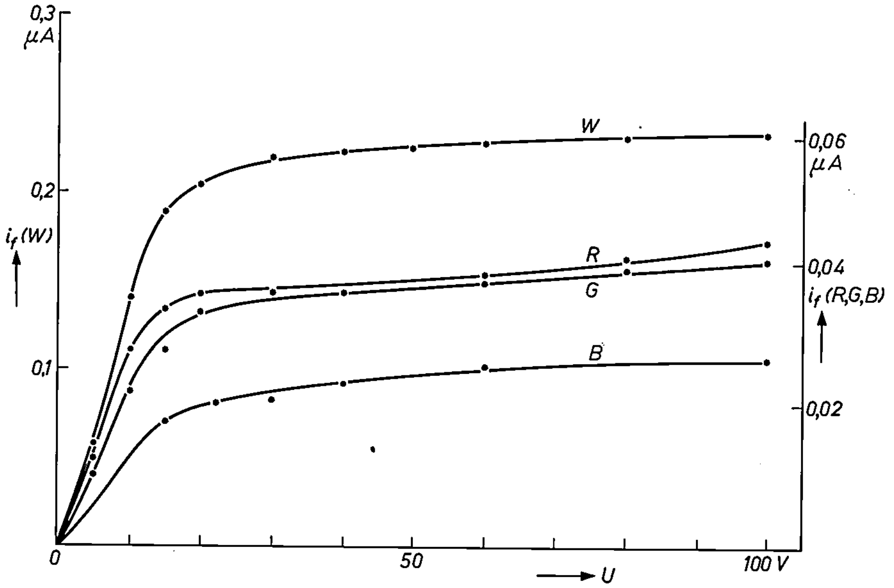


Fig.11. The variation of the photo-current i_f as a function of the signal-platevoltage V_u for different kinds of light, the intensity of illumination remaining the same (measured on a randomly selected tube). Curve *W*, which is to be read in conjunction with the i_f scale on the left, relates to white 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 R586, G537 and B488 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.

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.

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]).

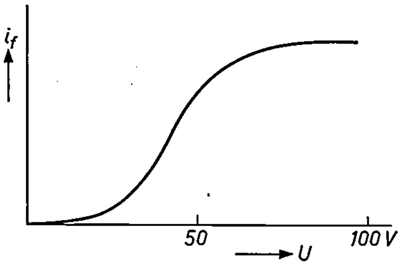


Fig. 12. When the I layer is too strongly N-type, the low-voltage end of the i_f,U characteristic will be depressed, and a high applied voltage will be required for i_f to attain its saturation value. This effect is strongest for blue light.

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.)

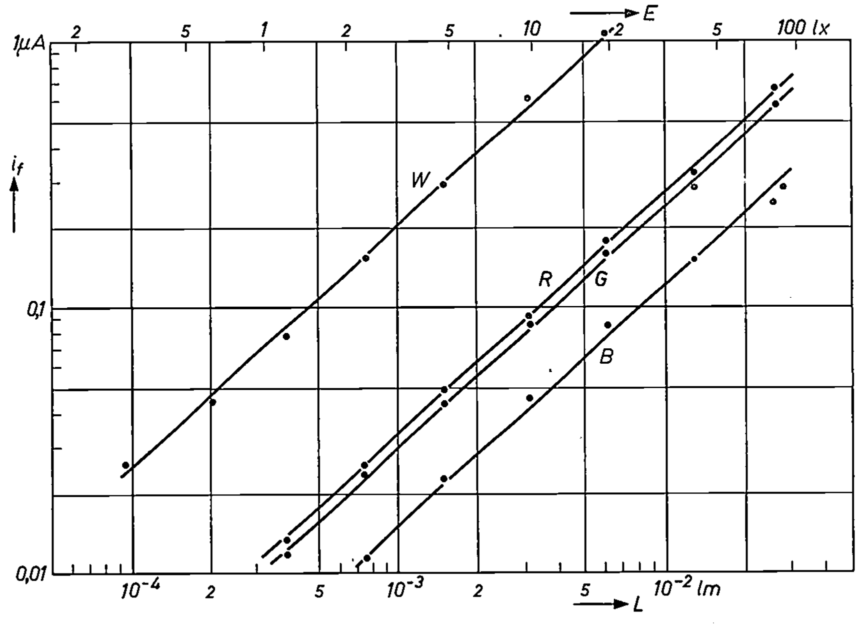


Fig. 13. The measured variation in photo-current i_f as a function of luminous flux L (or illumination E) for white, red, green and blue light (curves W, R, G and B), the applied voltage having been kept at a fixed value. In contrast to conventional vidicons, in the "Plumbicon" L and i_f are roughly 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.

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 10V and a layer thickness of 10μm the requirement becomes $\mu\tau > 10^{-7}$ cm²/V.

[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.

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.

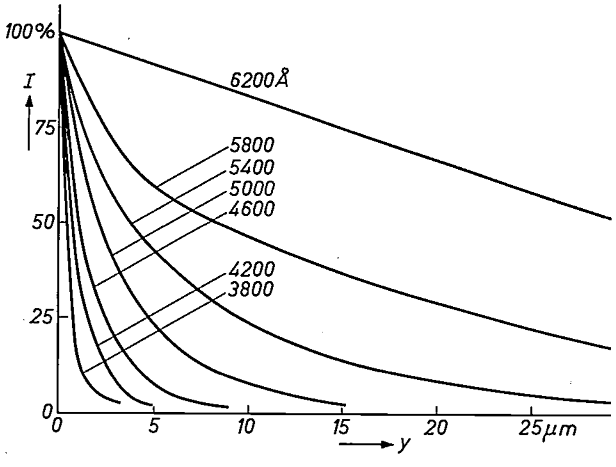


Fig. 14. The absorption of monochromatic light of different wavelength by PbO. The variation of the intensity *I* is plotted against the distance *y* the light has travelled.

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*).

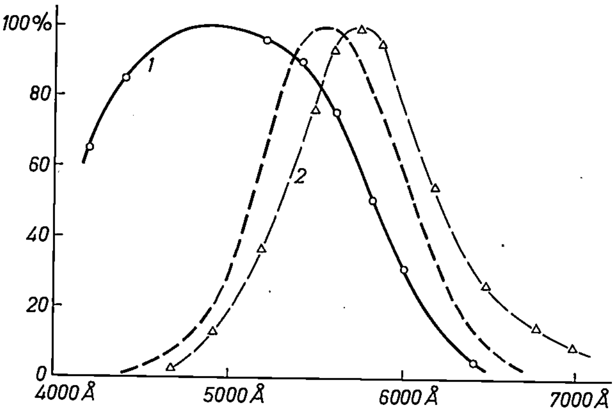


Fig. 15. The spectral sensitivity distribution of two tubes of the "Plumbicon" type. The thick dashed line represents the spectral sensitivity of the eye. Curve *1* is that of a tube whose PbO target was made by the standard process. The peak of the curve can, if desired, be shifted to the right (curve *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*).

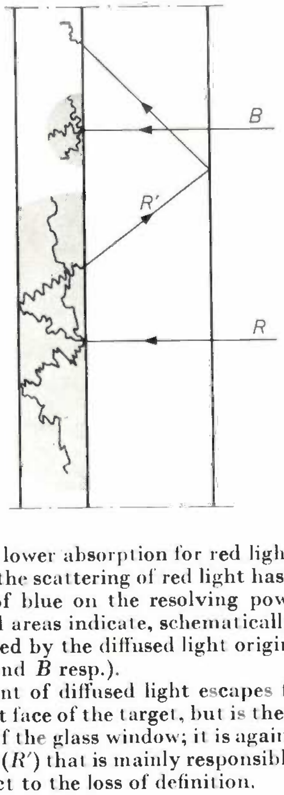


Fig. 16. PbO has a lower absorption for red light than for blue, and consequently the scattering of red light has a more adverse effect than that of blue on the resolving power of the PbO target. The shaded areas indicate, schematically, the extent of the zones penetrated by the diffused light originating 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

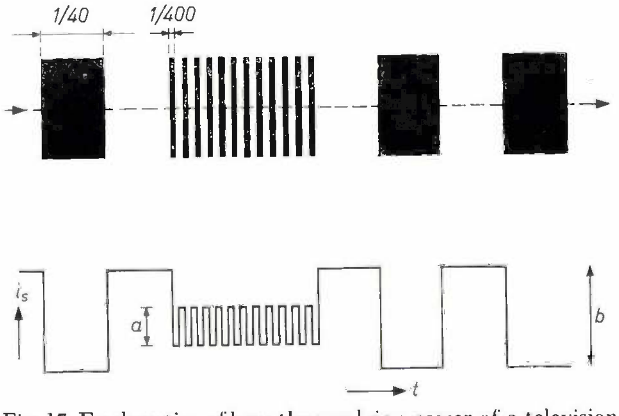


Fig. 17. Explanation of how the resolving power of a television camera tube is expressed in figures. A pattern of alternate light and dark vertical stripes, some having a breadth of 1/40th of the picture height and others a breadth of 1/400th of the picture height, is projected on the screen. This pattern 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%.

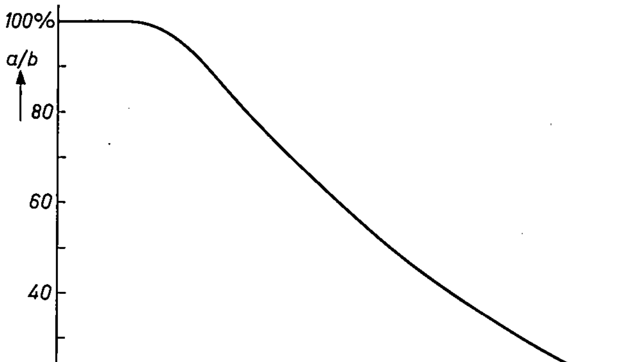


Fig. 18. A detailed description of resolving power can be obtained by plotting the variation in a/b against the breadth of the narrow stripes in the pattern appearing in fig. 17, the stripe breadth being expressed as the number n of black-white pairs of such stripes that would be reqnired to fill one 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 illustrated in fig.19b may be encountered. The names that we have given these inertial effects may be found in the caption to that figure.

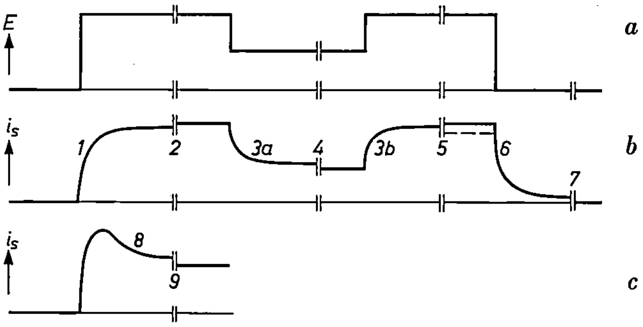


Fig. 19. The various forms of delayed response which, in principle, a "Plumbicon" may exhibit. In a good tube they are too slight to cause any trouble, and some are absent altogether. Diagram (a) shows the programme of varying intensity of illumination presented to the tube: the light is switched on, 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.

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 30V. 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.

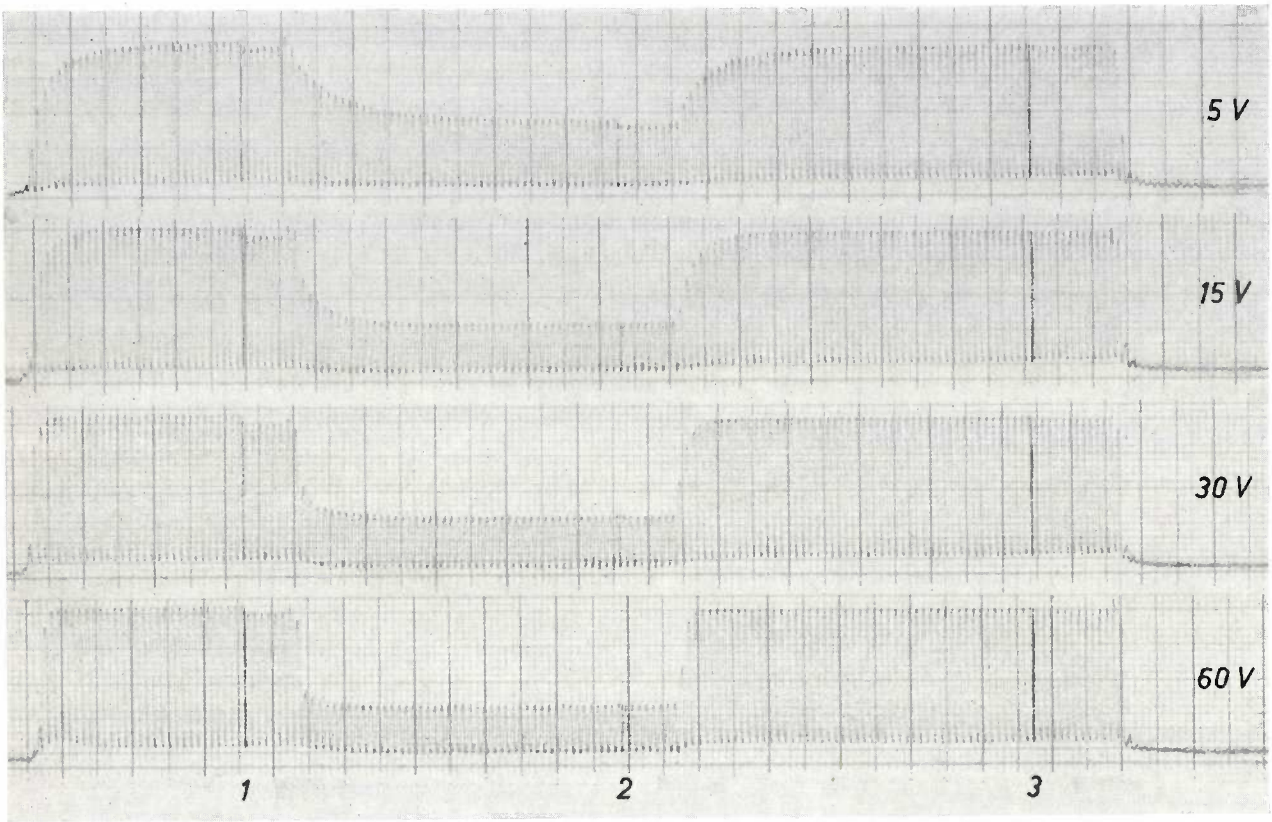


Fig.20. The response of a randomly selected "Plumbicon" tube at four values of U, the voltage across the PbO target. At the relatively low U value of 30 V the response of the tube is already quite fast, and there is scarcely any persistence. The traces were obtained by illuminating the top half of the 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 canbe 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 2000pF. 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 1500pF.

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

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \frac{1}{C}(i_f - i_a)\,.$$

(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:

$$i_a(t) = \frac{i_2}{1-(1-i_2/i_1)\exp(-tbi_2/C)}\,. \quad . \quad . \quad (6)$$

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 10V 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 800pF.

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. 21b). 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*).

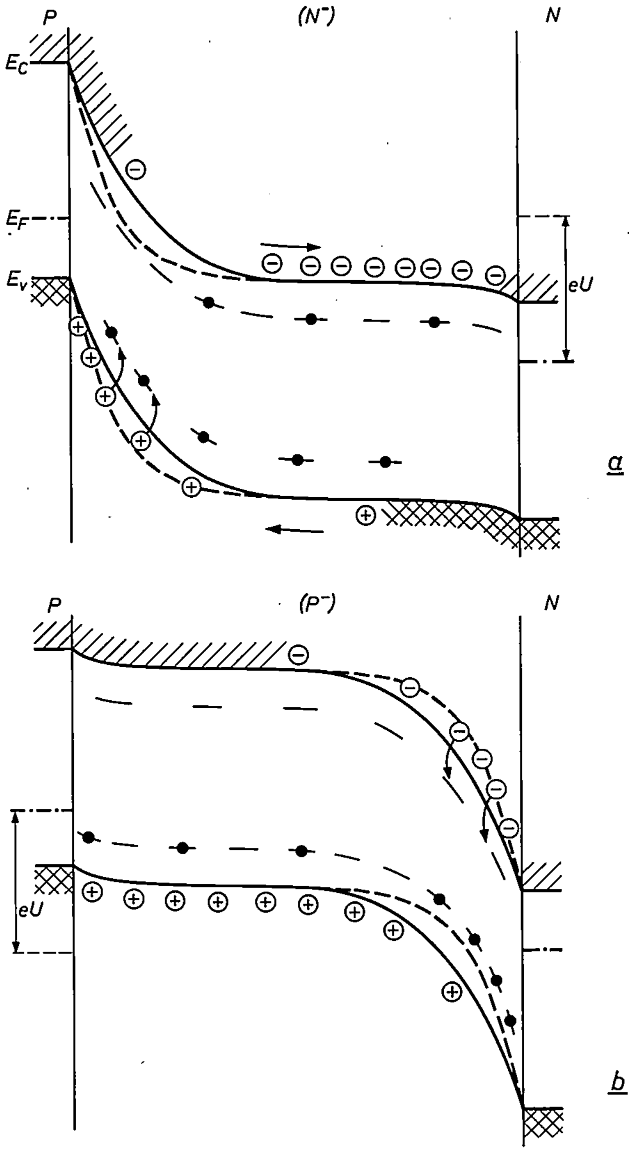


Fig. 21. To explain "fatigue".

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

· 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.

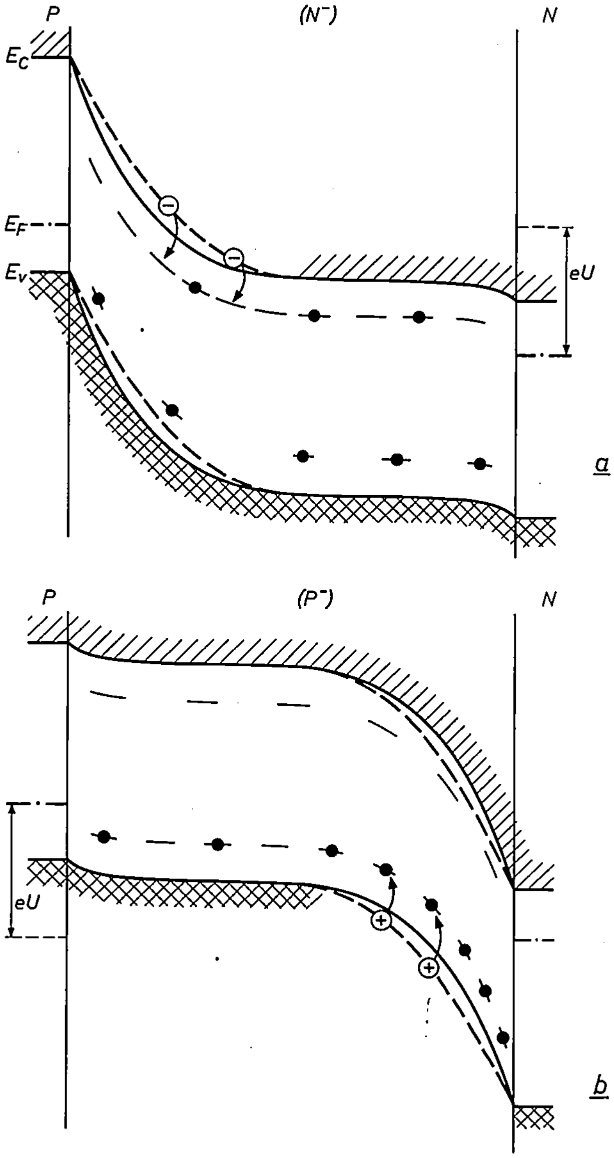


Fig. 22. To explain "black-white trailing".

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

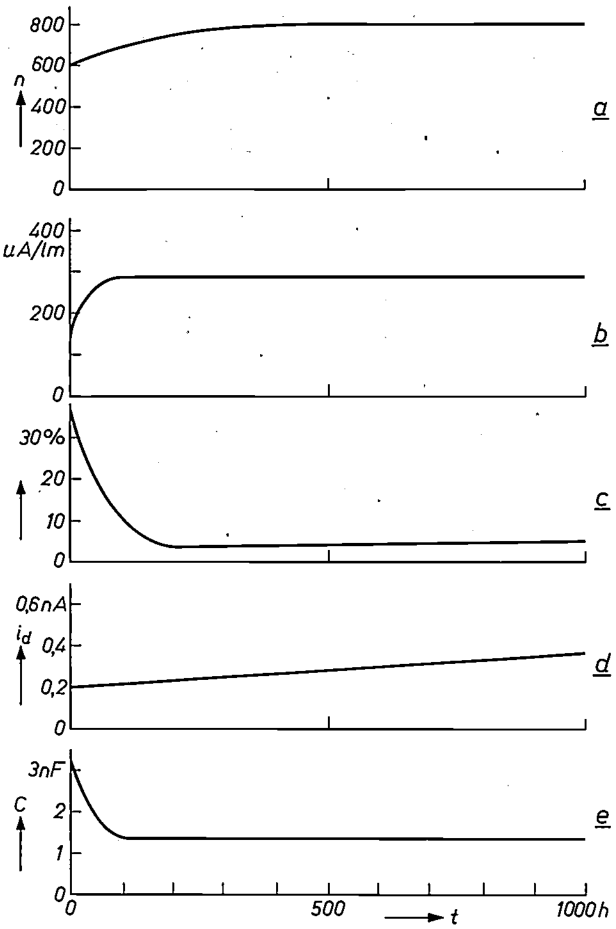


Fig. 23. Changes in various tube characteristics and in the capacitance of the PbO target, measured on a randomly selected tube of the "Plumbicon" type, during the first 1000 hours in operation. *a*) Resolving power, shown as the number of lines that can just be resolved by the eye (the corresponding *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.

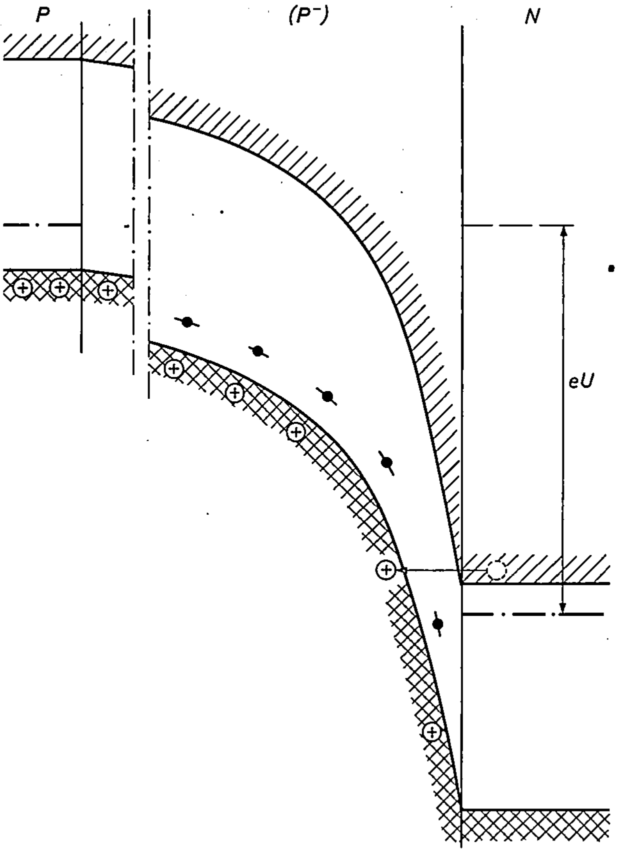


Fig. 24. To explain the occurrence of a strong dark current of holes, when the inner sublayer contains a limited region of high P-conductivity. The barrier can then be so narrow that tunnelling is possible (see arrow).

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.

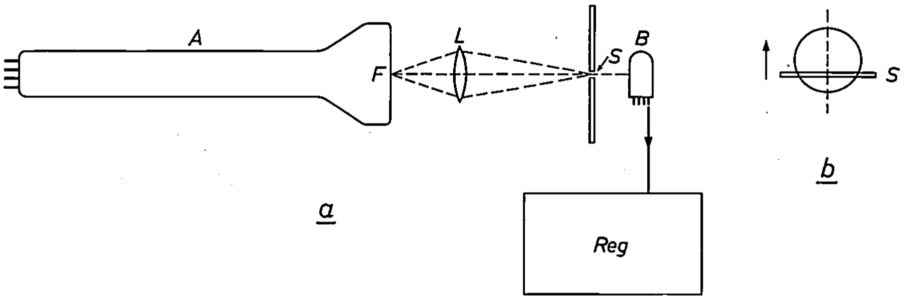


Fig. 1. *a*) Diagram of the set-up. The electron gun under investigation is mounted in a testing tube *A*. A magnified image of the focused spot *F* is produced on a screen by the lens *L*. The slit *S* cuts a narrow strip from the image (see *b*), the light from which falls on a photomultiplier *B*. The image is scanned by the relative movement of slit and image (see arrow in *b*). *Reg* pen recorder.

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 çontinuously, 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.

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).

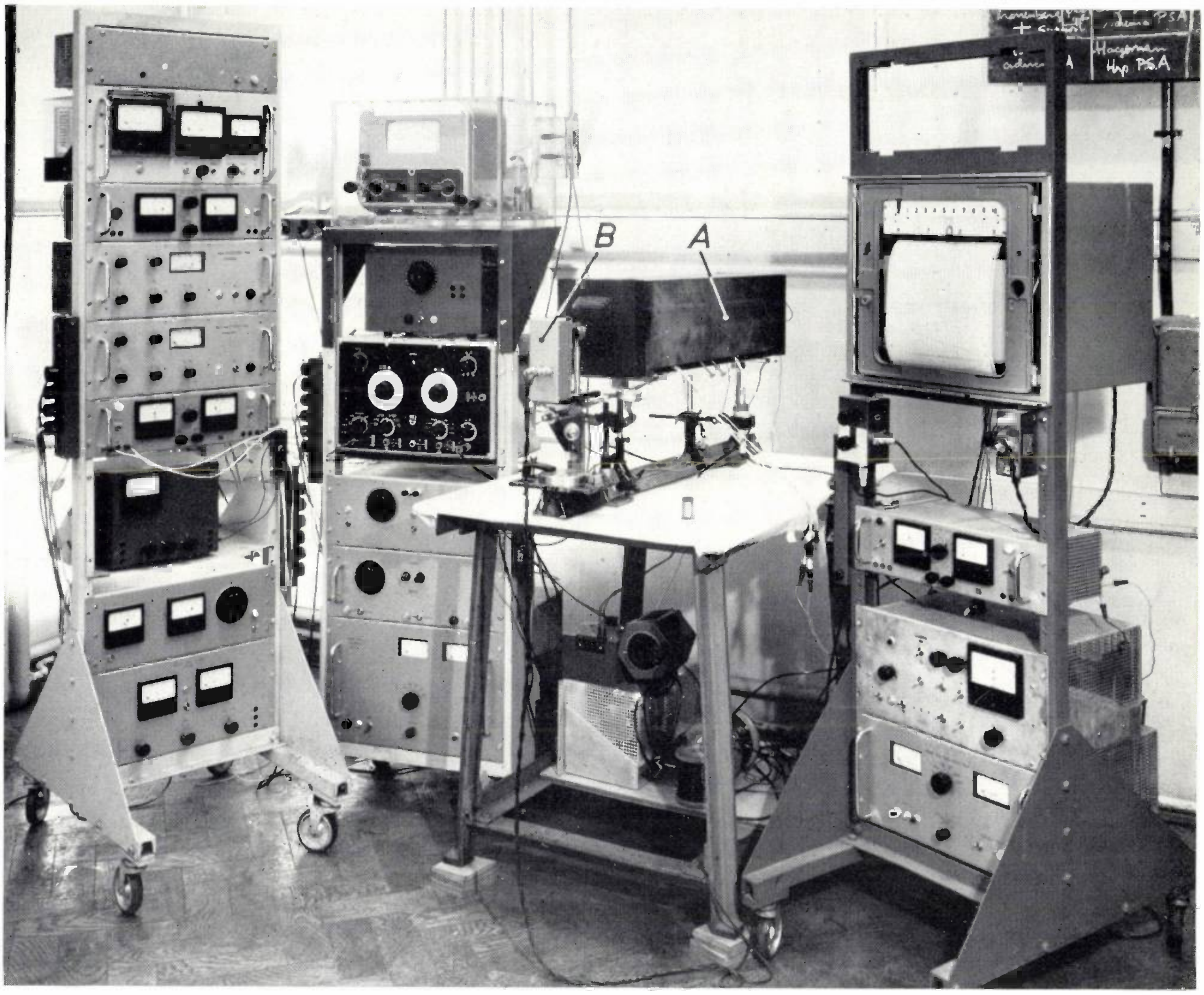


Fig. 2. Measuring set-up as in the diagram shown in fig. 1, used in the Philips Research Laboratories. The spot on the screen is scanned by slightly deflecting the beam. The testing tube is enclosed in a magnetically screened can of mu-metal (A). The photomultiplier is at B. The light-intensity distribution in the spot is approximately Gaussian (see the recording).

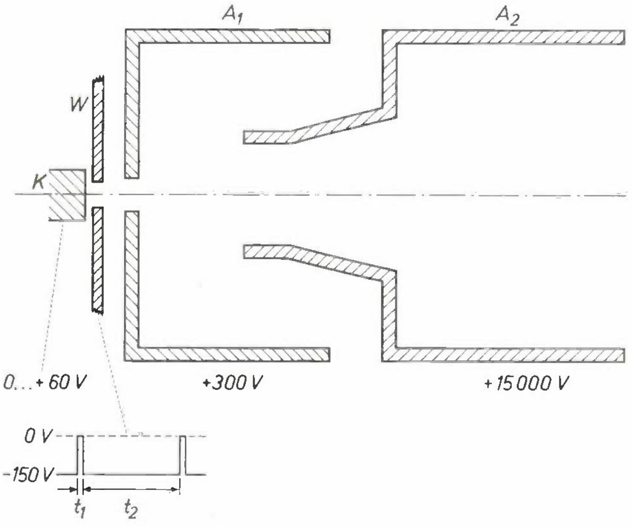


Fig. 3. Axial cross-section of the electron gun of a picture tube. K cathodc. A_1, A_2 anodes. During the measurement the beam is intermittently suppressed by applying pulses to the Wehnelt cylinder W; $t_1/t_2 = 10^{-4}$ to 10^{-5}.

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.

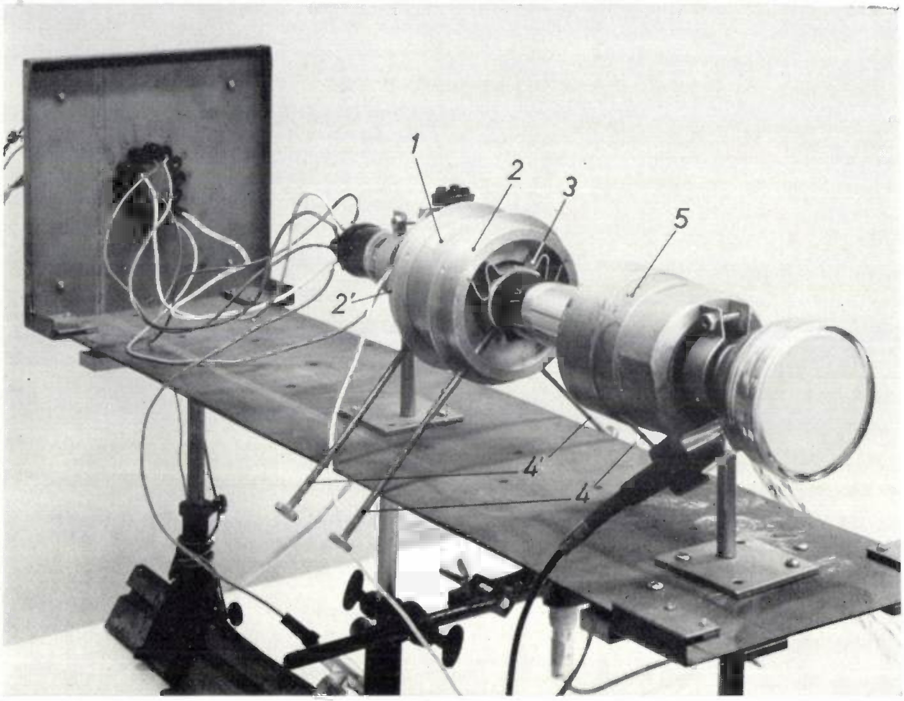


Fig. 4. The testing tube with its coils; screening can, optical system and photomultiplier have been removed. *1* focusing coil. *2*, *2'* clamping rings. *3* leafspring (ring *2'* contains an identical leafspring). *4*, *4'* adjusting screws. *5* coil used to deflect the beam slightly for scanning the 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

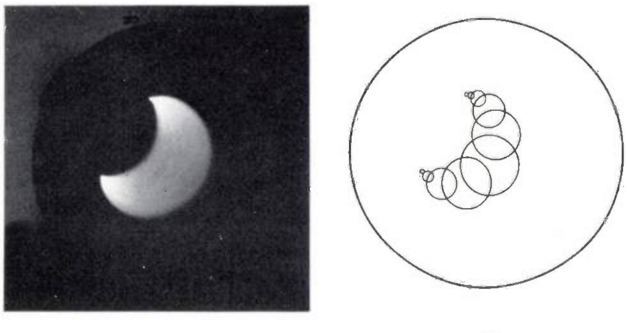


Fig. 5. *a*) Example of image that can appear on the screen when the non-aligned focusing coil is carrying an alternating current whose amplitude is equal to the focusing current (comma and asymmetrical forms can also occur, and also self-intersecting half-moons; see e.g. T. Soller *et al.*, Cathode 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 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.

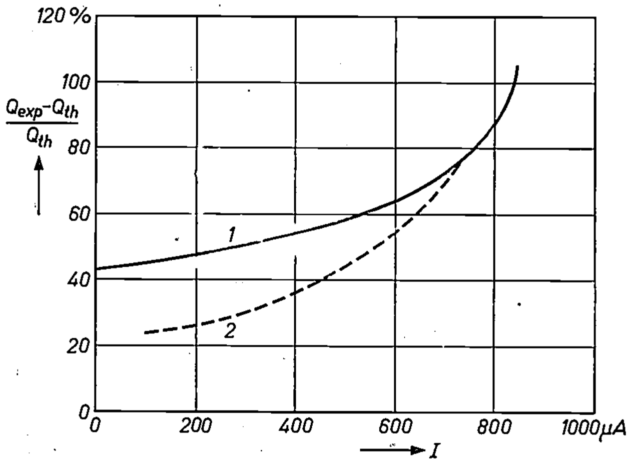


Fig. 6. Percentage deviation of the experimental figure of merit Q_{exp} from the theoretical value Q_{th}, for a certain type of electron gun to be used in a television picture tube. For curve *1* the beam current I was varied by means of the cathode voltage, for curve *2* by means of the cathode temperature. (Taken from [4]).)

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 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.*

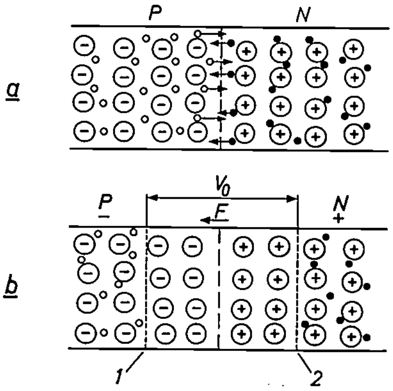


Fig. 1. Where a *P* region borders on an *N* region (*P-N* junction), holes diffuse from the *P* to the *N* region and electrons in the opposite direction (*a*). The charge carriers crossing the junction enter a region where they are minority charge carriers; they spread over this region and each ultimately 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.

[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 *p*ositive 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:

$$n(1) = n_{P0}\ e^{qV_u/kT}\ ,\ p(2) = p_{N0}\ e^{qV_u/kT}\ . \quad \ldots \quad (1\text{a and b})$$

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.

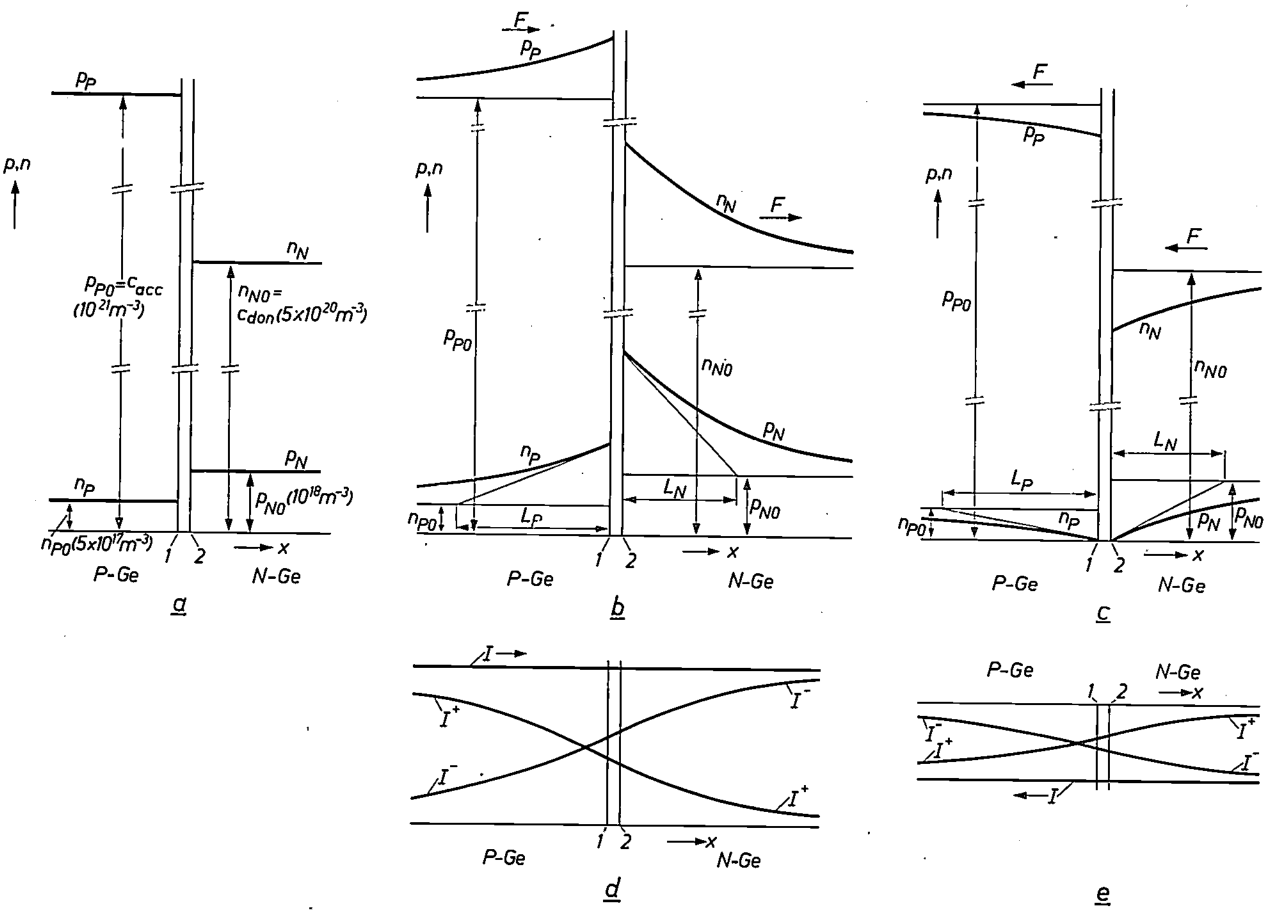


Fig. 2. Concentration pattern of electrons and holes at a *P-N* junction, *a*) without external voltage, *b*) with external voltage in the forward direction, and *c*) with external voltage in the reverse direction. The figures added in (*a*) indicate the order of magnitude of the concentrations involved. 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 consequently applies to the total current I. The current-voltage characteristic of a P-N junction is therefore given by

$$I = I_s \left(e^{qV_u/kT} - 1\right), \quad . \; . \; . \; . \quad (2)$$

which function is represented in *fig. 3*. The rectifying action of a P-N junction is immediately apparent 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.

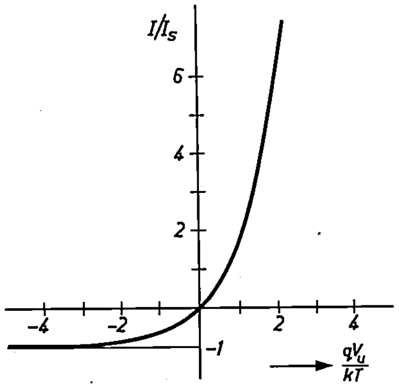


Fig. 3. Theoretical current-voltage characteristic of a P-N junction.

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-

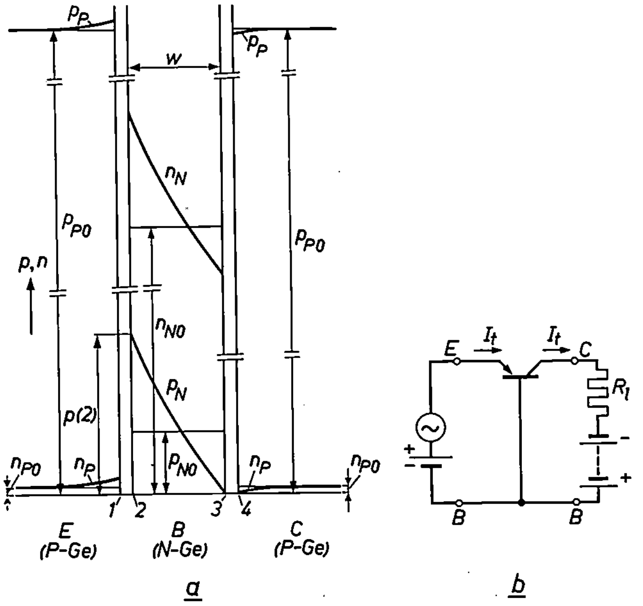


Fig. 4. *a*) The concentration pattern of holes and electrons in a P-N-P transistor, from which transistor action can be explained.
b) Common-base configuration in which the transistor can give voltage gain.

[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.

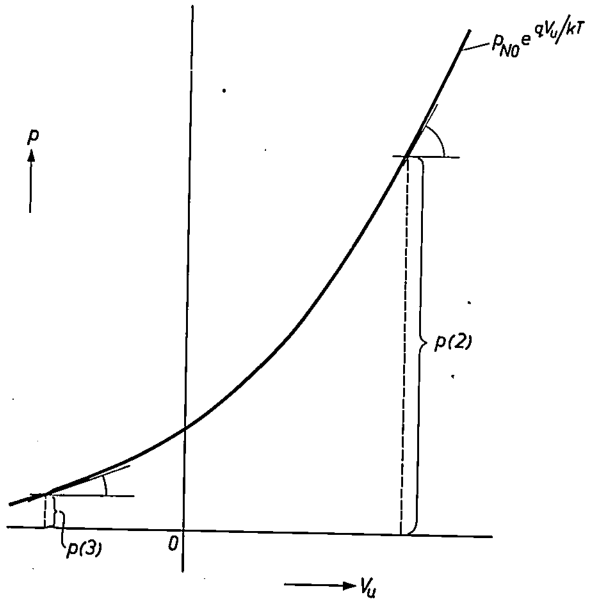


Fig. 5. The exponential relation between the hole concentration p and the external barrier voltage V_u, applicable to the boundary planes *2* and *3* in fig. 4*a*.

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 $V_{\mathrm{EB'}}$ and $V_{\mathrm{CB'}}$ 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: $V_{\mathrm{EB'}} = V_{\mathrm{E}} - V_{\mathrm{B'}}$.

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 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".

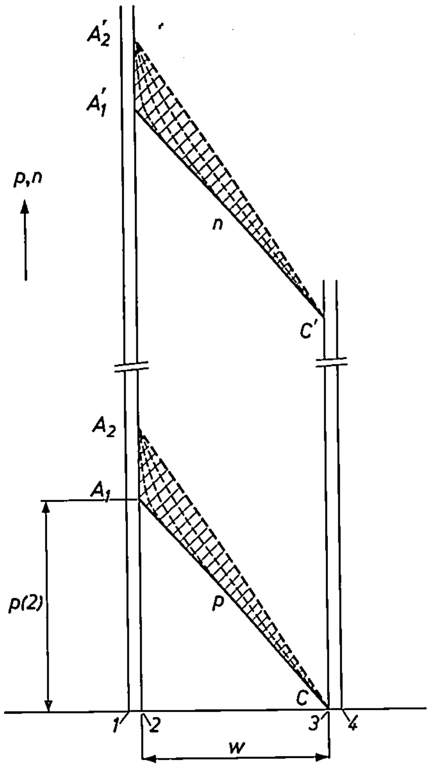


Fig. 6. After a suddenincrease in the concentrations of holes and electrons in boundary plane *2* between emitter barrier and base, the concentration lines A_1C and A'_1C' gradually change to A_2C and A'_2C'. The hole and electron contents of the base thereby increase by equal amounts which are represented by hatched triangles. w is the thickness of the base. The thin dashed lines represent some intermediate phases.

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 $V_{\mathrm{EB'}}$ across the emitter barrier changes so slowly that the concentration pattern remains constantly

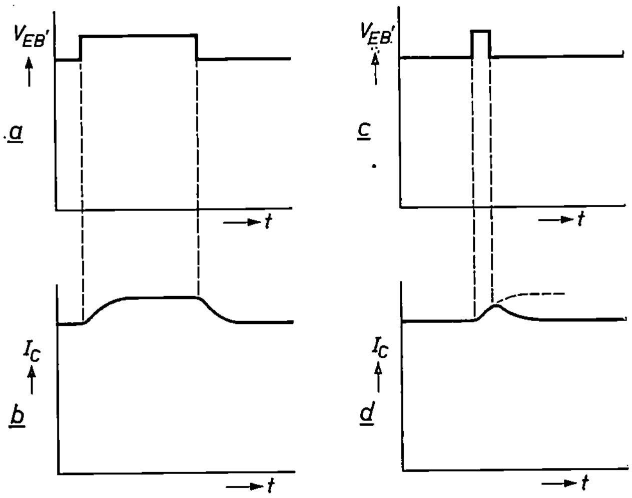


Fig. 7. Sudden changes in the voltage $V_{\mathrm{EB'}}$ across the emitter barrier (*a*) cause — as follows from fig. 6 — gradual changes in the current I_{C} across the collector barrier (*b*). A short voltage pulse (*c*) has relatively little influence on I_{C} (*d*).

[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. 4a). 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^{2}s^{-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, the region between the boundary planes constitutes an electrical double layer. The space-charge density ϱ in this layer is represented in fig. 8b. Since the total negative charge must be equal to the total positive charge, we can write

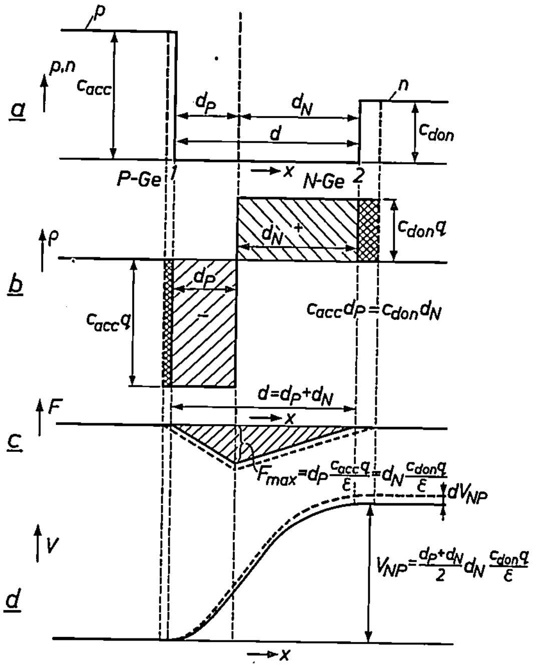


Fig. 8. Illustrating the barrier capacitance. The area of each of the hatched rectangles in (b), after division by the dielectric constant ε, yields the value F_{max} in (c); the area of the hatched triangle in (c) yields the potential difference V_{NP} in (d).

$$c_{acc}\;d_P = c_{don}\;d_N\,. \quad \ldots\ldots \quad (5)$$

(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":

$$C_{\text{b}} = \frac{O\varepsilon}{d} \quad . \; . \; . \; . \; . \; . \; . \; (6)$$

(where O is the cross-sectional area of the crystal), and the following expression holds for $\mathrm{d}Q$:

$$\mathrm{d}Q = C_{\text{b}}\,\mathrm{d}V.$$

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:

$$d_N = \sqrt{\frac{c_{\text{acc}}}{c_{\text{don}}\,(c_{\text{don}} + c_{\text{acc}})}\,\frac{2\varepsilon}{q}}\,\sqrt{V_{NP}}. \qquad (7)$$

According to (5) we can find d_P if we multiply (7) throughout by $c_{\text{don}}/c_{\text{acc}}$. 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:

$$C_{\text{b}} = O\,\sqrt{\frac{c_{\text{acc}}\,c_{\text{don}}}{c_{\text{acc}} + c_{\text{don}}}\,\frac{\varepsilon q}{2}}\Big/\sqrt{V_{NP}}. \quad . \; (8)$$

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. $O = 3\times10^{-7}\,\text{m}^2$, $c_{\text{don}} = 10^{20}\,\text{m}^{-3}$ and $c_{\text{acc}} = 10^{24}\,\text{m}^{-3}$. 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 higheenough 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 $v_{\text{EB}'}$, 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. 9b 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.10b). 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.

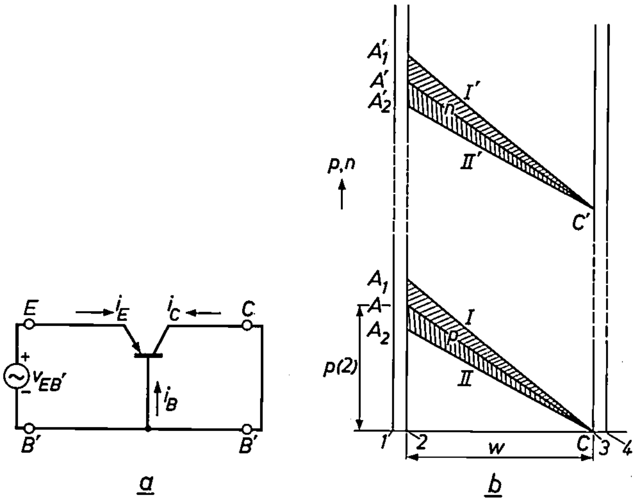


Fig. 9. *a*) The first fundamental case in an analysis of the behaviour of a transistor with increasing frequency: the collector barrier is short-circuited to alternating current; the biasing voltage across the emitter barrier has an alternating voltage $v_{EB'}$ superimposed on it. Biasing voltages 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'.

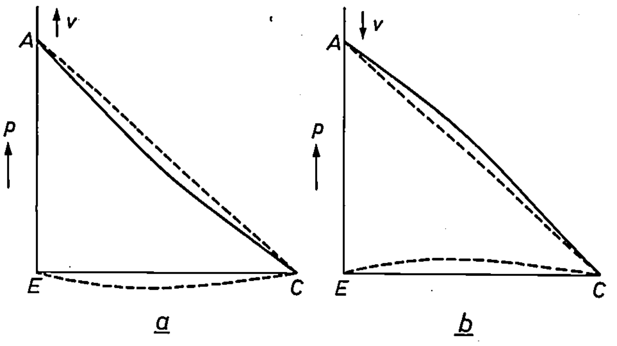


Fig. 10. When point A of the hole concentration line moves upwards at a velocity v, then AC must be slightly concave (a); when A moves downwards, then AC is convex (b). In both cases the concentration pattern can be regarded as the superposition of a moving straight line AC and a stationary curve EC (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

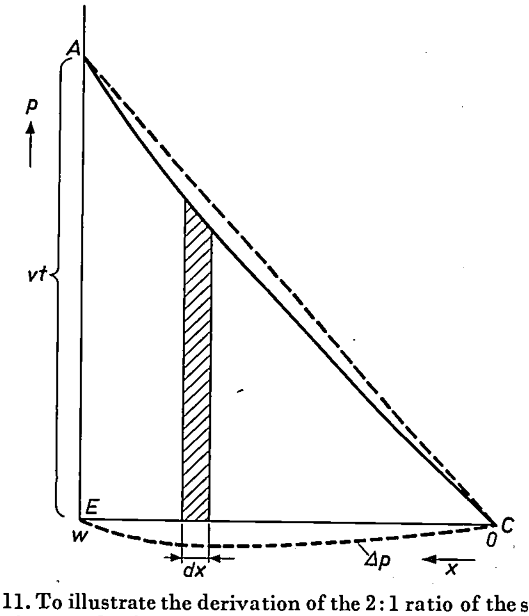


Fig. 11. To illustrate the derivation of the 2:1 ratio of the slopes at emitter and collector of the curve Δp as a function of position x.

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:

$$D_\mathrm{p}\frac{\partial^2 p}{\partial x^2}=\frac{\partial p}{\partial t}\,.$$

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:

$$D_\mathrm{p}\frac{\partial^2 \Delta p}{\partial x^2}=\frac{v}{w}\,x+\frac{\partial \Delta p}{\partial t}\,.$$

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:

$$\Delta p=\frac{vw^2}{6D_\mathrm{p}}\left[\left(\frac{x}{w}\right)^3-\frac{x}{w}\right].$$

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 $2vw/6D_\mathrm{p}$, and at the collector ($x = 0$) it is $-vw/6D_\mathrm{p}$. 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 $v_{\mathrm{EB}'} = \hat{v}_{\mathrm{EB}}\cos\omega t$ superimposed on the emitter DC voltage $V_{\mathrm{EB}'}$. 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.

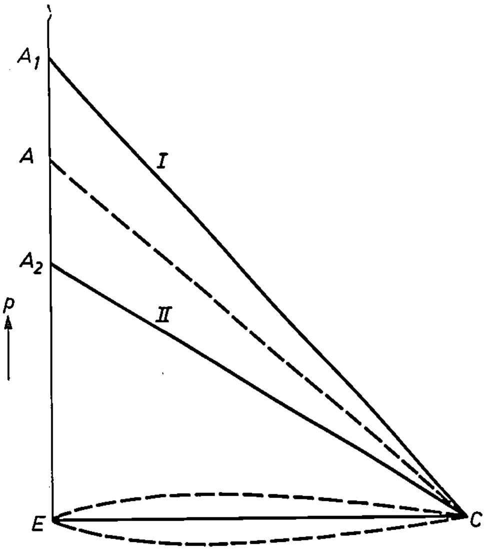


Fig. 12. If point A describes a slow oscillation, the hole concentration pattern can be regarded as the superposition of a straight line AC, oscillating around the fixed point C, and a curve EC oscillating in the manner of a standing wave. The extreme positions of EC correspond to the middle position of AC (dashed lines).

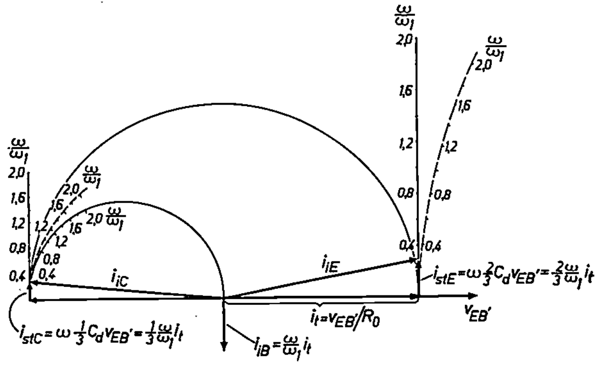

$\frac{\omega}{\omega_1}$
2,0
1,6
1,2
0,8
0,4
$\frac{\omega}{\omega_1}$
2,0
1,6
1,2
0,8
0,4
$\frac{\omega}{\omega_1}$
2,0
1,6
1,2
0,8
0,4
$\frac{\omega}{\omega_1}$
2,0
1,6
1,2
0,8
0,4
$\frac{\omega}{\omega_1}$
2,0
1,6
1,2
0,8
0,4
i_{iC}
i_{iE}
$i_{stE}=\omega\frac{2}{3}C_dv_{EB'}=\frac{2}{3}\frac{\omega}{\omega_1}i_t$
$v_{EB'}$
$i_t=v_{EB'}/R_0$
$i_{stC}=\omega\frac{1}{3}C_dv_{EB'}=\frac{1}{3}\frac{\omega}{\omega_1}i_t$
$i_{iB}=\frac{\omega}{\omega_1}i_t$

Fig. 13. Vector diagram of the "intrinsic" transistor when the reverse voltage across the collector barrier is kept constant, while the alternating voltage $v_{\mathrm{EB}'}$ 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 ($\omega_1 = 1/R_0C_\mathrm{d}$ 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 $i_{\mathrm{iC}} = -v_{\mathrm{EB}'}/R_0(1 + \mathrm{j}\omega/3\omega_1)$.

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:

$$\hat{i}_t = \hat{v}_{EB'}/R_0\,, \quad . \; . \; . \; . \; . \; . \; . \quad (9)$$

where

$$R_0 = \frac{kT}{qI_t}\,. \quad . \; . \; . \; . \; . \; . \; . \quad (10)$$

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):

$$\omega_1 = 1/R_0C_d = 2D_p/w^2\,. \quad . \; . \; . \; . \quad (11)$$

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

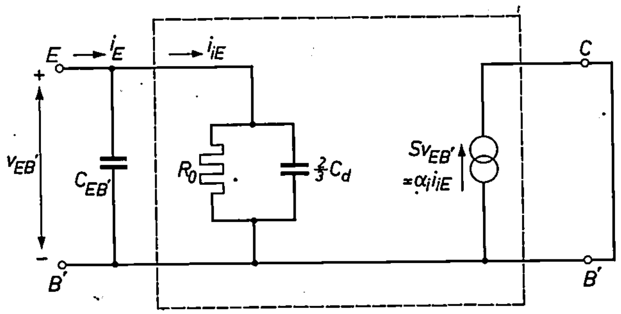


Fig. 14. Equivalent AC circuit for a transistor in common-base configuration with short-circuited output. The section inside the dashed lines relates to the intrinsic transistor and is based on the vector diagram of fig. 13.

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:

$$S = \frac{1}{R_0\left(1 + \mathrm{j}\,\dfrac{\omega}{3\omega_1}\right)}.$$

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.

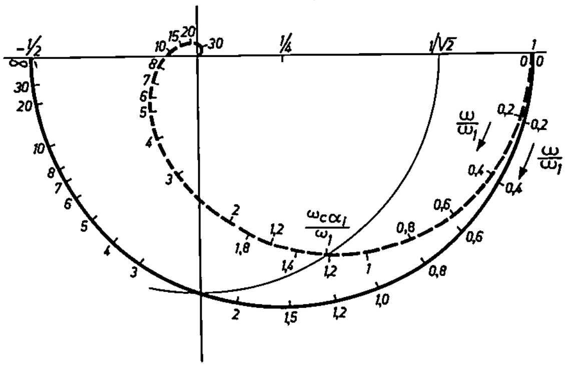


Fig. 15. Locus of the points which, in the complex plane, represent the current amplification factor α_i of the intrinsic transistor (really the transport factor β of the base [13])) when the frequency increases from zero to ∞. ω_1 (= $2D_p/w^2$, see formula 11) is a characteristic frequency. The semi-circle 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 (α_i = sech $\sqrt{2\mathrm{j}\omega/\omega_1}$), 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:

$$\omega_{c\alpha i} = 1.21\,\omega_1 . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

[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 $v_{\mathrm{CB'}}$ (*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 $v_{\mathrm{CB'}}$, and a diffusion capacitance gives rise to the flow of storage currents over the two barriers.

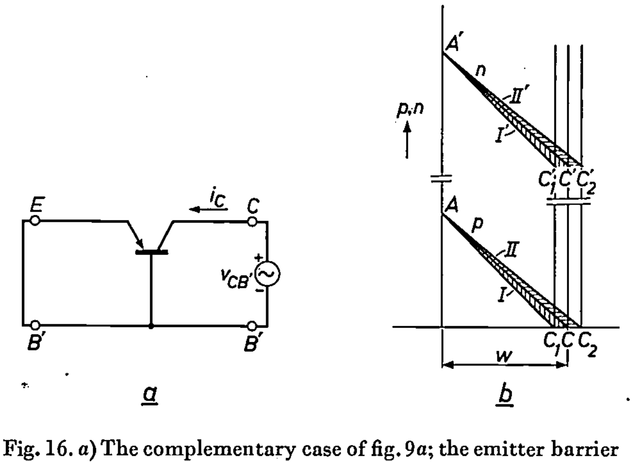


Fig. 16. *a*) The complementary case of fig. 9*a*; the emitter barrier is short-circuited to alternating current; an alternating voltage $v_{\mathrm{CB'}}$ is superimposed on the biasing voltage across the collector barrier.
b) The corresponding concentration pattern in the base: the concentration lines 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-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 $C_{\mathrm{CB'}}$ is again added to represent the barrier capacitance.

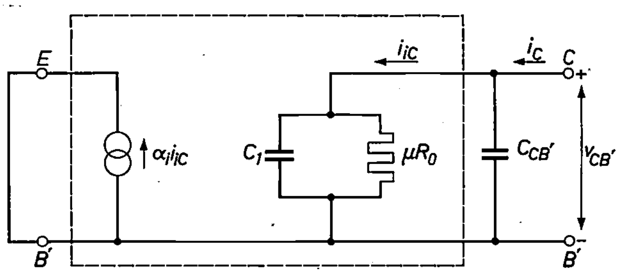


Fig. 17. Equivalent circuit for a transistor in common-base configuration with short-circuited input, drawn by analogy with the diagram in fig. 14 relating to short-circuited output.

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 $C_{\mathrm{CB'}}$, 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

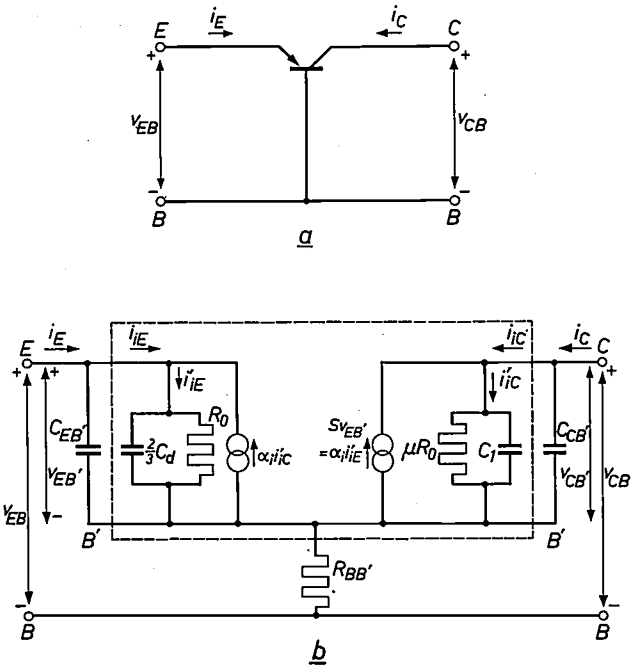


Fig. 18. *a*) A transistor in common-base configuration.
b) Equivalent circuit for (*a*), obtained by combining figures 14 and 17, at the same time adding the internal base resistance $R_{BB'}$. The correctness of the diagram follows from the fact that, afterputting $R_{BB'}$ at zero, the diagram changes 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.

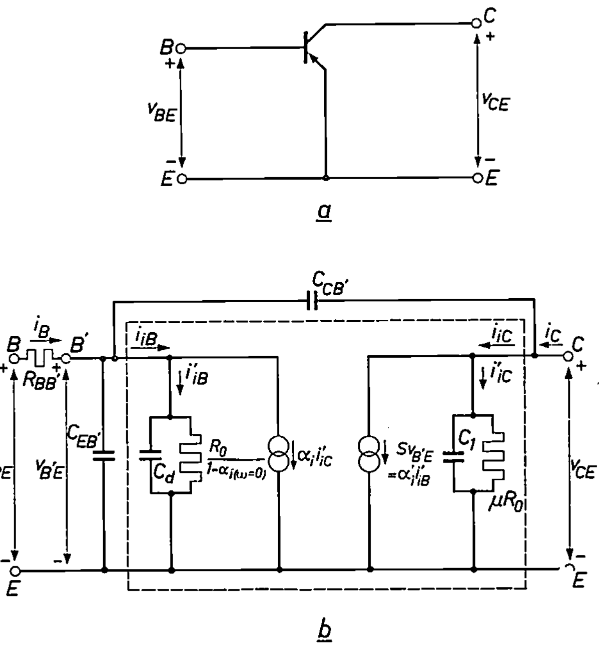


Fig. 19. *a*) The transistor in common-emitter configuration.
b) Equivalent circuit for (*a*).

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 concentration 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 $i_{iB(\omega=0)} \approx (1 - a_{i(\omega=0)})i_t$ in phase with the input voltage $v_{B'E}$. (Here $a_{i(\omega=0)}$ 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 ($\omega \ll \omega_1$).

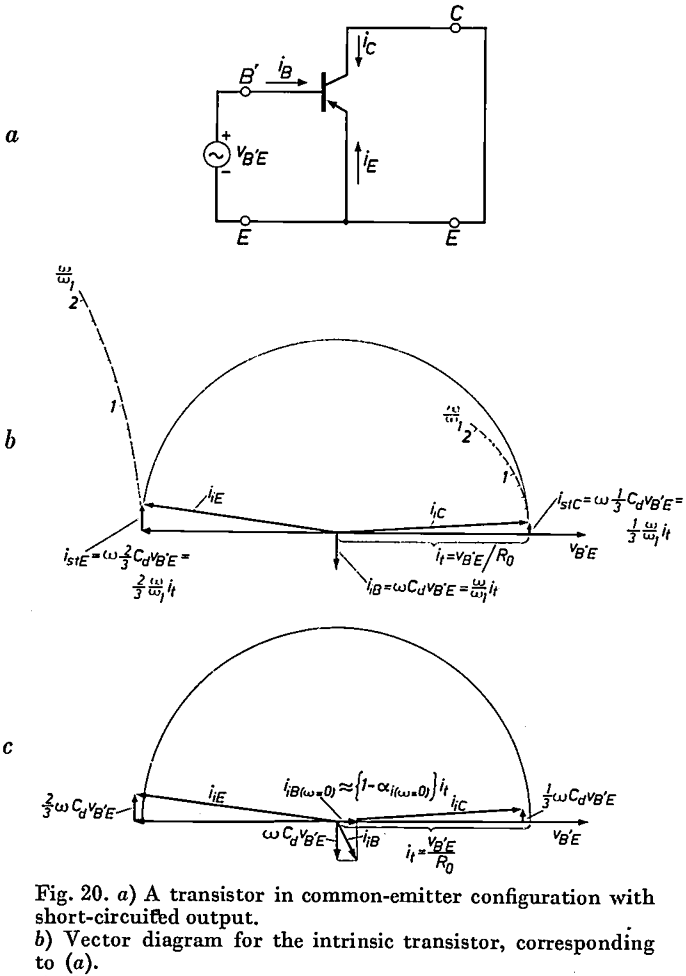

a
C
i_C
B'
i_B
+
$v_{B'E}$
-
i_E
E
E
b
$\frac{\omega}{\omega_1}$
2
1
$\frac{\omega}{\omega_1}$
2
1
i_{iE}
i_{iC}
$i_{stC} = \omega\frac{1}{3}C_d v_{B'E} = \frac{1}{3}\frac{\omega}{\omega_1}i_t$
$i_{stE} = \omega\frac{2}{3}C_d v_{B'E} = \frac{2}{3}\frac{\omega}{\omega_1}i_t$
$i_t = v_{B'E}/R_0$
$v_{B'E}$
$i_{iB} = \omega C_d v_{B'E} = \frac{\omega}{\omega_1}i_t$
c
$\frac{2}{3}\omega C_d v_{B'E}$
i_{iE}
$i_{iB(\omega=0)} \approx \{1-\alpha_{i(\omega=0)}\}i_t$
i_{iC}
$\frac{1}{3}\omega C_d v_{B'E}$
$\omega C_d v_{B'E}$
i_{iB}
$i_t = \frac{v_{B'E}}{R_0}$
$v_{B'E}$

Fig. 20. *a*) A transistor in common-emitter configuration with short-circuifed output.
b) Vector diagram for the intrinsic transistor, corresponding to (*a*).
c) The vector diagram (*b*) corrected for the base loss.

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 $R_0/(1 - a_{i(\omega=0)})$ 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.

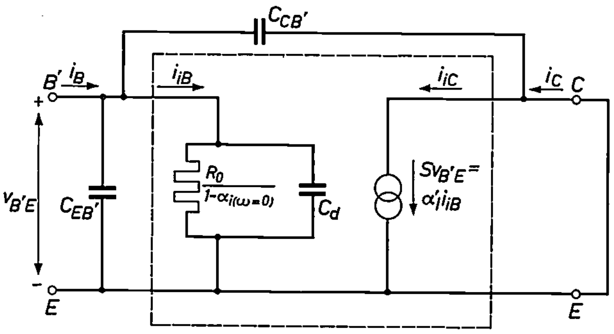

$C_{CB'}$
B'
i_B
i_{iB}
i_{iC}
i_C
C
+
$v_{B'E}$
$C_{EB'}$
$\frac{R_0}{1-\alpha_{i(\omega=0)}}$
C_d
$Sv_{B'E} = \alpha_i' i_{iB}$
-
E
E

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 a_i' and a_i of the intrinsic transistor in common-emitter and common-base configuration, respectively, there exists the relation $a_i' = a_i/(1 - a_i)$.

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 $a_i'i_{iB}$. The current amplification factor a_i' is provided with an accent to distinguish it from the current amplification factor a_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 $1/a_i' = (1/a_i) - 1$, that is $a_i' = a_i/(1 - a_i)$. 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 a'_i in the common-emitter configuration we can again define a cut-off frequency $\omega_{ca'_i}$ as the frequency at which the modulus of this factor has decreased by a factor $\sqrt{2}$. As we shall show, there exists between $\omega_{ca'_i}$ and ω_{ca_i} the simple relation:

$$\omega_{ca'_i} = \omega_{ca_i}/1.21a'_{i(\omega=0)} = \omega_1/a'_{i(\omega=0)} \,. \quad . \quad . \quad (13)$$

Since $a'_{i(\omega=0)}$ is much greater than unity in a good transistor, it follows from this that $\omega_{ca'_i}$ is much smaller than ω_{ca_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

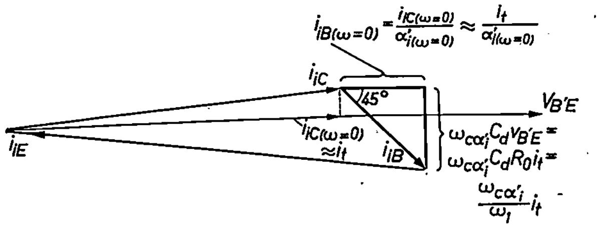

$i_{iB(\omega=0)} = \frac{i_{iC(\omega=0)}}{\alpha'_{i(\omega=0)}} \approx \frac{i_t}{\alpha'_{i(\omega=0)}}$
i_{iC}
45°
$V_{B'E}$
$i_{iC(\omega=0)} \approx i_t$
i_{iB}
i_{iE}
$\omega_{c\alpha'_i} C_d v_{B'E} = \omega_{c\alpha'_i} C_d R_0 i_t = \frac{\omega_{c\alpha'_i}}{\omega_1} i_t$

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 $R_{BB'}$.

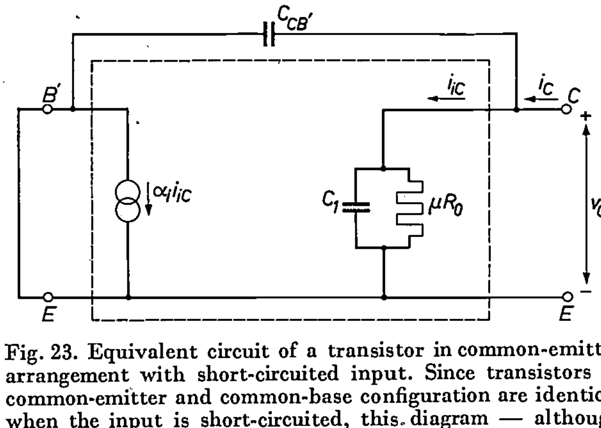


Fig. 23. Equivalent circuit of a transistor in common-emitter arrangement with short-circuited input. Since transistors in common-emitter and common-base configuration are identical when the input is short-circuited, this diagram — although drawn differently — is the same as that in fig. 17.

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

$$C_{CB'} = \frac{O\sqrt{\frac{1}{2}\varepsilon q c_{don}}}{\sqrt{V_0 + V_{B'C}}} . \qquad \ldots\ldots \quad (15)$$

To obtain a wide useful frequency range, $C_{CB'}$ must be small. According to formula (15) it is favour-

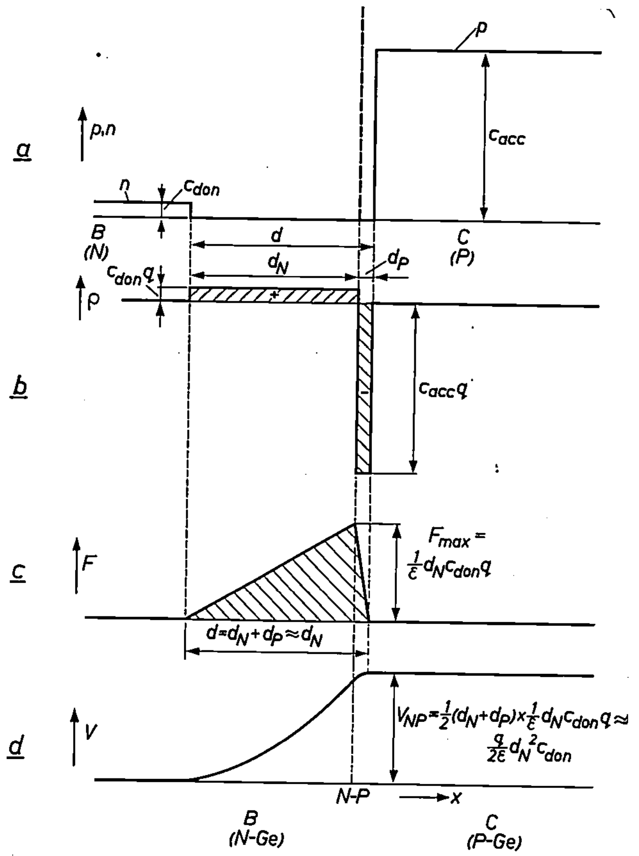

a
p,n
n
c_{don}
B
(N)
d
d_N
d_P
C
(P)
P
c_{acc}
$c_{don}q$
p
b
$c_{acc}q$
c
F
$F_{max} = \frac{1}{\varepsilon} d_N c_{don} q$
$d = d_N + d_P \approx d_N$
d
V
$V_{NP} = \frac{1}{2}(d_N+d_P) \times \frac{1}{\varepsilon} d_N c_{don} q \approx \frac{q}{2\varepsilon} d_N^2 c_{don}$
N-P
x
B
(N-Ge)
C
(P-Ge)

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 $V_{\text{B'C}}$ 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 $V_{\text{B'C}}$ cannot be raised ad libitum. Fig. 24*c* indicates that the maximum field strength $F_{\max}$ in the barrier has the value $qc_{\text{don}}\,d_N/\varepsilon$. Using formula (14) for d_N we find:

$$F_{\max} = \sqrt{2q/\varepsilon}\;\sqrt{(V_0 + V_{\text{B'C}})c_{\text{don}}}. \quad . \quad (16)$$

Thus, $F_{\max}$ increases with increasing $V_{\text{B'C}}$. 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):

$$V_{\text{br}} = \frac{\varepsilon F_{\text{cr}}^{2}}{2qc_{\text{don}}} - V_0 . \quad . \quad . \quad . \quad . \quad (17)$$

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 $R_{\text{BB'}}$. 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 $\sqrt{c_{\text{don}}}$. 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 extremely 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 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.

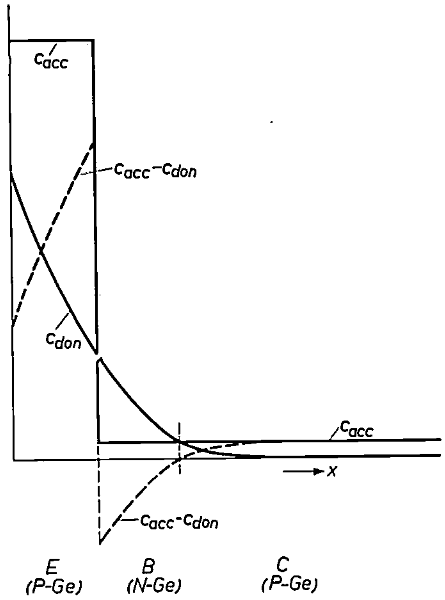


Fig. 25. The distributions of the concentrations of acceptors (c_{acc}) and donors (c_{don}) in a modern *P-N-P* high-frequency transistor (not to scale), where the donor concentration is obtained by diffusion. The dashed line represents the difference $c_{\text{acc}} - c_{\text{don}}$. Whether the region is *P*- or *N*-type depends on whether the acceptor or donor concentration is greater.

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:

$$I_F^+ = q\mu_p Fp \qquad I^+ = -qD_p\frac{dp}{dx}$$

$$I_F^- = q\mu_n Fn \qquad I_d^- = +qD_n\frac{dn}{dx}$$

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:

$$\left(\frac{I_F^+}{I_d^+} - \frac{1}{\frac{2n}{p}+1}\right)\left(\frac{I_F^+}{I_{tot}} - \frac{1}{\frac{2n}{p}+1}\right) = \frac{1}{\left(\frac{2n}{p}+1\right)^2} .$$

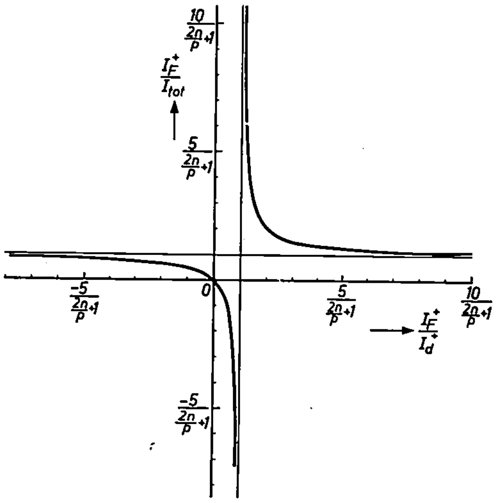


Fig. 26. Graphic illustration of the first fundamental theorem.

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:

$$p = p_0\,e^{-qV/kT} \text{ and } n = n_0\,e^{+qV/kT} . \qquad \text{(18a and b)}$$

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

$$V_0 = \frac{kT}{q}\ \ln\,(p_{P0}/p_{N0}) . \qquad (19)$$

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, 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.

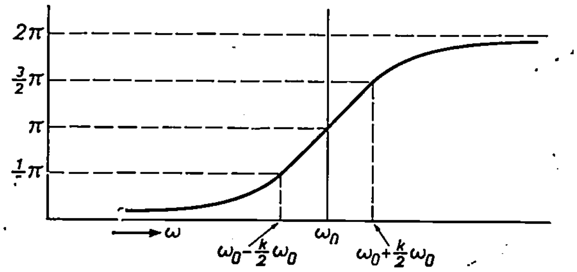


Fig. 1. Phase response characteristic used for audibility experiments. This curve which comprises a single phase "jump" of 360° near the angular frequency ω_0, is characteristic of the transfer function

$$\frac{1-\dfrac{\omega^2}{\omega_0{}^2}-\mathrm{j}k\,\dfrac{\omega}{\omega_0}}{1-\dfrac{\omega^2}{\omega_0{}^2}+\mathrm{j}k\,\dfrac{\omega}{\omega_0}}.$$

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.

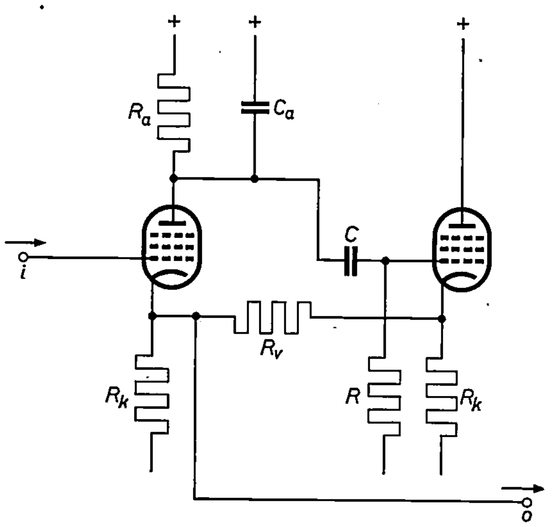


Fig. 2. Possible circuit for obtaining the transfer function mentioned in fig.1. Unlike the basically simple method indicated there, no *LC* circuit is used here, in fact not even a single inductance. For low values of ω_0 this is an advantage, not only because it is difficult to make the large inductances 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 $SR_k \gg 1$:

$$R_aC_a = RC = \frac{1}{\omega_0}, \qquad \frac{1}{k} = SZ_k = \frac{1}{2}SR_a,$$

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.)

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*).

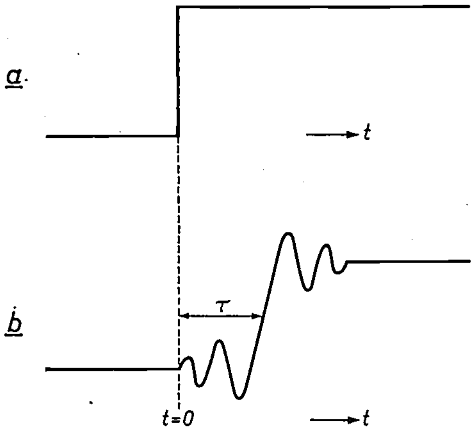


Fig. 3. Example of a phase-linear response (b) to a step-function signal (a) at the time $t = 0$. The response shows symmetry with respect to $t = \tau$. Physical reality of course demands that the transient should begin after $t = 0$, implying that there must necessarily be a delay τ.

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).

VOLUME 25, 1963/64, No. 8 pp. 181-216 Published 24th July 1964

Philips Technical Review

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

COMBINATIONS OF VALVES AND TRANSISTORS IN A STABILIZED 2000 V POWER SUPPLY

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

621.314.671:621.382.3

The article below — the third in a series on electronic circuits for special measuring equipment — describes a highly stabilized 2000 V DC power supply. A special feature is that in some parts of the circuit valves and transistors are used in combination. In this way results are obtained that can be achieved only with great difficulty using valves or transistors alone. An example given is a cascode consisting of an EL 34 pentode and an OC 139 transistor: the effective amplification factor of this cascode is greater than 10^5.

Thermionic valves and transistors are elements that do not at first sight seem likely partners in one and the same circuit. Closer examination shows, however, that it is possible to combine them to form circuits possessing unusual properties, enabling elegant solutions to be found for certain problems of circuitry.

Some years ago, for example, the need arose for a stabilized HT power supply to provide a variable DC voltage for the operation of scintillation counters. The photomultiplier, which is a major component of such counters, is extremely sensitive to fluctuations in the DC supply voltage. This voltage therefore had to be very highly stabilized.

The power-supply apparatus, delivering a current of 0 to 5 mA, had to meet the following specifications:

1) DC voltage variable in small steps from 400 to 2000V.
2) Optional earthing of either the positive or the negative terminal.
3) A 10% change in the mains voltage should cause no more than 0.01% change in the DC output voltage.

*) Philips Research Laboratories, Eindhoven.

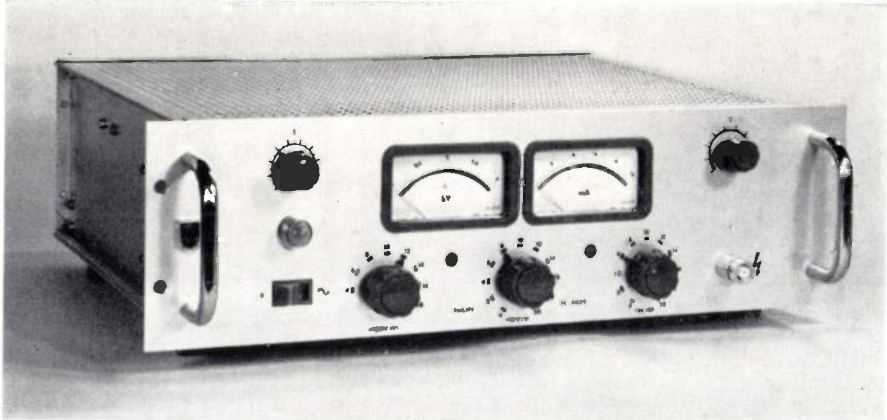


Fig. 1. The stabilized DC voltage source for 400-2000 V in a version that meets the same requirements as the apparatus described in this article, but which can supply a higher current, up to 10 mA.

4) A 10% change in the current drain should cause no more than 0.01% change in the DC output voltage.
5) The peak value of the sum of hum and ripple voltages should not exceed 10^{-5} times the chosen value of the DC voltage [1]).
6) In prolonged operation the chosen value of the DC voltage should remain constant to within 0.1%.
7) Short-circuiting of the DC terminals should have no adverse consequences.

Fig. 1 shows a later developed variant which meets the above specifications at currents from 0 to 10 mA.

Principle

Fig. 2 illustrates the principle of stabilized power supplies. A control system ensures that the instantaneous value of the output voltage U_o depends very little on the fluctuations present in the "uncontrolled" voltage U_i — irrespective of their cause — and on the changes in the load. The system consists of a valve T_1 (triode or pentode) which is in series with the load and is driven by an amplifier that amplifies the difference between a certain fraction k of the output voltage U_o and a fixed reference voltage U_{ref}.

To a good approximation the change ΔU_o in the output voltage produced by changes ΔU_i in the uncontrolled voltage and changes ΔI_o in the load current is given by:

$$\Delta U_o = \frac{1}{kA\mu}\,\Delta U_i - \frac{1}{kAS}\,\Delta I_o .$$

Here A is the gain of the control amplifier, while μ and S are respectively the amplification factor and the transconductance of the series valve T_1. The

$$P = kA\mu \text{ and } R_i = \frac{1}{kAS} \quad . . . \quad (1)$$

quantities will be called the stabilization and the internal resistance of the stabilized power supply, respectively.

Where the power is small and the requirements are not too rigorous as regards the constancy of the DC voltage, a convenient solution can sometimes be an RF oscillator whose output voltage is stepped up, rectified and stabilized. In the case under consideration, however, the power (10 W at the output) is on the high side for such a system, and would make it difficult to satisfy the stability and internal-resistance requirements.

In the case of stabilized power packs for *low* voltages the various capacitors can be generously dimensioned; this is particularly important as regards the buffer capacitor C_2 (fig. 2) across the output. In the case of *high* voltages, on the other hand, the volume, weight and price of the capacitors should not be unduly large compared with those of the other components. As will be shown, however, limiting the values of these capacitors can have far-reaching adverse consequences. For the capacitors C_1 and C_2 (fig. 2) a compromise value of 0.2 μF was chosen. From a simple calculation it follows that, using a smoothing capacitor of 0.2 μF and under full load, the ripple voltage of C_1 has an amplitude of 125 V.

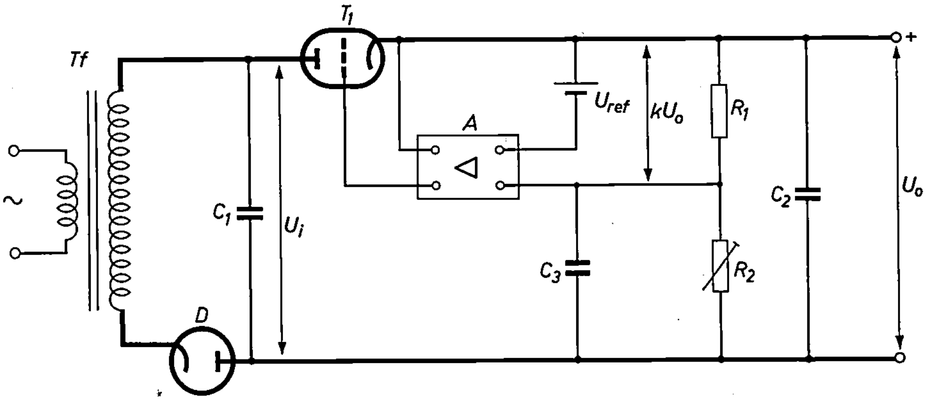


Fig. 2. Basic diagram of a stabilized power supply. *Tf* power transformer. *D* rectifying valve. C_1 smoothing capacitor across which the unregulated voltage U_i appears. T_1 series valve driven by the control amplifier *A*. The input voltage of *A* is the difference between the part kU_o of the DC output voltage U_o and a reference voltage U_{ref}; the fraction k is determined by the voltage divider R_1-R_2. Buffer capacitor C_2 is needed for the stability of the control system. C_3 bypasses R_2, so that the *full* ripple voltage of U_o is present at the input of *A*; in this way the ripple is minimized.

[1]) The disturbing voltages denoted as hum enter along stray paths, e.g. via the stray capacitance between the transformer coils. In the present case, screening in the transformer and appropriate assembly reduce the hum sufficiently for it to be disregarded in the following considerations.

The only type of valve suitable for use as T_1, while still belonging to the class of "receiving valves", is the EL 34 pentode. Since we preferred to avoid the use of tap switches on the HT winding of the transformer, a high voltage appears across T_1 in the case of a low output voltage: in the situation with the nominal mains voltage, the lowest output voltage (400 V) and full load, the series valve must take about 2500 V. At this voltage the anode dissipation (15 W) remains far below the permitted maximum, but the voltage itself exceeds the officially permitted maximum value, which is 2000 V for a cut-off valve. In the tests subsequently carried out on a number of EL 34 pentodes we did not, however, encounter any difficulties. *Fig. 3* gives a plot of the control-grid voltage V_{g1} and the screen-grid current I_{g2} versus the anode voltage V_a for constant anode current (5 mA) and constant screen-grid voltage (100 V). Both curves show that the valve still behaves quite normally at the exceptionally high anode voltages. We therefore preferred to accept the minor risk of the high tension rather than to adopt the alternative solutions of transformer tap switches, two EL 34 valves in series, or a transmitting valve (which would be much more expensive).

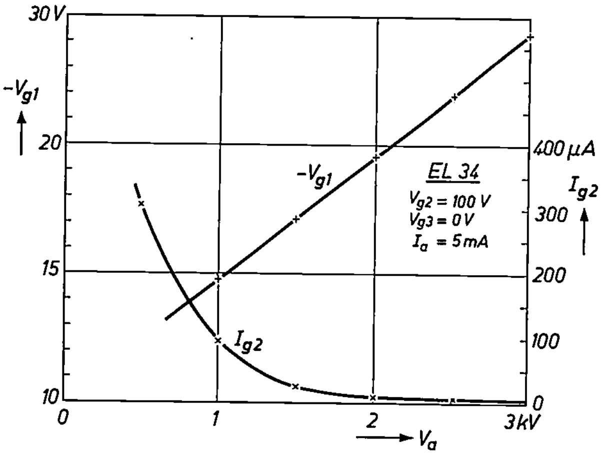


Fig. 3. Control-grid voltage V_{g1} and screen-grid current I_{g2} o the EL 34 pentode as a function of the anode voltage V_a, at constant anode current ($I_a = 5$ mA) and constant screen-grid voltage ($V_{g2} = 100$ V). The form of both curves indicates that the valve still behaves quite normally at anode voltages of 2 to 3 kV.

A screen-grid voltage of about 70 V is found to be sufficient for the EL 34 in our circuit. The slope of $V_{g1} = f(V_a)$ in fig. 3 corresponds to an amplification factor μ of 225. At 5 mA anode current the transconductance ranges from about 2.5 to 3.0 mA/V. We shall need these data presently.

Calculations relating to the gain

The specifications (3) and (5) above relate to the stabilization (of the mains-voltage fluctuations and the reduction of ripple resp.) and specification (4) concerns the regulation. In the following we shall calculate what the magnitude of the gain A of the control amplifier (denoted by A in fig. 2) should be to meet these specifications, taking the EL 34 as the series valve. We shall then consider what the permissible gain is, having regard to the transient response of the control system.

Stabilization against mains-voltage fluctuations

For *slow* changes the factor k in (1) is equal to $R_1/(R_1 + R_2)$; see fig. 2. Since the control system endeavours to keep kU_0 identical with U_{ref}, we can also write: $k = U_{\text{ref}}/U_0$.

At the nominal mains voltage and a current drain of $I_0 = 5$ mA, the value of U_i is about 2500 V. At a 10% deviation of the mains voltage, U_i therefore changes by approximately 250 V. According to the stabilization requirement, this should cause a change in U_0 not exceeding 10^{-4} U_0. The stabilization should therefore be at least $250/10^{-4}$ U_0:

$$\frac{\mu A\, U_{\text{ref}}}{U_0} \geqq \frac{250}{10^{-4}\, U_0}.$$

Given $\mu = 225$ and $U_{\text{ref}} = 83$ V, it follows that:

$$A \geqq 135. \qquad \ldots\ldots \qquad (2)$$

Smoothing of ripple

The ripple in the output voltage is P times smaller than the ripple in the uncontrolled voltage, provided that the factor k occurring in P (see eq. 1) can easily be made almost equal to unity for the frequencies of the ripple. For this purpose all that is needed is to bypass the resistor R_2 (fig. 2) with a capacitor C_3 of sufficiently high capacitance: almost the entire output ripple then appears at the input of the control amplifier A instead of only the fraction $R_1/(R_1 + R_2)$.

As mentioned above, under full load the amplitude of the ripple over C_1 is 125 V, the peak-to-peak value thus being 250 V. According to specification (5) the output ripple should be no more than 10^{-5} U_0. This specification is most severe when U_0 is set to the lowest value (400 V). One then finds:

$$\frac{250}{\mu A} \leqq 10^{-5} \times 400,$$

from which, given $\mu = 225$, it follows that:

$$A \geqq 270. \qquad \ldots\ldots \qquad (3)$$

Regulation

According to eq. (1) the internal resistance R_i of a stabilized power supply is: $R_i = 1/kAS$. The regulation specification (4) states that a 10% change in the current I_0 should cause a change of no more than $10^{-4}\,U_0$ in the output voltage. Hence:

$$\frac{1}{kAS} \leqq \frac{10^{-4}\,U_0}{0.1\,I_0}\,.$$

In this expression the maximum value (5 mA) should be substituted for I_0 and the minimum value (2.5 mA/V) for S. We then find, with $k = U_{\text{ref}}/U_0$ and $U_{\text{ref}} = 83$ V:

$$A \geqq \text{approx. } 25\,. \quad \ldots \quad (4)$$

From (2), (3) and (4) it follows that all three specifications are only met where $A = 270$.

Transient response

Control systems can respond to a step-function disturbance in various ways. In general, reactions with an oscillatory character are undesired: in other words, *the transient response should show little or no overshoot.* If this requirement is fulfilled, the *stability* of the control system will be assured, even with the conventional tolerance in the values of the components. The question is now whether, in the present case, a gain A of 270 is compatible with a transient response without overshoot.

For simple linear control systems with two time constants, τ_1 and $\tau_2 < \tau_1$, it is easily deduced that the condition for the required form of the transient response is: loop gain $kA \leqq \frac{1}{4}\,\tau_1/\tau_2$, hence:

$$A \leqq \frac{\tau_1}{4k\,\tau_2}\,. \quad \ldots\ldots \quad (5)$$

For simple linear control systems with two time constants (*fig.* 4) the following differential equation holds:

$$\left[\tau_1\tau_2\frac{\mathrm{d}^2}{\mathrm{d}t^2} + (\tau_1+\tau_2)\frac{\mathrm{d}}{\mathrm{d}t} + kA + 1\right] v = Av_{\mathrm{i}}\,.$$

For the transient response to be free from oscillations (including damped ones) the discriminant of the form between square brackets must be positive:

$$(\tau_1+\tau_2)^2 - 4\,\tau_1\tau_2\,(kA+1) > 0.$$

This condition may also be written as follows:

$$\frac{\tau_1}{\tau_2} + \frac{\tau_2}{\tau_1} > 4kA + 2.$$

If $\tau_1 \gg \tau_2$ and $kA \gg 1$, this reduces in good approximation to

$$\frac{\tau_1}{\tau_2} > 4kA\,,$$

from which (5) directly follows.

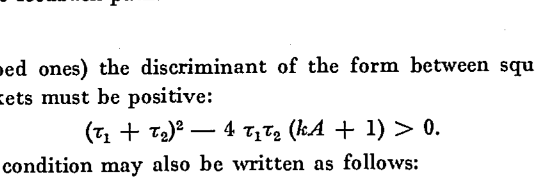


Fig.4. Schematic representation of a linear control system with two time constants, τ_1 and τ_2. The letter p denotes the differential operator d/dt, v the output voltage, and k the gain (< 1) in the feedback path.

In our case there are indeed only two principal time constants involved. One is the time constant of the control amplifier, the minimum value of which is determined by a resistance and a stray capacitance. If the gain A is to be of the order of 100, a time constant of at least about 5 microseconds should be taken into account. The other time constant is equal to the product of the differential resistance of the series valve (= $1/S$) and the buffer capacitance C_2. Either of the two time constants can be made the larger one by adding capacitance.

Which of the two should we now choose as the larger and which as the smaller time constant? In DC voltage supply apparatus it is favourable to shunt a buffer capacitance C_2 across the output; this capacitance smooths the effect of sudden changes in the load, and does so more effectively the larger it is. This apart, it remains of course desirable that the amplifier should respond as quickly as possible, i.e. that it should have the lowest possible time constant. Here, then, the obvious course is to make the latter time constant the smaller one: τ_2 = approx. 5 μs, and C_2/S the larger one. In the other case a sudden change in the load current would produce a jump in the output voltage which would only slowly be compensated by the amplifier: the power supply has a "recovery time". With many HT apparatus this is of the order of some tenths of a second. Given $k = U_{\text{ref}}/U_0$, eq. (5) now becomes:

$$A \leqq \frac{U_0}{4U_{\text{ref}}}\,\frac{C_2}{S\,\tau_2}\,. \quad \ldots\ldots \quad (6)$$

Substituting in this expression the minimum value (400 V) for U_0 and the maximum (3 mA/V) for S, and taking $U_{\text{ref}} = 83$ V, $C_2 = 0.2$ μF and $\tau_2 = 5$ μs, we find as condition for the transient response without overshoot:

$$A \leqq \text{approx. } 15.$$

The maximum permissible value $A = 15$ is thus far below the value 270 which, according to the foregoing considerations, is the minimum necessary to satisfy the requirements in regard to stabilization and regulation.

If it were a stabilized power supply for low voltage, one could easily get out of the impasse by choosing a larger capacitance for C_2: time constant τ_1 then increases, so that the gain A can be increased without the transient response showing overshoot. As remarked above, however, the voltage in the present case is so high that C_2 should preferably not be larger than 0.2 μF.

Cascode consisting of a pentode and a transistor

The conflict would disappear if the series valve had an amplification factor and a transconductance substantially greater than 225 and 3 mA/V respectively, for as appears from (1), the stabilization is proportional to μ and the regulation to S.

A high amplification factor might be obtained by combining the EL 34 with a second valve to form a *cascode*, the effective μ of a cascode being roughly equal to the product of the amplification factors of each of the valves [2]). If we make the EL 34 the "upper" valve in the cascode and the second valve the "lower" one, then the effective transconductance of the cascode is more or less identical with the transconductance of the second valve [2]). As regards transconductance, the cascode would thus be more favourable than the EL 34 if one could have a second valve possessing a transconductance substantially larger than 3 mA/V. No valves exist, however, with a transconductance of this order at currents from 1 to 5 mA.

By now using a *transistor* in lieu of the second valve, a solution is found which is satisfactory both as regards amplification factor and transconductance. Moreover, the power supply is simpler than with a cascode consisting of two valves. Because of the current direction, an *N-P-N* transistor is needed, e.g. an OC 139.

The circuit is shown in *fig. 5*. The control amplifier A drives the transistor between base and emitter, and the voltage between collector and emitter is used for driving the pentode (for the moment we shall disregard the resistance R_3). Since the latter voltage is —4 and —16 V in the extreme cases, the voltage between collector and emitter never rises above 20 V [3]), a value which must not be exceeded. A maximum of 0.1 W is dissipated in the transistor, which is far below the permissible maximum.

The effective amplification factor $\mu_{\rm tr}$ of transistors (this quantity is explained in the text in small print below) can easily attain a value of more than 500, so that the *effective amplification factor* $\mu_{\rm casc}$ of the cascode is greater than 10^5. The transconductance of transistors (defined similarly to that of valves) is also appreciably higher than the transconductance of valves at the same current, viz about 40 $I_{\rm e}$ mA/V, where $I_{\rm e}$ is the emitter current in mA [4]). This may be put in another way: whereas in valves the ratio $I_{\rm a}/S$ is of the order of 1 V at anode currents of a few mA, the corresponding ratio in transistors is of the order of 0.1 V, and even about 25 mV at very low collector currents.

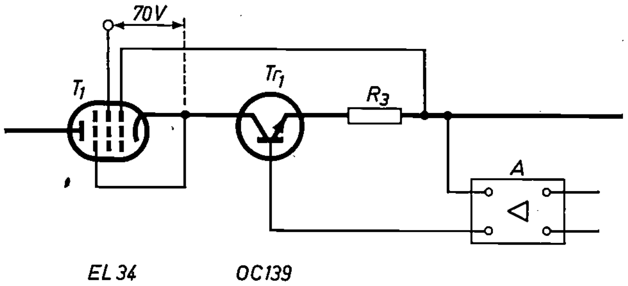


Fig. 5. Cascode formed by a valve T_1 (EL 34 pentode) and a transistor Tr_1 (*N-P-N* type OC 139). This cascode has an effective amplification factor of more than 10^5 and a high transconductance; it can therefore replace the valve T_1 in fig. 2 with advantage.

A control amplifier. R_3 (e.g. 68 ohms) limits the effective transconductance of the cascode (equal to the transconductance of the transistor, hence proportional to the emitter current); under full load the transconductance would otherwise become so high that the control system would show overshoot or even become unstable.

The effective amplification factor of a transistor circuit

Fig. 6 shows the simplified circuit containing the transistor Tr_1 of fig. 5. The resistances through which the emitter and base currents flow are denoted by $r_{\rm c}$ and $r_{\rm b}$ respectively, the amplified driving voltage $A(kU_{\rm o}$—$U_{\rm ref})$ by $v_{\rm s}$, and the current and voltage changes by i and v respectively; the subscripts e, b and c refer to the emitter, the base and the collector.

The changes of the collector and base currents are given by the following equations:

$$i_{\rm c} = S(v_{\rm b} - v_{\rm e}) + \frac{S}{\mu}(v_{\rm c} - v_{\rm e}), \quad . \; . \; . \quad (7)$$

$$i_{\rm b} = \frac{S}{a'}(v_{\rm b} - v_{\rm e}) - \frac{S}{a'\mu'}(v_{\rm c} - v_{\rm e}) \, . \quad . \; . \quad (8)$$

Eq. (7) is completely analogous with the equation for a triode, the cathode of which corresponds to the emitter, the grid to the base and the anode to the collector, and with a transconductance S and an amplification factor μ. In (8) a' is the amplification factor of the transistor in common-emitter circuits [5]), and μ' is a

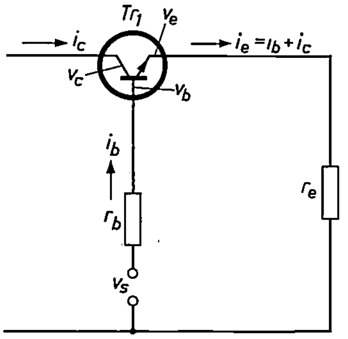


Fig. 6. For deriving the effective amplification factor $\mu_{\rm tr}$ of the transistor Tr_1 in fig. 5.

[2]) See e.g. Philips tech. Rev. **23**, 145, 1961/62.

[3]) Except when the DC terminals are short-circuited. We shall return to this point later.

[4]) F.H. Stieltjes and L.J. Tummers, Philips tech. Rev. **17**, 244, 1955/56.

[5]) See page 242 of the article under reference [4]).

factor that differs relatively little from μ. In addition the following relations hold:

$$v_e = (i_c + i_b)r_e \quad \ldots\ldots \quad (9)$$

and

$$v_s = v_b + i_b r_b . \quad \ldots\ldots \quad (10)$$

By adding (7) and (8) we find an equation from which, using (9), we can eliminate $i_b + i_c$. The result is:

$$\frac{1}{S}\frac{v_e}{r_e} = \left(1 + \frac{1}{a'}\right)(v_b - v_e) + \left(\frac{1}{\mu} - \frac{1}{a'\mu'}\right)(v_c - v_e). \quad (11)$$

Elimination of i_b and v_b from (8), (10) and (11) leads to an equation for v_e in which v_s and v_c occur with the following coefficients:

$$\frac{1 + \dfrac{1}{a'}}{1 + \dfrac{Sr_b}{a'}} \text{ and } \left(\frac{1}{\mu} - \frac{1}{a'\mu'} + \frac{\dfrac{Sr_b}{a'\mu'}}{1 + \dfrac{Sr_b}{a'}}\right), \text{ respectively.}$$

The first coefficient is a measure of the influence of v_s on v_e, the second is a measure of the influence of v_c on v_e. In the circuit of fig.5 they are therefore a measure of the respective influences of the driving voltage and the collector voltage (= cathode voltage of the EL 34) on the output voltage. The ratio of these coefficients may be called the *effective amplification factor* μ_{tr} of the transistor. Since in practice a' is much greater than unity and μ' is of the same order of magnitude as μ, we can write to a good approximation:

$$\mu_{tr} = \frac{1}{\dfrac{1}{\mu} + \left(\dfrac{1}{\mu} + \dfrac{1}{\mu'}\right)\dfrac{Sr_b}{a'}}. \quad \ldots . \quad (12)$$

If $r_b \leqq a'/S$ (of the order of 1 kΩ) and given $\mu' \approx \mu$, it follows from (12) that

$$\mu_{tr} > \tfrac{1}{3}\mu .$$

In most types of transistor μ is 1500 or more, so that μ_{tr} is greater than 500.

The quantity $(\mu^{-1} + \mu'^{-1})S$ which occurs in (12), and likewise $(\mu^{-1} + \mu'^{-1})$, are easily derived from the published characteristics of the transistors. Evidently, given a constant base current ($i_b = 0$), eq. (8) becomes:

$$v_b - v_e = \frac{1}{\mu'}(v_c - v_e).$$

From (7) we therefore find:

$$i_c = \left(\frac{1}{\mu} + \frac{1}{\mu'}\right) S\,(v_c - v_e),$$

hence

$$\left(\frac{1}{\mu} + \frac{1}{\mu'}\right) S = \frac{i_c}{v_c - v_e}. \quad \ldots\ldots \quad (13)$$

The last fraction corresponds to the slope of the characteristics which give the collector current as a function of v_c—v_e at a constant base current.

Plainly, using a cascode with such an exceptionally high effective amplification factor, we can accept a very low gain A (about 10). This can easily be obtained with a single stage. The number of time constants is thus kept to a minimum, aiding the simplicity and stability of the control system.

At 5 mA load the transconductance of the transistor (and hence of the cascode too) rises to about 100 mA/V. The time constant $\tau_1 = C_2/S$ therefore decreases to about 2 μs, which is of the same order as τ_2. To maintain the stability a resistor R_3 is therefore put in series with the emitter (fig.5). This makes $\tau_1 = C_2(R_3 + S^{-1})$, so that it is less dependent on S, in other words less dependent on the load. A resistance R_3 = 68 ohms is found to be large enough; the regulation requirement is for this value to remain amply fulfilled.

Other components

Reference-voltage source

The source of the reference voltage is a type 83 A 1 voltage stabilizer tube. The glow discharge in this tube has a sustaining voltage of about 83 V. Since the latter depends slightly on the current through the tube, the current is drawn from an auxiliary stabilized voltage source (—150 V with respect to the positive terminal). The temperature coefficient of the reference voltage is —3.5 mV/°C, i.e. roughly $-4 \times 10^{-5}\ U_{ref}$ per °C.

Control amplifier

The circuit diagram of the control amplifier is shown in *fig.* 7. It consists of a double triode T_2-T_3 (type E 80 CC) in push-pull, in which the voltage kU_0 is compared with the reference voltage. In this circuit the E 80 CC triode allows the use of an unstabilized heater voltage: a 10% change in this voltage has the same effect as a disturbing voltage at the input of no more than about 8 mV, i.e. $10^{-4} \times U_{ref}$. In most other types of tubes that might be used for this purpose the equivalent disturbing voltage is at least 3 times greater.

Since Tr_1 works single-ended, no anode resistor is needed for one half (T_2) of the E 80 CC triode. This half is fed from the —150 V auxiliary voltage source just mentioned. An auxiliary voltage of about +80 V is needed for the half with the anode resistor (T_3). These auxiliary voltage sources, which are both stabilized, will be discussed under the next heading.

The transistor Tr_1 must be driven by a source having a low impedance. In principle one might therefore insert a cathode follower between T_3 and Tr_1. Here, however, a valve has the drawback that a 10% change in the heater voltage causes the driving voltage to vary by 0.15 to 0.20 V, which corresponds to a disturbing voltage of 15 to 20 mV at the input of the control amplifier. For this reason, instead of a cathode follower a transistor (Tr_2) is used, which is circuited as an emitter follower, i.e. the emitter of Tr_2 (connected to the base of Tr_1) follows the voltage changes of the base of Tr_2, which is connected to the anode of T_3. The transistor Tr_2, like Tr_1, is a type OC 139.

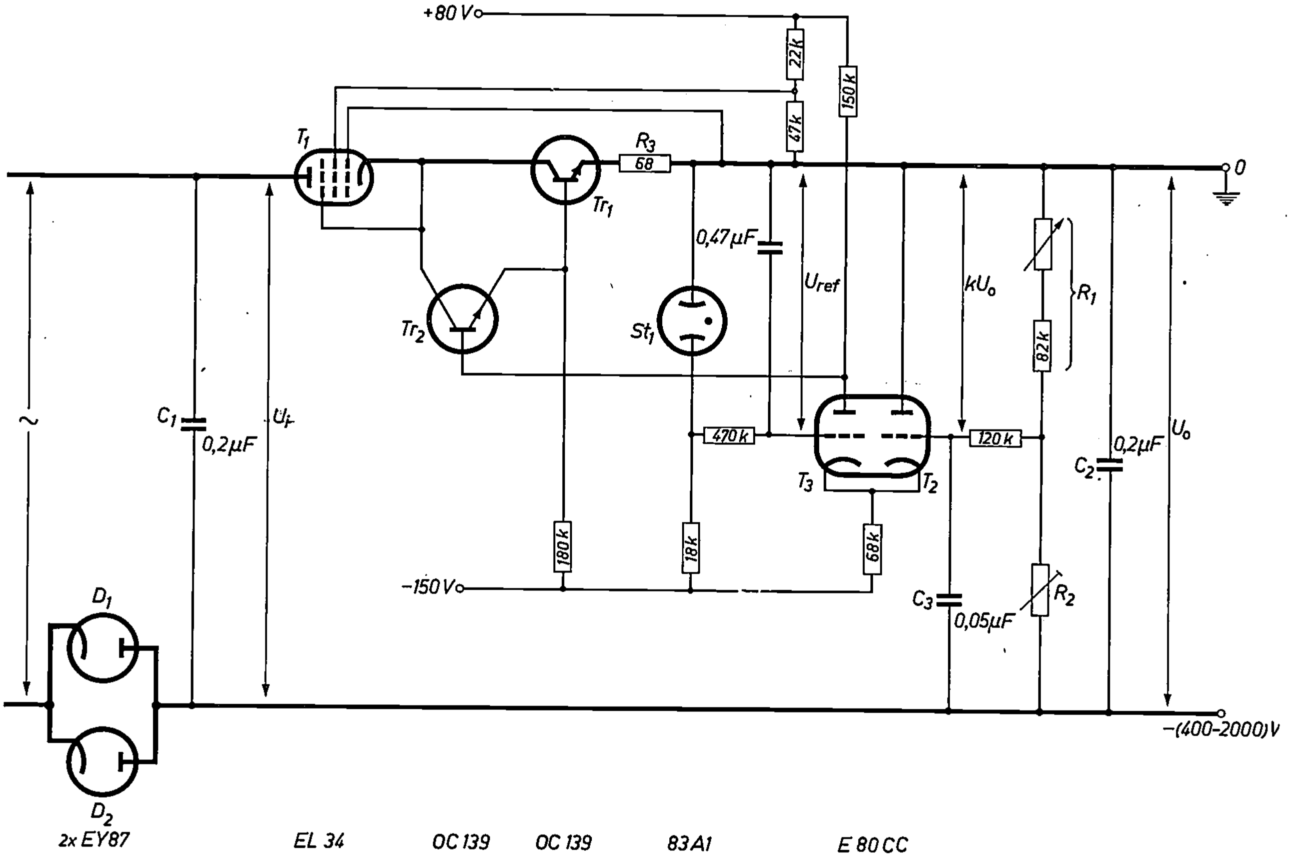


Fig. 7. The principal part of the circuit of the stabilized DC voltage source of 400-2000 V, maximum 5 mA, using the cascode T_1-Tr_1 of fig. 5. D_1, D_2 are two EY 87 rectifying valves in parallel. C_1, C_2, C_3, R_1 and R_2 have the same meanings as in fig. 2.

The control amplifier (A in fig. 2) contains a double triode T_2-T_3 in push-pull, with a common cathode resistor. Here the voltage kU_0 is compared with the reference voltage U_{ref} supplied by the voltage stabilizer tube St_1. The anode of T_3 drives the base of Tr_1 via a transistor Tr_2 (emitter follower).

T_2, T_3 and Tr_2 and the screen grid of T_1 are fed from stabilized auxiliary voltage sources of +80 and −150 V with respect to the positive terminal (here shown earthed).

Auxiliary voltage sources

As we have seen, two auxiliary stabilized voltage sources are needed, one of which delivers —150 V and the other +80 V with respect to the positive terminal. Both voltages are drawn from one auxiliary stabilized power supply. Here, too, a combination of a valve and a transistor is used to good advantage.

The circuit is shown in *fig. 8*. One half (T_4) of an ECC 82 double triode is used as the series valve of the auxiliary stabilized power supply, a stabilizer tube St_2 (type 150 A1) serves as the reference voltage source, and the other half of the ECC 82 (T_5) serves as the control amplifier. To obtain sufficient gain it is necessary to give T_5 a fairly high anode resistance. An ordinary resistor, however, would call for a high supply voltage. As such only the unregulated DC voltage is available, but if this were used it would considerably reduce the stabilization.

This difficulty is circumvented by employing a transistor (Tr_3) as the anode resistor. This can be done because a transistor with a fixed base voltage and a resistor in series with the emitter (fig. 8) has a very high differential resistance in the collector circuit. The DC resistance of the transistor is low, however, so that the stabilized DC voltage can be used (+80 V terminal) for feeding the valve T_5 and the transistor together. The only condition here is that the cathode of the series valve T_4 should be more than 3 V positive with respect to the grid, which is easily fulfilled.

Differential resistance of the transistor in fig. 8

The circuit containing the transistor Tr_3 of fig. 8 is shown in *fig. 9a*. Using equations (7), (8) and (9) and the relation $v_b = -i_b r_b$, we find for the changes in the collector and base currents the following expressions:

$$\left\{1 + \left(1 + \frac{1}{\mu}\right)Sr_e\right\} i_c + \left\{Sr_b + \left(1 + \frac{1}{\mu}\right)Sr_e\right\} i_b = \frac{S}{\mu} v_c,$$

$$\left(1 - \frac{1}{\mu'}\right)Sr_e\, i_c + \left\{a' + Sr_b + \left(1 - \frac{1}{\mu'}\right)Sr_e\right\} i_b = -\frac{S}{\mu'} v_c.$$

From this we can solve v_c/i_c, which is equal to the required

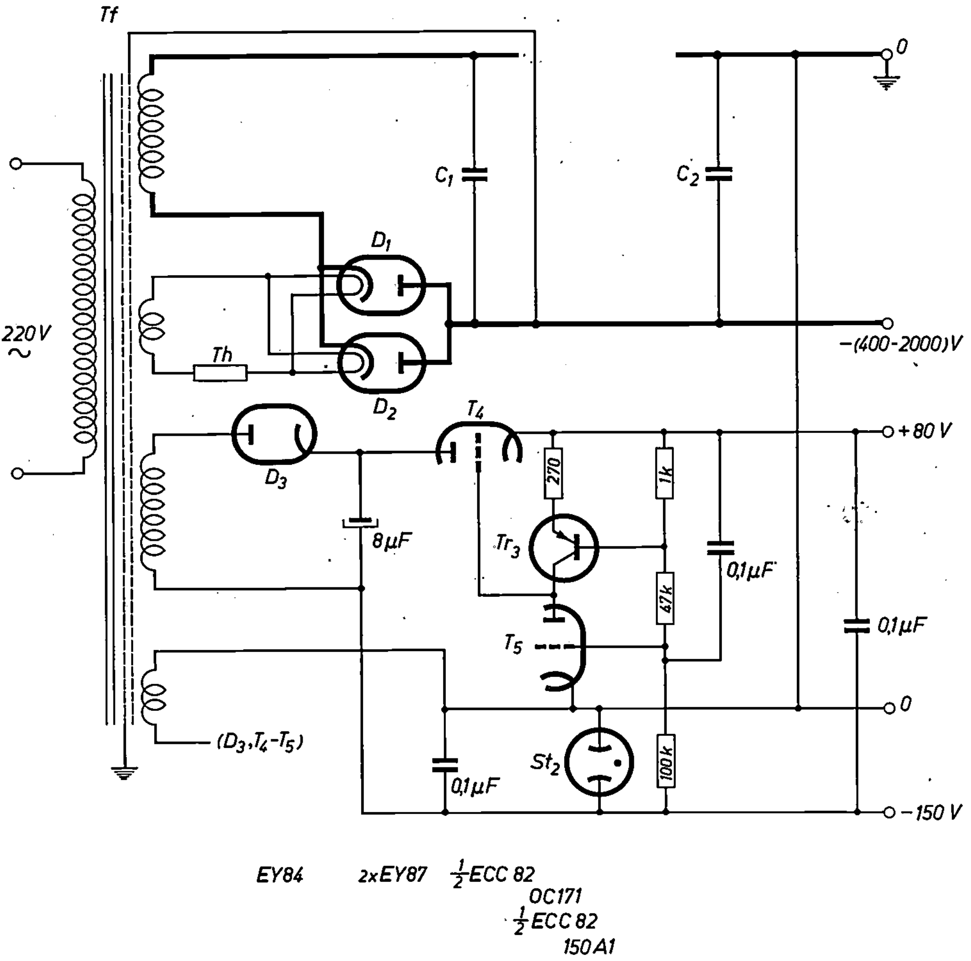


Fig. 8. Power-supply section. *Tf* power transformer with double screening. When the apparatus is switched on the thermistor *Th* in series with the heaters of the rectifiers D_1 and D_2 delays the heating-up of the cathodes so that the high tension does not come on until the control amplifier is already operating.

Rectifier D_3, series valve T_4, control amplifier T_5 and voltage stabilizer St_2 form a small stabilized power-supply circuit which delivers the auxiliary voltages of +80 and −150V with respect to the positive terminal. The anode resistor of T_5 is a transistor Tr_3, which combines a low DC resistance with a high differential resistance, so that sufficient gain is obtained without the need for extra supply voltage.

resistance R_d. The expression then found is complicated, but with some minor simplifying assumptions it can be reduced to the following formula:

$$R_d = \frac{\dfrac{\alpha'}{S} + \alpha' r_e + r_b}{\dfrac{\alpha'}{\mu} + \left(\dfrac{1}{\mu} + \dfrac{1}{\mu'}\right) S(r_e + r_b)} \quad . \quad . \quad (14)$$

We have found above that the expression $(\mu^{-1} + \mu'^{-1})S$ occurring in (14) corresponds to the slope of the characteristics which give the collector current as a function of the voltage between collector and emitter at constant base current; see eq. (13). For $(\mu^{-1} + \mu'^{-1})S$ we shall now briefly write $1/\varrho$.

To obtain high values of R_d it is necessary to ensure that r_e is preferably larger and at least not much smaller than r_b. If, moreover, $Sr_e \gg 1$, we can simplify (14) to:

$$R_d = \frac{1}{\dfrac{1}{\mu r_e} + \dfrac{r_e + r_b}{r_e}\,\dfrac{1}{\alpha' \varrho}} \quad . \quad . \quad . \quad . \quad (15)$$

From (15) it is seen that R_d may be regarded as the parallel arrangement of a resistance μr_e and a resistance $\alpha' \varrho r_e/(r_e + r_b)$; see fig. 9*b*. The differential resistance is larger the greater is the μ of the transistor, i.e. the less influence the collector voltage has on the collector current. In transistors made by the alloy-

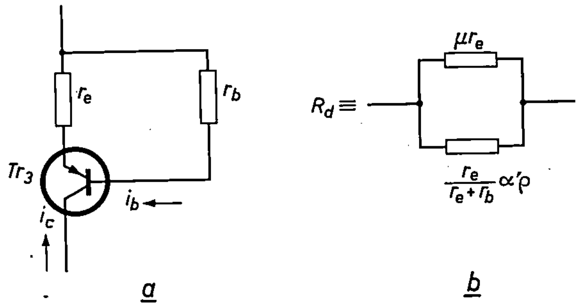


Fig. 9. *a*) For deriving the differential resistance of Tr_3 in fig. 8. *b*) The differential resistance R_d can be regarded to a good approximation as the parallel arrangement of a resistance μr_e and a resistance $\alpha' \varrho r_e/(r_e + r_b)$; see eq. (15).

diffusion technique ("p.o.b." transistors)[6]), such as type OC 171, this influence is much less than in other transistors, e.g. type OC 71. With an OC 171 transistor, therefore, exceptionally high differential resistances can be obtained.

If r_e is infinitely large and r_b is finite, R_d according to (15) attains its maximum value, which is equal to $\alpha'\varrho$. For the OC 171 this limiting value is about 10 MΩ, for the OC 71 about 0.5 MΩ. As already noted, in the present case a much lower value of R_d suffices.

Voltage divider

The DC output voltage can be varied by varying the resistance R_2 of the voltage divider (fig.7). In a particular case it was required that the voltage should be variable in steps of 200 V from 2000 V to 400 V, and moreover that there should be ten steps of 20 V and ten steps of 2 V available. For this purpose R_2 was built up from seven resistors of 180 kΩ, ten of 18 kΩ and ten of 1.8 kΩ; R_1 consisted of a fixed resistor of 82 kΩ in series with a continuously variable correction resistor of 0 to 15 kΩ. All these resistors are metallic and their temperature coefficient is lower than 5×10^{-5} per °C.

When varying R_2 measures are needed to guard against switching transients that could endanger the transistors. This is done by the circuit described below, which offers protection against the consequences of short-circuiting the output.

The transistors could also be damaged if the high tension came on too quickly after switching on the apparatus. For this reason a thermistor (*Th* in fig.8), which at room temperature has a 70 times greater resistance than at the working temperature, is connected in series with the heaters of the rectifying valves (2 EY 87 rectifiers in parallel). This thermistor delays the heating of the cathodes to such an extent that the high tension does not come on until the control amplifier is already in operation.

Protection against the consequences of short-circuiting of the output

It is particularly useful for power-supply apparatus — especially if intended for experimental purposes — to be proof against short-circuiting. It proved possible to give this property to the DC voltage source described here by the addition of only one transistor, one diode and a few resistors.

A short-circuited output presents two dangers:

1) During a short-circuit the voltage between the cathode and control grid of the EL 34 can rise above 20 V, thereby exceeding the permissible value for the transistor Tr_1.

2) The direct current becomes so high that various components are overloaded.

Both dangers are averted by using the *P-N-P* transistor Tr_4 (type OC 71) in combination with the resistor R_4 in the circuit of *fig. 10.* The emitter of Tr_4 carries with respect to the positive terminal a fixed voltage of +5 V, taken from a voltage divider across the stabilized auxiliary voltage source of +80 V; the collector of Tr_4 is fed via a resistor from the point carrying the stabilized voltage of —150 V. T_1 can be regarded as the series valve and Tr_4 as the control amplifier of a primitive stabilized power supply having a reference voltage of 5 V, which tries to keep the cathode of T_1 at a potential of 5 V with respect to the positive terminal. Since the control grid of T_1 is no longer connected to the emitter of Tr_1, but to a point of lower potential (the collector of Tr_4), the voltage between control grid and cathode can safely exceed 20 V without endangering Tr_1. Tr_4 is a transistor that can take a collector voltage of 30 V.

Since Tr_4 keeps the cathode of T_1 at a voltage of about 5 V, this voltage is also across the series arrangement of R_4, Tr_1 and R_3. The current through this configuration cannot therefore go higher than $5\,\mathrm{V}/(R_4 + R_3)$, though it may well be lower, de-

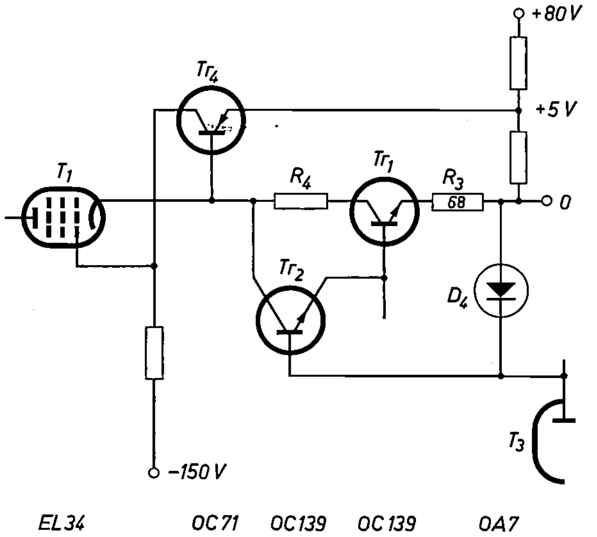


Fig.10. Circuit for safeguarding the apparatus against the consequences of short-circuiting.

T_1, Tr_1, R_3 and Tr_2 have the same meanings as in fig. 7. Since the control grid of T_1 is now at a point of lower potential than before, there is no longer any danger that Tr_1 will have to carry more than 20 V under exceptional conditions. Transistor Tr_4 can be regarded as a control amplifier which tends to keep the cathode of T_1 at a fixed potential with respect to the positive terminal, equal to the reference voltage of 5 V. A voltage of about 5 V is therefore across $R_4 + Tr_1 + R_3$, so that the current has an upper limit of $5\,\mathrm{V}/(R_3 + R_4)$ ohms. The short-circuit current is limited to approx. 7 mA by making $R_4 =$ 650 ohms.

Germanium diode D_4 — normally non-conducting — prevents a high negative voltage surge being applied to the base of Tr_2 upon the removal of a short-circuit.

[6]) P.J.W. Jochems, The alloy-diffusion technique for manufacturing high-frequency transistors, Philips tech. Rev. 24, 231-239, 1962/63 (No. 8).

pending on the drive for Tr_1. By the choice of $R_4 = 650$ ohms, the current concerned is limited to 5 V/(650 + 68) ohms = approx. 7 mA and the short-circuit current cannot therefore exceed this harmless value. The short-circuit current can be adjusted by making R_4 variable.

The foregoing considerations apply to the *stationary* short-circuited state. The occurrence and the removal of the short-circuit give rise to transients that could be dangerous to certain transistors. The removal of the short-circuit endangers transistor Tr_2. This can easily be avoided by inserting between the base of Tr_2 and the positive terminal a diode D_4 (type OA 7 germanium diode); see fig. 10. In the normal state diode D_4 passes no current, but upon removal of a short-circuit it becomes momentarily conductive, thereby preventing the voltage at the base of Tr_2 from going too negative.

The tendency of Tr_4 to keep the cathode of T_1 at a fixed potential (5 V) also contributes to the stabilization, which is thus an incidental advantage of the circuit.

Results

Measurements have been carried out under the following load conditions:

2000 V - 0 mA, 2000 V - 5 mA,
400 V - 0 mA, 400 V - 5 mA.

Stabilization against mains-voltage fluctuations. Tests on various E 80 CC double triodes (T_2-T_3) at the above settings showed that a 10% increase or decrease of the mains voltage caused a change in the output voltage ranging from 0.008 to 0.002%. The mutual disparities are attributable to the difference in the effect of changes in the heater voltage in the E 80 CC. Stabilization of the heater voltage can reduce these figures to about 0.001%. If even better stabilization of the auxiliary voltages (+80 and —150 V) were applied, the output variations would ultimately be caused only via the still unstabilized high tension; in that case the variations are smaller than 1 in 10^6.

Regulation. A change of 0.5 mA in the current drain caused a change of 0.005% in the output voltage.

Hum and ripple. At all settings, both with earthed positive and earthed negative terminals, the hum plus ripple was found to be lower than 2 mV peak-to-peak (0.0005% of the output voltage when this was 400 V).

Long-term drift. After the apparatus had been switched on for one hour, the drift in the output voltage during many hours of operation was less than 0.01%, in an ambient temperature which remained constant within a few °C. When the apparatus, after having been switched off, was again switched on, the output voltage returned after some time to within 0.01% of the original value. The magnitude of this deviation is principally determined by the quality of the E 80 CC valve and of the resistors in the voltage divider R_1-R_2. A limit of 0.1% can certainly be guaranteed.

Summary. Description of a DC voltage source for 400-2000V and 0-5 mA which meets very high demands: a 10% fluctuation in the mains voltage or in the current drain causes a maximum of 0.01% change in the DC voltage; hum + ripple (peak-to-peak) is no more than 10^{-5} times the output voltage; voltage drift maximum 0.1%. If an EL 34 pentode is chosen as the series tube, more gain is needed to meet the requirement than is compatible with the specified absence of overshoot in the control system. A favourable solution was found by combining the EL 34 with an OC 139 *N-P-N* transistor to form a cascode, the effective amplification factor of which is greater than 10^5. The control amplifier drives this transistor via a second transistor OC 139. In an auxiliary stabilized power supply, delivering two stabilized DC voltages, the anode resistor is a third transistor (OC 171), which has a high differential resistance and a low DC resistance. A fourth transistor (OC 71) safeguards the apparatus against short-circuits.

A PRECISION GRINDING MACHINE FOR TRACER DIFFUSION STUDIES

621.925

To investigate the diffusion of a particular element in a solid the following procedure is one often used. The element is deposited on a slice of the relevant solid — the specimen — which is then annealed for a suitable time at a controlled temperature in a furnace. To indicate the order of magnitude of the various quantities involved, we shall take as an example the results of an experiment on the diffusion of manganese in gallium arsenide, where the diffusion took place during 18 hours at 900 °C. At the end of the experiment the curve of the manganese concentration c_{Mn} versus penetration depth x was as shown in *fig. 1*. The method of deriving from this the diffusion constant and other required quantities does not concern us here [1]).

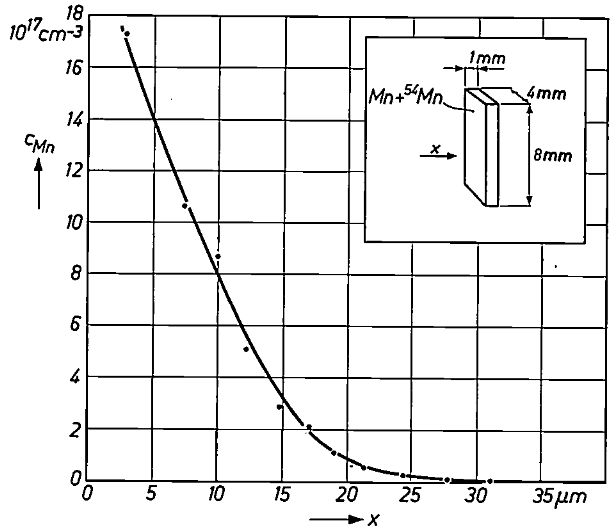


Fig.1. Diffusion of manganese in gallium arsenide. The curve gives the manganese concentration c_{Mn} (in atoms per cm^3) as a function of the penetration depth x. The sketch gives the dimensions of the specimen, the arrow indicating the direction of diffusion. The manganese, which contains a small fraction (about 0.1%) of the radioactive isotope ^{54}Mn, is applied to the specimen by moistening a side face with a solution of Mn*Cl$_2$ and letting the solvent evaporate. (Mn* means a mixture of normal and radioactive Mn.)

The problem is how to determine the concentration curve. For this purpose thin layers can be removed from the specimen by etching, grinding or turning, and the concentrations of the diffusing element in each successive layer determined. Chemical determination of these concentrations is time-consuming and moreover not always accurate enough, while for self-diffusion studies the method cannot be applied. Therefore use is made of a tracer diffusion method, employing a radioactive isotope of the element — provided a suitable one exists. The concentration curve can be established in two ways, either by measuring the radioactivity of the specimen after each layer is removed or by measuring the radioactivity of the removed layers. The latter method is the more accurate, and in the case of metals easy to apply. The layers are machined off on a precision lathe and the turnings collected. Where the materials are hard and brittle, however, grinding is necessary, and special steps are then required to ensure that all the removed material is collected. In such cases the measurements are often done on the specimen itself to save trouble, but at the expense of reduced accuracy.

The radioactivity involved in such experiments is always weak and no safety measures, such as lead shielding, are necessary.

In the literature descriptions have been given of several precision grinding machines specially designed for diffusion studies and allowing the complete collection of the material ground from the specimen [2]). We too have built a grinding machine of this kind (*fig. 2*), which serves its purpose very satisfactorily.

From the sketch (*fig. 3*) it can be seen that the machine has a horizontal grinding table *G*. The specimen *S* rests on a piece of brass foil *F*, 0.1 mm thick, which is held by suction to the surface of the table. To this end the table top contains concentric grooves which are connected by radial grooves to a central hole. Fig. 3 shows how the latter ultimately communicates with a vacuum line V_1. The shaft A_G about which the table rotates is mounted 30 mm off centre in the intermediate plate *E*, which rotates about the shaft A_E. When *E* rotates — driven by the pinion *P* — the grinding table rolls around the inside of the housing *I* and describes a planetary motion. To this effect the grinding table and housing are provided with gear teeth, the housing of course internally. The resultant rolling motion of the table ensures uni-

[1]) See e.g. P. G. Shewmon, Diffusion in solids, McGraw-Hill, New York 1963.

[2]) W. C. Dunlap, Jr., Diffusion of impurities in germanium, Phys. Rev. **94**, 1531-1540, 1954, in particular pages 1534 and 1536. H. Letaw, Jr., L. M. Slifkin and W. M. Portnoy, A precision grinding machine for diffusion studies, Rev. sci. Instr. **25**, 865-868, 1954. B. Goldstein, Precision lapping device, Rev. sci. Instr. **28**, 289-290, 1957. H. W. Schamp, Jr., D. A. Oakes and N. M. Reed, Grinder for sectioning solid diffusion specimens, Rev. sci. Instr. **30**, 1028-1031, 1959.

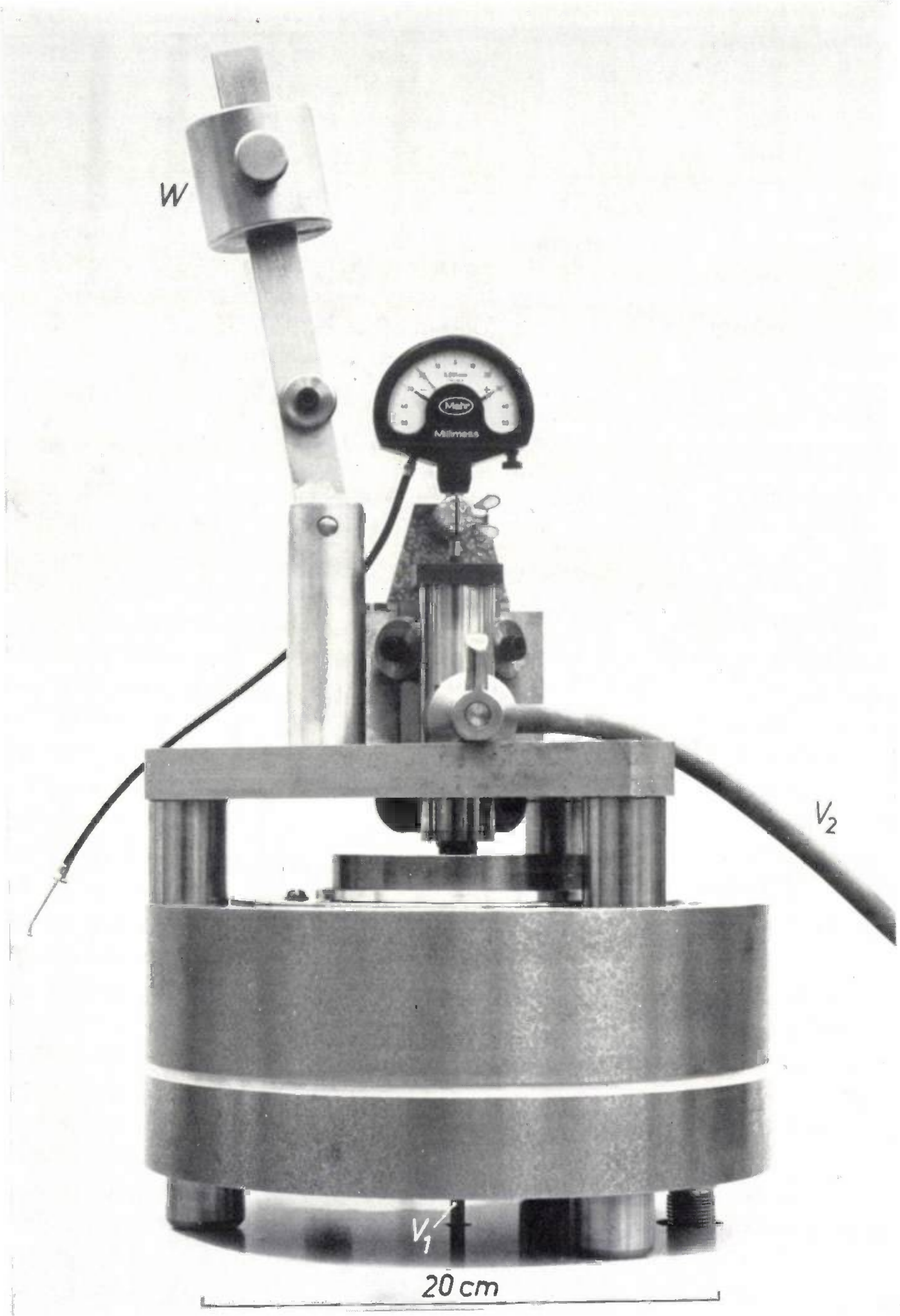


Fig. 2. The precision grinding machine for studying diffusion in solids.

form grinding. The ratio of the pitch circles of the teeth of table and housing is 5 : 8, so that E makes five and G eight revolutions before the table returns to exactly the same position. The components U, I, E and G are of pearlitic cast iron.

The specimen is fixed to the bottom of a carrier disc D by means of wax, the disc being held by suction to the vertical holder H. Both D and H are made of stainless chromium steel. The vacuum line V_2 can be seen in fig.2. The holder is pressed against the ball bearings K (a total of four) by the spring V and a fifth ball bearing K'. The holder is therefore able to move in the vertical direction without much friction, and can also readily be removed as a whole from the machine by pulling the bearing K' outwards against the action of spring V. The dead weight of the holder (approx. 600 g) is usually sufficient to press the specimen against the grinding table. The pressure can be increased by a sliding weight W on an arm (fig.2), which can be laid on the holder with a roller. So far, however, this has not proved necessary. The arm is not drawn in fig. 3.

Before grinding begins, the grinding plate is provided with a new brass foil F, to which the abrasive and a few drops of kerosene are applied. During grinding the abrasive distributes itself over a ring-shaped path on the foil (*fig.4*). The ground particles remain behind here, except for a fraction that adheres to the specimen and which is collected in the wadding used to clean the specimen. The wadding and the foil together thus contain all the material ground from the specimen. The abrasive used is diamond powder with grain sizes of 2-4, 4-7, 10-15 or 20-40 µm; the thicker the layer to be removed the coarser the abrasive.

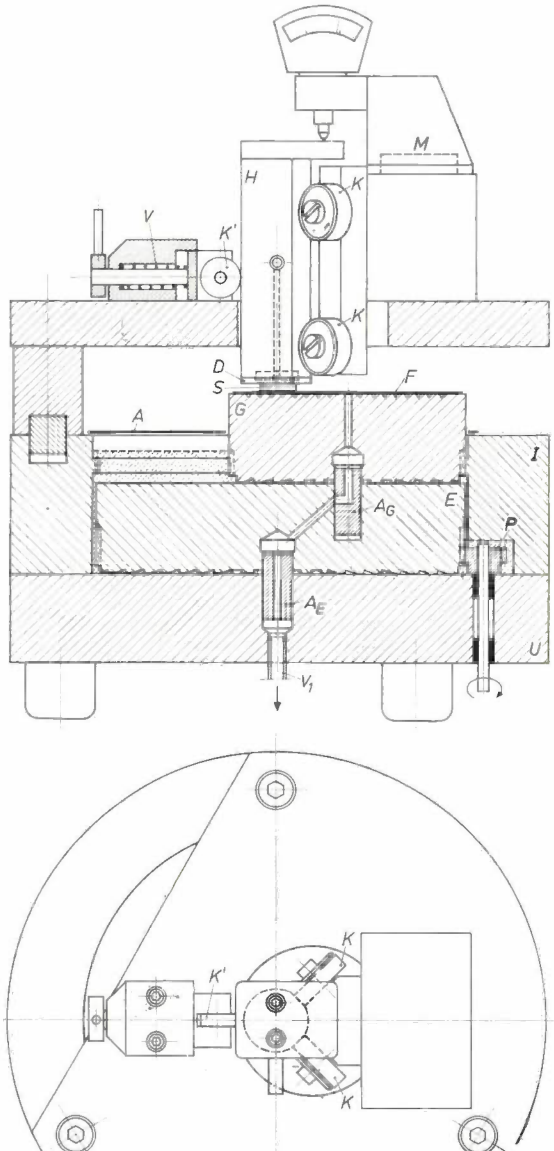


Fig. 3. Sketch of the grinding machine in fig. 2.

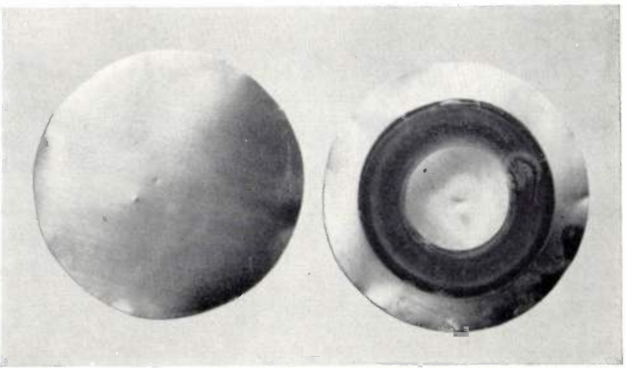


Fig. 4. Brass foils 0.1 mm thick, which are held by suction to the grinding table. Left, a new foil; right, a used one. It can clearly be seen on the latter where the specimen was last in contact with it.

During the grinding process the reduction of specimen thickness can be followed roughly on a dial gauge mounted on the holder H (figs. 2 and 3). For easy removal, which is necessary if the holder H has to be lifted out of the machine, the dial gauge is attached by means of a permanent magnet M (fig.3).

The exact thickness of the removed layer is determined by weighing together the specimen and the carrier disc D before and after a grinding run, or by measuring the thickness of specimen and disc together on a dial gauge before and after grinding. The fact that the specimen does not have to be detached from D for every measurement is a great advantage in connection with the successive grinding of numerous layers. The disc D weighs about 20 g, the specimen roughly 1 g or less. The combined weight of disc and specimen is therefore not too great for it to be determined with a precision balance. If the area of the specimen to be ground is very small (less than about 0.5 cm^2), the thickness measurement is more exact.

The concentration of the diffusing element in the removed material is easiest to determine by a gamma-radiation measurement. In most unstable atomic nuclei the radioactive decay is accompanied by the emission of gamma rays, so that this method can usually be adopted. For the measurement the brass foil is folded up — with the radioactive particles inside — and placed, together with the wadding used to clean the specimen, in a glass vial of e.g. 16 mm diameter and 50 mm length (*fig.5*). The gamma radiation is measured with a well-type scintillation crystal of NaI(Tl) (sodium iodide with 0.1% thallium); see *fig. 6*. We use a commercially available combination of crystal and photomultiplier for counting the scintillations. This method of counting is readily reproducible because the gamma rays are not significantly absorbed by the brass and the vial, and the crystal encloses the vial so well that it

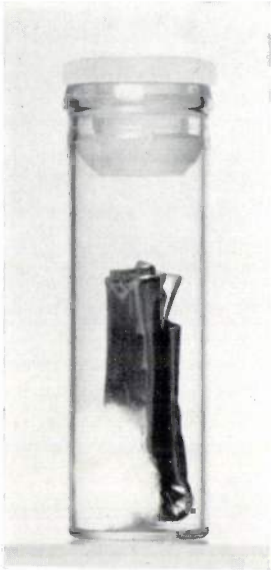


Fig. 5. A vial containing a folded brass foil with ground-off radioactive material and the wadding used for cleaning the specimen, ready for insertion in the scintillation counter.

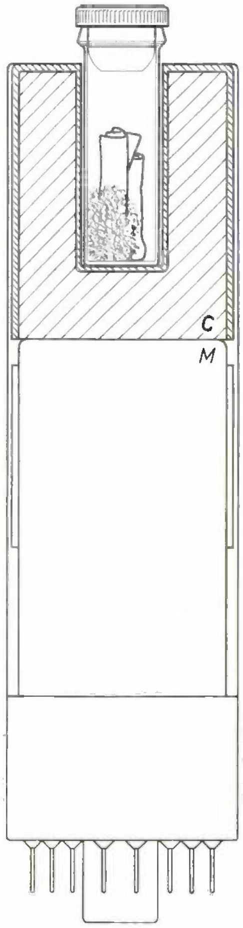


Fig. 6. For measuring the gamma radiation, the vial in fig. 5 is placed in a well-type scintillation crystal *C*, which receives a considerable part of the radiation and is mounted on a photomultiplier tube *M*.

Fig. 6

receives a substantial part of the radiation (usually between 20 and 70%, depending on the energy of the gamma quanta). If beta radiation is to be measured it is necessary — owing to the strong absorption — to prepare the specimens very carefully. Normally the brass foil has to be dissolved and the radioactive material chemically separated. For the actual measurement a Geiger counter suffices, which is much cheaper than the equipment needed for measuring gamma radiation.

The thicknesses of the removed layers vary from a few microns to several hundred microns. The preferred thickness is about 20 μm. From the example given in fig.1, however, which relates to measurements done with the apparatus described here, it can be seen that good results are also obtained with layer thicknesses of a few microns.

The thinner the layers the more important it is that each layer should have a uniform thickness, in other words be evenly ground. The base plate *U* of the machine should therefore be scrupulously flat and the grinding table *G* and the intermediate plate *E* accurately flat and plane-parallel. The carrier disc *D* and the specimen can be ground plane-parallel on the machine itself. As regards the specimen, this must of course be done before diffusion takes place. When fixing the specimen with wax, care is needed to ensure that no wax comes between specimen and carrier disc.

Finally, it is important that the oil films between *U* and *E* and between *E* and *G* should be plane-parallel. We have achieved good results by using thin oil and providing the grinding plate and the intermediate plate *E* underneath with a spiral groove, which pumps the oil inwards while the machine is turning [3]). These plates also contain radial grooves underneath, along which the oil returns and is thus kept circulating. The housing *I* is filled with oil up to the horizontal dashed line in fig.3. A cover plate *A*, containing a recess for the grinding table, protects the oil from dust. To prevent the oil being completely forced out when the machine is stationary, we use oil that adheres well to the metal, being the type used to lubricate the bed of planing machines.

After successive grinding operations the specimen has been found to remain flat and plane-parallel to within 1 μm; the accuracy is probably even better, but it is difficult to establish this with certainty. With the intermediate plate turning at 25 revolutions per minute the grinding time per layer is between one minute and half an hour, depending on the thickness to be ground away, the material of the specimen and the diamond powder used.

L. M. L. J. LEBLANS *).
M. L. VERHEIJKE *).

[3]) This kind of bearing will be dealt with in an article by E.A. Muijderman, which will shortly be published in this journal.

*) Philips Research Laboratories, Eindhoven.

INHOMOGENEITIES IN DOPED GERMANIUM AND SILICON CRYSTALS

by J. A. M. DIKHOFF *).

548.4:546.28

The demands made on the quality of doped germanium and silicon in science and industry are steadily increasing. This has stimulated research into the nature and genesis of inhomogeneities which are not normally detected by routine measurements on complete crystals. The article below gives a survey of the present state of knowledge on this subject, illustrated by a number of photos. Most of these show striation patterns which can occur in different cases, and which often give a clear record of what has happened during the growth of the crystal.

The technique of making rod-shaped germanium and silicon crystals containing desired amounts of impurities ("dope"), which are used for the production of transistors and other semiconducting switching elements, has in recent years reached a high state of perfection. It is now possible to make such doped crystals in which the concentration of impurities not only has the desired overall value but is also remarkably constant throughout the whole crystal [1]). The constancy of the concentration of dope is normally checked by measuring the resistivity ϱ. An example of the result of such a measurement is shown in *fig.1*, in which the resistivity ϱ is plotted against the distance x measured from one end of the crystal.

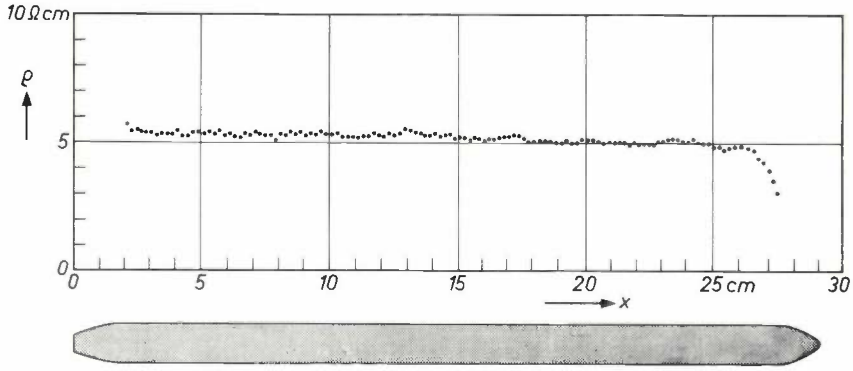


Fig.1. It is now possible to make doped rod-shaped crystals of germanium and silicon whose resistivity ϱ shows no systematic variation and only slight fluctuations in an axial direction. The graph shows measurements on a germanium crystal doped with indium.

It has been found that rods which give a very nice ϱ,x diagram may still exhibit considerable inhomogeneities, so much so in fact that they are sometimes unsuitable for their intended purpose. As an example of this, *fig.2* shows the variation of ϱ along a diameter of a transverse section of a germanium single crystal doped with antimony.

Since these inhomogeneities can give the manufacturer of transistors a great deal of trouble, a number of years ago we started an investigation into their nature and the way they are produced. In this article we shall discuss three of the most important sorts of inhomogeneities, and briefly indicate how they can be prevented (in so far as this is possible). These three sorts are: 1) striations, 2) cores and 3) inclusions [2]).

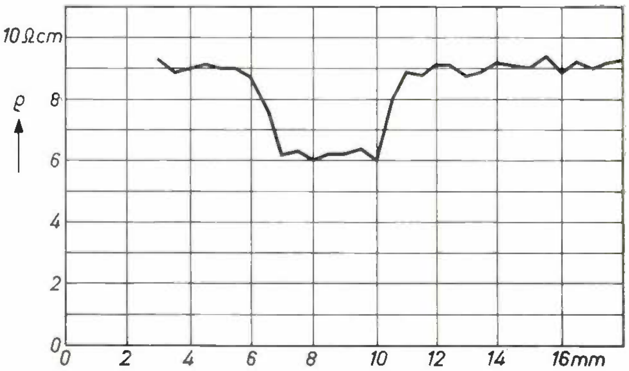


Fig.2. The variation of the resistivity ϱ along a diameter of a transverse section of an inhomogeneous germanium crystal doped with antimony.

*) Philips Research Laboratories, Eindhoven.

[1]) See e.g. J. Goorissen, Philips tech. Rev. 21, 185-195, 1959/60, or chapter 6 (by B. Okkerse) of the Handbook of semiconductor electronics, 2nd edn. (editor L. P. Hunter), McGraw-Hill, New York 1962.

[2]) Part of the results of this investigation, in particular with reference to cores, has already been published in: J.A.M. Dikhoff, Solid-state electronics 1, 202, 1960; references to previous investigations in this field are also given in this paper.

To begin with, we shall recapitulate some of the most important concepts of the theory of the solidification of binary mixtures, and give a brief description of the making of rod-shaped crystals by the "pulling" method.

The solidification of a binary mixture

In order to describe what happens at the solid-liquid interface during the solidification of a binary mixture, we make use of the phase diagram (*fig.3*).

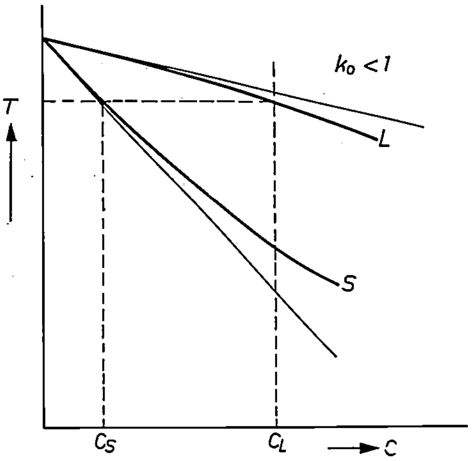


Fig.3. Part of the phase diagram of a binary mixture. At a certain temperature T the equilibrium concentrations C_S and C_L of the solute in the solid and liquid phases respectively are not equal. For low concentrations, the curves S and L which give the variation of C_S and C_L with T may be approximated to by straight lines. The ratio C_S/C_L, the distribution coefficient k_0, does not depend on the concentration in this case. When, as in the case represented in the diagram, T decreases with increasing concentration, $k_0 < 1$.

This diagram shows how the freezing point of a liquid solution (curve L) and the melting point of a solid solution (curve S) vary with concentration. The points on the two curves at a given temperature correspond to the concentrations C_S and C_L of the solute in the solid and liquid phases in equilibrium with one another at this temperature. As may be seen, C_S and C_L are not qual.

For small values of C_S and C_L, the curves S and L may be approximated to by straight lines; the ratio C_S/C_L is then equal to the constant k_0, which is known as the distribution coefficient. For the solutions we shall be dealing with in this article, the value of k_0 normally differs considerably from unity; for a solution of indium in germanium, for example, k_0 is about 0.001.

If a melt of concentration C situated in a long, narrow boat is allowed to solidify slowly from one end, the concentration in the first portion to solidify will be equal to k_0C, i.e. much less than in the liquid. A large amount of the impurity has thus been driven out of the volume in which solidification has occurred. In the extreme case that the excess impurity is immediately distributed throughout the liquid phase by vigorous stirring, the above remains true as solidification proceeds — except that C gradually increases as a result of this expulsion of impurity from the solid phase.

In practice, however, the stirring is never as good as this. There is always a layer of liquid near the solid-liquid interface in which transport of the impurity can only occur by diffusion. The concentration in this "diffusion layer" is thus greater than in the body of the liquid. As a result, the concentration C_S in the solid phase is also higher than would be the case with perfect stirring. The ratio k of C_S to the concentration C in the liquid *outside* the diffusion layer is thus greater than k_0. This ratio k is known as the segregation coefficient. Its value is determined by the thickness of the diffusion layer (i.e. by the efficiency of the stirring) and by the rate at which the solid-liquid interface moves; as this rate is greater, there will be a greater concentration of the impurity in front of the interface so that k will be greater. The dependence of k on the growth rate and the degree of stirring will keep on coming up in this article.

The pulling method

Rod-shaped germanium crystals are normally made by Czochralski's method, better known as the "pulling method" [1]. In this method, a seed crystal fixed on the end of a rotating shaft is lowered until it just touches the surface of a quantity of the molten material. The supply of heat to the melt is then reduced somewhat, and the seed crystal is slowly raised. In this way a single crystal is formed on the end of the seed. The value of the segregation coefficient k is here determined, apart from k_0, by the pulling rate f and the angular velocity ω of the shaft; the latter quantity (together with the kinematic viscosity of the liquid and the diffusion constant of the impurity in question) determines the thickness of the diffusion layer.

Striations

It was found in 1953 [3], during an investigation of the way in which the segregation coefficient k depends on the pulling conditions, that when a germanium crystal is rotated very slowly during pulling (e.g. 4 r.p.m.), periodic changes in concentration are produced along the length of the crystal. The variation of the concentration was investigated with the aid of radioactive impurities: as soon as a crystal had been pulled, it was sawn lengthways and

[3] J.A.Burton, E.D.Kolb, W.P.Slichter and J.D.Struthers, J. chem. Phys. 21, 1991, 1953.

an autoradiogram was made of the surface thus exposed. With crystals which had been rotated slowly, the autoradiogram showed transverse lines; crystals which had been rotated quickly showed no such lines.

It was found later that crystals which had been pulled with a fast rate of rotation also showed such lines ("striations"); but these lie so close to one another that they are not resolved in the autoradiogram. There is even less chance of finding them by routine measurements of the resistivity ϱ. A variation of ϱ as shown in fig.2 is due to another sort of inhomogeneity. The existence of striations very close to one another can only be demonstrated by methods with a high resolving power.

One of the most important of these is the method of pulsed copper-plating [4]). Here too the rod-shaped crystal is cut open lengthways, and the exposed surface is dipped in a solution of a copper salt. A current is then passed through the solution, the crystal acting as the negative electrode. Since the zones of differing concentration also have different conductivities, more copper is deposited on the zones which conduct better. The current is periodically interrupted, to prevent the desired effect from being disturbed by local exhaustion of copper salt in the solution; during the periods in which the current is cut off, the concentration of the copper in solution is evened out by diffusion. A crystal which has been treated in this way is shown in *fig.4*. The method of pulsed copper-plating can detect very slight concentration differences; the resolving power is such that striations can still be clearly observed when they occur at intervals of 10 μm.

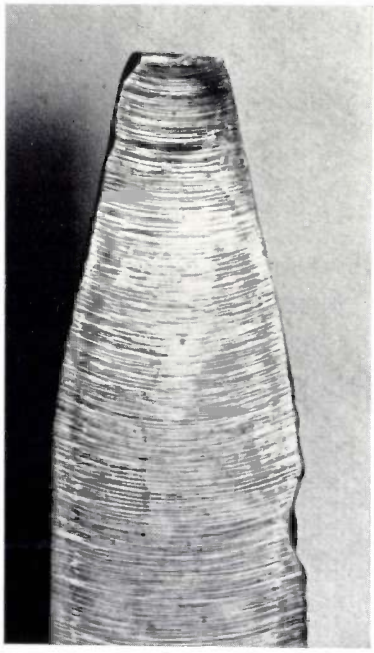


Fig.4. Longitudinal section of a crystal, with the striations made visible by pulsed copper-plating. (Growth direction from top to bottom; this is true of all the longitudinal sections reproduced in this article.)

It is also possible to make the striations visible by connecting the voltage source the other way round: more material is then dissolved from the zones with higher conductivity, so that grooves are produced in the surface of the crystal.

In heavily doped crystals, the striations can be made visible simply by etching (e.g. in a mixture of HF, HNO_3, and alcohol). The resolution is even better here, being about 1 μm (see *fig. 5*).

Two other, very elegant methods which also have a high resolution are based on certain aspects of the diffraction of X-rays. One of these makes use of the anomalous-transmission effect [5]), while the other has come to be known as Lang's method [6]). A good example of a photo made by the latter method is shown in *fig. 6*.

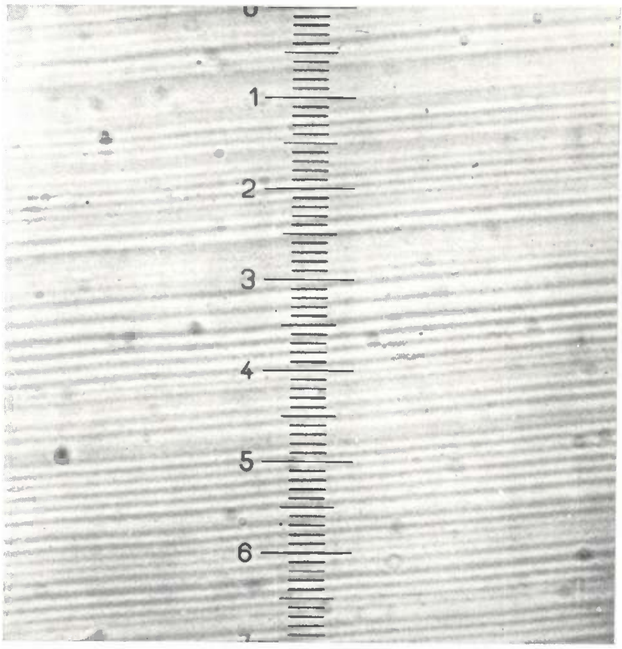


Fig.5. Detail of photo of striations in germanium heavily doped with gallium, which have been made visible by etching. The distance between successive striations is only 2 μm (magnification 600×).

The fact that the very fine striations occurring in crystals which were rotated quickly during pulling can be demonstrated by pulsed copper-plating is an indication that these striations, like those found in slowly rotated crystals, correspond to zones of different concentration. A second argument for this is that the striations can be made to disappear by heating the crystal to a high temperature, for a length of time which has been found to depend directly on

[4]) See P. R. Camp, J. appl. Phys. **25**, 459, 1954. The method has been further developed and modified by J. Bloem for *N*-type germanium (unpublished).

[5]) See e.g. B. Okkerse, Philips tech. Rev. **21**, 340-345, 1959/60.

[6]) See e.g. A. E. Jenkinson, Philips tech. Rev. **23**, 82-88, 1961/62.

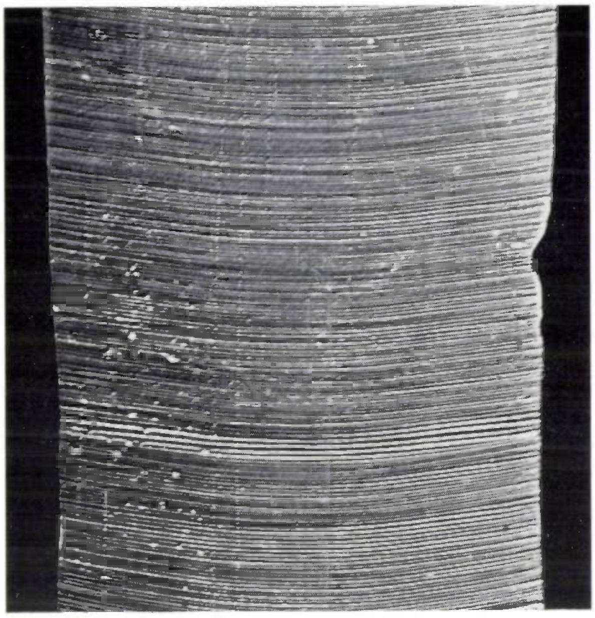


Fig.6. Striations in germanium crystal heavily doped with gallium (10^{20} atoms/cm^3), made visible by X-rays according to Lang's method [6]). (Photo: A. E. Jenkinson, Mullard Research Laboratories, Salfords, England.)

the diffusion coefficients of the impurities involved. A third argument is based on the results of the following experiment. Copper is allowed to diffuse into a crystal which has been doped with antimony in such a way that the Sb concentration in a certain direction increases linearly (or at least monotonicaly) with distance. Since Cu diffuses easily, it will soon be uniformly distributed throughout the crystal. One end of the crystal will then be *P*-type, owing to the excess of Cu there, and the other end will be *N*-type, owing to the excess of Sb. Where the excess of Cu gives way to an excess of Sb, one would thus expect to find a *P-N* junction. In fact, a region is found with a number of *P-N* junctions. This means that the concentration of Sb must exhibit a ripple (*fig.7*). The distance between *P-N* junctions is exactly equal to that between striations.

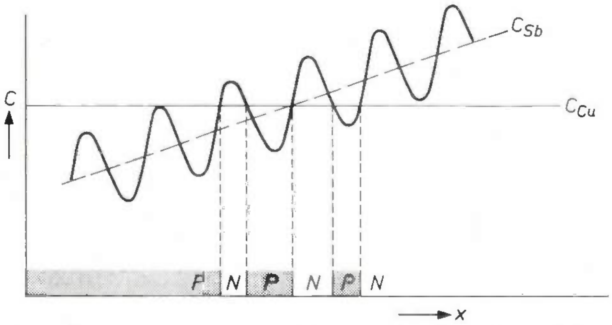


Fig.7. If a certain amount of copper is allowed to diffuse uniformly throughout a germanium crystal which has been doped with Sb in such a way that the Sb concentration increases linearly with the distance *x* (broken line), not one but a number of *P-N* junctions are found close together where the excess of copper gives way to an excess of antimony (concentrations C_{Cu} and C_{Sb} respectively). This proves that the concentration of antimony exhibits a ripple. The amplitude of this ripple can be calculated from the slope of the broken line and the width of the region in which the *P-N* junctions occur.

Form and origin of the striations

When a pulled crystal is sawn through perpendicular to its length instead of lengthways, a single spiral striation is found which fills the whole of the exposed surface (*fig.8*). This means that the striations found in a longitudinal cross-section (fig.6) must not be seen as the intersection of this surface with an equal number of curved surfaces, but as the intersection of the surface with one helical surface.

There is no doubt that the striations in a pulled crystal are usually caused by the action of pulling. The fact that the distance between striations is exactly equal to the distance the crystal grows per revolution already shows that the production of the striations has something to do with the rotation of the crystal during pulling. Both this fact and the fact that the striations are the intersection of the plane of the section with a helical surface, can be explained as results of the thermal asymmetry which afflicts nearly every pulling installation. What happens as a result of such asymmetry during pulling may easily be seen with reference to *fig.9*. Because the temperature distribution in crystal and melt is not symmetrical about the long axis of the crystal, the crystal grows e.g. more on the left than on the right. Since the crystal is rotated, each sector in turn passes the region of rapid growth. The crystal thus becomes a sort of helix of material which has solidified quickly, filled up with a helix of material which has solidified slowly. Since the segregation coefficient depends on

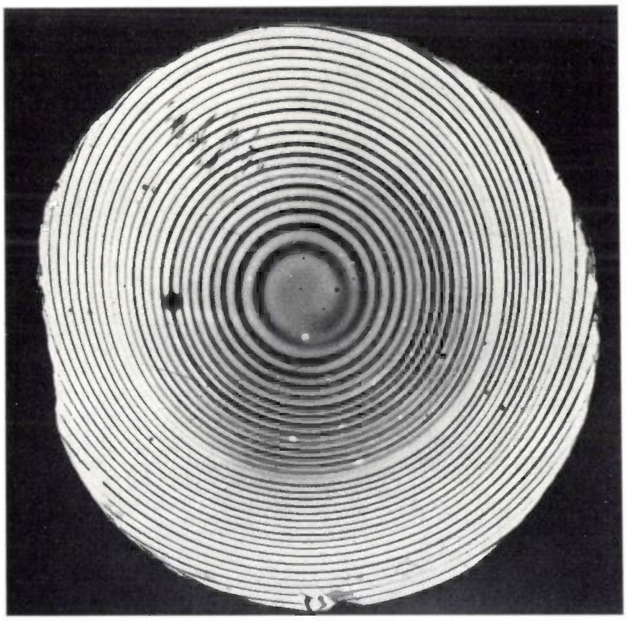


Fig.8. A transverse section of a pulled crystal shows only one striation, which has the form of a spiral.

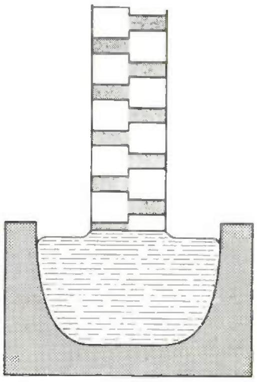


Fig.9. The helical surface which is responsible for the striations is produced because the crystal grows faster in one part of the crucible than in the rest, and is also rotated. The shaded parts of the crystal in the figure represents the parts which have grown more slowly.

the growth rate (see introduction), one of these helices contains more dope than the other; the one with the higher concentration is identical with the above-mentioned helical surface.

Since the above-mentioned thermal asymmetry of the pulling installation can hardly be avoided, the same can be said of the formation of striations. This form of inhomogeneity has no adverse effects on the semiconducting devices made from the crystal in question; when one works with a pulling installation whose thermal symmetry is as good as possible, the variations in the concentration are so slight that they play no role at all.

Although as we have seen, the striations found when a crystal is cut through lengthways are intersections of the plane of the section with a *helical* surface, they still give a good idea of the shape of the growth interface. If the crystal were cut through along a plane through the axis of the helix, one should in principle obtain a pattern like that shown in *fig.10*. Each of the lines situated at one side of the axis then gives an exact picture of the growth surface in a certain plane on the quick-growing side.

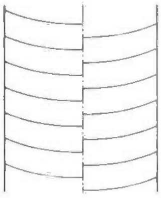


Fig.10. Sketch of the form which the striations should in principle have in a longitudinal section through the axis of the helical surface (cf. fig.9).

Successive lines show situations occurring at intervals of time equal to the period of revolution. The lines on the other side of the axis may be regarded as "exposures" each taken half a period later. The growth striations are thus very useful for studying the form of the growth interface. We shall meet examples of their use for this purpose in the following sections.

"*Fundamental*" *striations*

Striations are found not only in crystals which have been pulled in the normal way, but also in crystals which were not rotated during pulling [7]). They are also found in crystals made by horizontal zone melting [8]). Although one might imagine that these were due to a certain jerkiness in the operation of the pulling mechanism, the particular instrumental cause is not very clear. Sometimes a further fine striation pattern is found between the "normal" striations of a pulled crystal [7]). In particular these latter suggest that some striations may be due to some fundamental property of the growth process. One might for example imagine that the flow of heat and/or impurities away from the growth interface might not be continuous, but have the character of a relaxation vibration.

So far this problem has not been definitely solved. The result of the following experiments carried out by us is however a strong argument for the existence of "fundamental" striations. In these experiments during the pulling of a crystal, the high-frequency generator which warms the melt was switched off, and shortly afterwards the rotation of the crystal and the pulling were stopped. After all possible instrumentational causes of the production of striations had thus been eliminated, the crystal was rapidly withdrawn from the melt and the drop hanging on its end allowed to solidify (this takes about 15 s). It was found that these solidified drops also exhibited striations. *Fig.11* shows an example of this.

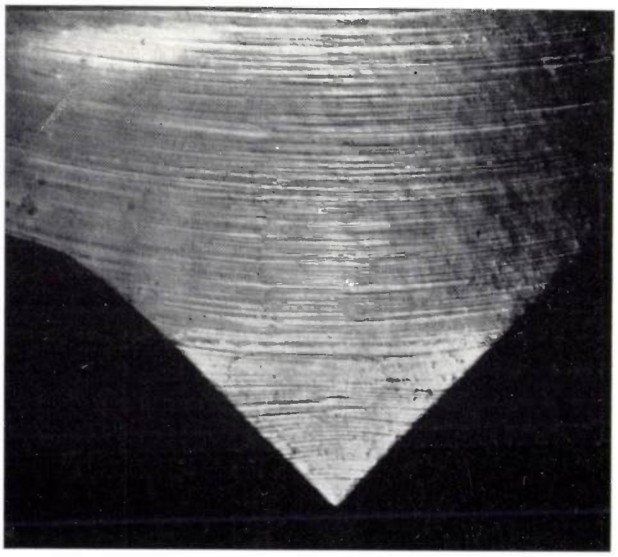
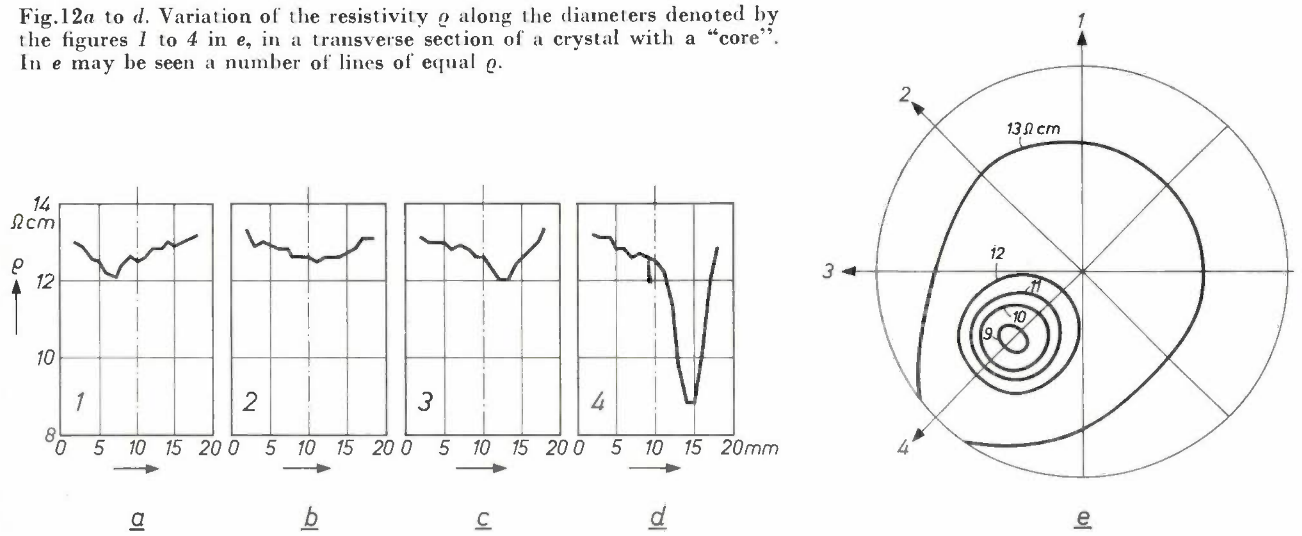


Fig.11. The drop which remains hanging on a crystal withdrawn from the melt (the point bounded by straight lines in the photo) also exhibits striations on solidification. This indicates that striations can also be produced by causes which have nothing to do with the instrumentation of the pulling method ("fundamental" striations).

[7]) H. C. Gatos et al., J. appl. Phys. **32**, 2057, 1961.
[8]) H. Ueda, J. Phys. Soc. Japan **16**, 61, 1961.

Fig.12*a* to *d*. Variation of the resistivity ϱ along the diameters denoted by the figures *1* to *4* in *e*, in a transverse section of a crystal with a "core". In *e* may be seen a number of lines of equal ϱ.



Cores

Sometimes a pulled crystal is found to contain a region whose conductivity deviates from the desired value and whose dimensions far exceed those of the striations. Such a region, or "core", is often roughly a cylinder parallel to the long axis of the crystal and extending throughout its whole length. This can be demonstrated by making transverse sections of the crystal at various points and measuring the variation of ϱ along various diameters in each section. Corresponding diameters from different sections are found in this way to give curves of similar form. For a diameter which passes through the core, the form of this curve is similar to that of fig.2. Along diameters which do not pass through the core the value of ϱ shows no deviation (*fig. 12*).

The core can very elegantly be made visible by making a longitudinal section of the crystal in a suitable plane and "developing" the striations e.g. by pulsed copper-plating. Two examples are shown in *fig. 13*.

Closer inspection of the striations shows that these are linear within the core (*fig.14*). This fact is directly connected with the way in which the core is formed. Investigation has shown that in crystals with a core, the growth interface exhibits a facet during pulling. Cores are only produced if the growth interface is convex (i.e. if the end of the crystal is convex; see e.g. the striation patterns in figs. 6 and 13) and if a ⟨111⟩ axis makes not too large an angle Θ with the long axis of the crystal; the facet is then part of a {111} face perpendicular to the ⟨111⟩ axis in ques-

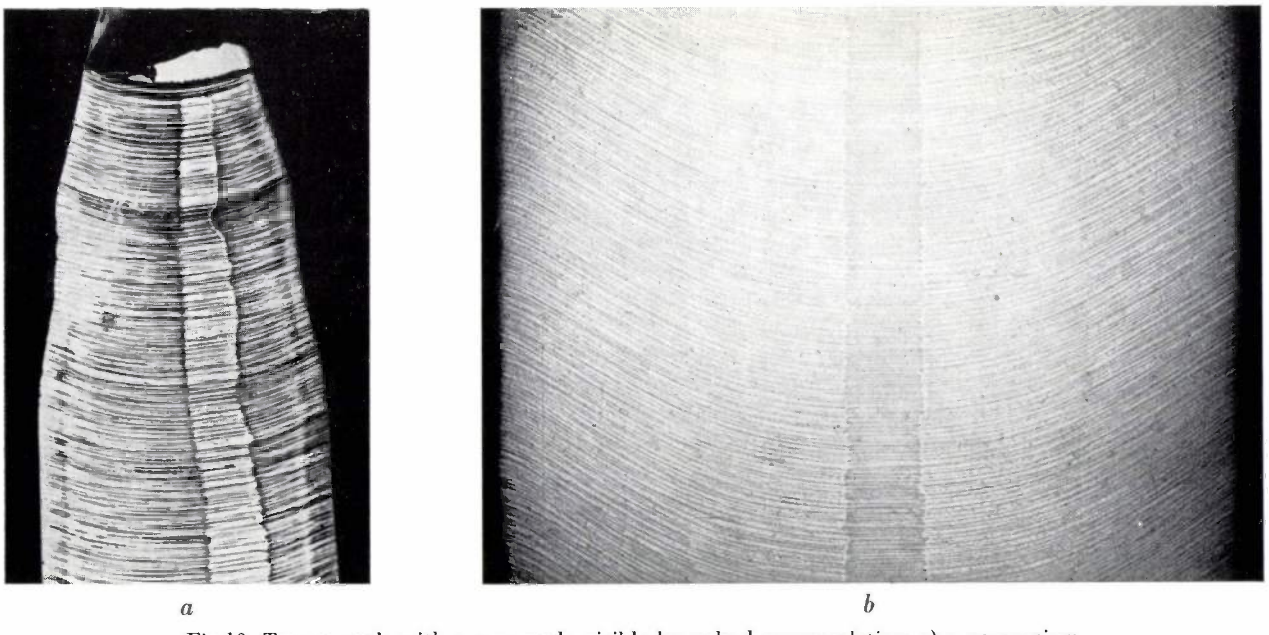


Fig.13. Two crystals with a core made visible by pulsed copper-plating: *a*) a germanium crystal and *b*) a silicon crystal. The germanium crystal was pulled, the other was made by the floating-zone technique [1]).

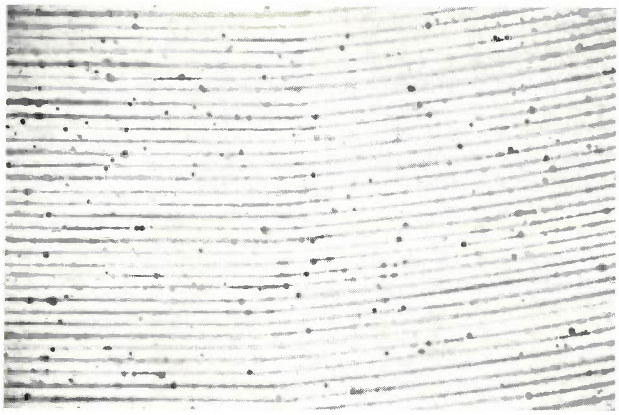


Fig.14. Detail of fig.13*a*. It may be clearly seen that the striations in the core (left-hand side of the figure) are straight lines.

tion [9]) (since germanium and silicon are both cubic). *Fig.15* shows the striations in a transverse section of a crystal with a core; here too the striations are linear within the core. Finally, *fig.16* shows the underside of a crystal which has been quickly withdrawn from the melt; the facet may be clearly seen.

The position of the facet is determined by the angle Θ; it is situated further from the middle of the crystal as Θ is larger. It appears that the solid-liquid interface becomes flat round about the spot where the curved interface in the absence of a facet would be tangent to a {111} face. The method of avoiding such cores follows directly from this fact; Θ must simply be made so large that a {111} face is nowhere tangent to the growth interface. It is not necessary to make Θ very large for this purpose; one can also try to make the temperature distribution such that the curvature of the growth interface is slight. If the temperature distribution can be made so favourable that the interface is naturally almost flat, it is also possible to make Θ very small: the whole growth interface then becomes one big "facet".

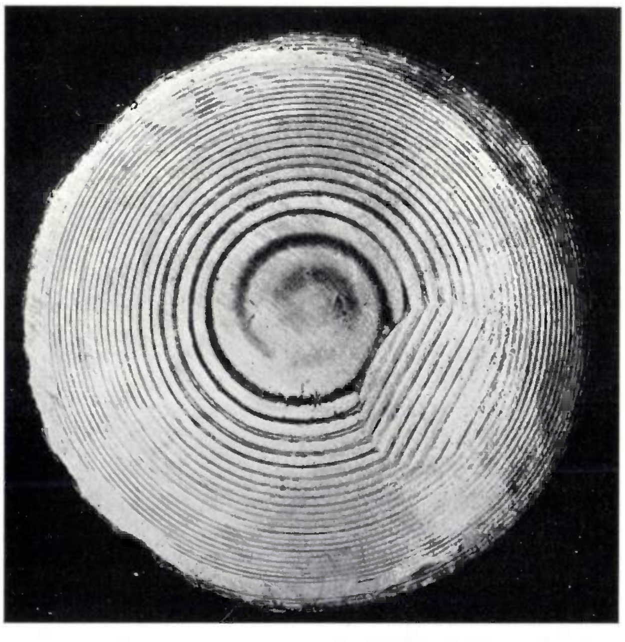


Fig.15. Just as in a longitudinal section of a crystal with a core, the striations in the core are also straight in a transverse section (cf. fig.14).

Measurements of the resistivity of germanium crystals doped with Sb have shown that the ratio of the resistivity in the core to that in the rest of the crystal is constant (about 2/3; see fig.2) over a wide range of concentrations (values of ϱ from 0.05 to 20 Ωcm). It thus appears that the value of the segregation coefficient is different for the facet. A similar phenomenon is found with other impurities; the quotient a of the segregation coefficient on the facet (k_f) and that on the rest of the growth interface (k_e) has a value which is characteristic of the impurity in question. A number of these values are shown in *Table I*.

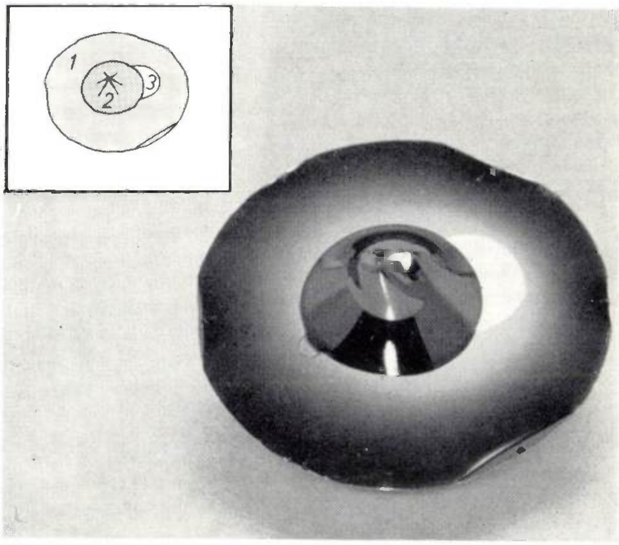


Fig.16. Underside of a crystal withdrawn from the melt whose growth interface had a flat part (facet). The inset helps to identify the various parts of the photo: *1* curved part of the growth interface, *2* solidified drop which remained hanging on the crystal when it was withdrawn from the melt, *3* the facet.

Table I. The values found for the ratio a of the segregation coefficient of several impurities on the facet (k_f) to that on the rest of the growth interface (k_e) for pulled germanium crystals. These values apply to a pulling rate of 1 mm per minute and a rotation rate of 50 r.p.m.

Dope	a $(=k_f/k_e)$
P	2.5
As	1.8
Sb	1.45
Bi	1.65
Ga	0.85
In	1.4
Tl	1.2

[9] See the article [2]) quoted above.

The same effect is also found in other materials. An example which is remarkable for two reasons is the case of indium antimonide doped with tellurium [10]). Not only is a unusually large in this case (about 8), but $k_f > 1$ while $k_e < 1$. No theory has yet been produced which can give a satisfactory quantitative explanation of the phenomena described.

The formation of a facet is a direct consequence of the fact that the growth of the crystal is not equally easy in all parts of the solid-liquid interface. If we consider a cubic crystal for the sake of simplicity, the lattice points near the solid-liquid interface will be occupied as shown in *fig.17a*. The growth of the stepped parts of this interface requires merely that atoms fill up the spaces indicated by × in fig.17*b*. This is quite easy, since the surface energy is hardly altered in the process. There are no such steps where atoms can be added on the flat part of the crystal; this part does not grow until such a step has been formed somewhere. A single atom is not enough for this purpose: a group of atoms (a "nucleus") is necessary. The formation of such a nucleus requires a certain amount of energy, whose value varies with the crystallographic orientation of the surface. In some cases this energy is quite large, so that the formation of a nucleus requires appreciable supercooling. The flat part of the solid-liquid interface cannot therefore coincide with the isothermal surface in which the rest of the interface is situated, but will lie somewhat behind this. It is thus larger than in the absence of supercooling, i.e. it forms a facet on the interface (fig.17*c*).

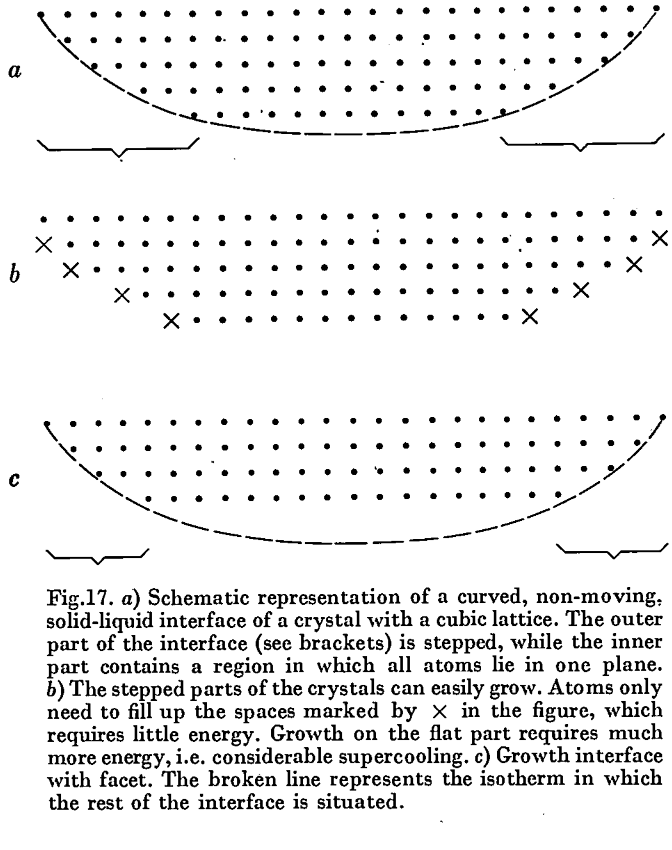


Fig.17. *a*) Schematic representation of a curved, non-moving, solid-liquid interface of a crystal with a cubic lattice. The outer part of the interface (see brackets) is stepped, while the inner part contains a region in which all atoms lie in one plane. *b*) The stepped parts of the crystals can easily grow. Atoms only need to fill up the spaces marked by × in the figure, which requires little energy. Growth on the flat part requires much more energy, i.e. considerable supercooling. *c*) Growth interface with facet. The broken line represents the isotherm in which the rest of the interface is situated.

Similar considerations make it likely that facet formation will occur in crystals of other lattice types when there are faces on which nuclei are not formed easily. In substances like Ge and Si with the diamond structure, the {111} faces are apparently such faces.

If a stable nucleus is once formed on a facet, it grows very fast laterally, because of the strong supercooling. Growth is then delayed until a new nucleus is formed, and so on. While the crystal thus grows uniformly on the curved parts of the solid-liquid interface, growth on the facet occurs in jumps; the average growth rate is naturally the same throughout.

The fact that the facet has a different segregation constant from the rest of the growth interface can be explained qualitatively by assuming that a certain adsorption (or desorption) of impurities occurs at the interface. This would have no effect on the final concentration of impurity in the solid phase when solidification was slow, as on the curved parts of the interface; but part of the adsorbed impurities could be frozen in on a part of the interface e.g. a facet where growth was fast. The concentration of impurity in the solid phase would then be greater here than in the rest of the crystal; with desorption it would be less [11]).

We may finally mention that it is not impossible that this adsorption or desorption should depend on the crystallographic orientation of the surface in question.

Inhomogeneities due to constitutional supercooling; inclusions

Heavily doped material, such as that used for the manufacture of tunnel diodes, sometimes contains regions consisting of nearly pure dope; we shall call such regions *inclusions*. It has been found that such inclusions are formed only in crystals which are pulled in such a way that "constitutional supercooling" occurs in the liquid just in front of the growth interface. We shall now briefly explain what constitutional supercooling is [12]), and how it can lead to the formation of inhomogeneities. Inclusions are

[10]) K. F. Hulme and J. B. Mullin, Phil. Mag. 4, 1286, 1959 and J. B. Mullin and K. F. Hulme, Phys. Chem. Solids 17, 1, 1960.

[11]) These ideas have been given quantitative expression by A. Trainor and B. E. Bartlett, Solid-state electronics 2, 106, 1961.

[12]) The existence of constitutional supercooling was first indicated by J. W. Rutter and B. Chalmers, Canad. J. Phys. **31**, 15, 1953. A theoretical treatment of this phenomenon with reference to the pulling process has been given by D.T.J. Hurle, Solid-state electronics 3, 37, 1961. An experimental test of this theory and a description of the form of the solid-liquid interface in Ge heavily doped with Ga may be found in W. Bardsley et al., Solid-state electronics 3, 142, 1961 and 5, 395, 1962.

the most extreme form of the inhomogeneities which can be produced in this way.

As has been mentioned in the introduction, when the distribution coefficient $k_0 < 1$, the concentration $C(x)$ of dope in the liquid just in front of the growth interface is higher than elsewhere. The variation of the concentration with the distance x from the interface is sketched in *fig.18a*; the concentration is maximum at the growth interface. If one determines with the aid of the phase diagram the equilibrium temperature T_E corresponding to $C(x)$ at each point, one obtains a curve like that shown in fig.18*b*. The equilibrium temperature $T_E(0)$ at the growth interface is lower than at some distance from this interface. If now the temperature gradient dT/dx in the liquid near the interface is less than the gradient dT_E/dx of the equilibrium temperature, then — even though T is everywhere greater than $T_E(0)$ — part of the liquid will be supercooled. This is what is known as constitutional supercooling. The minimum value of dT/dx necessary to avoid supercooling is higher the more concentrated the liquid, the higher the pulling rate, the thicker the diffusion layer (i.e. ω smaller) and the smaller the distribution coefficient k_0.

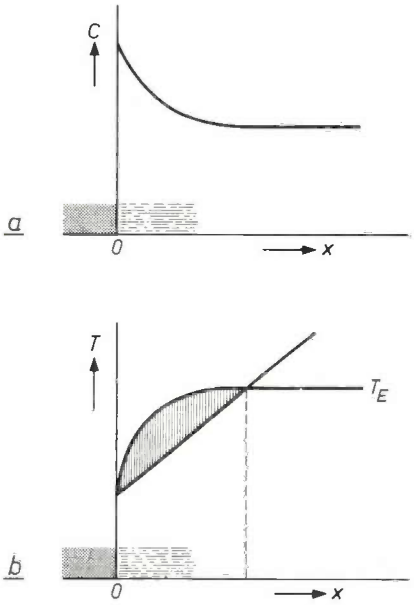


Fig.18. *a*) The variation of the concentration $C(x)$ in the melt near the growth interface as a function of the distance x from that interface. *b*) The variation of the equilibrium temperature T_E with x, which may be deduced from the above. If the temperature gradient dT/dx at $x = 0$ is less than the gradient of T_E, constitutional supercooling occurs in the layer where $T < T_E$.

A similar argument holds for $k_0 > 1$; here too constitutional supercooling is possible.

If the constitutional supercooling is not too strong, it does not lead to the spontaneous formation of crystal nuclei in the liquid. (Molten germanium can be supercooled by several tens of degrees without anything happening.) At the solid-liquid interface, however, something unusual does happen: the normal convex (or concave) shape of the interface is no longer stable in the presence of supercooling. This can be easily seen by considering what happens to a small projection which happens to form on the surface of the crystal. Since the supercooled layer has a certain thickness, such a projection grows still further instead of lagging behind until it disappears, as would occur in a stable situation. During this growth of the projection, the excess impurity diffuses out laterally as well as in the direction of growth of the crystal as a whole. The concentration of dope in the liquid in the immediate vicinity of the projection is thus increased, i.e. T_E is decreased. Since the heat of solidification is also led off laterally to a certain extent, the liquid round the projection will also be warmer than elsewhere at the same distance from the growth interface. These two effects, the decrease in T_E and the increase in T, combine to abolish the supercooling in the immediate vicinity of the projection. Other projections can thus only be formed a certain distance away.

In the long run, the growth interface acquires a "cell" structure (*fig.19*). This structure of the interface *is* stable, and can persist for some considerable time.

It is understandable that the projections from the growth interface of a Ge or Si crystal are just as likely to show {111} facets as is a normal interface. In fact, the strong curvature of these projections makes it practically certain that such a facet will occur. The shape of the projections (and thus of the growth interface as a whole)is therefore strongly

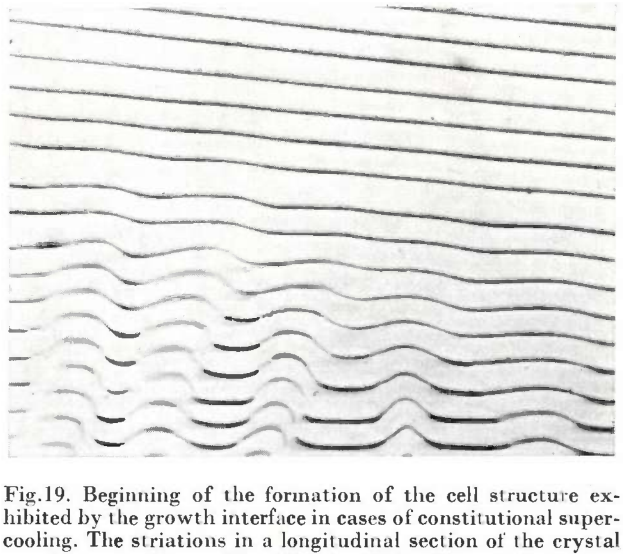


Fig.19. Beginning of the formation of the cell structure exhibited by the growth interface in cases of constitutional supercooling. The striations in a longitudinal section of the crystal are then wavy lines. (Magnification 200×.)

dependent on the orientation of the crystal in cases of constitutional supercooling.

The differences in temperature and dope concentration between the liquid at the tops of the projections and in the "valleys" will lead to considerable differences in segregation coefficient between these two spots. The portions of the crystal formed by freezing of the liquid in the valleys will have a considerably higher dope concentration than the rest of the crystal.

This process of cell formation is not restricted to crystal formation. It may be generally stated that where temperature inversion occurs, a plane surface of constant temperature will not be stable and cells will be formed. In 1901, Bénard found that if the bottom of a horizontal layer of liquid was kept warmer than the top, a cellular pattern of convection currents was produced. The liquid rises in the middle of the cells and falls at the edges [13]). A similar effect is found in horizontally enclosed layers of air, but here the flow is down in the middle and up at the edges [14]). In meteorology the formation of certain kinds of cumulus clouds has been suggested as being due to the production of Bénard cells.

As may be seen from fig.19, the course of the phenomena caused by constitutional supercooling can be followed very clearly by making the striations visible. *Fig.20* shows three more cases of cell formation in germanium strongly doped with gallium (about 10^{20} Ga atoms per cm^3). In fig.20*a* we see how cells can be formed by raising the pulling rate, and can be caused to disappear again by pulling more slowly; fig.20*b* shows a similar experiment, except that here the pulling rate was increased again before the cells were quite destroyed. It may be seen that the pattern remains nearly constant throughout the experiment, which demonstrates its stability. Fig.20*c* shows the cell formation in a crystal pulled in a ⟨100⟩ direction. Facet formation then causes the projections to be pyramidal (or tent-shaped).

Nature and origin of inclusions

The occurrence of inclusions may be regarded as an extreme form of inhomogeneous segregation due to a cellular structure of the growth interface. It is supposed that they arise when a number of projections fuse at a given moment, enclosing the very concentrated liquid situated between them. Since the temperature in the crystal decreases considerably with the distance from the (external) solid-liquid interface, the enclosed drop begins to solidify at the top. As we have seen, when solidification occurs in this way and the segregation constant k is small compared to 1, most of the impurities will be concentrated in the last part of the liquid to solidify (at the bottom of the drop); it is this part (often a eutectic mixture) which forms what we here call the inclusion.

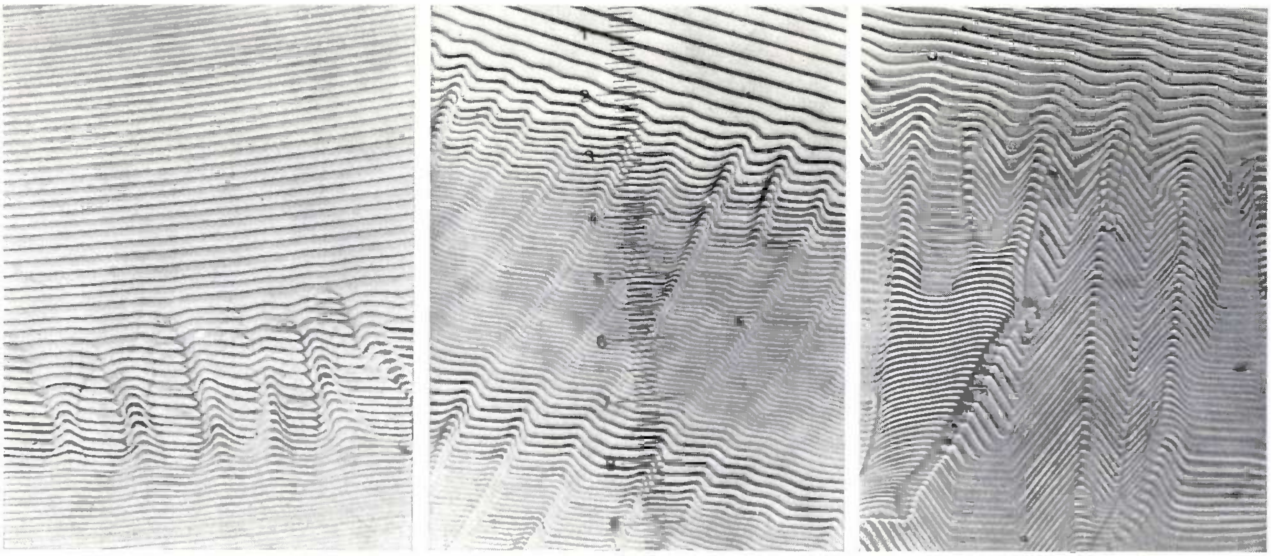


Fig.20. *a*) Example of a crystal where constitutional supercooling was produced in the melt by raising the pulling rate. The growth interface exhibits a ripple which steadily increases in magnitude. After some time the pulling rate was decreased again — visible from the smaller distance between striations; the amplitude of the ripple then decreases and it finally disappears altogether. *b*) As in (*a*), except that the pulling rate was here increased again during the experiment, before the ripples had completely disappeared. The same pattern is observed throughout the whole operation, which shows the stability of the ripple structure. Both these crystals were pulled in the ⟨111⟩ direction. The tops of the projections from the growth surface are flattened by facet formation. *c*) If a germanium crystal is pulled in a ⟨100⟩ direction under conditions of constitutional supercooling, the projections from the growth interface are given the form of a pyramid or tent by the facet formation.

[13]) See e.g. H. J. V. Tyrrell, Diffusion and heat flow in liquids, Butterworths, London 1961, chapter 11.

[14]) Further details may be found in J. G. A. de Graaf, thesis Utrecht, 1952, who studied the air flow in enclosed layers in connection with the thermal insulation of buildings.

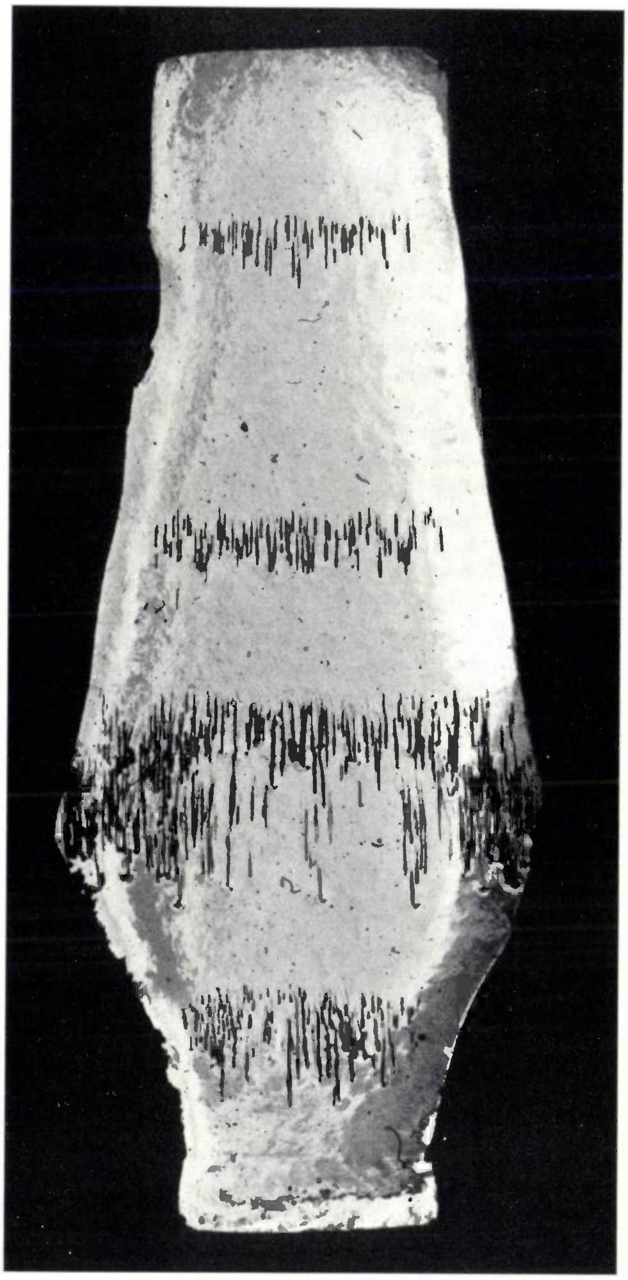


Fig.21. Inclusions in a Ge crystal pulled in a ⟨110⟩ direction. The dark vertical lines are the tracks made by enclosed drops. During the pulling of this crystal, the pulling rate was several times increased, so as to lead to constitutional supercooling and hence to the production of inclusions, and then decreased again after a short time. This is the reason why four groups of lines can be seen, each group beginning roughly in a horizontal plane.

We have implicitly assumed above, to simplify the argument, that an enclosed drop solidifies on the spot. In fact this is not so: it continues to move for some time in the direction of growth. This may be compared with the progress of the molten zone in "temperature-gradient zone melting": material is dissolved on the warm side of the zone, and is deposited on the cold side [15]). Finally, however, the drop comes to a stop and forms an inclusion as described above (*figs. 21* to *25*). This progress of the enclosed drop is the reason why the fusing together of the projections cannot be seen in the striation pattern.

The fact that the region through which the enclosed drop has travelled also shows striations has led to the suggestion that these are "fundamental" striations. Calculation has however shown that the small temperature variations at the solid-liquid interface which, according to the argument on p. 198, are the reason for the formation of the normal striations, penetrate some distance into the crystal — to be precise, the amplitude decreases exponentially, with a characteristic length of a few mm. They are therefore very well able to influence the solidification of the enclosed drops.

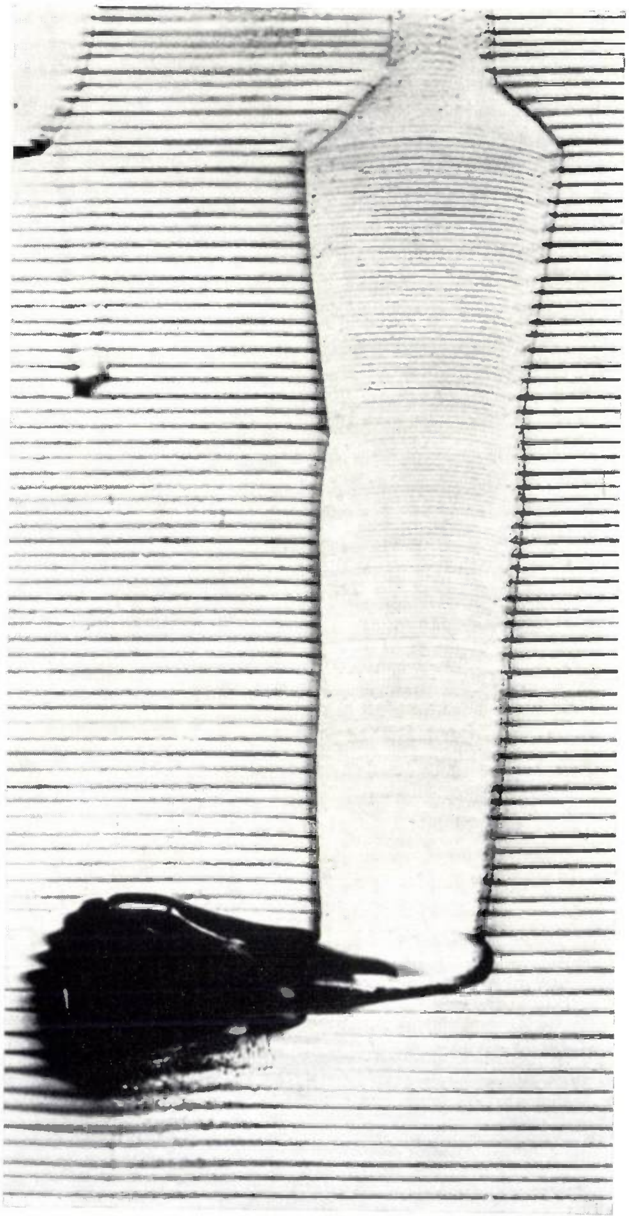


Fig.22. Highly magnified (200×) striation pattern of part of a crystal which has been traversed by an enclosed drop. The dark patch represents the inclusion formed by the drop at the end of its path.

[15]) See W. G. Pfann, Zone melting, Wiley, New York 1958, chapter 9.

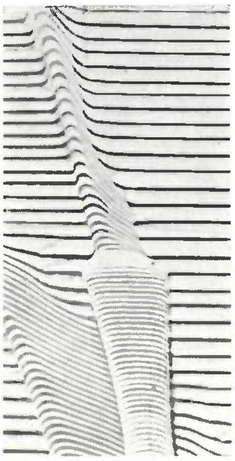


Fig. 23

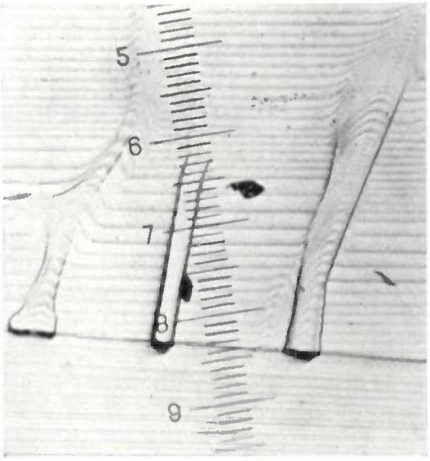


Fig. 24

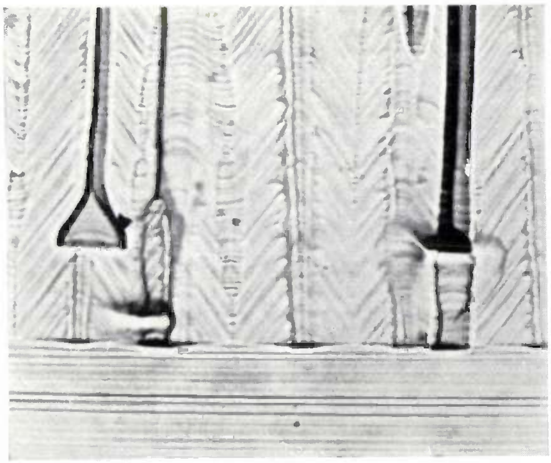


Fig. 25

Fig.23. Striation pattern of part of a crystal in which a valley of a cellular growth interface gives way to a zone traversed by an enclosed drop. In the valley (top left) the distance between striations is equal to that in the projections, while in the zone traversed by the drop the striations are closer together. (Magnification 110×.)

Fig.24. Striation pattern (magnification 120×) of part of a crystal grown with a cellular solid-liquid interface, in which the tracks of three enclosed drops can be seen. At a certain moment, the pull rate was decreased and the melt was given such a big heat pulse that the crystal melted a little at the end before growing further. The striations below this point are again completely regular.

Fig.25. As in fig.24. The striation pattern above the melt-back boundary is highly irregular. The fact that the pattern below this boundary is again regular indicates that despite the constitutional supercooling the material remains monocrystalline. (Pulled in ⟨100⟩ direction, magnification 80×.)

The striation patterns of fig.19 *et seq.* clearly show that constitutional supercooling, while giving rise to inhomogeneities, need not lead to the production of polycrystalline material. Even the crystal of fig.25 was a single crystal, as may be seen after being partly melted back. It is true however that regions where the growth interface had a cellular structure, and in particular regions in and around the zones traversed by an enclosed drop, have a great dislocation density. This can be demonstrated both by the etch-pit technique and with the aid of the anomalous-transmission effect for X-rays. The stresses caused by the differences in lattice constant corresponding to the concentration differences are apparently partly compensated by dislocations.

A remarkable property of these dislocations is that they do not normally extend beyond the region containing the inhomogeneities. If one carefully pulls a dislocation-free crystal [16]), then alters the conditions so that constitutional supercooling occurs, and finally returns to the original conditions, dislocations will be found mainly in the middle of the crystal. This shows that most of these dislocations form closed rings.

The occurrence of inhomogeneities produced by constitutional supercooling can also be prevented, even in very heavily doped crystals. The temperature gradient must be made very steep, and for the rest it is only necessary to choose suitable values for the pulling and rotation rates. For example, by pulling slowly we have succeeded in making homogeneous dislocation-free germanium crystals with a gallium content of about 4×10^{20} atoms/cm^3, i.e. nearly 1 wt.%. This content approaches the limit set by the solubility of the gallium.

Summary. Pulled doped germanium and silicon crystals which show no serious axial variation of the resistivity on routine examination can nevertheless be locally inhomogeneous. They can exhibit striations, cores and inhomogeneities due to constitutional supercooling, of which latter inclusions are an extreme form. Striations are produced by thermal asymmetry of the pulling installation, and are difficult to remove completely. In a longitudinal section they give a very good idea of the shape of the growth interface at various moments. Cores are produced when the growth interface contains a {111} facet; the segregation coefficient on this facet differs from that on the rest of the interface. When doping is heavy, constitutional supercooling easily occurs in the melt in front of the growth interface. The growth interface then acquires a cell structure owing to the growth of projections. The "valleys" in between these projections are filled with relatively concentrated liquid, which solidifies to give more strongly doped material. Enclosure of this liquid by fusion of projections leads to the formation of inclusions. Cores can be prevented by a suitable crystallographic orientation of the seed crystal, and inclusions by suitable pulling conditions. It has been found possible in this way to pull Ge crystals doped with Ga in amounts barely less than the solubility, which were still free from inhomogeneities and dislocations.

[16]) For the production of dislocation-free crystals and for various methods of investigating dislocations see the article quoted in [5]).

DETERMINING MAGNETIC QUANTITIES BY DISPLACEMENT MEASUREMENTS

by U. ENZ *) and H. ZIJLSTRA **).

621.317.44

For research on ferromagnetic materials, which has been in progress in the Philips laboratories for a considerable time now, the authors have built three instruments by which the magnetostriction, the magnetic anisotropy and the magnetization of these materials can be measured with a precision of approx. 1%. Up-to-date instrumentation can give a higher degree of precision, but wherever the above precision, which suffices for nearly all practical and many theoretical purposes, is acceptable, these measurements can be carried out with very simple equipment. In all three instruments the relevant magnetic property produces a displacement of a ferroxcube core or a coil, and this is measured with an established type of measuring bridge (PT 1200). These instruments, designed for use in our laboratories, have been used for many measurements and found to be satisfactory.

The *magnetization* of a sample of ferromagnetic material in a magnetic field is generally accompanied by *magnetostriction*, i.e. a change in its dimensions. If the sample is a single crystal, it will also tend to assume such a position that a preferred direction becomes parallel with the field, thus minimizing the *magnetic anisotropy energy*. The above three magnetic quantities together give a fairly complete picture of the magnetic properties of the sample. For measuring these quantities we have designed and built three instruments that can be conveniently operated and yet attain a precision of 1%. The measuring method of each instrument is based on the same principle: the relevant magnetic property (or rather its variation) is converted to a *displacement* of a ferroxcube core or to that of a coil. Because of this displacement there is a change in the coupling between two stationary coils forming a bridge circuit together with two other impedances. These two impedances are part of a commercially available device (measuring bridge PT 1200), which both feeds the bridge with an alternating current of 4 kc/s, and amplifies, rectifies and measures its output voltage. The magnitude of the displacement can either be read off an incorporated measuring instrument or be registered by means of an x-y-recorder as a function of e.g. the magnetic field strength or time. A further common feature of the three instruments is that the variation of the magnetic quantities as a function of temperature can be very easily determined. In all three cases the instruments are calibrated with the aid of samples whose magnetic properties are known.

Magnetostriction

When a non-magnetized rod of some ferromagnetic material is placed in a magnetic field that is parallel to the axis of the rod, then the length l of the rod will change by an amount Δl. The other dimensions will primarily change in such a way that the. volume of the rod remains constant. The magnetostriction λ is defined as $\Delta l/l$. In single crystals, λ assumes different magnitudes along different crystallographic directions. Most measurements, however, are made on polycrystalline material without texture, so that the same "average" value is found for all directions. The value of λ generally increases with increasing magnetic field strength up to the point of magnetic saturation. For most ferromagnetic substances its magnitude, either positive or negative, is then of the order of 10^{-5} and remains constant beyond that point.

In practice magnetostriction manifests itself in the humming noise produced by transformers and smoothing coils [1]): the dimensions of the ferromagnetic core alternate at twice the frequency of the electric current. One useful application of magnetostriction is the generation of ultrasonic vibrations [2]). Magnetostriction is also of theoretical significance for studying the atomic processes that cause magnetism.

Fig.1a shows a cross-section of the instrument for measuring magnetostriction. Test bar P, with a length l of about 5 cm, rests in the hollow cone in the bottom of quartz tube A. The variation in length Δl produced when the magnetic field is switched on is transferred by quartz rod Q to pin T. This pin is part of a commercially available type of dynamometer (PR9310). The dynamometer contains two coils L and L' and ferroxcube core F. The ferroxcube core is mounted on pin T, the latter being held by leaf-springs B and B' in a position where it can only move lengthways. If core F is displaced over a

*) Philips Research Laboratories, Eindhoven.
**) Philips Radio, Television and Record-Playing Apparatus Division, Metallurgical Laboratory, Eindhoven.

[1]) See e.g. E.W. van Heuven, The noise emission of ballasts for fluorescent lamps, Philips tech. Rev. 18, 110-119, 1956/57.

[2]) See e.g. C.M. van der Burgt, Ferroxcube material for piezomagnetic vibrators, Philips tech. Rev. 18, 285-298, 1956/57.

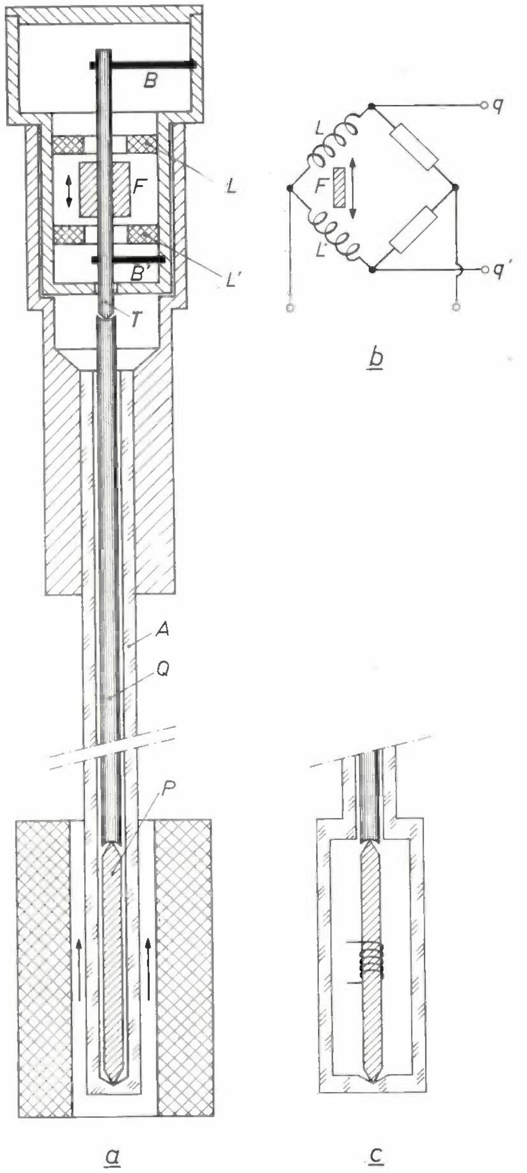


Fig.1. *a*) Schematic cross-section of the instrument for measuring magnetostriction. Test bar *P* rests in the hollow cone in the bottom of quartz tube *A*. The change in length Δl of test bar *P* caused by the magnetic field set up by a solenoid (see arrows) is transferred by quartz rod *Q* to pin *T*, the latter being part of a commercially available dynamometer (PR 9310). The pin, supported by leaf-springs *B* and *B'*, is fitted with a ferroxcube core *F* whose displacement over distance Δl changes the inductances of the two coils *L* and *L'*. The potential difference thus set up between the points *q* and *q'* (*b*) is measured by a measuring bridge (PT 1200), which is likewise commercially available. *c*) Modified version of quartz tube *A*, as explained under fig. 4.

distance Δl, the self-inductances of *L* and *L'* change, which unbalances a bridge circuit, producing a potential difference between points *q* and *q'* (fig. 1*b*). This potential difference is amplified and rectified in the measuring bridge, producing a deflection on the measuring instrument (*fig.2*). This deflection is proportional to Δl and hence to the magnetostriction λ. The smallest deflection that can be read off the measuring bridge is about 0.01 µm. For a value of λ of 2×10^{-5} and a test-bar length of 5 cm, this corresponds to a measuring precision of 1%.

The length of the quartz rod and tube must be such that the test bar and measuring probe are about 20 cm apart, allowing the rod to be placed within a large solenoid. By means of this solenoid (fig.2), field strengths up to 1.6×10^{6} A/m (20 000 Oe) can be produced, which makes it possible to measure into the saturation region of most ferromagnetic substances.

The pressure of the leaf-springs (a few grammes) is adequate to ensure a backlash-free contact between pin *T* and rod *Q*. Quartz rod *Q* and test bar *P* are cemented together to form a rigid unit. Since the leaf-springs do not allow lateral movement of the pin and the lower end of the test-bar rests in a hollow cone, *Q* and *P* are kept free of the wall, so that the whole assembly moves without friction.

The saturation magnetostriction is of the same order of magnitude as the thermal expansion at a temperature difference of 1°C. Therefore in order to retain a precision of 1%, the temperature during the measurement must remain constant within 0.01 °C. Since the measurement, which is simply a comparison of the meter readings before and after the field is switched on, takes very little time, the desired temperature stability can be easily realized.

The variation of the magnetostriction as a function of temperature can be measured from —196°C to 700 °C by surrounding quartz tube *A* with a bath of cold liquid or with a small furnace. The temperature of the test bar is then measured with the aid of a thermocouple. Because of the small dimensions of the bar and the short duration of each measurement, only a small amount of coolant or

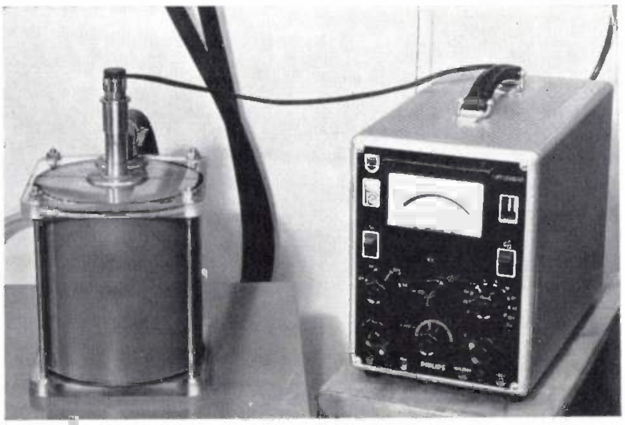


Fig.2. Set-up for measuring the magnetostriction. On the left is the solenoid for generating a field of known strength, into which the quartz tube holding the test bar has been inserted. The measuring bridge PT 1200 is shown on the right. The strongest field that can be set up by the solenoid (1.6×10^{6}A/m) is sufficiently strong to saturate most ferromagnetic substances.

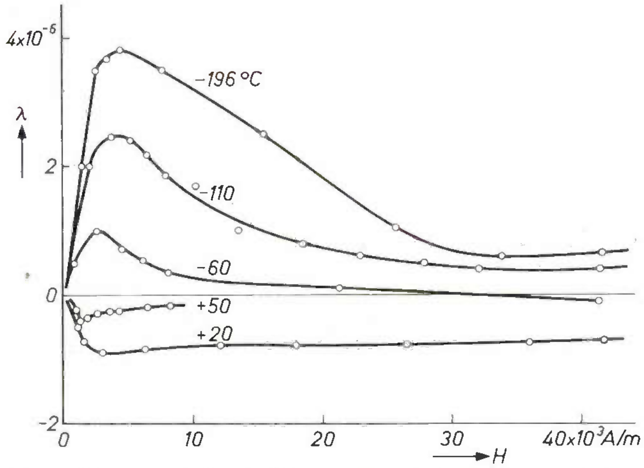


Fig.3. Magnetostriction λ of a manganese-ferro-ferrite as a function of the field strength H, at different temperatures. This ferrite is frequently used because of its high magnetic susceptibility in e.g. telephony in Pupin coils. From these and other measurements, this high susceptibility could be explained. (Cf. U. Enz, Proc. Instn. Electr. Engrs. **109 B**, suppl. No. 21, 246, 1962.)

a low furnace power is required. *Fig. 3* is an example of a few curves obtained with the instruments.

Fig. 4 finally shows three versions of the instrument. The one just described is shown on the left. The version shown in the centre allows free access to the test bar (see also fig.1*c*). A few turns of wire whose terminals are connected to a ballistic galvanometer can be wrapped round it. It is thus possible, whilst measuring the magnetostriction, to determine the change in flux and from this the magnetization of the test bar. With the right-hand version the magnetostriction of samples of annular form can be measured.

Magnetic anisotropy

If a monocrystalline sample of some ferromagnetic material is to be magnetized to saturation, we find that this does not require the same energy in every direction. There are some directions in which this energy is a minimum; these are the preferred directions of magnetization. In most cases they appear to be parallel to certain simple crystallographic directions, e.g. with ferroxdure ($BaFe_{12}O_{19}$) the preferred direction is parallel to the hexagonal axis. A crystal is usually divided into a large number of regions that have already been magnetized to saturation (the well-known Weiss domains), each being magnetized along one of the preferred directions.

The sample is placed in a magnetic field strong enough to bring it to saturation, whatever its crystal orientation with respect to the field. In order to rotate the sample about an axis perpendicular to the field, we must apply to it a couple with a moment M which supplies the energy ΔE required to turn the magnetization through an angle Θ with respect to the relevant preferred direction. This energy ΔE, called the *magnetic anisotropy energy*, is a periodic function of the angle Θ between the field and the relevant preferred direction. This function can be written in the form of a power series and the magnetic anisotropy of a specific material characterized with the aid of the coefficients of this series. With ferroxdure all coefficients but one are zero, which leads to the very simple formula:

$$\Delta E = K_1 \sin^2 \Theta .$$

The moment M is found by differentiating ΔE with respect to Θ. For ferroxdure this yields:

$$M = \frac{\mathrm{d}}{\mathrm{d}\Theta} \Delta E = K_1 \sin 2\Theta . \quad . \quad . \quad (1)$$

This shows that K_1, and in other cases also the remaining coefficients K_2, K_3, etc., can be determined by measuring M as a function of Θ. This is the prin-

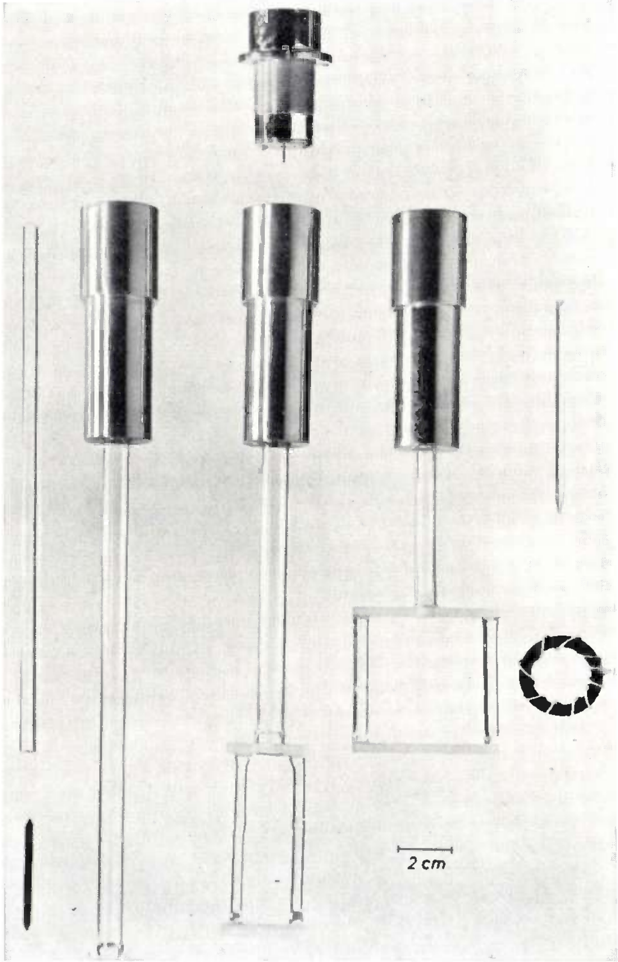


Fig.4. Various versions of the instrument for measuring magnetostriction. On the left is the version of fig.1*a*. With the one in the centre it is possible to measure at the same time the variation of inductance in the test bar and thus to determine the magnetization (see fig.1*c*). For this purpose a wire winding on the test bar is connected to a ballistic galvanometer. The magnetostriction of annular samples can be measured with the type shown on the right. In this case the magnetic field is produced by a current passed through the wire wound on the ring.

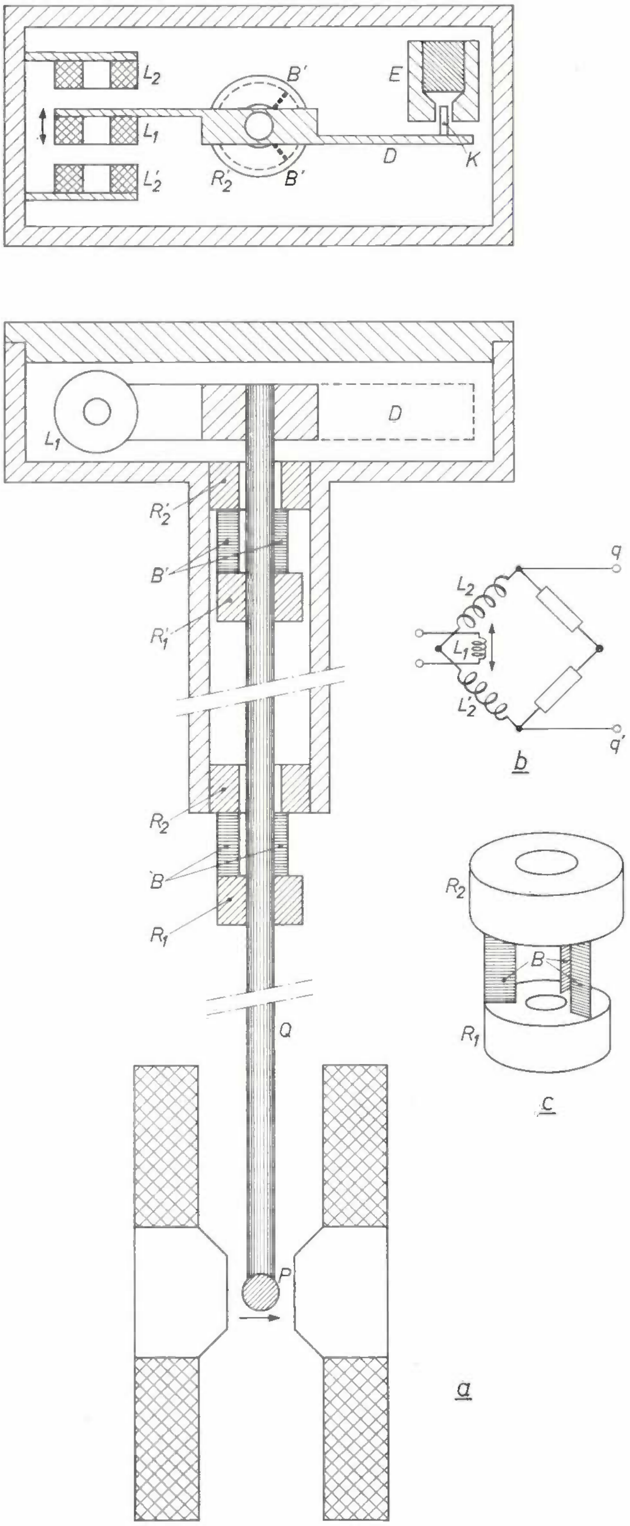


Fig.5. *a*) Schematic cross-section of the instrument for measuring the magnetic anisotropy. Quartz bar Q is held by means of rings R_1 and R_1' and leaf-springs B to the rings R_2 and R_2' of the housing in such a way that it can only turn about its longitudinal axis (see also *c*). If a magnetic field (see arrow) exerts a couple upon the spherical single-crystal sample P, then coil L_1 mounted on cross-bar D is moved out of its equilibrium position halfway between coils L_2 and L_2'. The resulting potential difference between points q and q' (*b*) can again be read from the measuring bridge PT 1200. Fitted to the other arm of the cross bar is a piece of copper K, situated in the field of permanent magnet E, for damping possible vibrations.

ciple on which the instrument shown in *fig. 5a* operates.

Sample P is a spherical single crystal, a few millimetres in diameter, cemented to the lower end of a quartz rod Q [3]). The bar is clamped into two rings R_1 and R_1', fitted by means of radially arranged leaf-springs B to immovable rings R_2 and R_2' (see sketch in fig. 5*c*). This allows the rod to turn through a moderate angle about its axis whilst motion in other directions is prevented. The top of the rod is fitted with a cross-bar D. Its left arm carries coil L_1, through which an alternating current of 4kc/s is passed, and which, together with two coils L_2 and L_2', forms a differential transformer. Coils L_2 and L_2' have the same specification and are circuited in such a way that L_1, when its position is halfway between the two coils, induces in them equal voltages of opposite sign, so that the potential between points q and q' (fig.5*b*) is zero. Any displacement of L_1 causes a potential difference between q and q', which is again amplified, rectified and measured by measuring bridge PT 1200. The measuring bridge also supplies the alternating current feeding the differen-

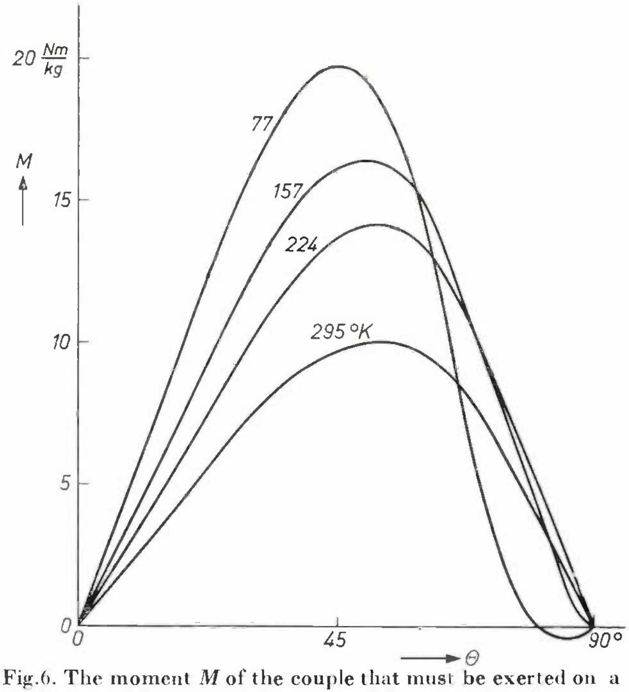


Fig.6. The moment M of the couple that must be exerted on a single-crystal ferromagnetic sample of 1 g to maintain a preferred direction at an angle Θ with the field, recorded as a function of angle Θ, at various temperatures. Values of K_1, K_2, etc. can be derived from this diagram. The measurements were carried out on the compound $Ba_3CoZnFe_{24}O_{41}$ (a ferroxplana) very suitable for HF applications. This compound owes its remarkable properties to the fact that it has a preferred plane of magnetization. (Graph reproduced from: F. K. Lotgering, U. Enz and J. Smit, Philips Res. Repts. **16**, 441, 1961. See further A. A. Aldenkamp, C. P. Marks and H. Zijlstra, Rev. sci. Instr. **31**, 544, 1960.)

[3]) A spherical shape was preferred as this eliminates the influence of the form anisotropy.

tial transformer. The right arm of cross-bar D is fitted with a small piece of copper K situated in the field of a permanent magnet E. Possible vibrations of the system are damped by the eddy currents induced by the magnet in the copper.

If the sample is situated in the field in such a position that one of its preferred directions is at an angle Θ to the field, the bar will rotate about its axis through a small angle φ where the couple exerted by the leaf-springs on the bar balances the couple exerted by the field on the sample. The displacement of coil L_1 corresponding to φ results in a deflection of the measuring instrument proportional to moment M of the relevant couples. Knowing M, the value of K_1 can be calculated from (1).

The variation of the anisotropy with temperature can be measured just as simply as with the first instrument. As an example, some measuring results have been plotted in *fig.6*. We see that the curves are only approximate sinusoidal functions of (1), which indicates that one or more of the coefficients K_2, K_3, etc. are not zero. In order to determine the magnitude of these coefficients as well, the measurement must be carried out over the whole region $0 < \Theta < 180°$. It is therefore of great advantage that this set-up is very suitable for automatic recording. To this end the measuring instrument rotates about its axis with respect to the field at the rate of about 1 r.p.m., by means of a simple auxiliary device. This rotation is transmitted via gears to the arm of a potentiometer connected to the measuring instrument. The voltage across the potentiometer, being proportional to the angle Θ, is applied to an x-y-recorder together with the output voltage of the measuring bridge, which is proportional to the moment M. In this way the recorder automatically plots the variation of M with Θ. The complete set-up is shown in *fig. 7*.

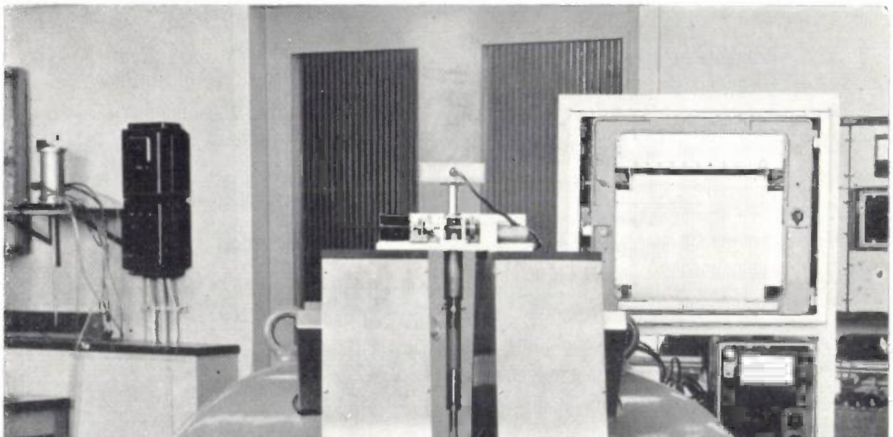


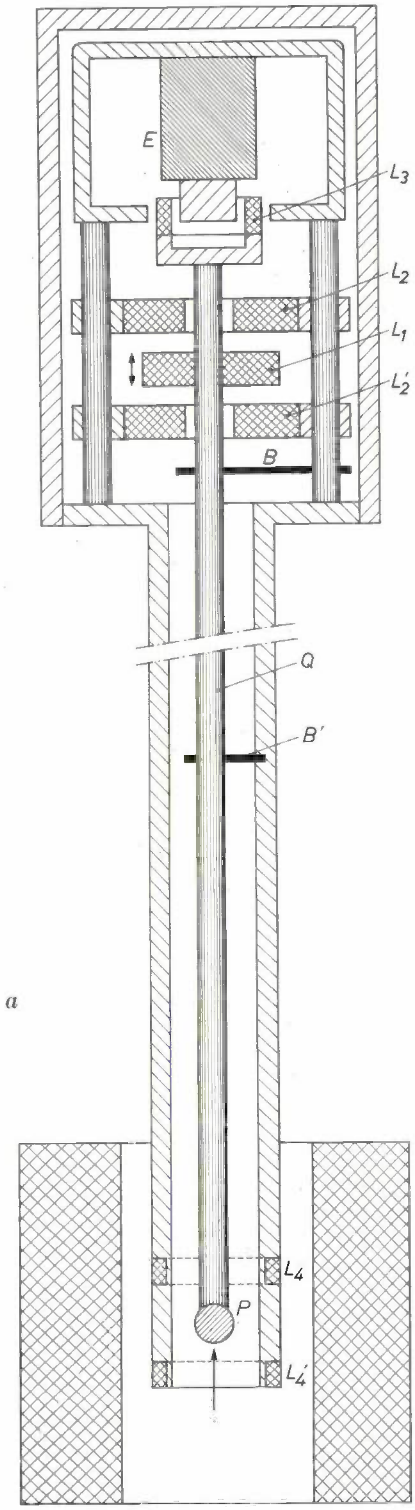
Fig.7. Complete set-up for measuring the magnetic anisotropy. The specimen is placed between the pole shoes of the large electromagnet. On top of the magnet rests the auxiliary device by which the measuring instrument is rotated at approx. 1 revolution per minute about its longitudinal axis. Behind the magnet can be seen the x-y-recorder by which moment M is recorded as a function of angle Θ. Below the x-y-recorder is the measuring bridge PT 1200.

Magnetization

A magnetic field non-homogeneous in a given direction, e.g. in the z-direction, will exert on a ferromagnetic sample a force F parallel to that direction, proportional to the mass m of the sample as well as to the gradient $\mathrm{d}H/\mathrm{d}z$ in this direction. We can therefore put:

$$F = m\sigma \frac{\mathrm{d}H}{\mathrm{d}z}. \qquad \ldots\ldots \quad (2)$$

The proportionality constant σ in this expression



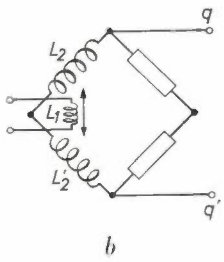


Fig. 8

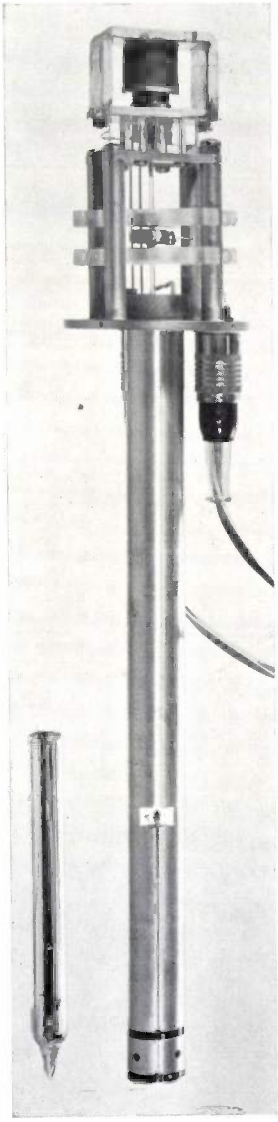


Fig. 9

Fig. 8. *a*) Schematic cross-section of the instrument for measuring the magnetization. Bar Q is held by the leaf-springs B and B' and is fitted at its lower end with the spherical sample P and at the top with coil L_3, which forms part of a loudspeaker system, with E as permanent magnet. Coils L_4 and L_4' produce a field with constant gradient $\mathrm{d}H/\mathrm{d}z$, which exerts a force upon the sample magnetized by the field of the solenoid. The resulting displacement, however, is reduced to zero by varying the current through L_3, the differential transformer (L_1, L_2 and L_2') incorporated in a bridge circuit (*b*), in combination with measuring bridge PT 1200, serving as zero indicator. The current through L_3 is a measure of magnetization.

Fig. 9. The instrument for measuring the magnetization with its cover removed. To the left of the instrument is a small Dewar flask which can be fitted into it, so that for measuring the magnetization at low temperatures only the sample and a small section of the quartz rod have to be cooled down.

is the *magnetization* per unit mass. Instead of $m\sigma$ we can also put in (2) the product of the volume V of the sample and the magnetization I per unit volume. The magnitude of I derived from this will not be very accurate as a rule, since most ferromagnetic materials are porous, so that their true volume cannot be precisely defined.

σ can be determined by measuring F when m and $\mathrm{d}H/\mathrm{d}z$ are known, and this is done with the instrument shown in *fig. 8a* and *fig. 9*. A long, thin quartz rod Q is supported by leaf-springs B and B' in the same way as indicated for the displacement meter in fig. 1, so that only lengthways movements are permitted. Cemented to the lower end of the bar is the sample P (usually a sphere a few mm in diameter). Fitted to the upper end of the bar is coil L_3 of a loudspeaker system, fitting very closely in the circular air-gap of a permanent magnet E [4]). Also fitted to the bar is a second coil L_1 which, together with two stationary coils L_2 and L_2', forms a differential transformer. Two grooves in the casing at the height of the sample hold two coils L_4 and L_4' by which a magnetic field of a known and constant gradient $\mathrm{d}H/\mathrm{d}z$ can be generated over a length of about 1 cm (see *fig. 10*).

In order to magnetize the sample, the narrow part of the instrument is placed inside a solenoid, the fields that can thus be generated being strong enough to saturate most materials. When the current through coils L_4 and L_4' is switched on, a field having a constant gradient is superimposed upon the field of the solenoid, and the sample is subjected to a small force of the same order of magnitude as its weight, causing a small displacement of coil L_1. This displacement is found just as in the anisotropy instrument,

[4]) The device built by Foëx and Forrer (J. Phys. Radium 7, 180, 1926) likewise employs a loudspeaker system in this place.

but the measuring procedure is somewhat different here. In this case we determine the current through coil L_3 necessary to compensate the force exerted by the field of magnet E upon coil L_3, using the measuring bridge as a null indicator. This method was preferred because the sample always remains in the same position relative to L_4 and L_4', thus avoiding any errors due to the gradient $\mathrm{d}H/\mathrm{d}z$ not being exactly uniform.

It is obviously desirable for the magnetization σ in the sample to be as homogeneous as possible, and since σ depends upon the field strength, this implies that the variation ΔH of H across the diameter of the sample should be small relative to H. This is fulfilled since with variations in H from 80×10^3 to 1600×10^3 A/m, ΔH amounts to approx. 3×10^3 A/m.

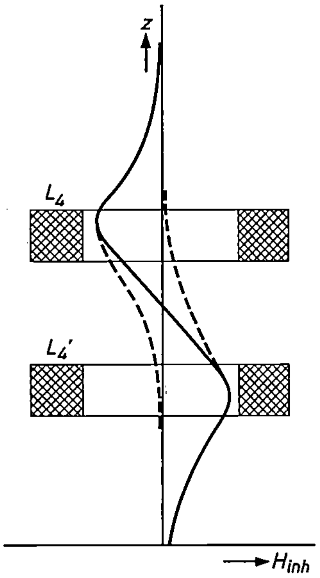


Fig. 10. A non-homogeneous magnetic field H_{inh} with a constant gradient $\mathrm{d}H/\mathrm{d}z$ is produced by means of two identical coils L_4 and L_4' through which currents are passed in opposite directions. The fields set up by the coils are shown by the dotted lines. The resultant of the two fields has a constant gradient over a distance of approx. 1 cm.

Another source of errors has been eliminated by using coil L_1 instead of a ferroxcube core as is used in the instrument for measuring the magnetostriction. The influence of the powerful stray field of the solenoid on the ferroxcube core could attain the same order of magnitude as the force on the sample produced by the gradient $\mathrm{d}H/\mathrm{d}z$. But since an alternating current of 4 kc/s is passed through coil L_1, the field of the solenoid, averaged over time, does not exert any force upon the coil.

The measuring method is particularly suitable for quickly establishing the relation between the magnetization and the field strength. Here again, the influence of temperature can be readily examined. The inner diameter of the casing surrounding the sample (16 mm) is sufficient to accommodate auxiliary devices and means for controlling the temperature. On the left in fig. 9 is shown a small Dewar flask fitting inside the casing. For a variable temperature below room temperature, a thin coiled copper tube is employed through which cold nitrogen is blown at a controlled rate. For the range above room temperature up to 700 °C, a small furnace is available. In *fig. 11* are plotted a number of curves measured with the instrument.

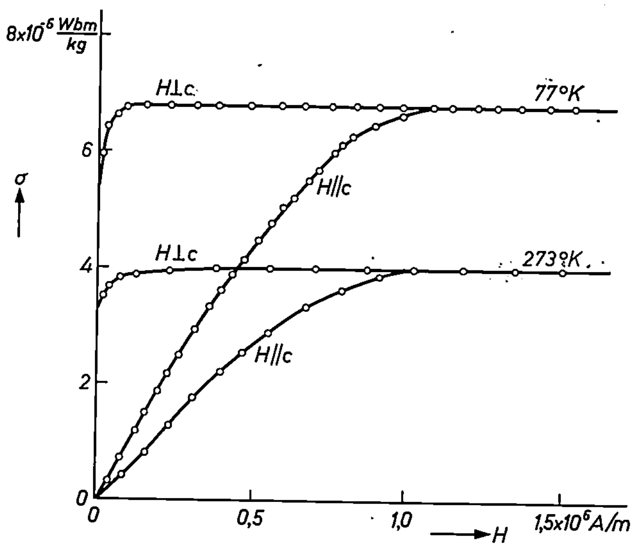


Fig. 11. Magnetization σ of a single crystal of $Ba_2Zn_2Fe_{12}O_{22}$ as a function of the field strength H at different temperatures. From similar measurements of the magnetization of single crystals along various crystallographic axes, in the present instance perpendicular or parallel to the hexagonal axis (c) of the specimen, conclusions regarding the magnetic structure of the substance can be drawn. (Graph reproduced from: U. Enz, J. appl. Phys. **32**, suppl. to No. 3, 22S, 1961. See also: H. Zijlstra, thesis, Amsterdam 1960.)

Summary. Three instruments are described for measuring magnetostriction, magnetic anisotropy and magnetization. In each, the magnetic quantity is determined by measuring the displacement of a ferroxcube core or a coil. This displacement varies the coupling between two coils, which is measured by connecting them to a measuring bridge PT 1200. The samples of the ferromagnetic materials on which the measurements are carried out are 5 cm long rods for magnetostriction measurements and small spheres of a few mm diameter for the other two measurements. For the desired precision of 1%, the design of the instruments has been kept as simple as possible, so that they can be conveniently and quickly operated. By a few simple additions to the equipment, the variation with temperature of the three magnetic quantities can be established.

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.

3176: C. Wansdronk: On the mechanism of hearing (thesis Leiden, Nov. 1961).

3177: H. M. Jongerius: Measurements of optical excitation functions of the mercury atom (excitation by electrons) (thesis Utrecht, Dec. 1961).

3178: H. L. Spier: Influence of chemical additions on the reduction of tungsten oxides (thesis Eindhoven, Nov. 1961).

3179: C. Z. van Doorn: Colour centres in potassium chloride (thesis Utrecht, June 1962).

3180: H. Elings: Experiences with 5-amino-3-phenyl-1-bis(dimethylamido) phosphoryl-1,2,4-triazole, a new fungicide controlling powdery mildew (Proc. Brit. Insecticide and Fungicide Conf. 1961, Vol. 2, pp. 451-459, 1962).

3181: A. J. Pieters: Triphenyl tin hydroxide, a fungicide for the control of Phytophthora infestans on potatoes, and some other fungus diseases (as **3180**, pp. 461-470).

3182: J. Meltzer and K. F. Jacobs: 2,4,5,4′-tetrachloro-diphenyl sulphide, an acaricide with ovo-larvicidal properties (as **3180**, pp. 499-505).

3183: P. A. H. Hart: Partition effects in transverse electron beam waves (J. appl. Phys. **33**, 2401-2408, 1962, No. 8).

3184: J. A. W. van Laar: The durability of paint coatings (Corrosion prevention & control **8**, No. 12, 32-42, 1961; **9**, No. 1, 35-42, 1962; **9**, No. 3, 57-60, 1962; **9**, No. 7, 31-32, 1962).

3185: J. E. Cook and W. J. Oosterkamp: Protection against radiation injury (Internat. Tables for X-ray Crystall., Vol. III, pp. 331-338, Kynoch Press, Birmingham 1962).

3186: W. Elenbaas: Dissipation calorifique par convection libre sur la surface interne de tubes verticaux (Journées internat. Transm. Chaleur, Paris 1961, Vol. I, pp. 379-384, publ. Inst. Français Combust. Energie, Paris 1962). (Heat dissipation by free convection on the inside surface of vertical tubes; in French.)

3187: H. J. G. Meyer: On the low temperature limit of mm-wave cyclotron resonance line widths (Physics Letters, Amsterdam, **2**, 259-260, 1962, No. 5).

3188: J. J. van Loef and W. van Dingenen: The influence of hydrogen on the magnetic scattering of neutrons by a nickel-palladium alloy (Physica **28**, 917-918, 1962, No. 9).

3189: A. Venema: Processes limiting the ultimate pressure in ultra high vacuum systems (1961 Trans. 8th Nat. Vacuum Symp. + 2nd Int. Congress on Vacuum Sci. and Technol., Washington D.C., Vol. 1, pp. 1-7, Pergamon Press, Oxford 1962).

3190: A. van Oostrom: A Bayard-Alpert type ionization gauge with a low X-ray limit (as **3189**, pp. 443-450).

3191: B. D. H. Tellegen: Magnetic-dipole models (Amer. J. Phys. **30**, 650-652, 1962, No. 9).

3192: J. H. Aalberts and M. L. Verheijke: The solid solubility of nickel in silicon determined by neutron activation analysis (Appl. Phys. Letters **1**, 19-20, 1962, No. 1).

3193: F. de Jager and P. J. van Gerwen: CO-modulation, a new method for high-speed data transmission (IRE Trans. on information theory **IT-8**, S 285-S 290, 1962, No. 5).

3194: H. J. Vink: Wechselwirkungen zwischen Störstellen in Halbleitern (Festkörperprobleme I, pp. 1-19, Vieweg, Brunswick 1962). (Interactions between lattice defects in semiconductors; in German.)

3195: J. Dieleman: Influence of solvent-shared association on the electronic spectra of aromatic hydrocarbon negative ions (thesis V.U. Amsterdam, Dec. 1962).

3196: J. W. L. Köhler and J. R. van Geuns: A small liquid-oxygen generator using the gas-refrigerating machine (Annexe 1961-5 au Bull. Inst. Int. Froid, pp. 65-70).

3197: W. F. Druyvesteyn and D. J. van Ooijen: Influence of cold-rolling at 78 °K on the superconducting transition of lead (Physics Letters, Amsterdam, **2**, 328-329, 1962, No. 7).

3198: W. van Gool and G. Diemer: Association of centers in zinc sulfide (Kallman and Spruch, Luminescence of organic and inorganic materials, pp. 391-401, Wiley, New York 1962).

3199: A. Bril: Absolute efficiencies of phosphors with ultraviolet and cathode-ray excitation (as **3198**, pp. 479-493).

3200: H. G. van Bueren, J. Haisma and H. de Lang: A small and stable continuous gas laser (Physics Letters, Amsterdam, **2**, 340-341, 1962, No. 7).
See also Philips tech. Rev. **24**, 95-97, 1962/63 (No. 3).

3201: R. Bleekrode, J. Dieleman and H. J. Vegter: Electron spin resonance on Mn in GaAs (Physics Letters, Amsterdam, **2**, 355-356, 1962, No. 7).

3202: H. B. G. Casimir: Technological advance: a stimulus to basic research (Cherwell-Simon Memorial Lectures 1961 and 1962, pp. 6-24, Oliver and Boyd, Edinburgh/London).

3203: N. V. Franssen: The mechanism of the human voice and wind instruments (4th Int. Congress on Acoustics, Copenhagen 1962, publ. No. G12).

3204: C. M. van der Burgt and A. L. Stuijts: Low-porosity ferrites for high-intensity ultrasonic radiators (as **3203**, publ. No. K22).

3205: E. de Niet, K. Teer and D. L. A. Tjaden: Magnetic recording of audio signals at low tape speeds (as **3203**, publ. No. N11).

3206: C. M. van der Burgt and E. E. Havinga: Finite-strain response of the piezoelectric transducer ceramics Pb-Ba and Pb-Sr zirconate-titanate (as **3203**, publ. No. N18).

3207: A. van Oostrom: Field emission cathodes (J. appl. Phys. **33**, 2917-2922, 1962, No. 10).

3208: A. Claassen and L. Bastings: A study of interferences in the iodometric determination of copper in fluoride medium (Z. anal. Chemie **191**, 401-408, 1962, No. 6).

3209: L. E. Vrenken: High-output fluorescent lamps with circular cross section — The influence of the lamp diameter (Illum. Engng. **57**, 683-687, 1962, No. 10).

3210: N. F. Verster, H. L. Hagedoorn, J. Zwanenburg, A. J. J. Franken and J. Geel: Some design features of the Philips AVF prototype (Nucl. Instr. Meth. **18/19**, 88-92, 1962).

3211: H. L. Hagedoorn and N. F. Verster: Orbits in an AVF cyclotron (Nucl. Instr. Meth. **18/19**, 201-228, 1962).

3212: N. F. Verster and H. L. Hagedoorn: Computer programs for an AVF cyclotron (Nucl. Instr. Meth. **18/19**, 327-335, 1962).
See also Philips tech. Rev. **24**, 106-120, 1962/63 (No. 4/5).

3213: H. L. Hagedoorn and N. F. Verster: Analogue computer studies for an AVF cyclotron (Nucl. Instr. Meth. **18/19**, 336-337, 1962).

3214: G. Meijer: Photomorphogenesis influenced by light of different spectral regions (Advng. Front. Plant Sci. 1, 129-140, 1962).

3215: J. H. Stuy: Transformability of Haemophilus influenzae (J. gen. Microbiol. **29**, 537-549, 1962, No. 3).

3216: C. Haas: On diffusion, relaxation and defects in ice (Physics Letters, Amsterdam, **3**, 126-128, 1962, No. 3).

3217: D. J. van Ooijen, J. H. N. van Vucht and W. F. Druyvesteyn: Superconductivity of $NbSn_2$ (Physics Letters, Amsterdam, **3**, 128-129, 1962, No. 3).

3218: J. J. Wilting: Die Entwicklung auf dem Gebiete der Transistorumformer für Fluoreszenzlampen (Bull. Schweiz. Elektrotechn. Ver. **53**, 1082-1091, 1962, No. 22). (Developments in transistorized converters for fluorescent lamps; in German.)

3219: M. L. Verheijke: Calculated efficiencies of a 3×3 in. NaI(Tl) well-type scintillation crystal (Int. J. appl. Rad. Isot. **13**, 583-585, 1962, No. 10).

3220: N. W. H. Addink and A. W. Witmer: Spectrochemical analysis of impurities present in semiconductors and semiconductor materials in the part per billion range (IX. Colloquium spectroscopicum internationale, Lyon 1961, Vol. II, pp. 340-354).

3221: A. W. Witmer and N. W. H. Addink: New graphite-backed electrode for spectrochemical analysis of metals independent of original sample shape (as **3220**, Vol. II, pp. 405-412).

3222: H. Groendijk: Amplification by longitudinal waves in beam-plasma systems (1962 Northeast Electronics Res. and Engng. Meeting, NEREM Rec. **4**, 184-185, 1962).

3223*: J. S. van Wieringen: Paramagnetic ions as impurities in crystals (and glasses) (Landolt-Börnstein, Zahlenwerte und Funktionen aus Physik, etc., 6th edition, Vol. II, Part 9, pp. 4-17 to 4-42, Springer, Berlin 1962). (Tables in German, introduction also in English).

3224: H. J. L. Trap and J. M. Stevels: Nouveaux types de verre présentant une conductibilité électronique (Verres et Réfract. **16**, 337-343, 1962, No. 6). (New types of glass showing electrical conductivity; in French).

3225: N. W. H. Addink, H. Kraay and A. W. Witmer: The putting to advantage of the absorbing qualities of diluting agents for obtaining 45 degree calibration curves in X-ray fluorescence analysis (as **3220**, Vol. III, pp. 368-384).

3226: S. van Houten: Investigation of the conduction mechanism in transition metal oxides by means of internal friction measurements (Rep. int. Conf. on the physics of semiconductors, Exeter 1962, pp. 197-201, publ. Inst. Phys./Phys. Soc., London 1962).

3227: S. M. de Veer and H. J. G. Meyer: Infra-red absorption by free electrons in germanium (as **3226**, pp. 358-366).

3228: A. C. Aten, C. Z. van Doorn and A. T. Vink: Direct and phonon-assisted optical transitions in zinc telluride (as **3226**, pp. 696-702).

3229: H. J. van Daal, W. F. Knippenberg and J. D. Wasscher: Mobility of electrons and holes in hexagonal silicon carbide (as **3226**, pp. 783-789).

3230: J. van Laar and J. J. Scheer: Influence of band bending on the photoelectric emission of silicon (as **3226**, pp. 827-831).

3231: H. Bremmer: Electromagnetic wave propagation around the earth (Summer school Varenna 1961, Centro Internazionale Matematico Estivo, Istituto Matematico dell'Università, Rome 1962).

3232: 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).

3233: Chr. Meyer: Einige Anwendungen von Entladungsblitzröhren in der medizinischen Photographie (I. Int. Kongress für medizinische Photographie und Kinematographie, Düsseldorf 1960, pp. 314-316, Thieme, Stuttgart 1962). (Some applications of flash-discharge tubes in medical photography; in German.)

3234: F. Th. Backers and J. H. Wessels: Magnetische Fernsehaufzeichnung mit einer Ein-Kopf-Anlage (Nachrichtentechn. Z. **15**, 644-649, 1962, No. 12). (Magnetic recording of television signals using a single head system; in German.)
See also Philips tech. Rev. **24**, 81-83, 1962/63 (No. 3).

3235: C. A. A. J. Greebe: The physical properties of grown *P-I-N* junctions in silicon carbide (thesis Eindhoven, June 1962).

3236: A. M. Kruithof, A. A. Padmos, J. de Vries, J. de Groot and A. L. Zijlstra: Spannungsoptische Bestimmung der Wärmeausdehnungen von Gläsern und Metallen (Sprechsaal für Keramik Glas Email **95**, 464-467, 484-487, 518-520 and 691-694, 1962, Nos. 17, 18, 19 and 24; **96**, 36-39, 1963, No. 2). (Determination of thermal stresses in glass and metals by means of stress birefringence; in German.)

3237: M. J. Sparnaay: Free energy calculations of diffuse double layer systems. The "secondary minimum" (Emulsion rheology, pp. 27-40, Pergamon Press, Oxford 1963).

3238: W. J. Witteman: On vibrational relaxations in carbon dioxide (thesis Eindhoven, March 1963).

3239: A. H. Boonstra, J. van Ruler and M. J. Sparnaay: On the influence of various ambients on the surface conductivity of germanium surfaces (Proc. Kon. Ned. Akad. Wetensch. **B 66**, 64-69, 1963, No. 2).

3240: A. H. Boonstra, J. van Ruler and M. J. Sparnaay: Surface states on cleaned and oxidized germanium surfaces (Proc. Kon. Ned. Akad. Wetensch. **B 66**, 70-75, 1963, No. 2).

3241: J. W. Steketee and J. de Jonge: Sensitized photoconductivity in anthracene (Proc. Kon. Ned. Akad. Wetensch. **B 66**, 76-79, 1963, No. 2).

3242: J. H. N. van Vucht: Interaction of Th_2Al and related getters with hydrogen (thesis Eindhoven, March 1963).

3243: C. Weber: A new resistor network for solving Laplace's equation (Proc. IEEE **51**, 252, 1963, No. 1).

3244: J. Haisma and H. de Lang: Mode patterns obtained by tuning a small gas laser (Physics Letters, Amsterdam, **3**, 240-242, 1963, No. 5).

3245: J. J. Scheer and J. van Laar: Photo-emission from semiconductor surfaces (Physics Letters, Amsterdam, **3**, 246-247, 1963, No. 5).

3246: B. Bölger, J. A. W. van der Does de Bye, H. Kalter and H. J. Vegter: Laser action in a GaAs junction (Physics Letters, Amsterdam, **3**, 252, 1963, No. 5).

3247: W. J. Oosterkamp and C. Albrecht: The evaluation of fluoroscopic screens and X-ray image intensifiers (Technological needs for reduction of patient dosage from diagnostic radiology, editor M. L. Janower, pp. 251-270, publ. Thomas, Springfield, Ill. U.S.A., 1963).

3248: H. B. G. Casimir: Über die statistische Begründung des Nernstschen Wärmetheorems (Z. Phys. **171**, 246-249, 1963, No. 1). (On the statistical basis of Nernst's heat theorem; in German.)

3249: P. J. Severin: Some dynamic aspects of the cathode fall region in a D.C. glow discharge (Physica **29**, 83-92, 1963, No. 1).

3250: H. J. van Daal, W. F. Knippenberg and J. D. Wasscher: On the electronic conduction of α-SiC crystals between 300 and 1500 °K (Phys. Chem. Solids **24**, 109-127, 1963, No. 1).

3251: J. R. van Geuns: Eine kleine Luftzerlegungsanlage mit zwei Gaskältemaschinen zur Gewinnung von flüssigem Sauerstoff (Kältetechnik **15**, 5-10, 1963, No. 1). (A small air fractionating installation with two gas refrigerating machines for the production of liquid oxygen; in German.)

3252: A. Claassen and L. Bastings: The determination of lead in carbon and low-alloy steel according to British Standard 1121 : Part 41 : 1960 (Analyst **88**, 67-68, 1963, No. 1042).

3253: A. van Oostrom: Temperature dependence of the work function of single crystal planes of tungsten in the range 78-293 °K (Physics Letters, Amsterdam, **4**, 34-36, 1963, No. 1).

3254: C. A. A. J. Greebe: On the frequency dependence of the acousto-electric effect in piezo-electric semiconductors (Physics Letters, Amsterdam, **4**, 45-46, 1963, No. 1).

3255: K. G. Srivastava: A magnetic study of some compounds having the K_2NiF_4 structure (Physics Letters, Amsterdam, **4**, 55-56, 1963, No. 1).

VOLUME 25, 1963/64, No. 9 pp. 217-252 Published 26th August 1964

Philips Technical Review

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

A PULSED MAGNETRON FOR $2\frac{1}{2}$ mm WAVES

by G. H. PLANTINGA *).

621.385.6:621.3.032.61

The development of tubes for operating at ever shorter wavelengths has not yet come to an end, as evidenced by the recent completion of a magnetron for $2\frac{1}{2}$ mm waves, described in this article. As far as we know, this is the first magnetron for this wavelength to give satisfactory results. It is capable of delivering a power of 2.5 kW in pulses of 0.1 μs duration, the mean power being 0.5 W.

Some years ago a range of pulsed magnetrons developed in our laboratories was described in this journal [1]). This development was mainly directed towards the attainment of shorter operating wavelengths. The magnetron with the shortest wavelength in this range was a 4 mm tube. A magnetron for $2\frac{1}{2}$ mm waves has now appeared, some aspects of which will be dealt with in this article.

The principal application of magnetrons of the above-mentioned range was said to be radar. Although $2\frac{1}{2}$ mm waves are less suitable for radar because of their severe attenuation in the atmosphere, due to the presence of an oxygen absorption line at this wavelength [2]), 4 mm is certainly not the shortest wavelength for radar purposes. Now that it has proved possible to build a $2\frac{1}{2}$ mm magnetron, there is nothing to stop the production of magnetrons for wavelengths between 4 and $2\frac{1}{2}$ mm, a band for which a need has recently arisen in the field of radar. The applications of the $2\frac{1}{2}$ mm magnetron, however, are mainly in the field of *physical measurements* [3]). Examples that come to mind are measurements in plasma research.

The smallest version of the above-mentioned range of magnetrons, the 4 mm tube, followed from an 8 mm type. After the development of this tube it seemed obvious to try to reduce the wavelength again by a factor of 2. Why the range of magnetrons cannot be extended in exactly this way will presently be explained. Some problems inherent in this scaling-down process will be discussed. Attention will also be paid to special details of construction, where the small dimensions of the essential components created problems. In conclusion a number of measurements will be described.

A photograph of the $2\frac{1}{2}$ mm magnetron (without the magnet) is shown in *fig. 1*.

Scaling down

Although the magnetron is one of the oldest of the microwave tubes [4]) and has been the subject of considerable research over the years, there is still no completely satisfactory theory for this type of tube [5]).

*) Philips Research Laboratories, Eindhoven.

[1]) J. Verweel and G. H. Plantinga, A range of pulsed magnetrons for centimetre and millimetre waves, Philips tech. Rev. **21**, 1-9, 1959/60.

[2]) J. H. van Vleck, Phys. Rev. **71**, 413-424 and 425-433, 1947. See also H. Bremmer, Philips tech. Rev. **15**, 150, 1953/54.

[3]) The same applies to the waveguide equipment for 2 mm microwaves, previously described in this journal: C. W. van Es, M. Gevers and F. C. de Ronde, Pbilips tech. Rev. **22**, 113-125 and 181-189, 1960/61.

[4]) See e.g. K. Posthumus, Principles underlying the generation of oscillations by means of a split anode magnetron, Philips Transmitting News **1**, No. 3, 11-25, 1934, or K. Posthumus, Oscillations in a split anode magnetron, Wirel. Engr. **12**, 126-132, 1935.

[5]) See Vol. I and II of Crossed-field microwave devices, edited by E. Okress, Acad. Press, New York 1961.

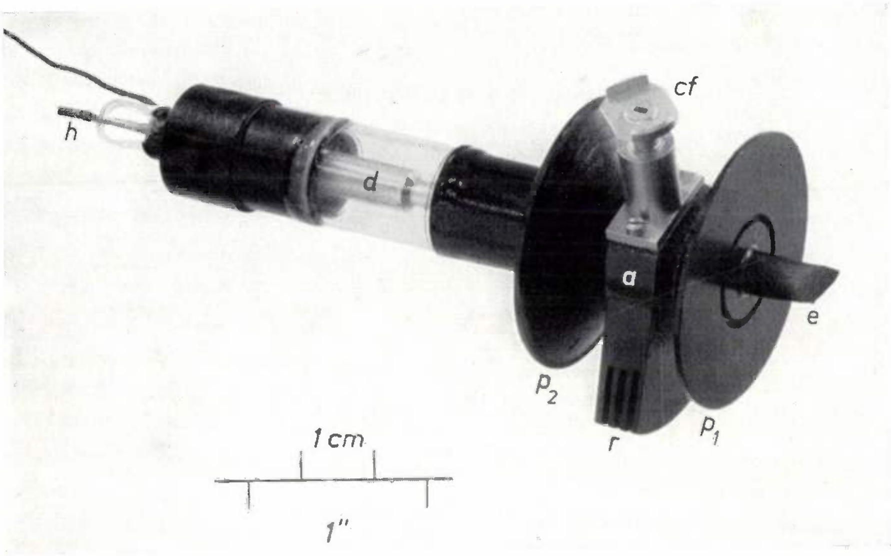


Fig. 1. Pulsed magnetron (without magnet) for $2\frac{1}{2}$ mm waves. It can deliver an output of 2.5 kW in pulses of 0.1 μs duration.

h heater connection. *d* cathode cylinder. *a* anode block. *cf* claw flange [3]) for connection to the waveguide through which the RF energy is extracted. p_1, p_2 pole pieces. *r* cooling fins. *e* pinch-sealed copper exhaust.

The design of a new magnetron is therefore governed to a very large extent by empirical knowledge of existing types. In general the aim will be to design a new magnetron as a modification of a type that has already proved itself in practice.

Magnetrons for shorter wavelengths can most simply be designed by modification on the principles of a *scaling law* [1]). Suppose that a given magnetron having a magnetic field with induction B delivers at a given anode current and anode voltage a certain power output with a wavelength λ. If we now scale up an existing magnetron so that its linear dimensions all become p times larger, and if we make the magnetic induction equal to B/p, then this magnetron at the same current and voltage will deliver the same power output at the wavelength $p\lambda$.

This principle applies exactly only provided the specific conductance at any given position in the new magnetron is $1/p$ times the specific conductance at the corresponding position in the old magnetron, and provided also that the same proportionality holds for the electric field-strength at the cathode. The first condition is never fulfilled, nor usually is the second one. Moreover, in the case of scaling down ($p < 1$), the saturation of the available ferromagnetic materials often makes it difficult to meet the requirement of a higher induction (B/p). Nevertheless a scaling law is very useful for gaining insight into the influence of the various parameters.

A scaling law was used in the development of the range of magnetrons described in the article under reference [1]). One can attempt to continue this range by scaling down to a tube for e.g. 2 mm waves. First, however, one should know the consequences of proportionately reducing the dimensions of the smallest tube in the series (the 4 mm magnetron).

Proportional scaling-down from the 4 mm magnetron

The following considerations are based on the operating region of the 4 mm magnetron. This region is characterized by the following values: magnetic induction $B = 1.5\text{-}1.9$ Wb/m^2, anode voltage $V = 11\text{-}16$ kV, anode current $I = 7\text{-}14$ A.

A 2 mm magnetron which is a scaled-down version of the 4 mm type possesses, according to the scaling law, a corresponding operating region with the same values of voltage and current; the corresponding induction values, however, must be $(4/2) \times (1.5\text{-}1.9) = 3.0\text{-}3.8$ Wb/m^2. The effective emissive area of the cathode of the 4 mm magnetron is 0.07 cm^2, so that the current density at the cathode reaches the appreciable value of $(7\text{-}14)/0.07 = 100\text{-}200$ A/cm^2. A current density of this magnitude is only possible as a result of the secondary emission caused by the electrons returning to the cathode, i.e. back-bombardment [6]). The cathode of the 2 mm magnetron would have to be capable of delivering an even greater current density, namely from 400 to 800 A/cm^2. It is questionable whether current densities in excess of 400 A/cm^2 are to be achieved with any known type of cathode.

With linear scaling-down, the specific dissipation increases in inverse proportion to the square of the scaling factor p. This need not in principle entail a higher average anode temperature, because the higher specific dissipation can be countered with better cooling. What cannot be compensated in such a relatively simple manner, however, is the greater temperature increase of the bombarded anode surfaces during the pulses. The energy conducted away from the bombarded anode surface during a pulse is

[6]) See J. Verweel, Philips tech. Rev. **14**, 44, 1952/53.

very small compared with the energy supplied with each pulse. During the pulses the temperature at the surface therefore rises rapidly. *After* the pulse most of the energy is transferred by conduction to other parts of the anode which, as we have said, can be prevented from getting too hot by more effective cooling. It is not possible, however, to improve the cooling of the bombarded area of the anode. An increase in the specific dissipation therefore means that the temperature of the effective anode surface will inevitably become higher during the pulses (we shall henceforth refer to this as peak heating).

The temperature increase at the surface of a body to which a specific power of W watts per cm² is supplied in a pulse of τ seconds duration can be calculated if a few simplifying assumptions are made. Assuming that the heat is conducted away only in the direction perpendicular to the surface, and neglecting radiation, we find that the temperature increase ΔT at the end of the pulse is given by the following expression [7]):

$$\Delta T = \frac{2W}{\sqrt{\pi k C}} \sqrt{\tau} \,. \quad . \quad . \quad . \quad . \quad . \quad (1)$$

Here k is the thermal conductivity of the medium (in W/cm °C) and C the heat capacity per unit volume (in Ws/cm³ °C). This formula is a reasonably good approximation for the peak heating of a magnetron anode. These anodes are made of copper, and the thermal conductivity and heat capacity of copper are respectively $k = 3.9$ W/cm °C and $C = 3.5$ Ws/cm³ °C. Using these values we can write eq. (1) in the following form:

$$\Delta T = \frac{31\, P\sqrt{\tau}}{\varphi \lambda^2} \,. \quad . \quad . \quad . \quad . \quad . \quad (2)$$

Here P is the power (in watts) which, during a pulse, is lost in the anode as heat, and $\varphi = A/\lambda^2$, where A is the effective anode surface (A and λ^2 expressed in mm², τ in seconds). According to the scaling law, φ remains constant upon a linear scaling. Consequently it follows from (2) that, for a scaling-down at a given value of P, the peak heating can only be kept unchanged by shortening the pulse duration τ in proportion to the fourth power of the scaling factor p; this soon sets a practical limit to the scaling-down process.

One of the specifications for the new magnetron was that it should be capable of operating at maximum pulse durations of about 0.1 μs. For the 4 mm magnetron $\varphi = 0.39$. Using this value as the basis for scaling down, we can calculate from (2) the peak heating to be expected for other tubes — assuming they can be made. *Fig. 2* was constructed for one operating point: $P/\varphi = 460$ kW, the peak heating thus calculated being plotted as a function of wavelength and for various values of τ. It can at once be seen from this figure that such an operating point is impossible for a tube of 2 mm wavelength, with $\tau = 0.1$ μs, for the peak heating would be more than 1100 °C, and the melting point of copper is 1083 °C. At a wavelength of 2 mm the graph only gives practical values at pulses of 0.01 μs or shorter.

The conclusion must be that a 2 mm magnetron, derived from a 4 mm version, is not a practical proposition.

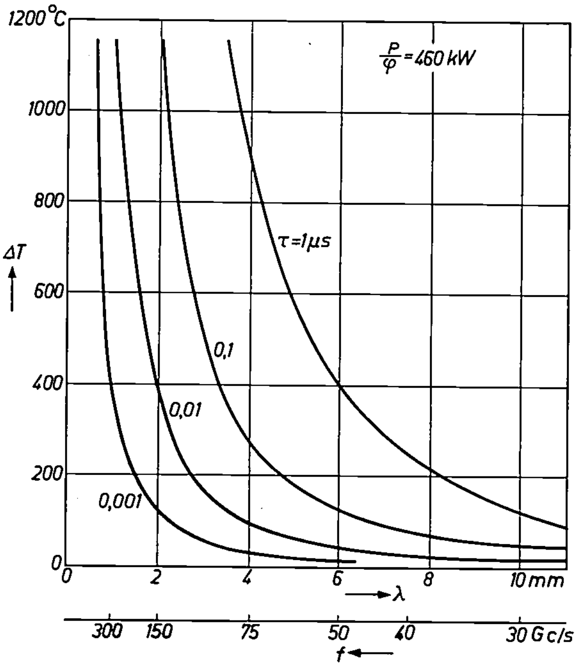


Fig. 2. The peak heating ΔT (calculated from (2)) at the surface of a magnetron anode as a function of wavelength λ, at constant P/φ (= 460 kW) and with the pulse duration τ as parameter.

Table I presents some data of the 4 mm magnetron and of 2½ and 2 mm designs derived from it (the characteristic values B_0 and V_0 at the foot of the table will be discussed later). The table indicates that a 2½ mm magnetron can be regarded as a borderline case, for the operating region of such a tube will be very small compared with that of the 4 mm tube. This is evident from the very high values of magnetic induction, current density and peak heating. Nevertheless, we decided to start on the development of a magnetron for 2½ mm waves. Our decision was prompted by the possibility of modifying the structure of the resonator system in a way that promised favourable results, while the very object

[7]) W. J. Oosterkamp, Problemen bij de constructic van technische röntgenbuizen, thesis Delft, 1939. (Problems in the construction of technical X-ray tubes; in Dutch.)

Table I. Data of the 4 mm magnetron and of scaled-down designs for a 2½ and 2 mm magnetron. In all cases there are 18 resonant cavities. Owing to the high values of magnetic induction, current density and peak heating, the 2 mm design is not a practical proposition, while the 2½ mm tube is a borderline case.

Wavelength (mm)	4	2.5	2.0
Magn. induction (Wb/m²)	1.5-1.9	2.4-3.0	3.0-3.8
Anode voltage (kV)	11-16	11-16	11-16
Anode current (A)	7-14	7-14	7-14
Current density at cathode (A/cm²)	100-200	260-500	400-800
Output power (kW)	5-40		
Peak heating (°C)	120-280	300-700	480-1100
Pulse duration (μs)	0.1	0.1	0.1
Charact. values B_0 (Wb/m²)	0.9	1.45	1.8
Charact. values V_0 (kV)	3.8	3.8	3.8

of the exercise was to determine the extreme limit. The modification referred to consisted in increasing the number of resonant cavities from 18 to 22.

Increasing the number of resonant cavities

Some insight into the effects of increasing the number of resonant cavities can be obtained by examining the equation formulated by Hartree as a starting condition for the oscillation of a magnetron. The Hartree equation gives the minimum value which the anode voltage V must have at a given magnetic induction B in order for oscillation to be possible, and may be written [6][8])

$$\frac{V}{V_0} = \frac{2B}{B_0} - 1 \,. \quad . \quad . \quad . \quad . \quad . \qquad (3)$$

V_0 and B_0 are magnetron characteristics, given by

$$V_0 = 2\,\pi^2 c^2 \,\frac{m}{e} \left(\frac{d_a}{N\lambda}\right)^2 \quad . \quad . \quad . \quad . \qquad (4)$$

and

$$B_0 = 8\,\pi c \,\frac{m}{e}\, \frac{1}{\left(1 - \dfrac{d_k{}^2}{d_a{}^2}\right) N\lambda} \,. \quad . \quad . \qquad (5)$$

Here N is the number of resonant cavities, d_a the diameter of the central aperture in the anode, d_k the diameter of the cathode, c the velocity of light, e and m the charge and mass of an electron. The condition (3) is valid only for magnetic fields with an induction greater than B_0 [8]).

The following remarks may be made on these characteristic quantities: if d_a is varied in proportion to λ there is no change in V_0 and if the ratio $d_k : d_a$ is constant, B_0 varies inversely with λ. This is what is to be expected from the scaling law.

From eq. (4) and (5) it follows that increasing the number of resonant cavities N results in a decrease in the characteristic values V_0 and B_0. Other relative values remaining equal, this also entails lower values of V and B at the relevant operating point. Now in practice there will never be only one parameter varied at a time. If the number of resonant cavities is increased, for example, the diameter d_a of the anode aperture around which the cavities are grouped will also be enlarged, since the thickness of the partitions between the cavities cannot be reduced ad libitum. Enlarging d_a would reduce the advantageous influence of the increased N on V_0 (see eq. 4), but at the same time, as a result of the larger surface area, it has the favourable effect of reducing the heating of the anode during the pulses (of the same power). An increase of N will also usually entail making the cathode diameter d_k larger, so that B_0 remains roughly inversely proportional to N.

There is a limit to the extent to which the number of resonant cavities can be increased; too many cavities would jeopardize the stable operation of the tube in the required mode of oscillation (the π mode) [9]). A sharply defined upper limit for the number of cavities of an open "rising-sun" system — as used in the 4 mm magnetron, with $N = 18$ — cannot be given. An experimental magnetron with 22 cavities for a wavelength of 12½ mm has long been known [10]). The resonator system of the 2½ mm magnetron was derived by roughly scaling down from the system used in the 12½ mm tube. *Fig. 3* shows a photograph taken along the axis of the resonator system for 2½ mm.

The scaling-down process obviously affects the construction of the tube. In the following we shall examine some of the principal constructional aspects of the 2½ mm magnetron, a schematic cross-section of which is shown in *fig. 4*.

Components of the 2½ mm magnetron

The anode block

The anode block of magnetrons with a rising-sun resonator system is made by a hobbing method [6][9]). The hob consists of a steel "negative" of the rising-sun system, which is forced into a block of copper. For the 2½ mm systems this is done at room temperature, whereas the hobbing of the larger resonator systems is done at a temperature of about 550 °C. The hob is made by a grinding process in which N

[8]) G. A. Espersen and B. Arfin, Philips tech. Rev. 14, 88 and 89, 1952/53.

[9]) G. B. Collins, Microwave magnetrons, M.I.T. Radiation Lab. Series, part 6, McGraw-Hill, New York 1948.

[10]) See p. 790 of the book cited in [9]). Not enough is yet known about the possibility of increasing N still further. See G. H. Plantinga, Proc. 4th int. Congress on microwave tubes, Scheveningen 1962, pp. 202-205, Centrex, Eindhoven 1963.

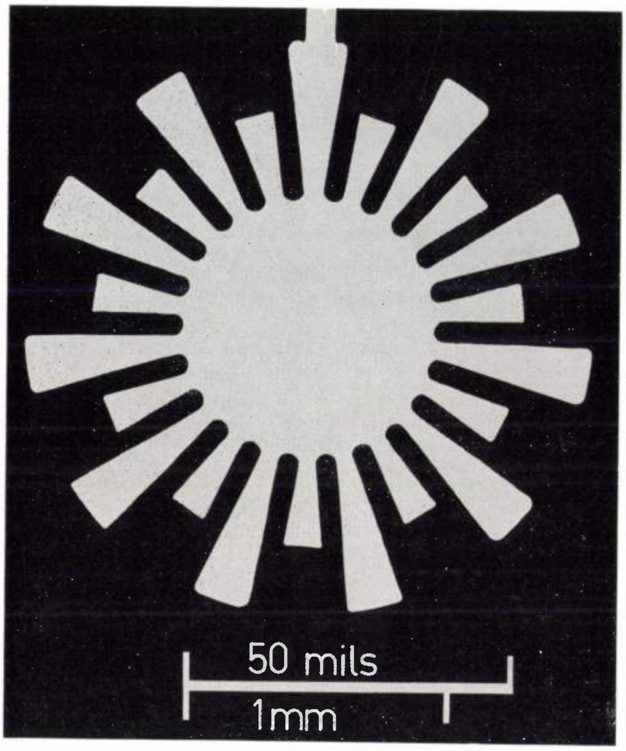


Fig. 3. Axial photograph of the resonator system of the 2½ mm magnetron. It is an open "rising-sun" system with 22 resonant cavities.

slots are ground into a solid steel cylinder; the width of the slots is equal to the required thickness of the cavity walls. Following this, the ½N ribs of the tool that correspond to the small cavities are ground to size.

While it was by no means simple to make a hob for a 4 mm magnetron, the grinding of a 2½ mm hob that meets reasonable demands of precision is even more of a problem. The diameter of the hob must be 2.15 mm, and into this 22 slots 80 μm wide and 0.60 mm deep have to be ground. Dimensions as small as this obviously call for precision-grinding of an exceptionally high order. The slots are ground out with a grinding disc which must be slightly thinner than 80 μm. Special care is needed in lining up the disc before grinding can start. The required alignment precision is achieved by adjusting the temperature of the cooling air supplied to the grinding machine as the last correction to the position of the disc. The same method ensures that, during the grinding operation, there can be no change in the position of the grinding disc in relation to the hob; displacements due to varying expansion of components of the grinding machine — as a result of slight fluctuations in ambient temperature — are compensated by adjustment of the temperature of the cooling air.

The dimensions of the 2½ mm rising-sun system are given in *Table II*. *Fig.5* gives an idea of the size of the 2½ mm hob, which is capable of making more than 100 anode blocks [11]); one of them can be seen in fig. 3.

Table II. Dimensions of cathode and resonator system of the 2½ mm magnetron.

Cathode	diameter 0.60 mm
Anode	diameter of hole 0.95 mm
	height 1.20 mm
Small cavities	depth 0.34 mm
Large cavities	depth 0.60 mm
Vanes	thickness 0.080 mm

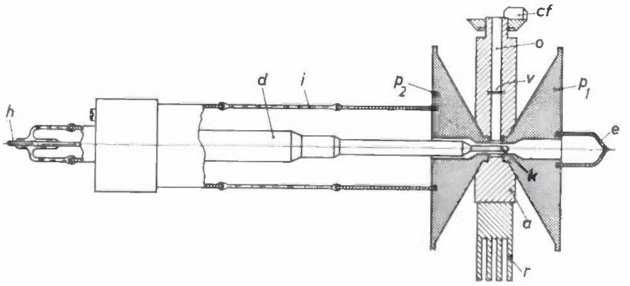


Fig. 4. Cross-section of the 2½ mm magnetron. *i* insulating glass sleeve. *k* cathode. *o* output waveguide. *v* vacuum-tight glass window. Meaning of other letters as in fig. 1.

Many details have been omitted here for the sake of clarity. See fig. 6 for the method of fixing and centering the cathode, and fig. 11 for the output system.

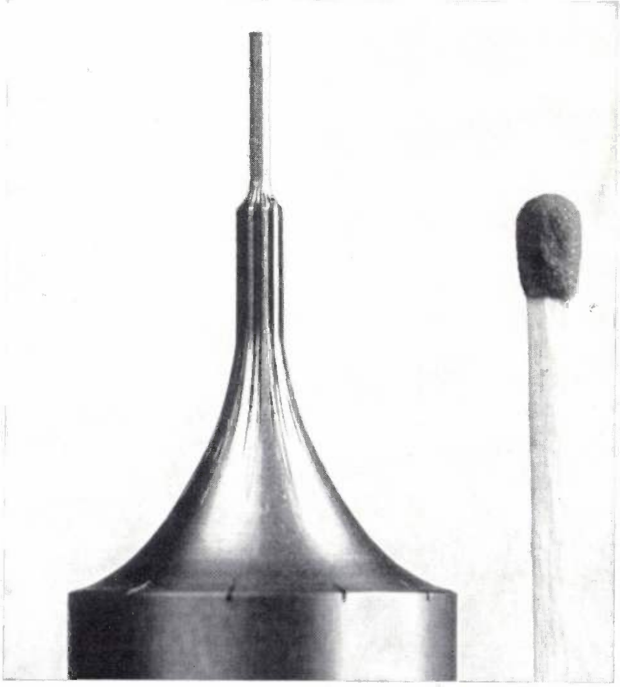


Fig. 5. Tool for hobbing 2½ mm magnetrons. With one such tool more than 100 anode blocks can be made.

[11]) This is a great advantage of hobbing compared with the method of spark machining sometimes used, where the tool is worn out after making only a few anode blocks.

The cathode

In most magnetrons the cathode is fixed to and insulated from the anode block by means of a small glass sleeve. Usually the cathode is mounted by sealing this sleeve to the anode. In the $2\frac{1}{2}$ mm magnetron the emissive part of the cathode has a diameter of 0.60 mm, and the diameter of the anode aperture is 0.95 mm. It is obvious that in this case sufficiently accurate centering of the cathode in the anode aperture is not possible using a sealing method. Centering was already a problem with the 4 mm tube. In that case the solution was found by a method of assembly in which the cathode is soldered to a sleeve which is sealed beforehand to a glass insulator[1]). This method was not sufficiently accurate for the $2\frac{1}{2}$ mm magnetron since the glass insulator, which is part of the cathode mounting, is never entirely free from stresses. These are produced, for example, when the cathode is soldered in. During the heating of the magnetron in the evacuation process, the stresses cause slight deformations, resulting in a displacement of the emissive part of the cathode in relation to the centre of the anode aperture. This is impermissible in this tube.

For the $2\frac{1}{2}$ mm tube a method of centering the cathode was therefore devised which made it possible *to correct the position of the emissive part after evacuating the tube;* once this has been done, however, there must be no chance of a spontaneous change in the position of the cathode. The method adopted is described below. During the correction operation an electrical signal indicates whether the cathode is in fact coaxially situated in the anode aperture.

The method of fixing the cathode is shown schematically in *fig. 6*. Sealed to the glass insulator i is a cylindrical metal body with two flanges (f_1 and f_2) which are interconnected by a very thin wall w. The flange f_3 of the cathode body d is clamped between the flanges f_1 and f_2 in such a way that slight radial movements are still possible. With the aid of a centering jig the position of the cathode in the anode aperture can now be corrected because the thin wall w permits sufficient local deformation. The correction is effected by pressing a steel ball b inwards by means of a lever l. The jig has four such balls and levers, so that the cathode, after setting up the jig, can be moved in two radial directions perpendicular to each other. Three screws s in the flange f_1 make it possible, after centering, to clamp the cathode flange f_3 by deformation of the base of the screw holes. After some time, presumably owing to the formation of a pressure joint, the flanges f_1 and f_3 seize together, so that no further displacement of the cathode is possible.

Unlike other magnetrons, in which the cathode can still be centered after assembly in the tube, the $2\frac{1}{2}$ mm magnetron is released from the jig after centering. This is an important practical advantage.

The fact that the cathode can still be displaced slightly would not of course be sufficient in itself for accurate centering if there were not at the same time a means of *indicating the eccentricity*. An optical indication here is ruled out. An electrical indication is therefore adopted, making use of a characteristic feature of the static — i.e. non-oscillating — magnetron. In a static magnetron the electrons, under the influence of the static electric and magnetic fields, describe curved trajectories between cathode and anode. When the magnetic field is increased and the electric field kept constant, the curvature of the trajectories increases until, at a certain critical magnetic induction, the electrons — putting the matter rather simply — no longer reach the anode, so that the anode current becomes discontinuously zero. The critical induction B_{cr} is easily calculated for a planar magnetron[6]). One finds:

$$B_{\mathrm{cr}} = \sqrt{\frac{2m}{e}}\,\frac{\sqrt{V}}{d}\,. \quad \ldots \quad (6)$$

Here d is the distance between cathode and anode. At a constant anode voltage V, then, B_{cr} is a measure of the distance d, and the latter follows from (6) as a

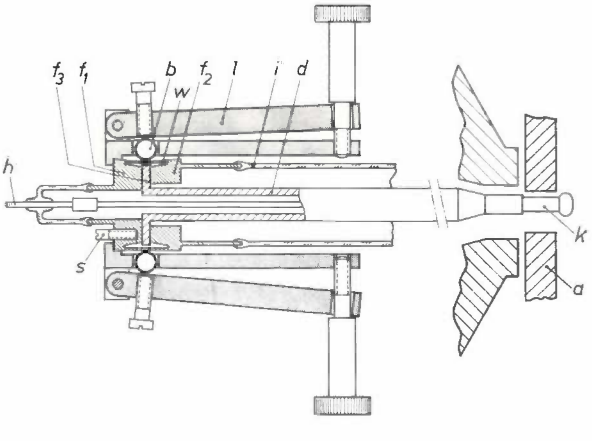


Fig. 6. Mechanism for fixing and centering the cathode in the $2\frac{1}{2}$ mm magnetron. a anode. k emissive portion of the cathode (here drawn off-centre in the anode aperture). h heater connection. The flange f_3 of the cathode body d is clamped between the flanges f_1 and f_2 of a cylindrical body which is sealed to the glass insulator i. Around this assembly is placed the aligning jig, shown shaded, consisting primarily of four levers l with which steel balls b can be pressed inwards. The balls thereby deform the thin wall w which interconnects flanges f_1 and f_2, and in so doing shift the flange f_3 (and hence the cathode) radially. The degree of eccentricity of the cathode follows from electrical measurements (see figs. 7 and 8). When the coaxial position is found, the cathode is fixed by tightening three screws s; this deforms the base of the screw holes so that the flange f_3 is rigidly clamped between flanges f_1 and f_2. This being done, the jig is removed. After some time flanges f_1 and f_3 seize together, making any further displacement of the cathode impossible.

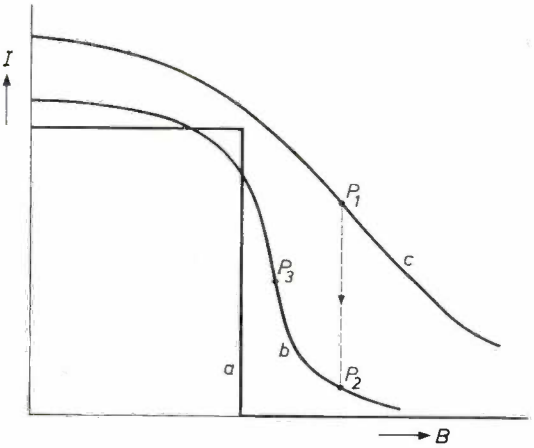


Fig. 7. Anode current I of a non-oscillating magnetron as a function of magnetic induction B at constant anode voltage (schematic). Curve a: cathode coaxially aligned in the anode aperture. Curve b: cathode off-centre. Curve c: cathode still farther off-centre.

function of B_{cr}. At a given anode voltage the value of B_{cr} can be used for determining the position of the cathode in relation to the anode.

In reality, of course, the magnetrons are cylindrical. The position of the cathode can still be determined on the same principle, but the position cannot be expressed so simply in terms of B_{cr}. For if the cathode is situated *eccentrically* in the anode aperture, then with increasing magnetic induction and constant anode voltage the electrons at the side of the cathode farthest away from the anode (where the electric field is weakest) will fail to reach the anode before those at the other side. We cannot therefore in this case distinguish one discrete value for B_{cr}; instead there is a region in which the current gradually decreases as the induction increases. This is represented for various degrees of eccentricity by curves b and c in *fig. 7*.

The shape of the curve $I = f(B)$ is therefore an indication of the measure of eccentricity of the cathode. For zero eccentricity the simplified theory again gives one discrete value for the critical induction, viz:

$$B_{\text{cr}} = \sqrt{\frac{32m}{e}}\;\frac{\sqrt{V}}{\left(1-\dfrac{d_{\text{k}}^2}{d_{\text{a}}^2}\right)d_{\text{a}}}\,. \quad . \quad . \quad (7)$$

For a coaxial cathode, then, the characteristic has the discontinuous form of curve a in fig. 7.

The procedure during centering is as follows. First the magnetic induction is set to a value at which the anode current is considerably lower (e.g. by half) than at $B = 0$. The cathode is then displaced slightly in one of the two directions allowed by the jig. If the result is a drop in anode current, this is an indication that the cathode has moved closer towards the coaxial position. The operating point in the graph $I = f(B)$ (fig. 7) has then shifted to another curve, e.g. from P_1 on curve c to P_2 on curve b. The cathode can now be moved farther in the same direction, but first it is useful to reduce the induction slightly: the effect of this is to move the operating point to a steeper part of the same curve (e.g. from P_2 to P_3), which improves the sensitivity of the indication. The optimum position in the direction of displacement is reached when the anode current is minimum. The same procedure is then repeated in the other direction of displacement. In this way highly accurate centering is achieved.

Fig. 8 shows two experimental $I = f(B)$ curves. The curve on the right was plotted before adjusting the cathode, the one on the left shows the result of measurements on the coaxial configuration. In both

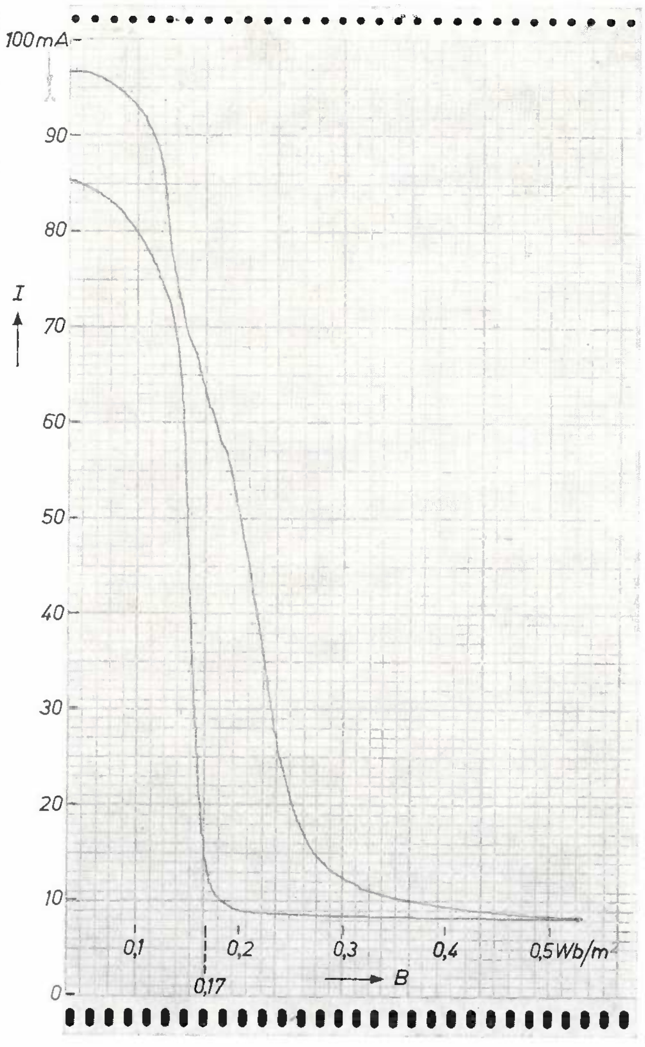


Fig. 8. Recording showing the anode current I of a non-oscillating 2½ mm magnetron versus magnetic induction B at 50 V anode voltage. The right-hand curve relates to an off-centre position of the cathode; the left-hand curve was obtained after careful coaxial alignment of the cathode. The critical induction calculated from eq. (7), $B_{\text{cr}} = 0.17$ Wb/m², fits the latter curve.

sets of measurements the anode voltage was 50V. At this value equation (7) gives $B_{cr} = 0.17$ Wb/m^2. This result is in good agreement with the latter curve.

There is, incidentally, a considerable discrepancy between experiment and theory, for according to eq. (7) one would expect a step-function curve with the discontinuity at $B = B_{cr}$. It would take us too far to go into the reasons for this difference. It will be enough here to comment on the "tail" of the curve. That the current does not drop to zero at high values of B can be explained by the fact that the emission from the cathode is not limited to the part referred to as the emissive area. There are other parts, in particular at the ends, that also emit electrons. These electrons traverse parts of the magnetic field where the induction is lower than that prevailing inside the anode aperture. If the induction here is higher than the critical value, this will not necessarily prevent some of the other electrons mentioned from reaching the anode.

In the experimental magnetron on which the curves in fig. 8 were recorded, no measures were taken to prevent end emission.

A photograph of the cathode used in the $2\frac{1}{2}$ mm magnetron — an *impregnated* cathode [12]) — can be seen in *fig. 9*. The favourable properties of this type of cathode are again evident in this tube, current densities being measured up to 400 A/cm^2.

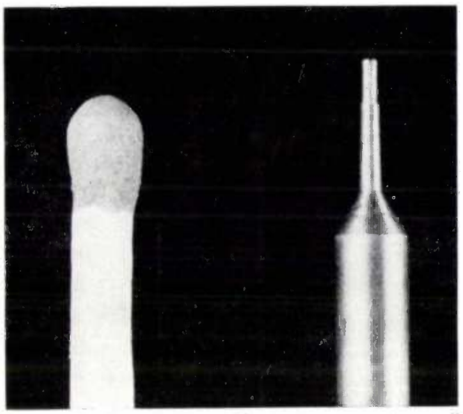


Fig. 9. The cathode of the $2\frac{1}{2}$ mm magnetron. It is an impregnated type [12]), capable of delivering 400 A/cm^2.

The output circuit

The high-frequency energy generated in a magnetron is extracted via a waveguide. The latter requires a vacuum-tight seal, designed so as to minimize power losses by reflection and absorption at the operating frequency (π mode). The seal should also reflect as little as possible at other frequencies, in particular the frequencies of some other modes of operation than the π mode. If there is any strong reflection at another mode, that mode is decoupled from the external load, and the only load is formed by internal (circuit) losses. It is known that oscillation can easily occur in a weakly damped mode [9]). If Hartree's condition for this mode is roughly equal to that for the π mode, eq. (3), the magnetron's operation in the π mode can be disturbed.

An output circuit more than capable of meeting these requirements has long been known for the centimetre range. Curve *I* in *fig. 10* shows for this circuit the modulus of the reflection coefficient r as a function of the frequency f. In the frequency range represented $|r|$ is nowhere greater than 16%. The attenuation does not exceed 0.07 dB. If this output circuit for the 3 cm range were to be scaled down for the $2\frac{1}{2}$ mm range, the result would again be curve *I* for $|r|$, but now with the lower frequency scale. For technical reasons, however, exactly proportional scaling-down was not possible. The resultant $2\frac{1}{2}$ mm circuit therefore has a slightly different frequency characteristic, represented by curve *II* in fig. 10. As can be seen, a reflection peak of about 50% occurs in the region of 135 Gc/s. This presents no difficulties, however. At the operating frequency of the magnetron the reflection is only about 10% and the attenuation here is less than 0.3 dB.

The construction of the output system to which curve *II* relates is shown, somewhat simplified, in *fig. 11*. The glass window providing the vacuum-tight seal for the waveguide is 100 μm thick.

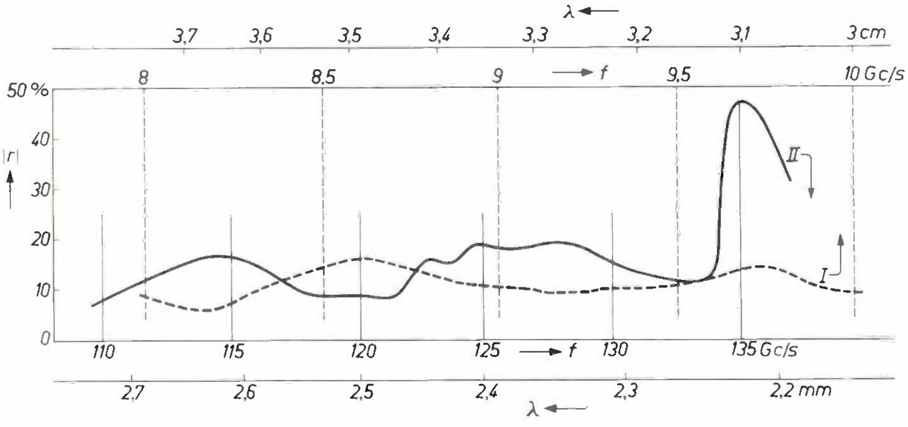


Fig. 10. Curve *I*: measured absolute values $|r|$ of the reflection coefficient of a 32 mm output system, as a function of frequency f (upper scale). Exactly proportional scaling-down of this circuit to $2\frac{1}{2}$ mm would show the same curve for $|r|$ on the lower frequency scale. Such a circuit, however, is not technically feasible. Curve *II* was measured on a slightly different circuit for $2\frac{1}{2}$ mm.

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

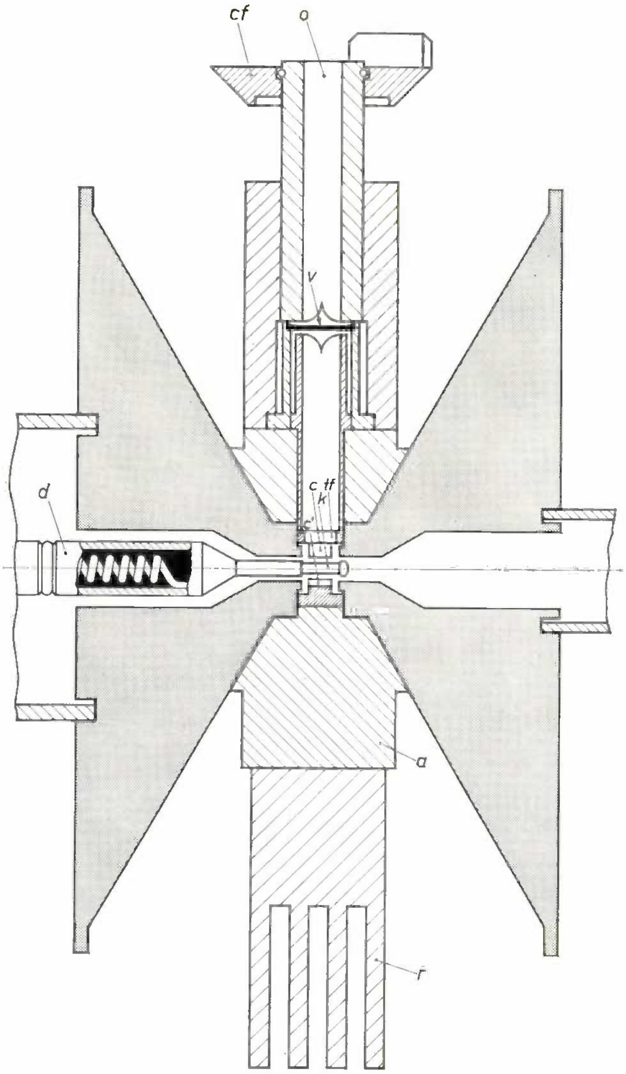


Fig. 11. Cross-section of the $2\frac{1}{2}$ mm magnetron showing the output system with the vacuum-tight window v in more detail than in fig. 4. One of the large resonant cavities can be seen (c) and one of the small ones (c'). tf is a quarter-wavelength impedance-matching transformer. A part of the heater inside the cathode is shown. Other letters as in fig. 4.

Measurements on the $2\frac{1}{2}$ mm magnetron

Some of the measurements necessary in the assembly of the tube have already been mentioned. We have shown how the measurement of the *I-B* characteristic (fig. 8) is used for the alignment of the cathode, and we have also referred to reflection measurements on the output system (fig. 10). As regards the latter measurements it may be noted that the transmission properties of the window can still be checked after the window has been sealed in, before the final assembly — unlike the method of assembly used for the larger predecessors of the $2\frac{1}{2}$ mm tube.

Another measurement required during assembly relates to the coupling of the resonator system to the output waveguide. The coupling used is a quarter-wavelength transformer (fig. 11). This is a $\frac{1}{4}\lambda$ section of waveguide (λ being the wavelength in the guide) the characteristic impedance of which is such that the low impedance offered by the resonator system is roughly matched to the high impedance of the output waveguide. The dimensions of the impedance-matching transformer, like those of the magnetron, are very small: the cross-section of the waveguide is 0.093×2.00 mm. The dimensions are so critical that, after the transformer has been fitted — even though it was made with the greatest precision — the coupling generally requires some additional correction. This is done by slightly deforming part of the transformer. Only after this has been done can the window be mounted. The coupling factor is measured with a bridge circuit, using a hybrid T and a variable impedance(seethe articles under ref. [3])); for the $2\frac{1}{2}$ mm magnetron the usual setting is 1.5. The quality factor Q_u of this type of tube unloaded is roughly 400.

The output power of the $2\frac{1}{2}$ mm magnetrons is measured by a conventional water-load, in which water passes through a glass pipe fitted in the waveguide. Dimensioning is such that the power delivered to the waveguide is completely dissipated in the water, so that the power can be calculated from the temperature difference between the water entering and leaving the pipe, together with the rate of flow. The dimensions involved in this method for $2\frac{1}{2}$ mm wavelength are again very small: in our arrangement the outside diameter of the glass pipe is 0.45 mm, the wall thickness about 0.07 mm. The flow rate of the water can be adjusted so that an average power of 0.2 W (or of 1 kW in pulses of 0.1 μs with a repetition frequency of 2000 per second) produces a temperature increase of 4 °C.

Table III below gives various characteristic values measured on a given $2\frac{1}{2}$ mm magnetron. The tube delivers a power of 2.5 kW in pulses of 0.1 μs at a magnetic induction of 2.5 Wb/m²,

Table III. Some characteristic values measured on a $2\frac{1}{2}$ mm magnetron.

Output power:	
peak value	2.5 kW
average value	0.5 W
Magnetic induction	2.5 Wb/m²
Anode voltage	9 kV
Anode current	14 A
Current density	400 A/cm²
Pulse duration	0.1 μs
Pulse repetition frequency	2000 s⁻¹
Peak heating	600 °C
Frequency	120 Gc/s
Coupling factor	1.5
Q_u	380

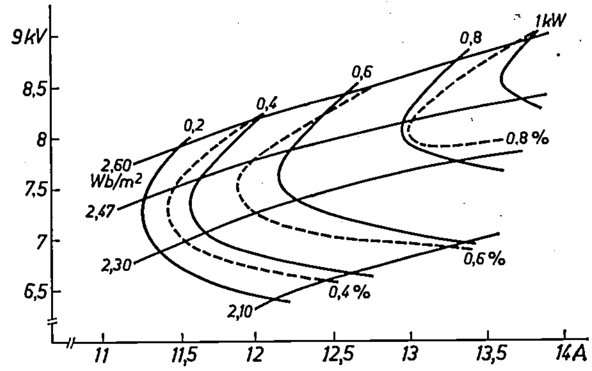


Fig. 12. Performance chart of a 2½ mm magnetron, for a frequency of 120.4 Gc/s, a pulse duration of 0.1 μs and a pulse repetition frequency of 2000 per second. The chart shows contours of constant output power, constant efficiency and constant magnetic induction.

an anode voltage of 9 kV and an anode current of 14 A. These values can be used for estimating the peak heating of the anode, giving a value of roughly 600 °C.

Note again the very high current density of 400 A/cm² delivered by the impregnated cathode employed in this tube.

Fig. 12 shows the performance chart for a 2½ mm magnetron. The chart gives contours of constant magnetic induction, constant output power and constant efficiency.

Several members of our laboratories have contributed to the realization of the magnetron described. The hob was ground by H. J. Ronde in cooperation with M. C. Verhagen, who supervised the assembly of the tubes; F. Lingers had an important part in the construction of the cathode centering device; the cathodes were made by J. R. Blatter, and all the measurements described were done by H. Tjassens.

Summary. A 2½ mm magnetron has now, been added to the earlier described range of pulsed magnetrons, which ended with a 4 mm tube. Unlike the previous tubes in the range, which were scaled-down versions of a 32 mm magnetron with 18 cavities, the 2½ mm tube has a resonator system consisting of 22 cavities. The article describes the anode block and the hobbing method used for making it, the cathode (an impregnated type) and the procedure by which it is aligned in the anode aperture after evacuation of the tube, and the output system. The 2½ mm magnetron can deliver a power of 2.5 kW in pulses of 0.1 μs. Some measurements on the magnetron are discussed and a performance chart is given.

FAST AND ACCURATE MAGNETIC MEASUREMENTS ON SAMPLES WEIGHING A FEW MILLIGRAMMES

by G. W. van OOSTERHOUT *) and L. J. NOORDERMEER **).

621.317.42

In research and development work on magnetic materials it is important to be able to measure quickly and accurately the magnetic moments of samples weighing only a few milligrammes. Such measurements provide the basis for determining important magnetic properties, such as saturation magnetization, coercive force and remanence. An apparatus designed for this purpose is described in this article.

In the development of magnetic materials it is important to know at an early stage whether a particular material is suited to the purpose for which it is intended. This applies especially to materials — usually in powder form — for magnetic recording [1]). At present it is not yet possible to decide from magnetic measurements on a powder whether the material is suitable for making e.g. magnetic tape. For this purpose it is still necessary to produce an experimental tape. It is, however, possible, by measuring the coercive force, saturation magnetization and remanence of the material, to see whether it is definitely *unsuitable*. The making of experimental tapes can thus be limited to materials that have at least some chance of success.

For experimental purposes it is useful if the measurements can be carried out on small samples. One of the advantages is that only small quantities need be made of the substances to be investigated, so that full benefit can be derived from the advantages of preparative chemistry on a small scale. Some ten years ago it was still a very difficult matter to perform magnetic measurements on samples weighing as little as a few milligrammes. One of the few properties that could be measured on small samples was the saturation magnetization [2]). In 1954 a method was developed in Philips Research Laboratories for rapidly and accurately measuring the magnetic moment of very small samples [3]). The method was intended at the time for measurements on galvanically deposited magnetic material. Since the properties of such materials depend on the thickness of the deposited layer, it was necessary to measure samples weighing a few milligrammes. The set-up designed to this end has gradually grown to be a useful tool for investigations concerning magnetic tape: firstly, for the above-mentioned selection of materials (powders), and secondly for measurements on finished tape. The apparatus is capable of measuring a magnetic moment as small as about 1.5×10^{-12} Vsm, which is the saturation moment of about 6×10^{-3} mg of iron.

After this apparatus had been used with good results for several years, the question arose as to whether its sensitivity could be increased sufficiently to permit the study of phenomena connected with the print-through effect. This effect occurs in a tape wound on a spool, and is due to the magnetization pattern in a winding being taken over by adjacent windings. Bound up with this effect are the weak magnetizations caused by the stray field of the magnetic tape itself, and which are therefore of the order of one thousandth of the remanence. In order to measure such weak magnetizations on a length of magnetic tape of e.g. 1.5 cm, a considerably higher sensitivity was needed.

In the following we shall first briefly describe the principle of the method of measurement and explain how the coercive force, the remanence and the saturation magnetization can be derived from measurements of a magnetic moment. We shall then consider the steps taken to increase the sensitivity sufficiently to permit the detection of a magnetic moment of about 7.5×10^{-14} Vsm (the saturation moment of 3×10^{-4} mg of iron). The article concludes with a few examples, including measurements concerning print-through.

Principle of the method

The principle of the method of measurement is illustrated in *fig. 1*. The sample under study — which, as mentioned, usually consists of a little powder or a piece of magnetic tape — is applied over a length of

*) Philips Research Laboratories, Eindhoven.
**) Formerly with Philips Research Laboratories, Eindhoven.

[1]) For a general introduction to magnetic recording, see: W. K. Westmijze, The principle of the magnetic recording and reproduction of sound, Philips tech. Rev. 15, 84-96, 1953/54.

[2]) G. W. Rathenau and J. L. Snoek, Apparatus for measuring magnetic moments, Philips Res. Repts. 1, 239, 1946.

[3]) G. W. van Oosterhout, A rapid method for measuring coercive force and other ferromagnetic properties of very small samples, Appl. sci. Res. B6, 101-104, 1956/57. A method based on the same principle is described by: S. Foner, Vibrating sample magnetometer, Rev. sci. Instr. 27, 548, 1956. For an improved version of this instrument, see: S. Foner, Versatile and sensitive vibrating-sample magnetometer, Rev. sci. Instr. 30, 548-557, 1959.

about 1.5 cm around a thin bar (diameter e.g. 2 mm) of non-magnetic material, preferably glass. The sample is magnetized in the direction of the bar axis. One end of the bar is connected to the speech coil of a loudspeaker. When the loudspeaker is connected to an alternating-current source, the sample moves to and fro axially with the bar. Identical search coils are placed around the ends of the sample, coaxially with the bar. When the sample is in vibration, an e.m.f. is induced in these two coils. The coils are connected in series opposition (astatic coils), so that interfering e.m.f.'s induced in the coils by fields from extraneous sources cancel each other. On the other hand the e.m.f.'s produced by the sample reinforce one another. The resultant alternating e.m.f. is amplified and measured with an electronic millivoltmeter, the deflection of which is a measure of the magnetic moment of the sample.

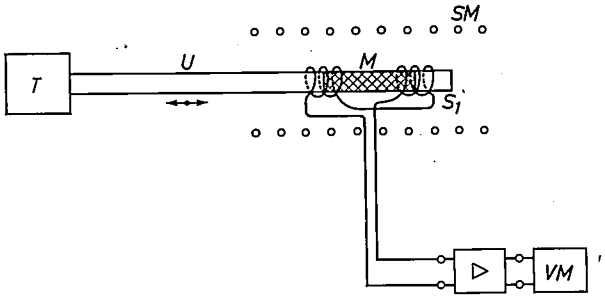


Fig. 1. The magnetic moment of a small sample M of magnetic material is measured by causing the sample to vibrate inside an astatic pair of coils S_1. The sample (usually a powder or a piece of magnetic tape) is applied over a length of about 1.5 cm to the surface of a bar U which, driven by a loudspeaker system T, moves to and fro in the axial direction. VM electronic voltmeter, preceded by an amplifier. SM solenoid which can generate a homogeneous external magnetic field at M.

Determination of saturation magnetization, coercive force and remanence

The saturation magnetization of a material is determined by measuring the magnetic moment of a sample of known weight in the magnetically saturated state. The measurement must of course be done in a strong external field, which has the direction of the vibrating bar and keeps the sample in the saturated state. This field is supplied by a solenoid, fitted coaxially with the bar and the search coils, or by an electromagnet, with holes in the pole pieces for the bar to pass through. As a rule the direct current through the solenoid or through the electromagnet shows a slight ripple, so that the external field contains an alternating component. This does not induce an e.m.f. in the astatic coils and therefore causes no trouble.

When the external field is reduced to zero and the magnetic moment measured again, one can determine from this the remanence of the material.

The coercive force is measured by letting the external field grow in the reverse direction until the e.m.f. induced in the search coils reaches zero. The magnetization of the sample is then zero, and the strength of the external field is equal to the coercive force.

When interpreting the measurements, account should be taken of the demagnetizing field of the sample itself. This applies in particular to remanence measurements. These complications of interpretation are not, however, characteristic of the method of measurement described here, and therefore will not be dealt with in this article.

During the measurement the sample is shifted in relation to the search coils until the induced e.m.f. is maximum. Since the signal is in the form of an alternating voltage that can readily be amplified, the sensitivity of this method is substantially higher than that of the ballistic methods formerly used.

For the purpose of absolute measurements the apparatus is calibrated, using, for instance, a nickel sample of known weight which is magnetized to saturation (for nickel 0.7×10^{-4} Vsm/kg). The nickel sample is fitted over the same length of the bar in order to eliminate the influence of the shape of the samples.

Measuring arrangement with compensating device

In the measurement under fig. 1 the e.m.f. produced depends on the amplitude and the frequency of vibration of the sample. These two quantities are not entirely constant, which limits the accuracy of the measurement. We have eliminated this limitation by adopting a compensating method. Mounted at another point on the bar carrying the sample is a small permanent magnet of platinum-cobalt[4]), surrounded by a second astatic pair of coils S_2 (*fig. 2*). The magnet and the sample are rigidly connected to each other, and therefore vibrate identically. The voltages induced in the two pairs of coils consequently depend in the same way on amplitude and frequency, so that changes in these quantities do not affect the ratio between these voltages.

This ratio is measured with a compensating circuit. The electronic millivoltmeter, again preceded by an amplifier, now serves as a null instrument. The setting q of the voltage divider R_2 at which the null instrument shows no deflection is a measure of the

[4]) Because of its low temperature coefficient and low sensitivity to extraneous fields, platinum-cobalt is eminently suited for use as a standard magnet. See: R. A. Mintern, Platinum alloy permanent magnets, Platinum Metals Rev. 5, 82-88, 1961.

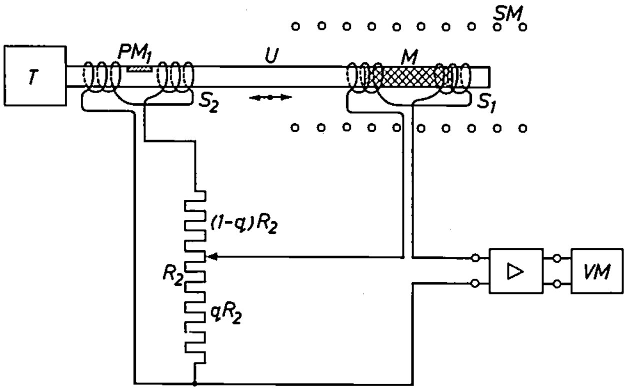


Fig. 2. Refinement of the method in fig. 1: the e.m.f. induced by the sample M in the pair of coils S_1 is compensated by an e.m.f. induced by a platinum-cobalt magnet PM_1 in the pair of coils S_2. This makes the measurement insensitive to changes in the amplitude and the frequency of vibration. The millivoltmeter VM now acts as a null instrument. The setting q of voltage divider R_2 is a measure of the magnetic moment. Meaning of the other symbols as in fig. 1.

required magnetic moment. Calibration can again be carried out using a sample of known magnetic moment.

The essentials of the circuit actually employed are shown in *fig. 3*. The coils S_1, containing the sample under measurement, are connected to the calibrated, stepwise variable voltage divider R_1, with which a part $E_x = pE_1$ of the voltage E_1 across R_1 can be taken off. In the successive steps of this voltage divider, p has the values 1, 1/3, 1/10, 1/30, 1/100, 1/300 and 1/1000. In this way E_x can be made smaller than the voltage E_2 produced by the permanent magnet vibrating in S_2. The setting q of R_2 is then adjusted until the null instrument shows no deflection, after which q can be read off to three decimal places. In this connection E_2 is chosen such that it is about 1000 times greater than the smallest voltage detectable with the null instrument. For this purpose R_2 is connected to the fixed voltage divider R_3-R_4.

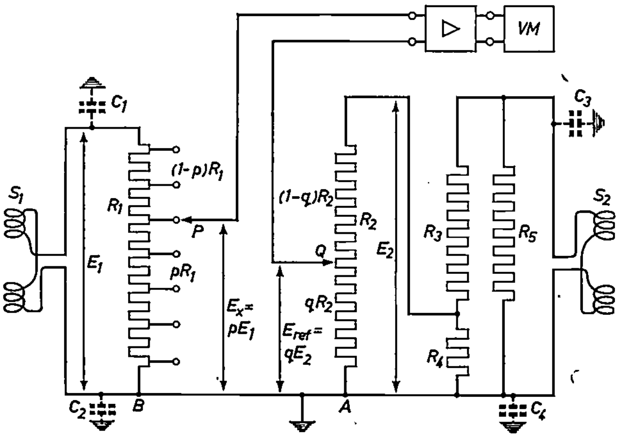


Fig. 3. Essentials of the compensating circuit. The pair of coils S_1, inside which the sample vibrates, is connected to a calibrated voltage divider R_1. The platinum-cobalt magnet vibrates inside the pair of coils S_2. A fraction E_2 of the e.m.f. induced in S_2 appears across the voltage divider R_2 via the fixed voltage divider R_3-R_4. R_5 is a resistor with which the phase difference between the compared voltages E_x and $E_{\rm ref}$ can be reduced to zero. When the circuit is earthed as shown here, no trouble is experienced from the four stray capacitances C_1, C_2, C_3 and C_4, which are produced by the electrical screening of the cables interconnecting the pairs of coils and the compensating circuit. Now, however, points P and Q, between which the voltage must be reduced to zero, both have a certain potential with respect to earth. For this reason it is necessary to use a difference amplifier [5]).

Owing to the inductances of the pairs of coils S_1 and S_2, the currents through R_1 and R_2 will show phase differences with respect to the vibration of the bar. Only when these differences are identical will the currents be in phase, which is necessary for exact compensation. To make the phase shifts equal, a resistance R_5 is put parallel with the coils S_2. For given pairs of coils S_1 and S_2, the resistance R_5 only has to be adjusted once.

The combination of amplifier (about 1000×) and electronic millivoltmeter as null instrument makes it possible to observe voltages of 20 nanovolts. Special measures have been taken to turn this high sensitivity to good advantage. The compensating circuit and the amplifier are magnetically screened. The amplifier is fitted with low-noise valves at the input, and is made selective by means of a bandpass filter for the vibration frequency of the sample (125 c/s) with a pass band of 30 c/s. Moreover, the mains harmonics at 100 and 150 c/s are additionally suppressed. A photograph of the measuring arrangement is shown in *fig. 4*.

Stray capacitances

Care was taken in the design of the circuitry to avoid interference from stray capacitances. The pairs of coils S_1 and S_2 are connected to the compensating circuit by cables with earthed screening. These cables therefore have a capacitance to earth of a few hundred pF. At a frequency of 125 c/s a capacitance of 100 pF represents an impedance of more than 10^7 ohms. This is very high compared with the resistances, of the order of 1000 ohms, used in the measuring circuits. The stray capacitances therefore have no significant influence on the amplitudes of the voltages that have to compensate each other. On the other hand the slight phase shifts which they give rise to in these voltages result in a perceptible difference voltage (*fig. 5*). This causes no trouble provided the phase shifts are independent of the settings of the voltage dividers R_1 and R_2 (fig. 3), for the phase correction by means of R_5 relates to a phase shift of this kind. The phase shifts will be independent of the settings of the voltage dividers when no currents flow through the two control contacts P and Q. If the null instrument has an infinitely high input impedance — or if we imagine it to be removed for a

[5]) See e.g. G. Klein and J. J. Zaalberg van Zelst, General considerations on difference amplifiers, Philips tech. Rev. **22**, 345-351, 1960/61, and by the same authors: Circuits for difference amplifiers, I and II, Philips tech. Rev. **23**, 142-150 and 173-180, 1961/62.

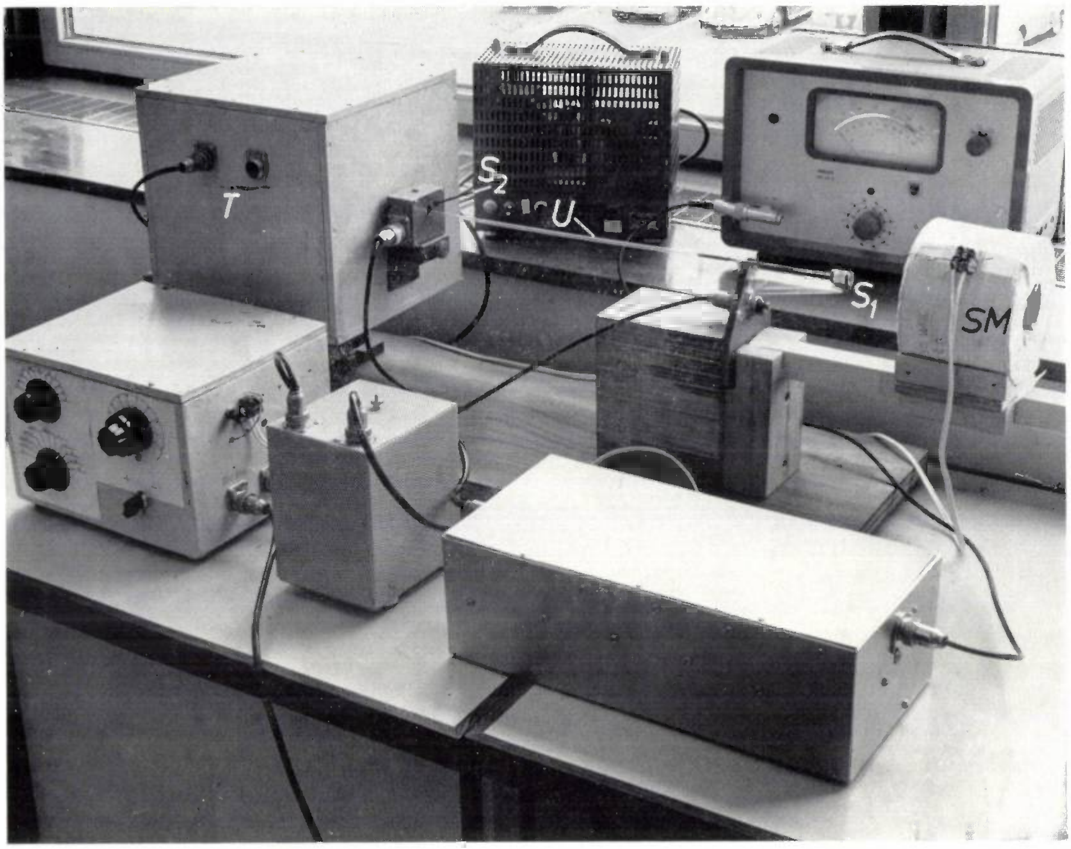


Fig. 4. The apparatus for measuring magnetic moments by the compensation method. S_1 and S_2 are the astatic pairs of coils, inside which, respectively, the sample and a small platinum-cobalt magnet (both fixed to the bar *U*) are caused to vibrate. Left, the loudspeaker system *T* for the drive. In the foreground, from left to right: the compensator (including the difference amplifier), a standard amplifier (1000 ×) and a bandpass filter. In the background: the amplifier that drives the loudspeaker system, and the electronic millivoltmeter (the null instrument). The solenoid *SM* is mounted on a carriage and can be slid around the coils S_1.

moment — then the control contacts in a circuit as in fig. 3 will certainly carry no current. No trouble is then experienced from the stray capacitances. The same then applies when the input impedance of the null instrument is not infinitely high, for once compensation has been effected it makes no difference whether the null instrument is connected or disconnected.

In fig. 3 none of the terminals of the null instrument are earthed, in other words the first amplifying stage must be designed as a difference amplifier [5]). It is in fact impossible to earth the circuit at one of the terminals of the null instrument, for in that case a current could flow via this earth connection and the stray capacitances through the control contact connected to the earthed terminal. Movement of this contact would then not be entirely without influence on the current distribution in the circuit, and would in general cause slight phase shifts, making exact compensation impossible.

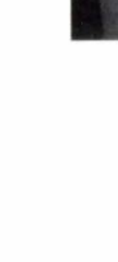


Fig. 5. A small phase difference φ between the voltages E_x and E_{ref}, which are to compensate each other, gives rise to a relatively large voltage difference ΔE.

The vibration frequency

In the original measuring arrangement (with no compensating device) the sample was set in motion by a loudspeaker system operating on 50 c/s. Since the sensitivity was relatively low, little inconvenience was experienced from interfering fields of 50 c/s and multiples of it. When, however, the sensitivity was increased by introducing the compensation method, these fields became troublesome. Little could be done with filters because of the fact that the frequency of the desired signal was also 50 c/s. It was therefore necessary to choose a different vibration frequency.

A second source of interference that influenced the choice of frequency was the noise of the first amplifier tube. This noise contains a component which is inversely proportional to the frequency [6]). From these considerations it is evident that the frequency should be chosen as high as practicable, avoiding as far as possible multiples of 50 c/s. A limit is set to the height of the frequency, however, by considerations of noise nuisance and by the acceleration forces, which are proportional to the square of the frequency and threaten the connection between sample and bar. The frequency finally decided upon was 125 c/s.

The vibrator

The mechanical part of the vibrator is sketched in *fig. 6*. The bar, which carries the sample and the platinum-cobalt magnet, is fixed to a spring of a type that combines high lateral stiffness with a long life [7]). On the left in fig. 6 can be seen the loudspeaker coil and the loudspeaker magnet. The current through the loudspeaker coil is supplied by an amplifier which receives its input signal from a second platinum-cobalt magnet fixed to the vibrating bar, and contained in a third pair of coils S_3. Because of this feedback the system is able to oscillate at its natural (mechanical) frequency (*fig.* 7). The amplitude of the vibration is limited to about 1 mm by shunting across the amplifier output a small incandescent lamp, mounted opposite to a photo-resistor [8]) shunting the input of the amplifier.

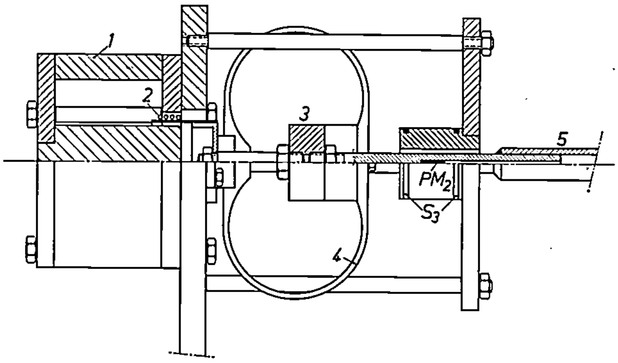


Fig. 6. The vibrator. *1* loudspeaker magnet, *2* speech coil, *3* mass of the vibrating system with which the frequency is adjusted, *4* spring (the vibrator contains two of these springs), *5* coupling piece for the bar carrying the sample and the magnet for the compensation voltage. PM_2 platinum-cobalt magnet fixed to the same bar and contained inside a third astatic pair of coils S_3, connected to the input of the amplifier which drives the current through the speech coil (see fig. 7).

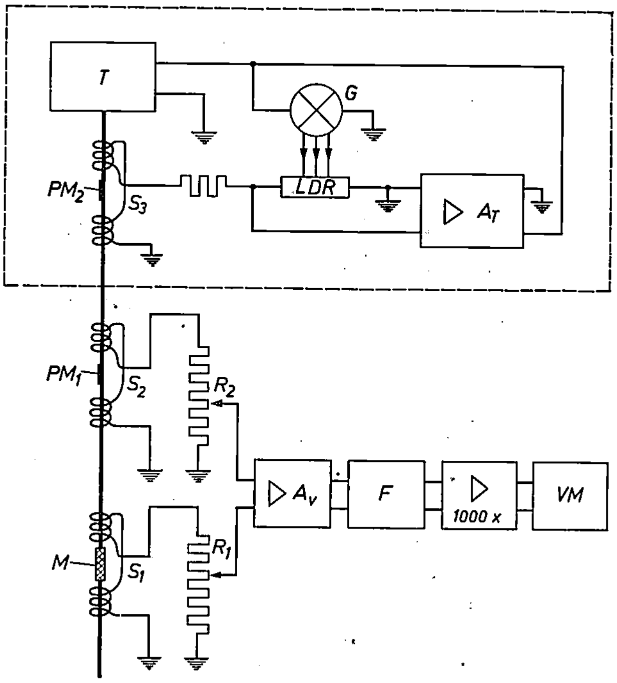


Fig. 7. The pair of coils S_3, in which the magnet PM_2 vibrates, is connected to the input of an amplifier A_T, which feeds the speech coil of the vibrator T. Due to this feedback the system vibrates at its natural (mechanical) frequency. The photo-resistor LDR, connected to the input terminals of A_T, and the electric lamp G connected to the output terminals of A_T ensure that the vibration amplitude is limited to about 1 mm. A_v is the difference amplifier, F the bandpass filter.

Examples of applications [9])

Fig. 8 gives an example of coercive-force measurements. The coercive force here is determined as a function of the thickness of a galvanically deposited layer of a Co-Ni-P alloy. The shape of the curve is

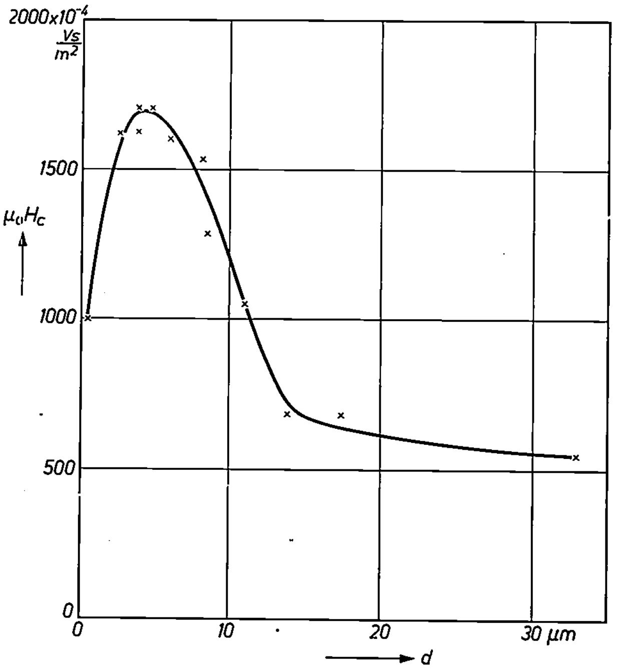


Fig. 8. Coercive force of a galvanically deposited layer of Co-Ni-P as a function of layer thickness *d*. The quantity of material varies from 0.4 to 50 mg.

[6]) See K. S. Knol, Noise; a general survey, Philips tech. Rev. 20, 50-57, 1958/59.

[7]) Another application of a spring of the same type is described in: B. Bollée and F. Krienen, The CERN 600MeV synchro-cyclotron at Geneva, III. The tuning-fork modulator, Philips tech. Rev. 22, 162-180, 1960/61, especially p. 169.

[8]) N. A. de Gier, W. van Gool and J. G. van Santen, Photo-resistors made of compressed and sintered cadmium sulphide, Philips tech. Rev. 20, 277-287, 1958/59, esp. p. 286.

[9]) These measurements were made by B. J. G. Hamer.

most remarkable, the coercive force showing a pronounced maximum at a layer thickness of about 5 μm. With increasing layer thickness the quantity of material increased from 0.4 to 50 mg. These measurements were done with the relatively insensitive arrangement, without compensating device.

As a second example *fig. 9* shows two remanence curves A and B measured on one sample weighing 1.5 mg, using the set-up with compensating device. The procedure for measuring a remanence curve is as follows. First, the sample is demagnetized. Next, it is exposed for a short time to a magnetizing field. The field is then switched off and the residual magnetization measured. This being done, the sample is again demagnetized, and the measurement is repeated for a different value of the magnetizing field, and so on. The measured values of remanence plotted against the magnetizing field producing them result in a remanence curve. As the magnetization field increases in strength the remanence curve of course approaches saturation remanence. Of the two remanence curves in fig. 9 curve A was obtained after conventional demagnetizing in an alternating field, the initially high amplitude of which is slowly reduced to zero (anhysteretic demagnetization). Curve B, with branches B_1 and B_2, relates to the same sample, but in this case the non-magnetic state was reached along the path marked by s in the inset. The marked difference between curves A and B demonstrates the fact that the state of a magnetic material is in general by no means unambiguously defined by the point corresponding to that state in the I-H plane, but depends on the manner in which this point is reached. Only in the saturation state this is not the case. In order to obtain a material in a well-defined magnetic state, it is therefore necessary to start from saturation. This evidently also applies to the non-magnetic state.

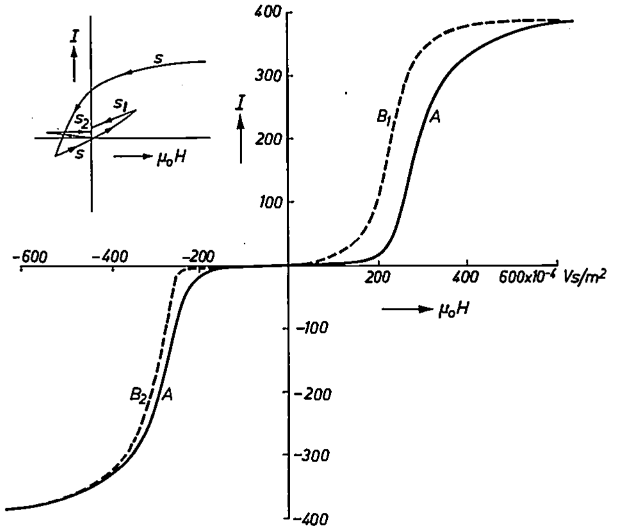


Fig. 9. Remanent magnetization produced in a piece of magnetic tape which — starting from a non-magnetized state — was magnetized by a field of the magnitude indicated on the abscissa (remanence curve). The two curves A and B (with branches B_1 and B_2) relate to different non-magnetic initial states: the symmetric curve A relates to conventional anhysteretic demagnetization, curve B to a demagnetizing process following the path indicated by s in the inset. The difference between branches B_1 and B_2 is due to the fact that the magnetizing field for these branches was respectively equal and opposite in direction to the field in which the sample was saturated (paths s_1 and s_2 in the inset). Quantity of material approx. 1.5 mg. Vertical scale not calibrated.

The third example is the remanence curve of a piece of magnetic tape, measured with the sensitive version of the apparatus, where the accurate measurements were made even in the region of very low remanence (*fig. 10*). These low remanence values are important in connection with the print-through effect mentioned at the beginning of the article.

The fourth example relates to a study of the effect of grinding upon the magnetic properties of a powder [10]), the magnetic properties being investigated as a function of the grinding time. Since small samples could be used, it was possible to take ten samples from a charge of only 10 g in the ball mill without

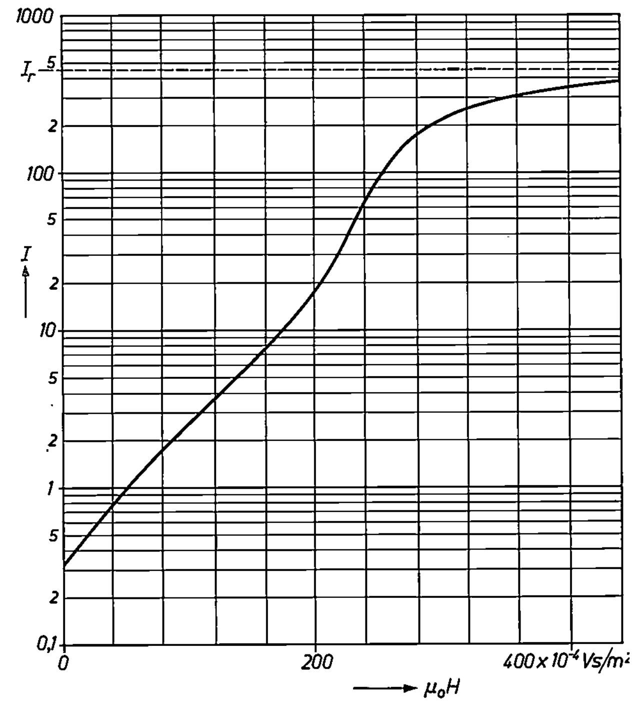


Fig. 10. Remanence curve for a piece of magnetic tape over a wide range of remanence values I (log scale for I). I_r is the saturation remanence. The processes responsible for print-through take place at values of μ_0H smaller than roughly 40×10^{-4} Vs/m². Quantity of material approx. 1.5 mg. Remanence values of one thousandth of the saturation value can still be measured on samples as small as this. Vertical scale not calibrated.

[10]) 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. B9, 288-296, 1962 (No. 4/5).

appreciably changing the size of the charge. This size has a considerable influence on the grinding conditions.

Finally, it may be mentioned that the method of measurement described here is also used in geological investigations for studying the "frozen-in" magnetism of rocks. From these measurements conclusions can be drawn about the earth's magnetic field at the time when the magnetism was frozen in (palaeomagnetism).

Summary. When a sample of a magnetic material — which possesses a magnetic moment — is caused to vibrate inside a coil, an alternating e.m.f. is induced in the coil. Since this e.m.f. can easily be amplified, a basis is thus provided for magnetic measurements which are much more sensitive than ballistic measurements. In this way magnetic measurements can be done on samples weighing only a few milligrammes. The use of an astatic pair of coils eliminates interference from extraneous alternating fields. The accuracy is increased by compensating the induced e.m.f. with an e.m.f. generated in a second astatic pair of coils, in which a small permanent magnet, rigidly fixed to the sample, vibrates. Small variations in the amplitude and the frequency of vibration then do not influence the results. With a set-up based on this principle, developed and used in Philips Research Laboratories for the investigation of magnetic-recording materials, it is possible to measure a magnetic moment of 7.5×10^{-14} Vsm, i.e. the saturation moment of 3×10^{-4} mg of iron. Methods of suppressing interference are discussed.

ULTRASONIC DELAY LINES AND THEIR APPLICATIONS TO TELEVISION

by C. F. BROCKELSBY *) and J. S. PALFREEMAN *).

534.321.9-14/-16

Ultrasonic delay lines and their applications have been studied at the Mullard Research Laboratories for the past 15 years. The subject has long outgrown the space that can normally be allowed in a journal and in fact its theory and practice, together with many design details of delay lines, have recently been dealt with in a book [2]) published by the above authors in conjunction with R. W. Gibson. In the present article the limitation to one special though noteworthy field of application has offered the opportunity to consider the main problems of ultrasonic delay lines and to discuss some of the developments by M.R.L., while at the same time an idea is given of the variety of types of delay lines that exist and of the characteristics that can be achieved.

In many applications it is necessary to delay an electrical signal for a given period and to recover the signal after this time without appreciable distortion. The need for such a delay arose in the early days of telephony, and electrical delay circuits were developed to provide the comparatively short delays needed.

The delay systems used consisted of electrical transmission lines using lumped or distributed components, e.g. a coaxial cable; the required delay was achieved at the expense of attenuation and often some distortion of the signal. Such systems are practicable for obtaining delays up to a few microseconds in length, and where the bandwidth of the signal is not large. In modern electronics, however, applications occur for delay systems having delays of a few milliseconds with bandwidths of several Mc/s. It is to satisfy these requirements that ultrasonic delay lines have been developed.

In an ultrasonic delay system the electrical signal (oscillation of electric potential) to be delayed is converted into a corresponding mechanical vibration (i.e. described by the same function of time) and launched into a suitable solid or liquid delay medium. The designation "ultrasonic" simply stems from the fact that the electrical signals will usually have frequencies higher than 20 kc/s, putting the corresponding mechanical vibrations into the ultrasonic range. The velocity of mechanical waves in liquids and solids lies in the range 1-6 km/s, a factor approximately 10^5 lower than the velocity of an electrical signal along a coaxial cable. Thus a long delay can be obtained using a comparatively short path length in the medium, for example if fused quartz is used as the delay medium; a delay of approx. 2.5 ms can be obtained in a path length of 10 metres. After the mechanical wave has travelled a distance such that the vibration has undergone the required delay, it is converted back into an electrical signal.

The first application of ultrasonic delay line techniques was in the pre-war "Scophony" television receiver where use was made of the variations in density of the water in a system resembling a delay line to provide optical readout of the video signal (see page 243). Ultrasonic delay line systems first came into prominence, however, during the Second World War, and pioneer work at the British Telecommunications Research Establishment (now known as the Royal Radar Establishment) later resulted in the development of a water delay line for use as an information storage device in Doppler radar. Meanwhile, in 1942, the first ultrasonic delay line to be used for a wartime application was produced at the Bell Telephone Laboratories, a mixture of water and ethylene glycol being used as the delay medium. It was soon discovered that mercury was a more suitable delay medium and many radar systems were developed using mercury delay lines.

Early work at the Telecommunications Research Establishment (1943) showed that vitreous fused quartz was likely to prove a valuable solid delay medium at frequencies at least up to 10 Mc/s. Much basic work on the properties of solid delay media was carried out at the Massachusetts Institute of Technology and the Bell Telephone Laboratories. Solid delay lines are extensively used nowadays and suitable designs enable long delays to be obtained in a compact space (see page 242). In 1949 Bradburd showed that magnetostrictive wire could also be conveniently used as an ultrasonic delay medium.

Although ultrasonic lines were originally developed as information storage devices in the radar field, their field of application has widened considerably in the last few years. They are now used extensively in digital and analogue computers, in communications networks, and in instrumentation based on pulse

*) Mullard Research Laboratories, Salfords (Surrey), England.

timing [1]). Several applications of ultrasonic delay techniques occur in the field of television and the discussion of ultrasonic delay lines in this article, after a general introduction to such lines, will concentrate on this group of applications (page 243 ff.). They are found to occur in the television studio, in the research laboratory, and in some possible kinds of domestic receiver.

I. THE DIFFERENT TYPES OF DELAY LINES, AND GENERAL CONSIDERATIONS [2])

From the above it follows that an ultrasonic delay line consists basically of three components. First there is a device known as a transducer which converts the electrical signal into a mechanical vibration. Second there is the delay medium through which the mechanical signal travels and undergoes the required delay. Finally there is a second transducer which converts the mechanical vibration into the required electrical signal.

Ultrasonic delay lines may be divided into three main categories, using *wires*, *liquids*, and extended *solids* respectively as the delay media. These three types of line differ markedly in the mode of vibration employed for transmission and hence in the form of transducer required. Some principles of delay lines will now be explained using the comparatively simple wire line as an example. General considerations applying to liquid and solid lines will be given afterwards, followed by some details of these lines.

Wire delay lines

In a wire delay line the electrical signal is converted into a mechanical oscillation of a magnetostrictive wire or tube. A magnetostrictive material will undergo reversible expansion or contraction in the direction of an applied magnetic field, owing to internal stresses arising as the magnetic dipoles are diverted from their preferred orientations. Conversely a mechanical deformation of such a material will give rise to a magnetic flux. Thus by passing a fluctuating current through a coil surrounding the wire a longitudinal mechanical wave can be launched down the wire, and this wave can be detected at the output end of the line by another coil surrounding the wire, the change of magnetic flux through this coil giving rise to the desired electrical signal.

Magnetostriction is a square law effect, i.e. both positive and negative currents through the input coil will cause deformation of the wire in one sense only. In order to obtain a more linear characteristic, bias must be provided by means of a current or by a magnet as shown in *fig. 1*.

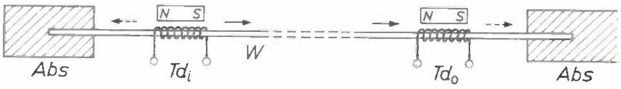


Fig. 1. Wire delay line. The electrical input signal fed to the input transducer Td_i launches an acoustic wave down the wire W; this wave induces a voltage in the coil of the output transducer Td_o. Reflections from the ends of the wire are prevented by the use of absorbent terminations *Abs*.

The delay obtained may be adjusted if necessary by sliding the coil along the wire, and this is a useful facility. Since signals may travel in either direction along the wire, absorbent terminations must be placed at each end of the wire to prevent reflections.

If a direct current is passed down the wire in addition to the alternating signal current in the coil, a helical magnetic field with alternating pitch results. Under these circumstances the signal is propagated as a *torsional* vibration. This has the advantage that a greater delay is obtained using the same length of wire. Another advantage is that the line may be coiled into a spiral for compactness of layout, as shown in *fig. 2*. This is not possible with a

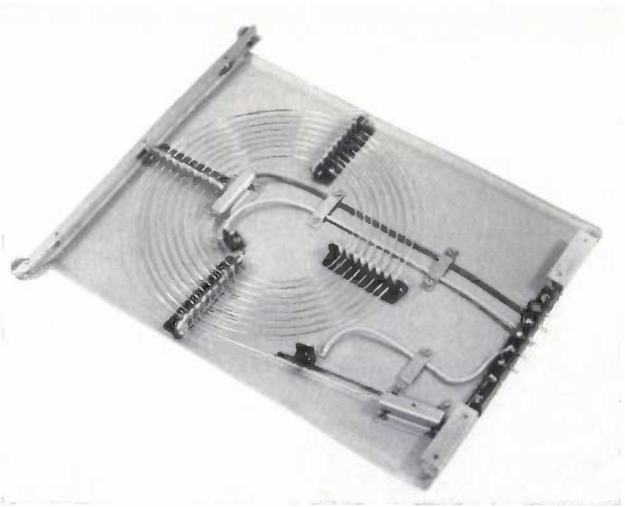


Fig. 2. Practical wire delay line (Mullard). A torsional wave is propagated down the line by passing a direct current along the wire itself in addition to the signal current in the coil. The wire itself is coiled into a spiral for convenience of layout.

[1]) See for example H. A. Dell, D. S. Hobbs and M. S. Richards, An automatic particle counter and sizer, Philips tech. Rev. **21**, 253-267, 1959/60.

[2]) A detailed account of most of the information contained in the first part of this article is given in the book: Ultrasonic delay lines, by C. F. Brockelsby, J. S. Palfreeman and R. W. Gibson, published by Iliffe, London 1963. A list of references to the literature on the history of delay lines and on many developments mentioned in this article is also given in this book.

compressional vibration since bending of the wire would cause excessive frequency dispersion (differences in velocity of travel along the wire for different frequency components of the signal) resulting in excessive distortion. This effect never becomes serious for a torsional vibration even if the wire is bent.

Efficient conversion of electrical to mechanical energy is achieved only at frequencies near the mechanical resonance of the transducer: the resonance frequency is the centre frequency of the frequency band transmitted by the delay line. In the conventional designs of figs.1 and 2 the transducer consists of the coil together with that portion of the wire in which it produces an appreciable magnetic field. This means that for high frequency applications the excited length of the wire, enclosed by the coil, must be extremely short, and the difficulty of making very short coils thus imposes a practical limit of about 1Mc/s on the centre frequency. Because of this limitation wire delay lines are little used in television applications. In addition wire lines always degrade the signal owing to the existence within the material of inhomogeneities and grain boundaries at which reflection and refraction occur, thus limiting the usable bandwidth; the degradation is enhanced if the wire is coiled. Despite these limitations, however, wire delay lines are used in the field of digital computing where frequencies are lower and a lower bandwidth is acceptable.

The output signal of a wire delay line (and for that matter, of other delay lines too) is basically the second differential of the input, since the input and output transducers respond to changes in flux only. For sinusoidal operation this is of little consequence, since the shape of a sine wave is unchanged by differentiation. For arbitrary signals the differentiation will result in a linear distortion, and a compensation (de-emphasis of higher frequencies) may be necessary in the output. In digital computing, however, linearity is not important.

The choice of the delay line material depends chiefly on the intensity of the magnetostrictive effect. The velocity of propagation and the damping per unit length (or rather per unit delay time) should also be taken into account, but the first requirement does not leave much choice: nickel, nickel alloys or a special iron-cobalt alloy ("Permendur") are used for most applications of wire delay lines.

Liquid and solid delay lines: propagation patterns, transducer behaviour

Liquid and solid delay lines have a number of important problems in common, especially those relating to the patterns of wave propagation and to the energy flow between external circuit, transducer and delay medium and vice versa. These problems will be considered first and separate sections will be devoted to the design of transducer and choice of delay medium for a) liquid and b) solid delay lines.

Both for liquid and solid lines the transducers are based on the piezoelectric effect. A piezoelectric material is one which undergoes reversible deformation on application of an electric field and which gives rise to a field when it is strained. Crystalline quartz exhibits this phenomenon, and some polarised ferroelectric ceramic materials behave in a very similar manner. A thin circular slice of crystalline quartz cut in a suitable orientation from a monocrystal is often used as transducer. The electrical input signal is applied to thin metal film electrodes coated on each face of the crystal wafer. The field produced causes a varying deformation of the crystal, thus launching a mechanical wave in the delay line medium which is in contact with one of the faces. The resultant wave travels through the delay medium along a path that will be discussed presently, ultimately arriving at the output end of the line where it produces a mechanical deformation of an identical crystal which forms the output transducer. This gives rise to an electric field in the crystal, and the signal is detected as a voltage signal on the electrodes.

Closer consideration of the process reveals that the size and orientation of the transducer used are of considerable importance in determining the properties of the line. Consider a circular disc transducer of diameter d radiating ultrasound of wavelength λ into an unbounded medium. If the transducer were a point source it would radiate spherical waves and the fraction of energy received by a receiver would depend only on the solid angle which it subtended at the source. The transducer not being a point source, it may be shown that the energy is emitted as a beam parallel to the transducer axis up to a distance approximately d^2/λ from the source. The region where the wave is defined by this beam is known as the "near field" (or Fresnel zone), and if the output transducer, also of diameter d, is placed within this region it will receive substantially all the energy radiated. Ideally, the transducers should always be chosen of such a size that the path traversed by the wave in the medium lies within the near field. Since the size of the transducers is limited, however, this is not practicable if the delay required is long.

At distances greater than d^2/λ from the source the energy is distributed in a diffraction pattern, constituting a number of lobes when plotted in a polar diagram. This region is known as the "far field" (or Fraunhofer zone). Most of the vibration energy is contained in the principal lobe of this pattern, and the output transducer must be positioned to receive this lobe. It will then receive a fraction of the transmitted energy which is directly proportional to the area of the transducer and inversely proportional to the square of its distance from the source; the fraction of energy received is thus proportional to the solid angle which the receiver subtends at the source, but it is still much larger than that which it would receive from a point source.

In order for the incident signal to produce a disturbance which is in phase at all points across the face of the output transducer, this device must be aligned approximately normal to the incident beam. This situation, however, is not strictly necessary. More generally, a polar diagram can also be drawn for the response of the output transducer, making the situation entirely symmetrical: in fact it may be demonstrated that the energy transfer through a delay system is independent of which transducer is the input and which is the output.

In order to obtain the required delay, the path traversed by the wave in the medium may need to be several metres long; to obtain a straight path of this length is often difficult and sometimes impossible. For this reason a "folded" path must be used, in which the signal is deliberately made to undergo reflections from the boundaries of the medium during its passage from input to output. At each reflection the shape of the diffraction pattern is slightly altered but in practice such changes may usually be neglected. Each reflection may be regarded as specular, and the delay system may be approximated by a model in which the two transducers are separated by a distance equal to the total path traversed by the wave in the medium.

Since d^2/λ is the governing parameter, the transition from near field to far field will be seen to occur at larger distances the higher the frequency. In other words the principal lobe of the propagation pattern will be narrowest for the highest frequencies contained in a signal and this will cause a relative loss of *low frequency* intensity received at the output when the wave has travelled a long distance in the delay medium. On the other hand a certain amount of the wave energy is dissipated in the medium owing to viscous damping and thermal effects, and in solids losses also occur due to scattering of the signal by inhomogeneities in the medium. These losses in all cases increase with frequency, in some cases even according to a square law (see page 240), entailing a relative loss of *high frequencies* at long delays. Both effects combined are responsible for a fundamental limitation of delay lines: long delays can only be achieved at the expense of useful bandwidth. However, the bandwidth is also limited by several other factors, which will be discussed below.

Since in the far field some of the energy transmitted is contained in the minor lobes of a diffraction pattern, it is possible for signals to arrive at the output having traversed paths different from that defined for the main signal. If these "secondary" signals happen to be incident at an angle corresponding to one of the minor lobes of the output transducer polar diagram they will be detected; these signals have in general undergone delays different from that experienced by the main signal and thus constitute a source of interference which may be particularly noticeable in lines where many reflecting surfaces are present, and in lines where the attenuation is appreciable: secondary signals which have travelled on a much shorter path than the main signal will then have a relative advantage. One task of the delay line designer is to minimise these spurious signals: they should be at least 50 dB weaker than the wanted signal. This may be achieved by using transducers of the largest lateral dimensions conveniently practicable, in order to confine most of the energy transmitted and received to the principal lobe of the polar diagram of each transducer; some improvement may also be obtained by putting absorbent material on walls of the delay line which cause particularly troublesome reflections, or by grinding these faces away to alter the direction of the reflection wave.

One commonly encountered unwanted signal is the "third-time-round" signal which is reflected by the output transducer, returns by the same path to the input, and is reflected back to the receiver again. This signal can be reduced to negligible amplitude at any particular frequency by tilting the output transducer through a small angle Θ. This angle is chosen so that, while the main signal still falls within the principal lobe of the output transducer polar diagram, the angle of incidence of the third-time-round signal ($3\,\Theta$) corresponds to the first minimum of this polar diagram. This will only be effective for a narrow frequency band since the angular spacing of the lobes of the diffraction pattern depends on the frequency.

Other important considerations common to both liquid and solid delay lines, as stated above, refer to the behaviour of the transducers and their coupling to the delay medium. This will be very roughly described in the following paragraphs. The subject is rather complex and a full treatment would require much more space than can be allowed here.

In the static or quasistatic case (i.e. at low frequencies) a constant fraction k^2 of the electrical energy supplied to a piezo-electric crystal is stored as elastic energy in the deformed crystal; k is called the electromechanical coupling coefficient of the transducer material. On first sight it would seem that for efficient energy conversion at any frequency, k should be as large as possible. This condition, however, is neither sufficient nor necessary. That it is not sufficient is seen by considering the mechanical response of the crystal wafer to a given electrical input. The response will be greatest (i.e. resonance will occur) when the wafer thickness, measured in the direction of propagation, is equal to half the wavelength (or an odd multiple of it) of the ultrasound in the transducer material. If the thickness is an even multiple of it, the response is zero and thus a transducer chosen to give maximum response at e.g. 10 Mc/s will give zero response (and zero energy conversion) at 20 Mc/s — whatever its value of k. On the other hand it can be seen that a high k is not strictly necessary. A transducer crystal may be represented very approximately by a capacitance in parallel with a resistance representing the mechanical strain energy flowing out from the crystal (radiation resistance). No energy is *dissipated* in the system even when k is low, since the portion of the electrical energy which is *not* converted into mechanical strain energy is stored in the transducer capacitance and released when this capacitance is discharged.

Thus, k is not a direct measure of the efficiency of a transducer, this depending in addition on the internal losses and external loading of the transducer crystal, and should be discussed in terms of *energy flow*. Nevertheless, the coupling coefficient k is of fundamental importance and controls some aspects of transducer performance.

In the first place, if k is low (i.e. $k^2 \ll 1$) then the effect of the electrical terminations on the behaviour of the transducer may be neglected. If the voltage drive is constant then the mechanical response of the transducer at the resonant frequency is determined by the ratio of the "acoustic impedance" of the medium to that of the transducer. For a progressive wave the acoustic impedance of a material is defined as the complex ratio of the stress in the medium at any point to the "particle" velocity, and is directly analogous to impedance in electrical systems. The specific acoustic impedance may be shown to be equal to the product of the density of the medium and the velocity of the acoustic wave [3]). Thus, if the coupling coefficient k is low, the acoustic response at the resonant frequency is determined by the relative densities of the transducer and delay medium and the relative velocities of the acoustic wave in the two materials.

For a low-k transducer, loaded on one face only, it may be shown that the *fractional bandwidth* $\Delta f/f$, where $\frac{1}{2}\,\Delta f$ is defined by a decrease in response of 3 dB, is very approximately equal to $2/\pi$ times the ratio of the specific acoustic impedance of the medium (Z_1) to that of the transducer (Z_0) [2]). Thus, if the medium has a much lower accoustic impedance than the transducer, the fractional bandwidth is small, and it will be higher the more the impedance of the medium approaches or surpasses that of the transducer; see *fig. 3*. The most suitable condition, however, is obtained if these impedances are equal ($Z_1/Z_0 = 1$). The transducer and delay medium are then said to be matched acoustically and in that case the amount of energy reflected from the boundary

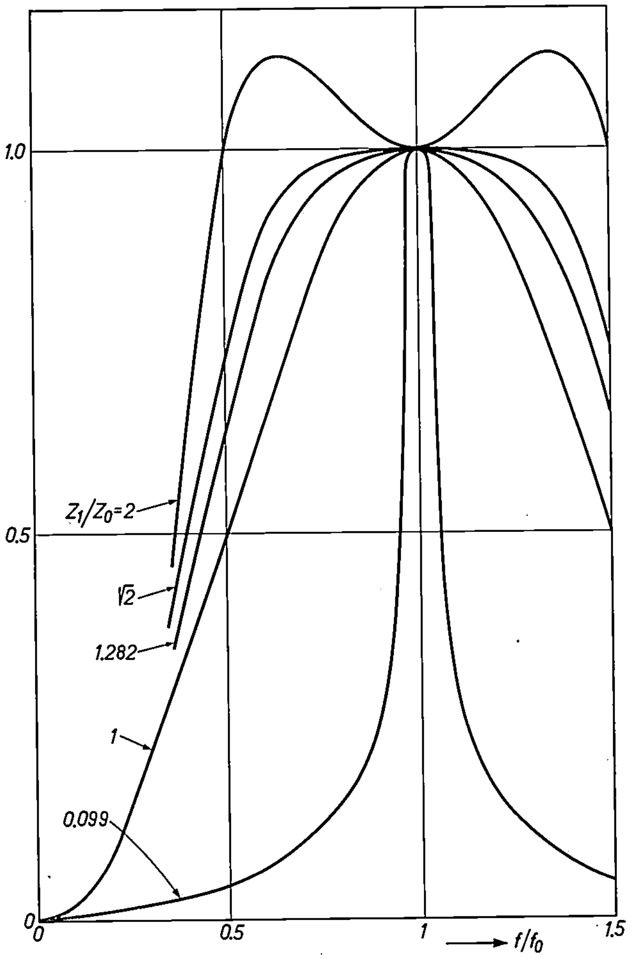


Fig. 3. Response of an X-cut quartz crystal transducer, unbacked, in a liquid medium, with various ratios Z_1/Z_0 of the specific acoustic impedance of the medium to that of the transducer. For water $Z_1/Z_0 = 0.099$; for mercury $Z_1/Z_0 = 1.282$.

[3]) T. F. Hueter and R. H. Bolt, Sonics, Wiley, New York 1955.

between them is zero. Any energy which is reflected from the face of the output transducer produces no output and is not detected at the output except as an unwanted signal (the "third-time-round" signal mentioned above). This troublesome effect, therefore, is minimised by the acoustical matching. The fractional bandwidth in this case evidently is given by $2/\pi$, which is a sizeable portion of the ideal value 2 that would be permitted by the resonance behaviour of the vibrating wafer considered above (cf. the example of maximum response at 10 Mc/s, when the response can differ from zero in the band between 0 and 20 Mc/s).

When using the concept of energy flow the mechanical wave entering the output transducer should also be considered. Part of its energy will be used for producing the output signal, but in a low-k material a relatively large part of the wave energy will not be made use of in this way and will proceed to the back of the transducer. If this face is loaded with a material of negligible acoustic impedance, such as air, then energy which is incident on this face is almost totally reflected and may again give rise to a third-time-round signal. In a practical delay line it is therefore often desirable to "back" the transducer with an absorbent medium of acoustic impedance similar to that of the transducer.

When the coupling coefficient k is high, the electromechanical response of the transducer depends not only on the ratio of the acoustic impedances but also on the electrical terminations of the system. In this case acoustical matching is not so important since an appreciable part of the mechanical power is usually converted into electrical output power and for this reason is not liable to reflection.

To conclude these general considerations the importance of a low attenuation of the signal in the medium should be stressed. As will be shown later, the electrical terminations which must be used to obtain an electrical bandwidth comparable to the acoustic bandwidth result in a voltage loss of perhaps 40 or 50 dB, if low-k transducers are used. The total voltage transfer ratio $V_{\text{in}}/V_{\text{out}}$ should not exceed 60 or 70 dB, since a convenient level for the input signal is of the order of one volt and the output signal should not be reduced to much less than 1 mV, lest the effect of electrical noise in the output circuit should become appreciable. Thus, the loss in the medium should not amount to more than 20 dB. When long wide-bandwidth delays are required, this low loss value is only possible if the attenuation per unit delay is very small indeed. This will normally confine the choice of media to an extremely limited number of materials.

Liquid delay lines: delay medium and design

Mechanical vibrations in liquids can only be supported in a compressional mode, torsional or shear oscillations being impossible. An X-cut crystalline quartz wafer undergoes a compressional change under the influence of an electrical field and may therefore be used as a transducer in a delay line employing a liquid delay medium.

The choice of the liquid delay medium is determined by its acoustical properties and by those of the transducer. The attenuation of most liquids is too high; apart from the liquefied monatomic gases, only mercury, water, and the lower alcohols have a low enough attenuation. In addition, since the value of k, the electromechanical coupling coefficient, is only 0.1 for crystalline quartz, a large fractional bandwidth can only be obtained if the specific acoustic impedance of the transducer is similar to that of the delay medium. In this case this requirement means that the transducer must be loaded with a medium of high density. It is found that mercury and crystalline quartz are approximately matched acoustically.

Since mercury is a conductor, it is only necessary for the back of the transducer to be coated with a gold film electrode. The other electrode of the transducer consists of the mercury in contact with the front face.

Ferroelectric ceramic transducers have higher ~~values of the~~ electromechanical coupling coefficient, typical values being 0.45 for compressional waves and 0.65 for shear waves. However, these materials have the disadvantage that it is very difficult to obtain the extremely thin samples necessary for use at very high frequencies. For this reason the practical limit on the use of ceramic transducers is about 20 Mc/s.

If the mercury delay medium is in contact with a steel surface, then a wave striking the liquid/solid interface at an angle of more than 10° to the normal will undergo total reflection. This property is employed in the variable mercury delay line illustrated in *fig. 4*. The signal from the input transducer travels through the mercury and falls on the steel "corner reflector" as shown. Here the signal suffers two reflections at approximately 45° and is thus reflected back to the output transducer which is mounted beside the input. The reflector is mounted on a piston, so that the path-length in the medium can be changed. Thus a liquid delay line can be made continuously variable in length, and this is of considerable importance in some television applications.

This property of total internal reflection is also employed in the fixed path length delay line shown

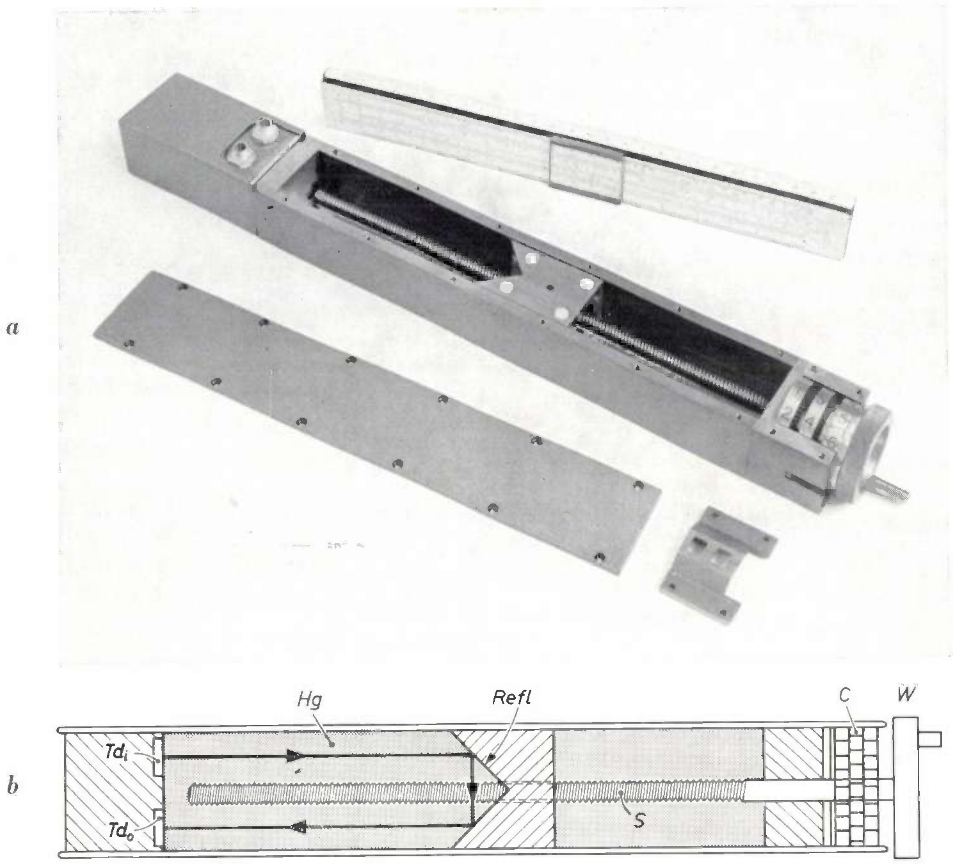


Fig. 4. *a*) Variable mercury delay line, with lid removed. *b*) Path of waves in the line. Td_i input transducer. Td_0 output transducer. The delay is varied by changing the position of the corner reflector *Refl*, mounted on a sliding piston driven by a precision lead screw *S*. This screw is cut with the exact pitch to make 1 revolution correspond to 10 µs change of delay. *W* driving wheel. *C* revolution counter. (Photograph from: C. F. Brockelsby, Ultrasonic mercury delay lines, Electronic and Radio Engr. **35**, 446-452, 1958.)

in *fig.* 5. In this system many reflections occur as the signal traverses its "billiard table" path from input to output and a comparatively long path can be obtained in a reasonably small space.

The classical theory of absorption of sound in liquids predicts an attenuation constant which is proportional to the square of the frequency. The molecules of most liquids have rotational and vibrational degrees of freedom which are neglected by classical theory, and which result in attenuation constants higher than those predicted. Mercury however is monatomic and agrees well with the theoretical predictions at frequencies below 50 Mc/s.

It is the quadratic variation of the attenuation constant which limits the frequency at which a given delay can be obtained in a practical system using mercury as the delay medium. As the transducer resonant frequency is raised, the fractional bandwidth of the line at first remains constant. As the frequency is raised further, however, the mercury attenuation at the high frequency extremity of the passband becomes increasingly significant; this both limits the bandwidth and depresses the frequency of maximum response to a value below the crystal frequency. Thus at television frequencies long delays are difficult to obtain if the required bandwidth is to be preserved. The characteristics of some typical mercury delay lines, including those illustrated in figures 4 and 5, are given in *Table I* on page 241.

Solid delay lines: delay medium and design

Solid materials are, in general, capable of supporting two types of vibration, namely compressional waves and shear waves. In a solid ultrasonic delay line shear waves are chosen for two reasons. First, shear waves travel more slowly through a solid medium than compressional waves, and thus a longer delay may be obtained in a given path length. Second, and more important, when a dilatational wave is reflected from the boundary of the delay medium then in general shear waves are also generated. These waves are propagated in a direction different from that taken by the reflected compressional wave, and travel through the medium with a different velocity; they may thus give rise to spurious signals at the output transducer. If shear

Table I. Physical and electrical characteristics of some typical liquid and solid delay lines suitable for use in television applications.

Delay medium	Transducers	Delay µs	Band-centre Mc/s	Bandwidth Mc/s	Insertion loss V_{in}/V_{out} dB	Input and output capacitance pF	Largest spurious signal, dB below wanted signal
Mercury	X-cut quartz crystal	25	15	6	65	31	46
,,	,,	30-330 *)	14.3-15.5	6.8-7.6	61-65	31	35
,,	,,	1000 **)	7.5	3	69	44	33
Fused quartz	Y-cut quartz crystal	33.3	59	28	48	80	50
,,	,,	2500	29	7	38	180	40
Lime soda glass	Lead zirconate-titanate type Piezoxide 3	64	4.4	2.5	10-20 ***)	1000-2000 ***)	40

*) Illustrated in fig. 3.
**) Illustrated in fig. 4.
***) The insertion loss of this line depends on source and termination impedances; the capacitances vary widely over the passband.

waves are used initially, however, and are polarised parallel to the reflecting surface (i.e. normal to the plane of incidence of the wave on the reflecting surface), then they are simply reflected at each impact with no such "mode conversion". The geometrical design of the delay line configuration is thereby greatly simplified.

The piezoelectric transducer used to generate the shear vibration normally consists of a Y-cut crystalline quartz wafer; alternatively a polarised ferroelectric material can be used. The transducer is coated on both faces with metal film electrodes as before and bonded on one face to the delay medium by means of a material which should be acoustically matched to both the transducer material and the delay medium. Indium is often used to make the bond because it has good adhesive properties and its acoustic impedance has a suitable value.

Consider now the choice of the medium. Single crystals are difficult to use since their elastic constants are not normally the same in all directions. Polycrystalline materials are unsuitable for wide-passband delay lines since the ultrasonic waves interact with the grain structure of the medium at high frequencies causing scattering. The material currently considered most suitable as a delay medium is vitreous silica (fused quartz), which has an extremely low attenuation. For short delays mixed oxide glass may be used, although the attenuation in this medium is much greater and it cannot be made as homogeneous as fused quartz.

If large pieces are called for, even fused quartz is difficult to make with the required homogeneity, and the cost will be very high. For producing a long delay in a single piece of quartz, the path traversed by the beam is therefore folded, as in the case of the

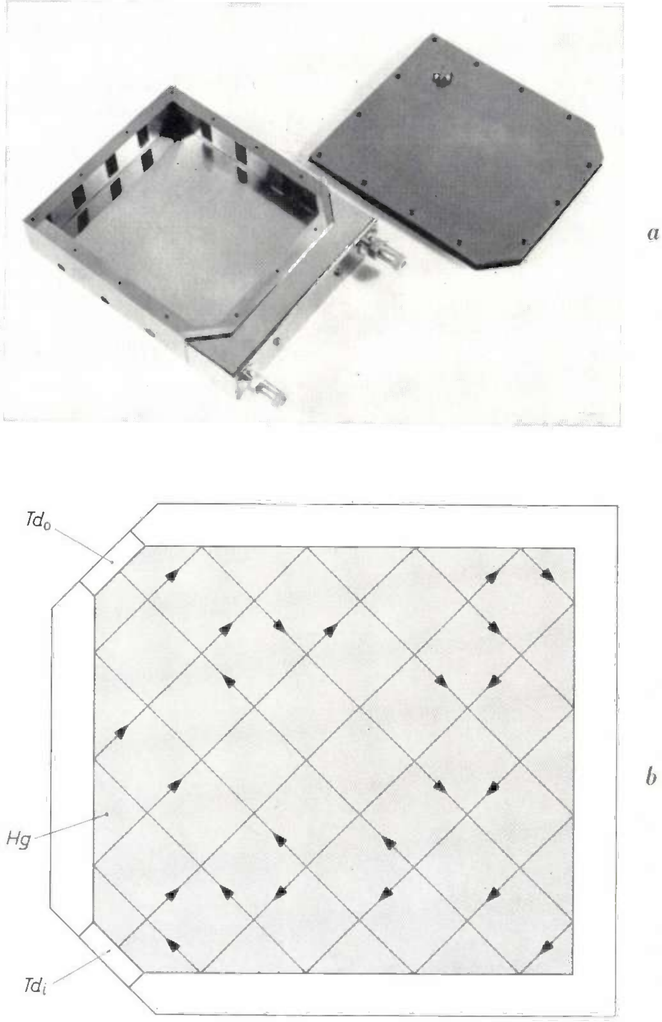


Fig. 5. *a*) "Billiard table" type of mercury delay line, with lid removed. *b*) Path of waves in the line. The signal propagated by the input transducer travels through the mercury by the billiard table path shown, and thus undergoes a long delay in a comparatively small volume of mercury. (Photograph from: C. F. Brockelsby, Ultrasonic mercury delay lines, Electronic and Radio Engr. **35**, 446-452, 1958.)

mercury delay line shown in fig. 5. In *fig. 6* some practical delay line geometries of varying complexity are illustrated. The longest delays are obtained by using the complex fifteen-side line shown in fig. 6*d*. The delay may then be doubled by using the double-decker configuration shown in the photograph of *fig. 7*. In this system the input transducer launches a signal into the lower half of the line; after traversing the path shown in fig. 6 the signal strikes a corner reflector which transfers it to the upper half of the line. The signal then follows an identical path (in reverse) in this portion of the line, and finally arrives at the output transducer where it is reconverted into an electrical signal. This system has the additional advantage that the probability of secondary signals travelling on different paths from the input to the output transducer is reduced to practically zero.

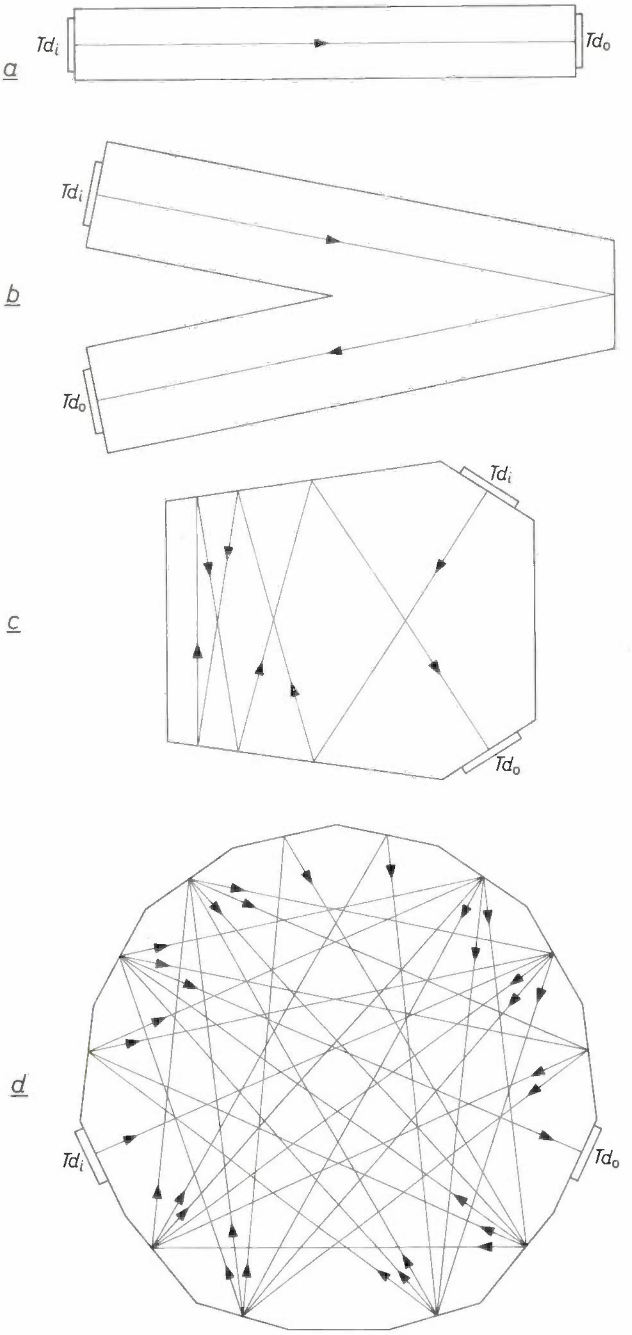


Fig. 6. Different configurations of solid delay lines, e.g. in fused quartz. Large pieces of quartz of suitable homogeneity are expensive and difficult to make. When long delays are required the configuration is made such that the signal must undergo many reflections in its passage through the line.

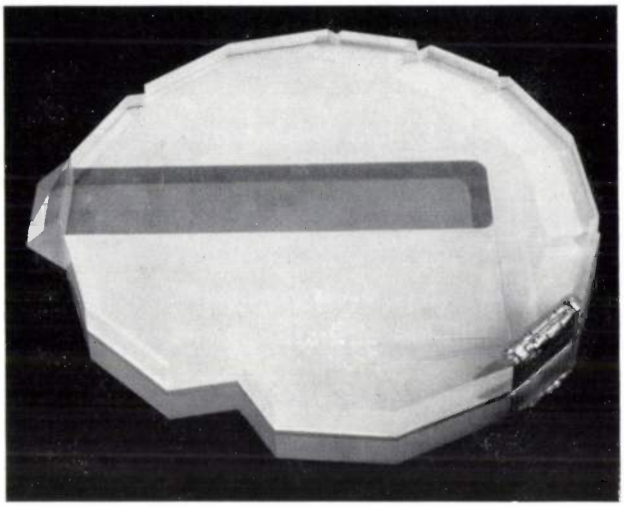


Fig. 7. "Double-decker" quartz delay line. In this configuration two of the complex fifteen-sided delay lines shown in fig. 6*d* are placed on top of each other and are connected by means of a corner reflector, in order to double the delay obtained. Edges of the line which give rise to troublesome reflections are ground away.

Fused quartz has a lower attenuation per unit delay than mercury at a given frequency and thus much longer delays may be achieved in this medium. Delays of several milliseconds may be obtained at a centre frequency of 30 Mc/s with a bandwidth of 8 Mc/s. The characteristics of some typical solid delay lines are also given in Table I.

It is worth mentioning at this point a metal strip delay line which has recently been developed in the United States [4]). A ceramic transducer is bonded to the end of the strip which is made from a metal having a low acoustic attenuation, and is used to propagate shear waves down the interior of the strip. It may be shown that if the thickness of the strip is less than half the wavelength of the highest frequency component of ultrasound present, then the signal travels down the line without dispersion. The width of the strip may be made large compared with the wavelength to provide mechanical rigidity and to enable transducers of convenient size to be bonded to the end. Ultrasonic energy striking the edges of the strip is absorbed by means of adhesive tape. The advantage of this form of delay line is that it may be bent or rolled up without loss of performance, and although its attenuation is greater than that of fused

[4]) A. H. Meitzler, IRE Trans. **UE-7**, 35, 1960.

quartz, a delay of 10 milliseconds has been achieved with a video bandwidth of 2 Mc/s.

Input and output circuits

In order to obtain optimum performance from an ultrasonic delay line it is necessary that the electrical source and load impedances should be chosen correctly. This choice is determined by the nature of the transducers used and by their coupling to the delay medium.

It has already been mentioned that a transducer for use with solid or liquid delay lines may be represented very approximately by a capacitance in parallel with a resistance (in reality this is valid only at the centre of the passband). If the transducer is a quartz crystal which has a low value of the coupling coefficient k, this effective parallel resistance is very high compared with the reactance of the transducer capacitance within the passband, and, as previously stated, the acoustic response of the line is not significantly dependent on the electrical terminations. A satisfactory *mechanical* bandwidth (fractional bandwidth $2/\pi$) is then achieved by acoustical matching of transducer and delay medium. The electrical circuit of which the trandsucer forms a part evidently should have at least the same bandwidth. To this end it is necessary to damp the circuit with a shunt resistance (alternatively a suitable four-terminal matching network may be used). A typical quartz transducer for a solid delay line may be represented electrically by a 10 kΩ resistance in parallel with a 200 pF capacitance. The damping resistor required to terminate this system might be as low as 75 Ω. (The capacitance of course must be tuned to the crystal frequency by means of a parallel inductance.) The output of the delay line system may then be regarded as a constant current generator, and the effect of the damping resistor is to produce a voltage insertion loss $V_{\text{in}}/V_{\text{out}}$ which depends largely on the ratio of termination impedance to source (transducer) impedance, and in this case is of the order of 40 dB. To reduce this voltage loss it is important to keep stray capacitance to a minimum and thus maximise the value of the damping resistance necessary to produce the required electrical bandwidth. A similar electrical bandwidth is, of course, necessary for the driving circuit.

In the case of ceramic transducers, which have *high* coupling coefficients, the effect of a parallel damping resistor is by no means so simple. The shunt radiation resistance of the transducer is now comparable with the reactance of the capacitor and both vary appreciably over the passband. The optimum driving and receiving circuits can then be predicted exactly only by laborious calculations; a good account of the effect of electrical and mechanical terminations on the loss and bandwidth of delay systems using ceramic transducers is given by Thurston [5]).

The electrical characteristics of wire delay lines, where coils are used as input and output transducers, are entirely different from those of solid and liquid lines. However, as previously mentioned, wire delay lines are seldom used in television applications; for this reason no further treatment of the relevant electrical circuitry will be given.

II. APPLICATIONS TO TELEVISION

1) *The "Scophony" receiver*

The earliest application of ultrasonic techniques in the field of television was in the pre-war "Scophony" mechanical scanning receiver, in which use was made of the *optical* properties of a liquid delay line when an ultrasonic signal was present in the line.

A television picture is always transmitted as an array of horizontal lines. During one "field" period the video signal corresponding to every alternate line in the picture is transmitted, during the next field period the transmission consists of those lines omitted during the previous scan. The two sets of lines are "interlaced" by the receiver to describe the complete picture or "frame".

In the "Scophony" receiver [6]), the video information corresponding to each line of the picture was modulated onto a 10 Mc/s carrier and applied to the input transducer of a delay line, which used water as the delay medium. Absorbent material was placed at the output end of the line, since no electrical output was required. Under the influence of the carrier alone the periodic variations in the density and hence in the refractive index of the water cause it to behave as a diffraction grating, the spacing of the lines being equal to the wavelength of the ultrasound in the

[5]) R. N. Thurston, IRE Trans. UE-7, 16, 1960.

[6]) J. H. Jeffree, Television 9, May 1936, p. 260, and British Patent No. 439 236.

water. If a parallel beam of light is allowed to pass through the medium, then the amount of light diffracted away from the zero order maximum into subsidiary maxima by any part of the grating is proportional to the amplitude of the periodic variations in the refractive index of the medium at that point.

The system used is illustrated in *fig.8*. Light from an illuminated slit is collimated to pass through the delay line as a parallel beam. The 10 Mc/s carrier, modulated with the picture information, is applied to the delay line, which is of such a length as to accomodate a wave train corresponding to one complete line of the picture. The signal corresponding to each picture element causes an amount of light proportional to the amplitude of this signal to be diffracted away from the zero order maximum for that section of the delay line. The undeviated zero order beam which is thus modulated in intensity across its width, is focussed onto the screen of the receiver by a lens and rotating mirror system as shown. A given progressive picture point on the delay line is arrested by the rotating mirror system to produce a stationary picture point on the screen; this point will be present for the complete line duration, and the same applies to other points of the picture line. A further rotating mirror system, not shown in fig.8, is used to provide the "frame scan", i.e. to combine the successive lines of the picture to produce the complete display.

The "Scophony" receiver suffered from all the usual problems of mechanical scanning systems; the chief of these was the difficulty in synchronising the extremely high speed motor (30 000 r.p.m.) which drives the rotating mirrors providing the stationary picture. For this reason the "Scophony" receiver was soon superseded by the electronic scanning receivers in use today. However, the principle of using ultrasonic techniques in order to provide optical readout of an electrical signal is still of interest as a means of high speed data processing, for certain special applications. Bandwidths of 40 Mc/s may be achieved, together with the facility of simultaneous display or inspection of bits of information fed in sequentially over a period of many microseconds. An application of such a technique will be described later in the section on systems conversion.

2) *Inertia compensation for a vidicon tube*

Ultrasonic delay lines have been used by Hughes [7]) (1961) to correct for effects due to moving objects in the well known vidicon television camera tube. In this tube the light from the object in the field of view of the camera falls on a photoconductive layer of antimony trisulphide producing a pattern of conductivity which at any point corresponds to the brightness of the picture observed. In order to convert this pattern into an electrical signal the photoconductive layer is scanned by a beam of electrons, in two fields with interlaced lines as previously described.

Difficulty arises in the vidicon tube when the image on the photoconductive layer is not erased completely by each scan. Owing to the finite decay time of the layer, the signal obtained on scanning consists of the new information plus a certain percentage of the previous information. This effect, which is only noticeable under conditions of poor illumination, results in the "smearing" of the image of a moving object.

Hughes has used the following method in order to correct for this effect. Part of the signal from each line in a field was delayed by one field period (16.651 ms on the US system) and subtracted from the signal due to the corresponding line on the next field, which is the one *adjacent* to it in the picture. The fraction of the signal used for the correction was 10-40%. Since this signal is merely a correction to the main signal, an appreciable improvement may be obtained even when using a low bandwidth delay line, and the line used was of the torsional magnetostrictive type with a bandwidth of 600 kc/s on a 800 kc/s carrier.

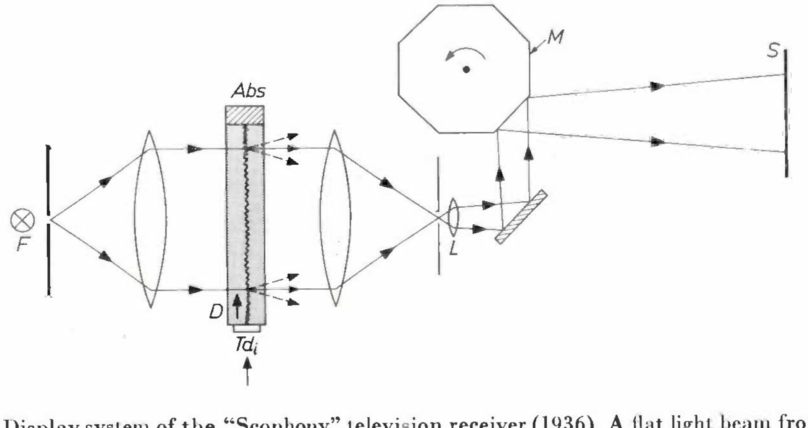


Fig. 8. Display system of the "Scophony" television receiver (1936). A flat light beam from a source *F* passes through the delay line *D* and is locally affected in intensity by the wave originating from the video signal. The transmitted beam is focussed by lens *L* on the screen *S*. Any progressive picture point in the line is arrested by the rotating mirror *M* to produce a stationary picture point on the screen.

[7]) W. L. Hughes, IRE Trans. **PGBC-7**, No. 3, p. 8, 1961.

The correction was inadequate in that the signal should ideally have been delayed by *two* field periods and applied to the *corresponding* line on the next frame. This would have been more difficult to achieve owing to the very long delay required. However, although interlace was ignored in this way, and the bandwidth was low, a useful reduction of smear was achieved with this experimental system.

An alternative form of television camera tube, the image orthicon, employs a photoemissive cathode and does not exhibit "smear". For this reason the vidicon is little used in television broadcasting except under conditions of high illumination or in applications where its comparatively small size is of advantage. In addition, in a more recently developed tube of the vidicon type, the "Plumbicon" [8]), lead oxide is used as the photoconductive material; in this case the "smearing" effect is hardly noticeable. Thus it appears that smear correction will seldom be necessary in future television applications.

3) *Vertical aperture correction*

Ultrasonic delay line techniques may also be used to correct for another fault of a television camera tube, viz vertical aperture distortion.

The electron beam in the pick-up tube as well as in the picture tube must scan the discrete lines from which the television picture is formed (say 405 lines per picture). The effective diameter of the scanning spot formed by the beam of electrons in the pick-up tube, however, cannot be made as small as would be required for the ideal line width. Thus the signal generated when scanning one line is diluted with information from the adjacent lines. This effect is known as vertical aperture distortion and will obviously reduce the definition of the picture produced in the receiver. A *horizontal* aperture distortion also exists: whereas the television transmitter is designed to have a bandwidth such that frequencies corresponding to $\frac{5}{4}\times 405$ changes from black to white along one picture line can be coped with, the diameter of the scanning spot overlapping several of these changes will prevent the resolution of all these picture elements in the receiver. This horizontal distortion, which strongly resembles the effect of a finite slit width in scanning sound film or magnetic tape recording, can be approximately corrected by use of electronic circuits. Correction of the *vertical* aperture distortion, however, has to deal with the admixture of information pertaining to picture elements scanned one complete picture line period before or afterwards and is possible only by using delay line techniques.

An extra difficulty arises from the way in which the picture is scanned. Since two fields are interlaced in order to describe one complete picture, or frame, it follows that in order to correct one line of a frame for the admixed information of adjacent lines, it would be necessary to store the information not for one line (96 μs on the 405 line 50 c/s system used in Great Britain) but for one field period (20 ms).

For simplicity it is proposed first to ignore interlace and to correct each line with the information stored in the adjacent lines of the same field. To perform this correction at all adequately, each line must be corrected with the signal due to the subsequent line in the field as well as the preceding one. The system used is illustrated in *fig. 9*. The main signal is delayed by one line period, i.e. 96 μs; from this is subtracted a part of the signal from the *preceding* line of that field, which has undergone a delay of twice the line period, and part of the signal from the *subsequent* line, which is made available in advance by bypassing the main delay line. Dependent on the vertical aperture distortion in the signal, the fraction of the signal from the adjacent lines needed for the correction is 10-20%.

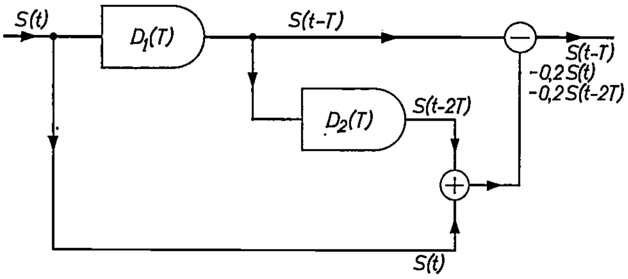


Fig. 9. Vertical aperture correction using two "one-line" delay lines D_1 and D_2. The input signal $S(t)$ delayed by one line period T is used as the main signal, $S(x—T)$. This signal is then combined with 10-20% of the signal from the previous line, $S(t—2T)$, which has been delayed by $2T$, and the same proportion of the signal from the subsequent line, $S(t)$, which has undergone no delay.

Since this correction involves delaying the main signal, the delay line used must be one with a bandwidth sufficient to accomodate the video information. For the 405-line, 50 c/s system in which double sideband modulation is used, the delay line must have a bandwidth of 6 Mc/s. In order to prevent appreciable distortion, the frequency characteristic of the line should be flat to within 1 dB over the whole band. In addition, the delay line must be variable in length if the line system is synchronised to the mains, since the frequency of the mains may drift appreciably. A variable mercury delay line is suitable for this application and it is possible to use

[8]) E. F. de Haan, A. van der Drift and P. P. M. Schampers, The "Plumbicon", a new television camera tube, Philips tech. Rev. **25**, 133-151, 1963/64 (No. 6/7).

a servo mechanism to adjust the length of the line and to provide automatic control of the vertical aperture correction. If, however, the line system is controlled by a crystal-locked oscillator then no adjustment should be necessary and a solid delay line may be used.

Using the system described, a very marked improvement in picture quality may be obtained. Gibson and Schroeder [9]) (1960) and Howorth [10]) (1962) have described systems of this kind. However, in order to perform this correction in the most effective manner interlace should not be ignored, and each line should be corrected with information from the adjacent lines in the *frame* instead of the field. The system used should then be essentially the same as that shown in fig. 9 with the exception that the delays involved are now equal to one field period (20 ms).

Since this long delay is required for the main signal, which however should not suffer my loss of definition, a rather difficult problem has to be solved. As previously mentioned, the most suitable delay medium for providing long delays at large bandwidth is fused quartz. However, using existing techniques, it is not possible to construct a single delay line of 20 ms delay even using fused quartz, since the path length required is 80 metres. A piece of quartz of the required size and homogeneity cannot at the moment be obtained; in any case the attenuation in this path length would be too great. It is therefore necessary to use a number of shorter delay lines to achieve this delay, with repeating amplifiers to boost the signal after passage through each section. A 20 ms delay which has been made by Mullard consists of eight lines each of 2.5 ms delay. This system has a bandwidth of 8 Mc/s with a centre frequency of 30 Mc/s.

As previously stated, if the line system is synchronised to the mains frequency then the delay used must be variable. In order to achieve this, the cascade of quartz delay lines is made with a delay slightly less than 20 ms. The remainder of the delay is provided by means of a short mercury delay line which may be fitted with a servomechanism to keep the delay matched to the mains frequency, as described above. This servomechanism might also be used to compensate for variations in temperature, although it is also possible to stabilise the temperature at a constant value by use of a thermostat. This should preferably be set to a value between 50 °C and 70 °C, since the attenuation in fused quartz is lower in this temperature region than at room temperature.

It would, incidentally, be possible to feed the mechanical signal from the final section of the quartz delay line directly into the mercury line without the use of an intermediate repeater amplifier and the two associated transducers. This might be done as shown in *fig. 10*. The shear wave in the quartz undergoes a reflection at a plane such that the plane of incidence is not normal but parallel to the polarisation of the shear wave, and at such an angle of incidence that the wave is converted entirely into a compression wave. The resultant signal may now be used to propagate a compressional wave in the mercury which is held in a steel container bonded to the end of the quartz.

Since the delay line required for this ideal form of vertical aperture correction is so complex, it would be convenient if a single delay line could be used to provide both the 20 millisecond delays needed. In fact, this may be done by using two separate carrier frequencies in the 8 Mc/s pass-band of the delay line. A possible system is illustrated in *fig. 11*. If the centre frequency of the delay line is 30 Mc/s, the main video signal is passed through the delay line as a single-sideband modulation (width 4 Mc/s) on a 26 Mc/s carrier. This is then combined with a fraction of the signal from the subsequent line in the next field which has undergone zero delay, and with that from the previous line of the preceding field which, having traversed the delay line once on a 26 Mc/s carrier is then passed through the delay line once more on a 34 Mc/s carrier. Filters are used to prevent mixing of the two signals.

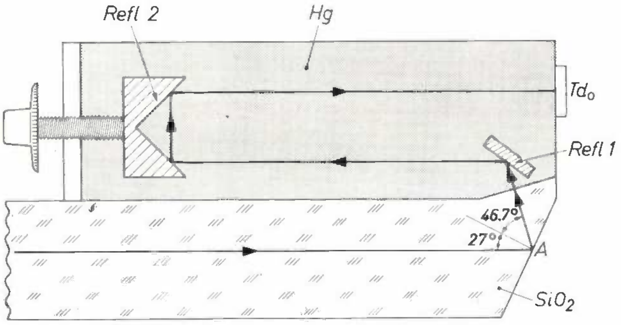


Fig. 10. Combination of a fused-quartz delay line with a variable mercury delay line. At the point *A* the plane and angle of incidence of the vertically polarised shear wave in the fused quartz delay line (S_iO_2) are such that the signal is converted entirely into a compression wave. This wave then passes directly into the mercury delay line, with reflector *Refl 1* and adjustable corner reflector *Refl 2*, by means of which the delay of the composite line may be adjusted.

Owing to attenuation in the delay lines, and noise in the amplifiers used, the improved definition

[9]) W. G. Gibson and A. C. Schroeder, J. Soc. Motion Picture and Television Engineers **69**, 395, 1960.

[10]) D. Howorth, B.B.C. Research Department Report **T-085**, 1962.

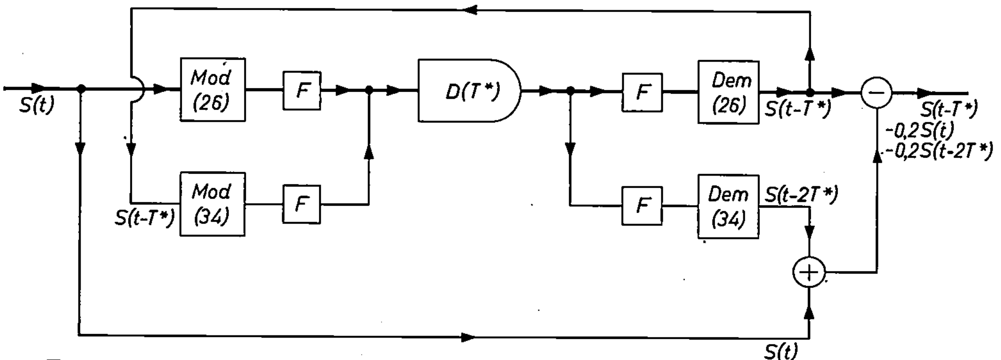


Fig. 11. Vertical aperture correction using signals delayed by one *field* period. This system is similar to that shown in fig.9, except that a single delay line D is used to provide *both* the one-field-period delays required. This is done by passing the signal with a bandwidth of 4Mc/s once through the line as single-side-band modulation on a 26Mc/s carrier, by means of a modulator *Mod* (*26*) and a demodulator *Dem* (*26*), with suitable filters F, and then again on a 34 Mc/s carrier; the centre frequency of the 8 Mc/s passband of the delay line is 30 Mc/s.

resulting from both these methods of vertical aperture correction is paid for by some small deterioration in the signal-to-noise ratio. However, since the signal-to-noise ratio of the delay system is normally of the order of 50 dB, the overall result is a considerable improvement in picture quality.

4) *The "Secam" and the "PAL" colour systems*

In the previous applications described in this article an ultrasonic delay line was employed to reduce the effect of one line of a television picture on the adjacent lines, and thus to improve definition. In the "Secam" system [11]) of colour television, which was proposed a few years ago, a similar process is used for a purpose opposite to that of the previous case: a delay line is used to store one line of the picture so that it may *add* information to the subsequent line.

Every colour television system should preferably be "compatible" i.e. it must be possible to receive the total signal as a black-and-white picture on a normal black-and-white receiver. Three signals are transmitted; one is a luminance signal which defines the brightness or luminance of each element of the picture just as in a black-and-white system. The other two signals contain the colour information and define the hue and saturation of the colour at each point of the picture. These are transmitted as the red and blue colour difference signals, one describing the amount by which the red signal differs from the luminance signal and the other similarly describing the blue. The green signal is that remaining when the red and blue signals have been subtracted from the luminance signal. In the "Secam" system these colour signals are not transmitted simultaneously, as in the N.T.S.C. system, but on alternate lines of each field. The receiver displays each line of the picture by using directly the information being received at that moment (say the blue) whereas for the other colour difference (red) it employs the information which was transmitted on the previous line of that field. This sharing of half the colour information between adjacent lines of each field entails some loss of precision in the colour detail since the mixing process has introduced errors similar to those which the vertical aperture correction previously described sought to remove. However, these errors to a first approximation are in the colour of the picture rather than in its luminance, and the eye is less sensitive to loss of colour definition.

In order to effect this combination of colour information from adjacent lines of a field, a "Secam" receiver requires a device which is capable of storing the colour signal for one line period (64 µs on the 625 line 50 c/s system which will be used for colour television in Europe). Since the transmitter is to be crystal controlled rather than locked to the mains frequency, this delay need not be variable and thus it is not necessary to use a mercury delay line. In addition, the delay is comparatively short, the centre frequency is low (4.43 Mc/s), the required bandwidth is only 2 Mc/s and some attenuation can be tolerated at the edges of the passband. Thus a suitable delay line can be made using *glass* as the delay medium. Glass has the disadvantage that the velocity of shear waves in the delay medium varies slightly from sample to sample, and that each line must be individually checked and adjusted to give the required delay, which must be correct to within 0.05 µs. Such adjustment involves grinding away one of the reflecting edges of the line, and this is normally difficult since errors may be introduced in the direction of the reflected beam. However, by use of the delay path geometry shown in *fig. 12* this

[11]) P. Cassagne and M. Sauvanet, Ann. Radioélectricité **16**, 109, 1961.

adjustment is simplified. If the longest edge of the delay line body, on which two reflections occur, is ground away, then any variation in the direction of the wave in the plane of the paper, introduced at the first reflection by errors in the angle of grinding, is compensated at the second reflection. Such compensation is important if the received signal is to fall within the central lobe of the polar diagram of the output transducer.

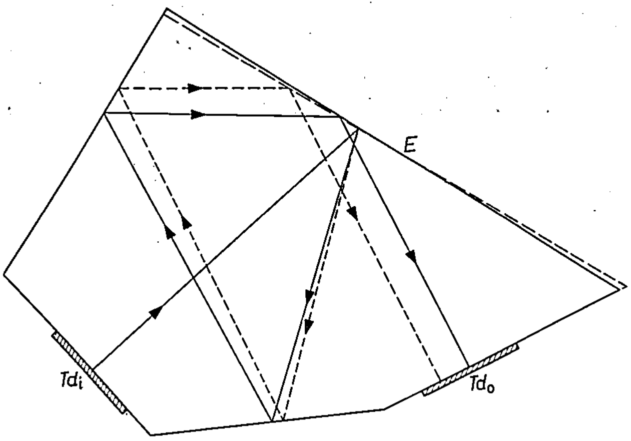


Fig. 12. Path geometry in a glass delay line for use in "Secam" and "PAL" colour television systems. This diagram shows how deviation of the ray, caused by an error in the angle at which the first reflecting surface E is ground away, is compensated by the second reflection at this surface. The error in grinding has been considerably exaggerated for convenience of drawing. (J. S. Palfreeman and R. W. Gibson, British patent applied for.)

It is necessary that the delay obtained should not vary significantly over the temperature range in which the receiver normally operates. In some recent experimental delay lines made from Philips type 18 glass, the temperature coefficient of the delay is about 0.001 μs/°C, so the temperature can rise or fall by about 50 °C before the picture is degraded.

It is important in this application that the amplitude of signals arriving at the output transducer of the delay line at the wrong time should be kept to a minimum, since these would produce areas of colour in the wrong place. In the design shown in fig. 12 these spurious signals are kept at a level of —35 to —40 dB, which could be extended to —50 dB if necessary.

The insertion loss of these delay lines is between 20 and 25 dB if symmetrical terminations are used (i.e. if the internal resistance of the signal source is equal to the load resistance) and if the resistance value is optimized for maximum flat bandpass characteristics. The insertion loss may be up to 12 dB better for asymmetrical terminations. This loss has to be made up by an amplifier coupled to the delay line. The component values and circuit parameters are not critical.

In the N.T.S.C. system the colour difference signals are transmitted simultaneously but separated in phase. In another system for colour television, the "PAL" system [12]), the phase of one of these signals is *inverted* on alternate lines. It has recently been proposed for the receiving equipment which decodes and separates the two signals to store the colour information in a particular picture line for one line period and to add or subtract it to that received for the next line. However, since it is important that the two signals are mixed in the correct phase, the delay line used must be accurate to within a fraction of the subcarrier period, i.e. a few nanoseconds. In addition, the delay must maintain this value over the frequency band of the colour signal and over the temperature range of the receiver.

5) *Study of multipath effects*

The picture displayed by a television receiver frequently suffers from what are known as multipath effects. These effects are due to the arrival at the receiving antenna of two (or more) signals, one direct from the transmitter and the other by a different path, perhaps by reflection from a passing aircraft. If the reflected signal is received at appreciable strength in comparison to the direct signal a "ghost" picture is observed as shown in *fig. 13*. As the aircraft moves, the path difference, and hence the time delay between the two signals, varies. Thus, the phase and position of the reflected signal is constantly changing and this may give rise to a beat effect or "flutter" on the screen of the receiver.

When a new receiver is being designed, or when a new system (e.g. colour television) is being developed in the laboratory it is important to study the effects of reflected signals on the received picture. Since it is inconvenient to use real aircraft to produce the reflected signal, and in any case the exact reflection is not readily repeatable, the reflected signal must be produced by artificial means.

A variable mercury delay line can be used for this purpose [13]). In the system illustrated in *fig. 14*, the test signal, from either a live television transmitter or from a local test pattern generator, was modulated onto a carrier of a suitable frequency, viz 15 Mc/s, and applied to two transmission paths. One of these consisted of a short fixed delay of 30 μs and the other included a variable delay line, the delay of which could be varied continuously from 25 μs to 300 μs. In series with the variable delay was an attenuator.

[12]) W. Bruch, Farbfernsehsysteme — Überblick über das NTSC-, SECAM- und PAL-System, Telefunken-Z. **36**, 70-88, 1963 (No. 1/2).

[13]) C. F. Brockelsby, Electronic Design, Oct. 1957, p. 36.

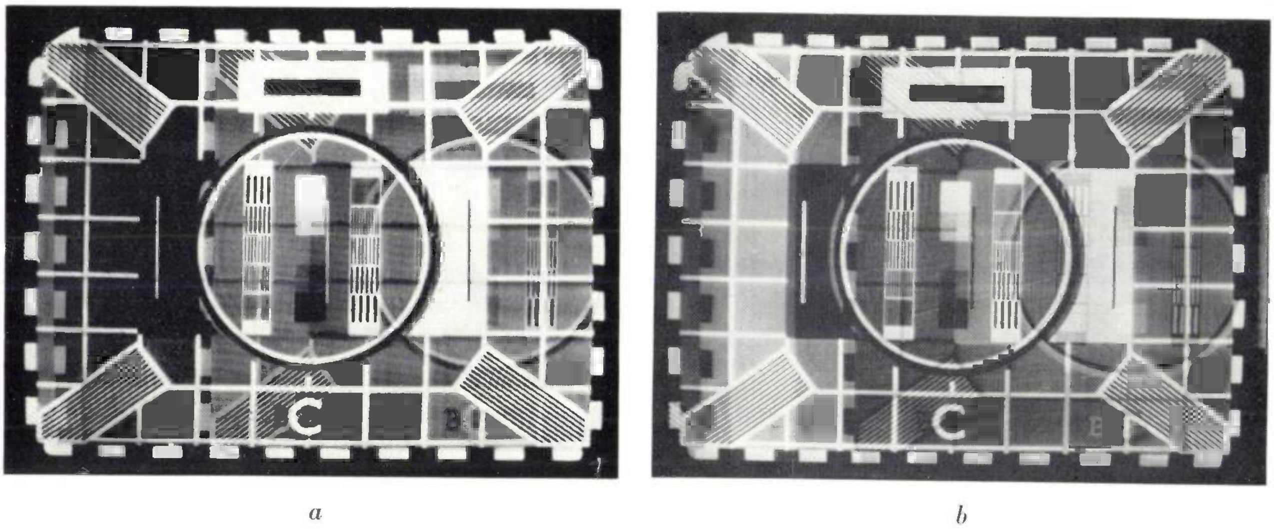


Fig. 13. Ghost images produced by signal reflections, e.g. at a passing aircraft.
a) Reflected signal in phase with main signal.
b) Reflected signal out of phase with main signal.

After passing through the delay lines the signals from the two paths were combined and their frequency changed back to the original carrier frequency, which was applied to the receiver under test. By altering the variable delay the effective path difference between the two signals could be changed, and by adjusting the attenuator the strength of the reflected signal could be controlled.

The variable delay line used was a mercury delay line similar to the line illustrated in fig. 3. In order to simulate the effect of a moving aircraft the lead screw driving the sliding piston was driven from a Velodyne motor. The input to the Velodyne amplifier was a voltage to which the motor speed was proportional. The delay could then conveniently be programmed by supplying the Velodyne control voltage from a potentiometer.

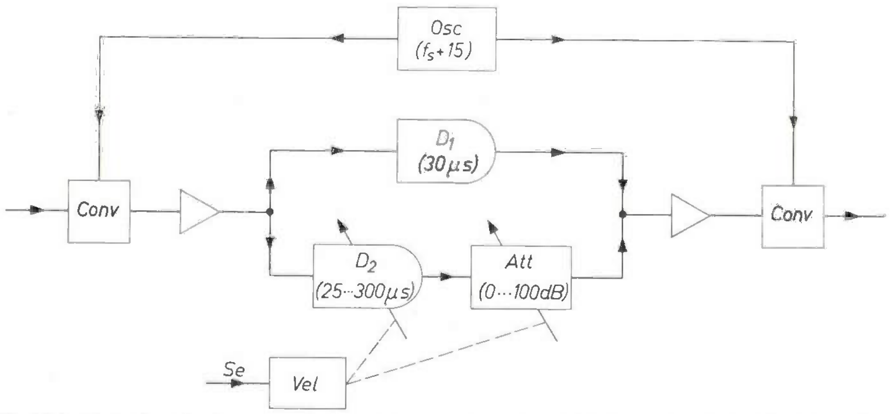


Fig. 14. Aircraft reflection simulator using a variable liquid delay line. The signal passing through the fixed 30 µs delay line D_1 represents that arriving at the receiving antenna directly from the transmitter. The signal passing through the variable delay line D_2 and attenuator *Att* represents that reflected from a passing aircraft; the amplitude and phase of this signal may be varied. Both the delay line and the attenuator are driven by a Velodyne motor *Vel*, whose control voltage can conveniently be programmed (servo system *Se*).

At a carrier frequency of 15 Mc/s the delay line system had a bandwidth of 8 Mc/s for a 300 µs delay. Working at a received signal level of 3mV a signal-to-noise ratio of 50 dB was achieved. The insertion loss of the line was around 50 dB, so the input transducer required a drive of about one volt which could readily be provided with adequate linearity by a small receiving tube.

Gander and Mothersole [14]) (1958) have given a detailed description of such a system. They found this application of delay line techniques particularly useful when appraising a new television system such as the N.T.S.C. colour system.

6) *Systems conversion*

It is frequently required to convert a television video signal from one scanning system to another, e.g. from 625 to 405 lines per picture. Such conversion is a regular requirement when programmes are relayed between countries using differing line systems. The conventional technique is to display the input signal on a cathode ray tube of relatively long persistence, and to examine this with a television camera tube working on the line system

[14]) M. C. Gander and P. L. Mothersole, Electronic Engng. **30**, 408, 1958.

required [15]. This process may introduce noise in the conversion of video signals into a display and vice versa, and may produce loss of definition and smearing of moving images due to the persistance of the phosphor. Moreover, moiré-effects are virtually unavoidable with this technique. Other possibilities for a systems conversion are therefore of interest.

In most countries the field period of the television system is equal to the period of the mains. It therefore follows that in general systems conversion involves a change in the field period as well as in the number of lines. However, in many cases it appears that the field periods of the input and output signals may be equal and that conversion may involve only a change in the number of lines per field. For example, in Great Britain it is proposed to use a 625-line camera, but to transmit the picture simultaneously on both 625 lines and 405 lines using the same field period. Conversion without change of the field period is known as synchronous conversion and Lord and Rout [16] (1962) have recently described a standards converter which uses ultrasonic delay line techniques in order to effect such a conversion.

In order to effect the change from one line system to another, two processes must be performed. First, by selective rejection or repetition of information the number of lines in the field must be changed to suit the new standard. Secondly, the information in each line must be redistributed to the time scale of the new line system. This would involve, for instance, stretching a 64 μs "625" line into a 96 μs "405" line.

As a simple example the reduction of the number of lines by a factor $\frac{3}{4}$ may be considered as shown in *fig. 15*. This conversion is done simply by discarding every fourth line of the field. However, when the lines are now displayed on the new system they will be distributed as shown by the dotted lines. An originally straight inclined row of picture elements *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H* will thus appear as a discontinuous line *a*, *b*, *c*, *e*, *f*, *g*.

Moreover the process of discarding the fourth line, which contains picture element *D*, must not result in the loss of the information contained in this line. This is avoided by correcting each of the remaining lines with information from the adjacent lines, suitably weighted in magnitude as shown in *fig. 16*. Output line 1 is reproduced as before; output line 2 however is displayed between input lines 2 and 3 and thus contains information from line 2 (e.g. *B* displayed at point b_1) and information from line 3 (point c_1). Output line 3 similarly contains both information from input line 3 (c_2) and from input line 4 (d_1). In this way interpolation is achieved which results in a reduction of the original distortion of the inclined row of points at the expense of some loss of horizontal definition.

In the process described by Lord and Rout it is proposed that a one-line delay should be used to permit the required interpolation to be performed. This system is illustrated in *fig. 17*. The input signal is split into two parts. One part is fed into a one-line delay; the other is multiplied by an "interpolation function" *F* and is fed into one input of an adder.

[15] J. Haantjes and Th. G. Schut, A line converter for the international exchange of television programmes, Philips tech. Rev. 15, 297-306, 1953/54.

[16] A. V. Lord and E. R. Rout, I.E.E./I.R.E. Int. Telev. Conf., London, June 1962.

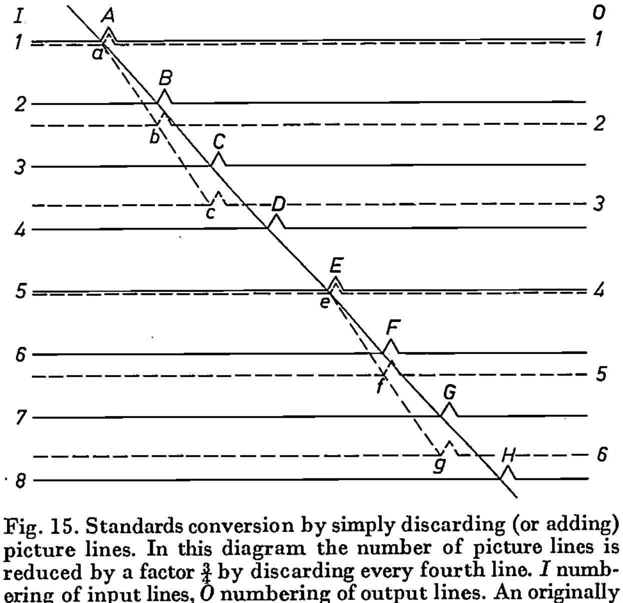


Fig. 15. Standards conversion by simply discarding (or adding) picture lines. In this diagram the number of picture lines is reduced by a factor $\frac{3}{4}$ by discarding every fourth line. *I* numbering of input lines, *O* numbering of output lines. An originally straight diagonal row of picture points *ABCDEFGH* is thus transformed into the zigzag row *abcefg*.

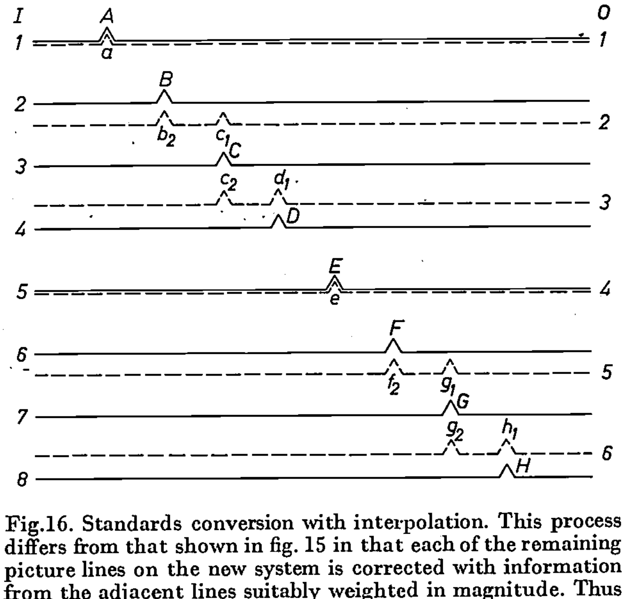


Fig. 16. Standards conversion with interpolation. This process differs from that shown in fig. 15 in that each of the remaining picture lines on the new system is corrected with information from the adjacent lines suitably weighted in magnitude. Thus no picture information is lost and the resultant distortion is reduced.

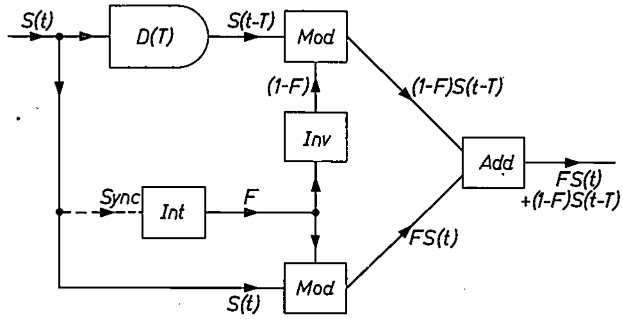


Fig. 17. The interpolation described in fig.16 may be performed using a delay line as shown. The delay is equal to one input line period T. The value of the interpolation function F produced in the generator *Int* (which is synchronised with the line frequency of the input video signal) at any time determines the proportion of information from each input line present in the resultant output line. *Mod* modulators, *Inv* inverter, *Add* adder.

In this unit the signal is combined with that corresponding to the previous line of the field which, after passing through the delay line, is multiplied with the complement of the interpolation function ($1-F$). The output of the adder is the interpolated signal on the new line system which must now be converted to the new line period.

The value of the periodic interpolation function F at any instant determines the proportions in which the information from each of the two adjacent input field lines is combined to produce the resultant line. For example, when F is 0.3 then 30% of the resultant signal consists of information from the input line $S(t)$ and 70% of information from the previous line $S(t-T)$. In the simple example under consideration F must have a period equal to four input lines since every fourth line is discarded. In practice the interpolation function will be of sawtooth form.

The one-line delay proposed may conveniently consist of a variable mercury delay line of the type previously described if it is required to adjust the delay to compensate for variation of the mains frequency, or a fused-quartz line if the system is synchronised to a crystal controlled frequency.

It is possible, though by no means certain in the case of moving images, that a more satisfactory result would be obtained if the interpolation were performed using, not adjacent lines in the field, but adjacent lines in the picture. For this case a one field period delay would be required and the 20ms quartz delay line system previously described could be used in this application.

The remaining process in the conversion is the expansion of each of the selected lines to the line period of the new system. This may be achieved by feeding the information contained in the picture line into a storage system, such as an analogue computer store, and reading it out again at a rate determined by the output line period. The computer store might employ magnetic tape as the storage medium. Alternatively, this store might consist simply of a series of capacitors, one for each element of the picture line[17]. For reading-in and reading-out, two electronic rotary switches are provided, each with a number of contacts equal to the number of picture elements (and capacitors). During one input line period the input switch arm rotates steadily to charge each capacitor with a charge proportional to the signal amplitude of the appropriate picture element. The output switch arm "samples" the charge on all the capacitors in a total time equal to the output line period.

This line stretching may, however, also be performed by a combination of ultrasonic delay line techniques and optical picture handling such as the "Scophony" receiving system previously described. In that system a rotating mirror was provided in order to prevent a continuous progression of picture points across the screen. By adjusting the speed of rotation these points may be made to pass a photosensitive device, e.g. a photomultiplier, at any desired rate, and thus the input signal may be reconverted into an electrical signal, the duration of the output signal corresponding to one input line being adjustable to any required value.

In practice, the system used would be somewhat different from the "Scophony" receiver: a more linear conversion of electrical signal into optical modulation may be obtained if use is made of the photoelastic properties of fused quartz to effect the required conversion [18]). A fused quartz bar, under the influence of a stress, will produce rotation of the plane of polarisation of polarised light passing transversely through it; this property is frequently used as a method of stress analysis. If such a bar is placed between two crossed polarisers then, in the absence of stress, no light will pass through the system. If stress is present in the bar, however, then light which has passed through the bar will have a component polarised in the plane of the analyser, and this will be received by the photomultiplier. Thus a video signal applied to a transducer bonded to the bar will result in a video frequency modulation across a beam of light transmitted.

The system just described would however have a square law response, because a stress of either sign would allow light to pass through. To give the required linear characteristic, optical bias to a state

[17]) P. Rainger, I.E.E./I.R.E. Int. Telev. Conf., London, June 1962.

[18]) C. F. Brockelsby, J. S. Palfreeman and R. W. Gibson, British patent applied for.

mid-way between extinction and full transmission is provided by including an optical quarter-wave plate. A certain amount of light is then transmitted when there is no stress in the bar, and it is this amount which is modulated when a video signal on a carrier is applied to the transducer.

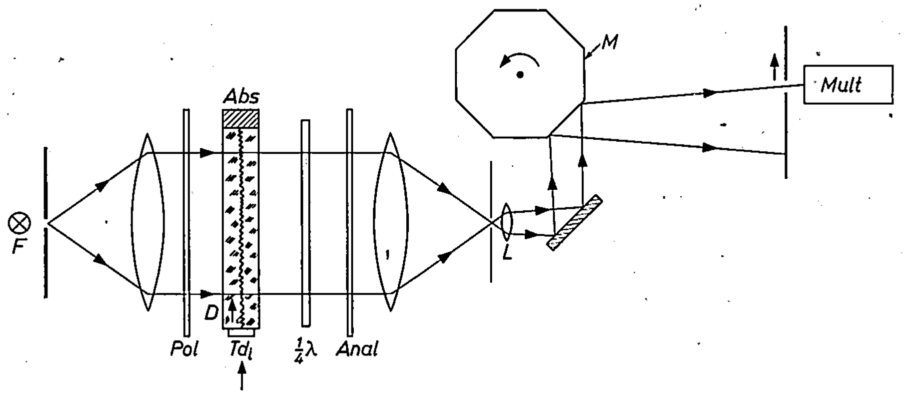


Fig. 18. Line stretching in systems conversion. The system illustrated is similar to the "Scophony" receiver shown in fig. 8. A solid photoelastic delay line is used in conjunction with a polariser (*Pol*), analyser (*Anal*) and $\frac{1}{4}\lambda$-plate, and the speed of rotation of the mirror *M* is adjusted so that the picture points corresponding to each input line move across the slit of the photomultiplier (*Mult*) in a time equal to the desired output line period.

A complete line stretching system using a photoelastic delay line is shown in *fig. 18*. The input signal applied to the delay line causes an amount of light to be transmitted through the system whose local intensity depends on the signal amplitude present in the delay line at any point. As in the "Scophony" receiver, the delay line is imaged on a screen by a lens system; at one point on this screen a slit allows the incident light to fall on a photomultiplier. The speed of the rotating mirror is adjusted so that the picture elements of a complete picture line will pass across the photomultiplier slit in a time equal to the output line period. In this way the required time expansion of the picture line may be achieved. In view of the very high rate on information handling required, such a system may prove more attractive than one using a computer store.

It would be possible to dispense with the rotating mirror and to use instead a moving light source. This source might conveniently consists of a spot of light on the screen of a cathode ray tube, although in practice a vertical line of light would be used to reduce phosphor noise and screen burn. By electronically sweeping the spot (or the vertical line) across the screen in a horizontal direction at an appropriate speed, the picture points may be made to pass the photocell at the required rate.

Alternatively the moving light source might be produced by using a second photoelastic delay line, in which a single pulse is travelling, to "gate" part of a parallel beam of light. By using an optical systen of appropriate magnification, an image of this moving source may be made to scan the original delay line at the correct rate and thus produce the required change of time scale.

Since rotating mirrors are difficult to synchronise at high speed, both these electronically controlled scanning systems offer considerable practical advantages.

The rather sophisticated combination, in this last instance, of several delay lines for interpolation and for a mechanical change-of-rate via optical scanning seems to us a fitting conclusion of this review of delay-line applications in television.

Summary. In ultrasonic delay lines use is made of the low velocity of mechanical waves in solids and liquids to effect a delay of wide-bandwidth electrical signals for times ranging from a few microseconds to several milliseconds. An ultrasonic delay line consists of three components: an input transducer which converts the electrical signal into an identical mechanical wave; the delay medium; and an output transducer which converts the mechanical wave back into an identical electrical signal (whose frequency is usually much higher than 20 kc/s — hence the name "ultrasonic"). Ultrasonic delay lines fall into three categories, using wire, a liquid and an extended solid respectively as the delay medium. In general only solid and liquid lines are used in television applications. When a delay of one picture line period is required a mercury delay line can be used; when a delay of one field period is needed a fused-quartz delay line is better. Wafers of crystalline quartz or of a ferroelectric ceramic are used as transducers for solid and liquid delay lines. It is important that the transducers should be of sufficient size to transmit and receive mechanical signals mainly over narrow ranges of angle. Low-*k* transducers should be matched in mechanical impedance to the delay medium. For high-*k* transducers electrical adaptation to the input and output networks is more difficult to calculate. A table of the characteristics of some typical delay lines is given in the text.

Applications of ultrasonic delay lines in the field of television are described. The earliest application was in the "Scophony" receiver where the variations in density of water accompanying a mechanical wave were used to modulate the intensity of a light beam across its width and to provide the display. A delay of one field period has been used to correct for the effects of smear in the vidicon television camera tube. A delay of one line or one field period may be used to provide correction of "vertical aperture distortion" in television camera tubes. A delay line of one line period delay is incorporated in a receiver for the "Secam" and "PAL" systems of colour television. A variable mercury delay line has been used to simulate the effects of reflections from aircraft on television reception. Finally, it has recently been proposed to use one or more delay lines in a system providing synchronous conversion between two television line systems, e.g. from 625 to 405 lines per picture.

VOLUME 25, 1963/64, No. 10 pp. 253-288 Published 2nd October 1964

Philips Technical Review

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

NEW FORMS OF BEARING: THE GAS AND THE SPIRAL GROOVE BEARING

by E. A. MUIJDERMAN *).

621.822.5:621.892.9

I. THE CONTACTLESS BEARING PRINCIPLE
II. GAS BEARINGS
III. SPIRAL GROOVE BEARINGS

Although the use of oil to counteract friction was known to the Ancient Egyptians forty centuries ago, the idea of substituting air for lubricating oil was not conceived until the second half of the 19th century. This suggestion, put forward by Hirn, was not taken up immediately, and could not be taken up until certain prerequisites had been fulfilled. One necessary condition was the development of a new form of bearing, the contactless bearing. Part I of the article which follows explains the principles on which the contactless bearing is based.

For more than half a century all practical types of contactless bearing were lubricated with oil or other liquids. The development on a large scale of air or gas bearings (Part II) which worked on the same principles, and which were first demonstrated by Kingsbury in 1897, began round about 1950. They are now being employed in nuclear reactors, turbo-jet engines and gyroscopic compasses, in the textile and food industries, and in dentists' drills.

In Part III of the article, having sketched in a background of these quite recent developments, the author goes on to describe a further innovation — the spiral groove bearing. This opens up new possibilities not only by reducing coefficients of friction but also by permitting a reduction in the size of the bearing. Amongst the topics discussed are a spherical-type spiral groove bearing which has a diameter of only 3 mm.

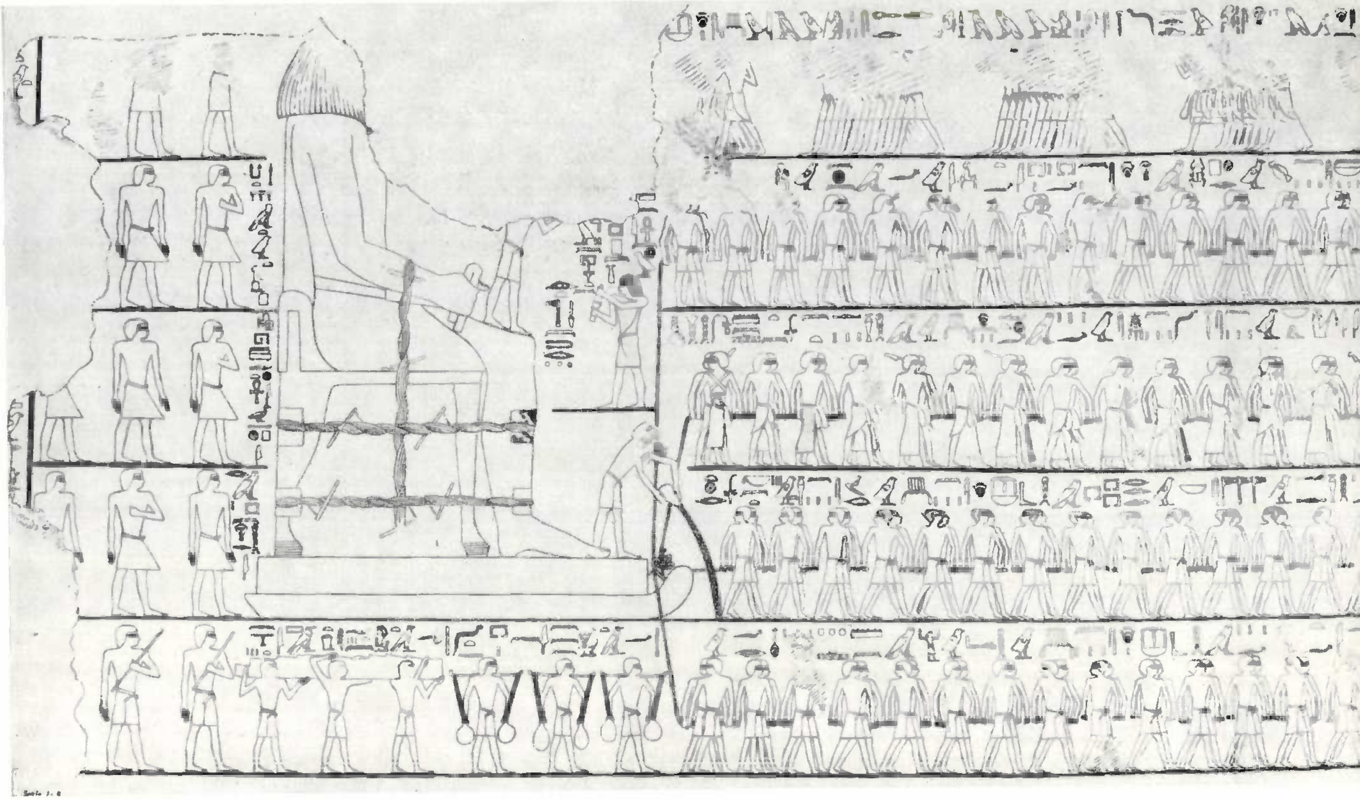
I. THE CONTACTLESS BEARING PRINCIPLE

The rotation of a shaft or spindle in bearings involves friction and wear which, for obvious reasons, must be kept within bounds by some means or other. A common practice is to use oil as a lubricant (*fig. 1*); alternatively, the shaft can be allowed to turn in ball-bearings (*fig. 2*). This does not exhaust the possibilities, and in the present article we shall be concerned with bearings which, essentially, rely on a clearance being maintained between the shaft and the surface of the bearing oil or any one of a whole range of other fluids being used as separating medium.

The oldest representative of this class is the oil-lubricated *journal bearing*. In 1883 Tower discovered that its good properties were not just a consequence of oil lubrication, as had been thought. One can regard the journal bearing as a forerunner satisfying the conditions for maintaining the clearance between the surface of the shaft and the bearing before these conditions were recognized. Present-day understanding of these conditions makes it possible to design a great variety of bearings in which separation of the shaft and bearing surfaces is a fundamental feature. Our proposed name for the whole class is "contactless bearings".

Magnetic bearings, in which the spinning shaft is supported by magnetic fields, must also be included in the contactless class. These bearings, which can be used for certain specialized applications, have

*) Philips Research Laboratories, Eindhoven.



already been dealt with in this review [1]). Here it is only proposed to discuss contactless bearings in which a viscous fluid separates the bearing from the shaft or spindle (full-film lubrication). Either a liquid or a gas can be used for this purpose; a gas, after all, also possesses viscosity, if only to a small degree. In the present article a particular point will be made of going rather more fully into gas lubrication, which is being employed on an increasing scale in all kinds of applications [2]).

"Hovercraft" in motion are supported on an air-cushion that separates them from the surface of the earth or the sea. Separation is achieved by virtue of inertial forces set up in the medium, and does not depend on viscosity forces as in the case of the contactless bearings just referred to. Surface-separating effects in which viscosity forces play the major part can only be obtained with very small clearances.

The *spiral groove bearing* is a recently developed type in which either a liquid or a gas, or even grease, will act as separating viscous fluid. The merits of the new type can only be stated very summarily in this introductory account. Its outstanding features are the low coefficients of friction that can be achieved with it, and the small dimensions it can be given. For example, cheap, highly wear-resistant and almost frictionless spiral groove bearings have been made that have a diameter of only 3 mm. To get a really clear idea of the properties of the spiral groove bearing, we must consider it against the background of contactless bearing properties in general.

Two categories will now be discussed in turn: contactless bearings with an external pressure source ("pressurized bearings"), and "self-acting" contactless bearings. Since spiral groove bearings fall into the latter category, it will be accorded all the greater attention.

Contactless bearings with an external pressure source

At the Industrial Exhibition held at Paris in 1878, it was shown how a heavy object could be moved almost frictionlessly over a smooth steel base-plate. The object had four legs which rested on the base-plate in the manner shown schematically for one leg in *fig. 3a*. To create a clearance between the base-plate and the sliding block, which was carrying its share of the overall weight of the object, it was necessary to build up sufficient pressure in the lubricant between the block and the plate by means of a pump. Creating a strong enough *upward thrust* (or "lift") is no problem, but the requirement of *stability* also has to be satisfied: if the clearance (or layer thickness) h changes, then forces opposing the

[1]) F. Th. Backers, A magnetic journal bearing, Philips tech. Rev. **22**, 232-238, 1960/61.

[2]) One of the first publications having an important bearing on present-day gas lubrication practice was G. W. K. Ford, D. M. Harris and D. Pantall, Proc. Inst. Mech. Engrs. **171**, 93, 1957.

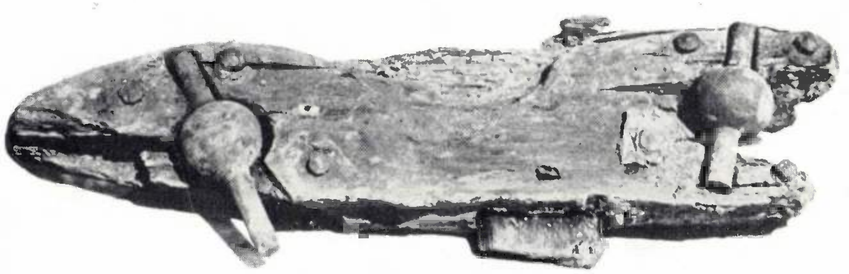


Fig. 2. Remains of a ball bearing found in one of two Roman ships that were raised round about 1930 out of the lake of Nemi. On the evidence of coins and inscriptions discovered in and near the ships it is fairly certain that they were built by the emperor Caligula (about 40 A.D.) and that they sank during the time of Nero. This 4.5 cm diameter ball bearing seems to have helped to support a turntable. The load was at the same time borne by the shafts, so strictly speaking this was a forerunner of the modern ball bearing rather than an early example of it.

The ships, which were housed in a special museum, were entirely destroyed by fire during the fighting that took place in the outskirts of Rome in 1944. See G. Ucelli, Le navi di Nemi, Libr. dello Stato, Rome 1950, p. 191 et seq., from which book the above photograph has been reproduced, with the permission of the author.

←

Fig. 1. Egyptian tomb painting dating from about 1650 B.C. A slave is pouring oil in front of a sledge bearing a statue of Pharaoh Tehuti-Hotep. Three porters carrying flasks of oil may be seen behind the slaves who are pulling the sledge. This is the oldest known picture representing the use of oil as a lubricant. (Reproduced from P. E. Newberry, Archaeological Survey of Egypt, El Bersheh, Part I, The tomb of Tehuti-Hotep.)

change must arise, with the result that the clearance remains more or less constant.

Stability might be obtained by employing a pump giving a constant output (e.g. a gear pump). When the clearance h decreases, the flow resistance of the gap increases, and if the pressure remains the same, less of the fluid will be able to escape from the gap (under laminar flow conditions, leakage is proportional to h^3). But this is impossible because the delivery rate is constant, so the pressure in the gap will increase and the gap will widen again.

Often, particularly where the fluid is a gas, pumps are employed that supply a constant pressure; that is the case we are concerned with in fig. 3*a*. Here the same stabilization effect is obtained by inserting a flow restrictor r between the pump and the gap. Fig. 3*b* shows pressure distributions over the sup-

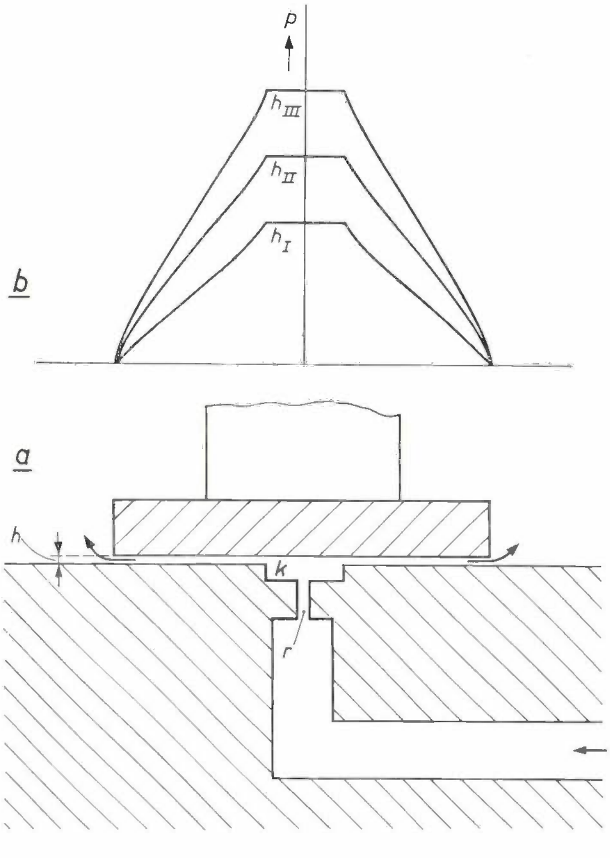


Fig. 3. (*a*) Principle of a contactless bearing with an external pressure source, the first application of which was in an exhibit called "*le chemin de fer en glace*" demonstrated at the 1878 Industrial Exhibition in Paris. This was a vehicle on "lion's feet" that was able to skim over a smooth steel surface. A pump working at a constant pressure forces a viscous fluid (which was oil in the "skating railway") through a constriction r, a chamber k and the gap (of height h) between the two surfaces to be separated, the fluid then escaping to the exterior. The constriction has a stabilizing effect, as will be clear from diagram (*b*).

(*b*) The pressure p in the gap as a function of place for three gap height values ($h_{\rm I} > h_{\rm II} > h_{\rm III}$). Any reduction in the gap height gives rise to an increase in pressure over the whole of the base-line, and thus to forces opposing the decrease in h. (The above curves relate to the case where the supporting fluid is compressible, but the conclusion is equally valid for an incompressible fluid.)

porting surface at various values of the gap height h. As before, if the gap shrinks, the pressure inside it will increase. This can be explained on the analogy between the pressure fall-off over the base-plate surface and the potential drop along an electrical circuit consisting of a constant voltage source, a fixed resistor (the flow restrictor) and a variable resistor (the gap) connected up in series: if the variable resistance is increased, more of the available potential difference will appear across its terminals.

Self-acting bearings

In the second category of contactless bearings, the self-acting kind, the pressure is built up from *within* the bearing. There are various ways of generating this but we shall only deal with the most important, which is exploited in self-acting bearings based on the *convergence* or *wedge effect*.

Between the flat block and base-plate in *fig. 4* is a wedge-shaped space; if the block is kept at an angle to the base-plate and moved through a viscous fluid in direction U, the fluid forced under the block will build up a pressure on it. We shall assume for the time being that the fluid is incompressible and that the flow caused by the movement of the block is steady and laminar, and we shall neglect the effects due to the fact that the block has a finite breadth L. Suppose that the pressure in the gap is everywhere the same; if the block is moving through a fluid of density ϱ at a velocity of U relative to the base-plate, the extreme layer thicknesses being h_1 at the front and h_2 at the back, then fluid will pass under the front edge at the rate of $\varrho ULh_1/2$ volumes per unit time, and leave the wedge-shaped space via the rear edge at the rate of $\varrho ULh_2/2$ volumes per unit time. But this does not satisfy the condition for continuity — more fluid is entering the space than is leaving it. Pressures between the block and the base-plate must therefore build up in such a way as to produce a velocity distribution that will ensure that fluid enters and leaves the gap at the same rate (*fig. 5*).

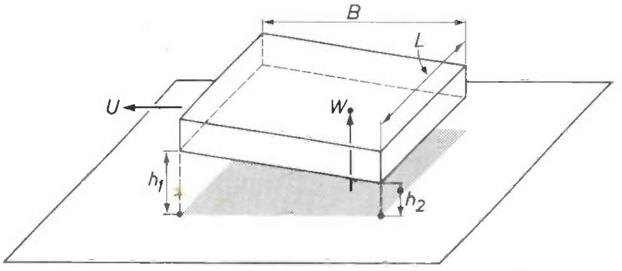


Fig. 4. To illustrate the principle of a self-acting bearing, based on the convergence or wedge effect. A flat block inclined to a base-plate in such a way that their opposed surfaces converge (in the case shown here, the gap in between is strictly cuneiform — a right triangular prism) is displaced through a viscous fluid at a velocity of U. Pressure thus builds up between the block and the base-plate with the result that a lifting force W is exerted on the block.

Movement of the block in the opposite direction would result in a fall-off in pressure in the gap, with the result that the block would be acted upon by a force opposite to W.

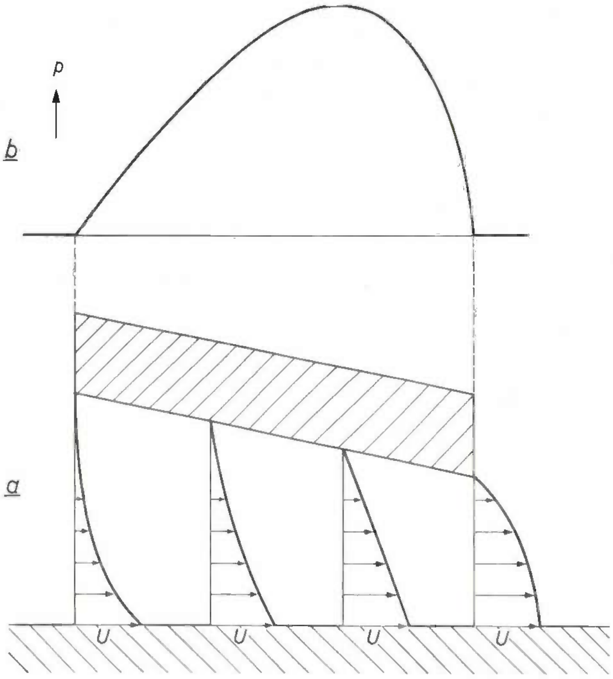


Fig. 5. (*a*) Velocity distribution and (*b*) pressure distribution in the incompressible viscous fluid filling the wedge-shaped space in fig. 4. The flat block is at rest and the base-plate is moving at a velocity U, dragging with it the fluid in the neighbouring boundary layer. The pressure is at a peak roughly half-way along the block, and at this point there is a linear increase in the velocity along a perpendicular drawn to the base-plate.

As has already been observed, the journal bearing was being used long before its essential mode of functioning was understood. We now know that the load on the spinning shaft causes it to lie somewhat eccentrically inside the journal bearing, so that the gap converges (*fig. 6*) in much the same way as the wedge-shaped space between block and base-plate. Provided the plain bearing is adequately lubricated the convergence effect will prevent the shaft from rubbing against the bearing bush. For a long time it was not understood why 19th-century journal bearings should be less subject to friction and show less wear than thrust bearings of the same period.

Tower cleared the way for an explanation of this fact when he discovered that the pressure in a journal bearing could (at points near the convergence) attain a local value considerably higher than that calculated on the assumption that the load was evenly distributed over the superficial area of the bearing. He became aware of this when he attempted to use a cork for stopping up a hole that had been drilled in a journal bearing. Observing that the cork repeatedly worked itself out of the hole, Tower conceived the idea that it was under high pressure and that this might be the reason for the journal bearing's good properties. Experimental data on the pressure distribution in a journal bearing investigated by him, which were published in 1883/5, attracted wide notice. Amongst those whose interest was aroused was a theoretician, Reynolds, who succeeded as early as 1886 in explaining Tower's data in terms of the wedge effect we have just been discussing [3].

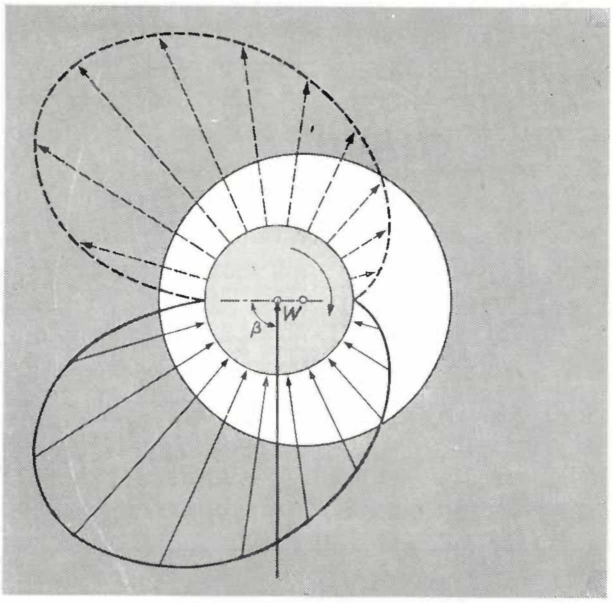


Fig. 6. Build-up of pressure inside a journal bearing lubricated with an incompressible fluid (schematic representation with bearing clearance greatly exaggerated). The fully-drawn arrows represent pressures above atmospheric, the dashed arrows pressures below it. Excess pressure arises where the bearing surfaces converge in the direction of shaft rotation; a fall-off in pressure occurs where they diverge.

The case to which the above diagram relates is the purely theoretical one in which the two pressure patterns are mirror images. In practice, the excess pressures generated are often much greater than the fall-off in pressure the latter being limited by the vapour pressure of the lubricating fluid. The non-symmetrical pressure build-up obtained in practice leaves the shaft in a different equilibrium position from that shown above, an angle β that is smaller than 90° being enclosed between the gravity vector and the line through the shaft centre and the points of closest convergence. The equilibrium position of the shaft is also dependent on the compressibility of the fluid. For a more detailed account of how pressure builds up in a journal bearing, reference should be made to the literature [2].

This newly acquired knowledge was to be applied about ten years later to thrust bearings. At that time attention was being devoted to the problem of taking up the reaction of the water on a rotating ship's screw, transmitted through the propeller shaft to the ship itself. The arrangement employed prior to 1898 consisted, in principle, of a flat circular plate parallel to and bearing upon a similar fixed plate, the rubbing surfaces being lubricated with oil (*fig. 7*). Wear was considerable, and the load-carrying capacity was very poor. The idea of fitting one of the circular plates with a ring of pivoted blocks (*fig. 8*) occurred independently to Kingsbury (1898) and Michell (1905). Each

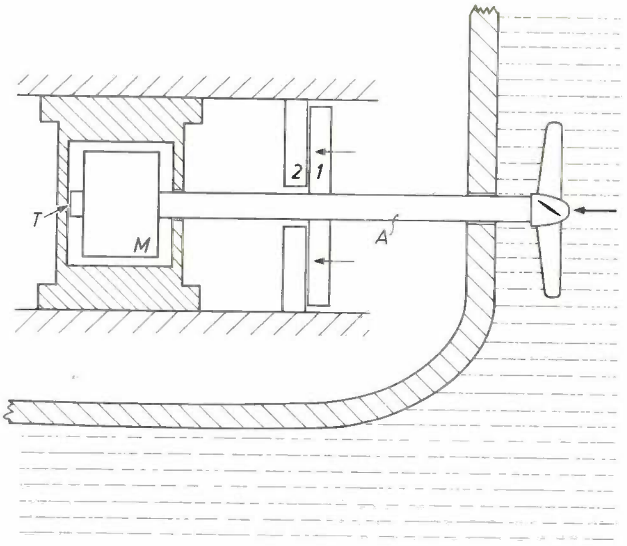


Fig. 7. A marine engine M drives the ship's screw via a long shaft A passing through a "tunnel" in the hull. The thrust exerted by the water on the screw, when this is turning, cannot of course be passed on to the hull of the ship by the bearing T supporting the crankshaft end in the engine housing. The thrust may amount to 150 or 200 tons, and it has to be taken up by a special bearing consisting essentially of a collar 1 which is fixed to the shaft and which presses against a stationary circular plate 2 built into the frame of the ship. The severe wear occurring in this type of thrust bearing led to the adoption of the Michell bearing.

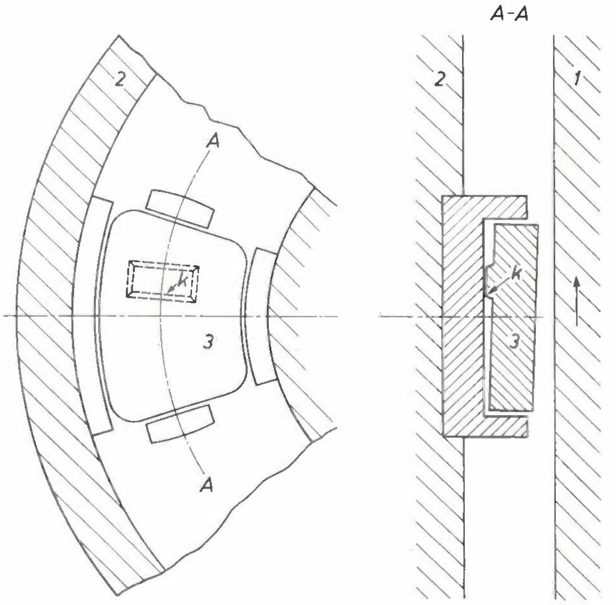


Fig. 8. The Michell bearing consists of a plane collar 1 and a circular plate 2 around whose edge are fitted slipper blocks 3 able to tilt about a pivoting edge k. The pivoting edge is given a position such that each block automatically adjusts itself for maximum load-carrying capacity (see the literature [4][20]) referred to in the text).

[3] B. Tower, Proc. Inst. Mech. Engrs. **34**, 632, 1883; **35**, 29, 1884; **36**, 58, 1885. O. Reynolds, Phil. Trans. Roy. Soc. **177A**, 157, 1886.

block introduces a wedge-shaped constriction into the space between the plates. If the pivoting ridge is correctly placed the pad will automatically take up a position such that, under the conditions prevailing, the bearing has optimum load-carrying capacity. In this way it is possible to create a complete clearance between the plates, even when they are subject to strong compression. Known as a Michell bearing [4]) in Europe, this type is generally referred to in America as a Kingsbury bearing.

The *step bearing*, another contactless thrust type, is of later date and rather simpler construction. Here rectangular ridges serve to bring about the constriction that is an essential feature of contactless bearings. Although the space between the plates is not wedge-shaped in the true sense of the word, any more than it is in the spiral groove bearing that will be discussed below, the convergence effect is achieved just the same. These differences in construction involve no fundamental change in the principle. We can therefore quite reasonably study the wedge configuration in fig. 4 for the purpose of gaining some insight into the basic properties of the self-acting bearing.

By integrating the pressure built up in the gap over the inclined surface, whose length is B, we obtain the lift W exerted on the inclined block:

$$W = \frac{C_1 \eta U B^2 L}{h_2{}^2}, \quad \dots\dots \quad (1)$$

where η is the viscosity and C_1, to a first approximation, is a function of h_1/h_2 alone. The quantities h_1, h_2, U and L have already been defined. When a set of such blocks is incorporated in a thrust bearing, the total lift or load-carrying capacity of the bearing can be arrived at by treating B as the sum of the average tangential length of all the blocks. (If end effects are taken into account, which means allotting L a finite value, it will be found that C_1 also depends to some extent on the $L:B$ ratio.) The frictional force F in the same bearing is:

$$F = \frac{C_2 \eta U B L}{h_2}, \quad \dots\dots \quad (2)$$

where C_2 is likewise a function of h_1/h_2 alone. We shall refer to C_1 and C_2 as the bearing coefficients. The bearing will adjust itself to a value of h_2, the smallest layer thickness, that is dependent on operating conditions, i.e. the speed U and the load or required load-carrying capacity W. If the blocks are rigidly mounted on the shaft or bearing the difference between h_1 and h_2 will remain constant, so that if h_2 changes, the h_1/h_2 ratio and the bearing coefficients C_1 and C_2 will change too. If the blocks are pivoted, as in the Michell bearing, shown in fig. 8, h_1/h_2 can remain constant whatever the inclination of the block.

The ratio between frictional force and load-carrying capacity, $f = F/W$; is called the coefficient of friction of the bearing because it has much the same significance as the familiar coefficient of dry friction. It follows from (1) and (2) that

$$f = \frac{C_2}{C_1}\frac{h_2}{B}, \quad \dots\dots \quad (3)$$

which indicates, perhaps rather surprisingly at first sight, that the friction coefficient of a bearing does *not* depend on the viscosity of the lubricant used e.g. oil or air.

The following formulae, analogous to (1) and (2), can be derived for a *journal* bearing, of radius r, in which a shaft is spinning with an angular velocity ω:

$$W = \frac{K_1 \eta \omega r^3 L}{(\Delta r)^2}, \quad \dots\dots \quad (4)$$

$$P = \frac{K_2 \eta \omega^2 r^3 L}{\Delta r}. \quad \dots\dots \quad (5)$$

The quantity P given by the second formula is the rate at which energy is dissipated by friction; Δr is the uniform clearance around the shaft when it is in the exact centre of the bearing bush. Since the situation is slightly different from that in the thrust bearing, we have different bearing coefficients to deal with, K_1 and K_2. These depend, amongst other things, on the degree to which the shaft sinks into the lubricating film, or in other words on its eccentricity with respect to the bearing bush. The eccentricity of the shaft is analogous to gap height h_2 in the thrust bearing. (If end effects are taken into account, L being allotted a finite value, K_1 and K_2 are found to be dependent, in addition, on the ratio r/L.) Like h_2, the shaft eccentricity adjust itself to a value dependent on operating conditions; again depending on operating conditions, the bearing coefficients vary over ranges whose extreme values, in practice, may differ by a factor of about 5.

Some important conclusions can be drawn directly from the above formulae.

From either (1) or (4) we can see that the shaft speed must exceed a certain minimum if a self-acting bearing is to develop any load-carrying capacity. Inevitably, then, starting from rest will involve wear, an important factor that will be discussed at greater length in Part II. We can further infer from (4)

[4]) See for example G. Vogelpohl, Betriebssichere Gleitlager, Springer, Berlin 1958,

that in so far as it is possible to reduce the diametral clearance ($2\Delta r$), doing so will allow the desired load-carrying capacity to be achieved at a lower speed. The implication is that very narrow bearing clearances can be a vital requirement, particularly where the lubricant is air (which has a small η). For one special application, American engineers went so far as to design and build a journal air bearing in which the diametral clearance around the shaft was only 0.6 μm [5]). That the demands on technology and materials are extreme in such a case (the only suitable materials are ceramics) will be obvious. It is partly on that account that work on air bearings did not really get into its stride until about 1950.

Formula (1), as it stands, is not entirely valid for very high shaft speeds. In a thrust bearing under a certain constant load, an increase in shaft speed will not in the first instance involve any change in the pressure built up in the bearing because the smallest layer thickness h_2 will increase accordingly, automatically adjusting itself to the required load-carrying capacity. But in practice h_2 can only widen to a limited extent, seeing that normally a shaft is held in *two* thrust bearings. Once h_2 has reached this limit, any further increase in shaft speed will involve an increase in bearing pressure and give a surplus of load-carrying capacity, which is of no use, but which will result in increased friction in accordance with formula (2).

Much the same applies when the speed of a shaft in a journal bearing is increased, though the implications are not quite so evident in formula (4). As has already been explained, the shaft adapts itself to a given load by assuming a certain degree of eccentricity with respect to the bearing bush, this being associated with a K_1 value such that the bearing develops exactly the required load-carrying capacity. The higher the speed, the smaller will be the degree of eccentricity and the value of bearing coefficient K_1, but a limit is reached once the shaft has come to lie dead-centre in the bearing. If the speed is further increased, K_1 will remain constant and surplus load-carrying capacity and increased friction will again be the result.

The themes developed so far will be taken up again at an early point in Part II of this article.

II. GAS BEARINGS

Air versus oil

In Part I an attempt has been made to explain the principles which enable an air or gas bearing to function. Before these forms of bearing are examined more closely it is natural to inquire what are the advantages of using air or some other gas as lubricant in a contactless bearing.

Oil was the lubricant in an application of the contactless bearing principle exhibited in Paris in 1878. The oil pumped under the sliding blocks had to be collected in a tank and recirculated. It is interesting to compare this contrivance with a very similar one now being used for a purpose of which nobody in 1878 is likely to have had any inkling. The allusion is to a space flight simulator (*fig. 9*). Here the lubricant is air which is pumped under the supporting pads. One convenient feature of air as a lubricant is immediately obvious: it can simply be allowed to escape into the surrounding atmosphere; it does not need to be collected or circulated; and the user has no need to worry about the lubricant gradually getting dirty, and about the complications this normally gives rise to. This property of air was pointed out by Kingsbury as early as 1897, when he was demonstrating an air-lubricated journal bearing [6]): "The atmosphere furnishes a bath of tbe lubricant, thus maintaining a constant supply of constant quality". Contamination of the lubricant is one problem, contamination of its environment by the lubricant is another: the latter danger, which persists despite all the ingenuity the designer may devote to making his bearings leak-proof, rules out the use of oil in machinery for processing food, for example, or for manufacturing textiles or other materials that would be spoilt by contact with oil. In such cases air lubrication is clearly the method of choice.

There are other cases in which oil lubrication is unsuitable for purely practical reasons, and in which the gas bearing offers a way out of the difficulty. Examples are machines whose bearings must work at very low or very high temperatures, temperatures at which oil would congeal or break down chemically. Cases of this kind will be referred to in the section on typical applications of gas bearings.

In addition, however, we may point out that a big advantage of gas as compared with oil lubrication, predictable on purely theoretical grounds, is the fact that gas bearings are inherently more suitable for

[5]) The bearing was developed by the Minneapolis-Honeywell Regulator Company; see Machine Design **32**, no. 13, p. 39, 1960.

[6]) A. Kingsbury, Experiments with an air-lubricated journal, J. Amer. Soc. Naval Engrs. **9**, 267-292, 1897.

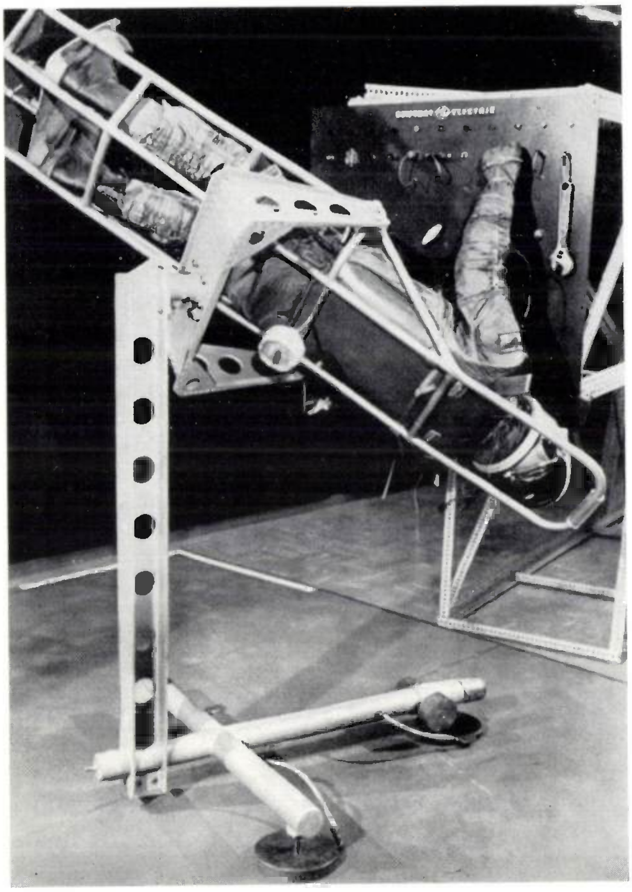


Fig. 9. Space travel simulator built by General Electric. The trainee astronaut is strapped into a frame mounted in a universal joint and supported on air-lubricated pads which glide almost frictionlessly over the floor (the principle being exactly the same as that of the "skating railway" exhibited at Paris in 1878). The simulator is used for investigating one of the consequences of weightlessness, namely the absence of friction. Normally, friction is an aid to locomotion and other bodily movements. In the appliance shown here, movement of any part of the body results in displacement of the rest of the body in the opposite direction. It appears to be possible to perform maintenance routines in these circumstances, but the work raises the subject's oxygen consumption to 30% or 40% above normal.

very high shaft speeds. This property is bound up with the limitations on eccentricity and smallest layer thickness that were pointed out in the final paragraphs of Part I. A shaft whose speed is continuously increased will finally take up a position exactly central with respect to the bearing (similarly, in the case of a thrust bearing, a point will be reached at which h_2 cannot widen any further). If the speed is still further increased the bearing will develop excess load-carrying capacity, and the frictional losses P will soon become prohibitive. An obvious remedy would be to reduce the radius r and/or the length L of the bearing by an amount such as to cancel out the surplus load-carrying capacity. But if the bearing cannot be redimensioned without a dangerous loss of mechanical strength, the only way of overcoming dead-centre running will be to switch to a lower-viscosity lubricant. In a manner of speaking, oil is too good a lubricant, since its load-carrying capacity at high speeds is too great. Formula (5) reveals that by using air, whose η is roughly 10 000 times smaller than that of oil, it is possible to increase the shaft speed by a factor of 100 or the dimensions of the bearing by a factor of 10 without any change in frictional losses, provided other conditions likewise remain the same. In reality, comparing air with oil is not quite so straightforward because it is normal to design gas bearings for smaller film thicknesses; but this does not detract from the general validity of the above conclusion and we shall see that fast shaft speeds are a prominent feature in typical applications of the gas bearing principle.

Comparison between self-acting and externally pressurized gas bearings [7]

The self-acting gas bearing differs from the externally pressurized type in being of simpler design and having greater reliability (since the supply of gas under pressure cannot break down). Another big difference between the two categories concerns their starting characteristics. As has already been pointed out, formulae (1) and (4) show that a self-acting bearing cannot develop a load-carrying capacity at zero shaft speed; the lift increases with speed, linearly to begin with. Shaft speed, in the first instance, has no influence on the load-carrying capacity of bearings of the other type; this is merely a matter of the pressure provided by the pump. In the self-acting bearing, then, starting produces wear, and it is only when the shaft has attained a certain speed that the clearance between the bearing surfaces is large enough to reduce friction and wear to the desired low levels. This means, firstly, that self-acting bearings must be made of hard, wear-resistant materials [8]), and secondly, that they must be designed in such a way that separation between the bearing surfaces occurs at the lowest possible shaft speed. As already stated, American engineers, with these aims in mind, have made bearings fitted with ceramic bushes and having a diametral clearance of only 0.6 μm.

Obviously, because their load-carrying capacity depends on shaft speed, self-acting bearings will be less suitable in cases where a low speed is desired, or where the shaft speed is subject to wide variations.

[7] In dealing with liquids, we have to remember that they are far less compressible than gases. With that reserve, much of what is said here would apply equally well to contactless bearings lubricated with liquids.

[8] This consideration is not of such great importance in oil-lubricated bearings since, under starting conditions, there is always a thin film of oil to provide "boundary lubrication" and so reduce friction and wear to some extent.

The two classes cannot be further compared without taking the *compressibility* of the working fluid into account.

Limits to load-carrying capacity

Formula (1) was derived for a wedge-shaped gap through which, it was assumed, an *incompressible* fluid was passing. The condition for continuity, we said, was that no difference $\varrho ULh_1/2 - \varrho ULh_2/2$ should exist between the quantity of fluid entering and the quantity of fluid leaving the gap. We also assumed that such a difference was prevented from arising by a build-up of pressure in the gap, this build-up being responsible for the lift exerted on the underside of the block. Any change in the density ϱ of the fluid was not taken into account. Now, what happens if there is in fact such a variation? Then the extra lift obtained by increasing the relative speed of the surfaces will fall off steadily with each increment of velocity because of the density variation involved, till a limit is reached at which it is *only* the density variation that is preventing a difference from arising between the quantities of fluid entering and leaving the gap. In this limiting case, then, speed has ceased to play any part in building up pressure (and so has viscosity). *Only the pressure p_i of the surrounding atmosphere is of importance in the limiting case.* A straightforward calculation [9]) shows that in these circumstances the pressure in the gap approaches

$$p_{\max} = p_i \frac{h_1}{h_2}. \qquad \ldots\ldots \quad (6)$$

Thus the load-carrying capacity of a self-acting bearing is limited; but that of an externally pressurized bearing can always be increased, simply by stepping up the pump pressure.

For reasons like these, self-acting bearings are most likely to be used, in practice, for taking up relatively light loads, whereas a bearing with an external pressure source will often be the best solution to the problem of taking up a relatively heavy load.

Compressibility is not the only factor limiting the load-carrying capacity of a self-acting bearing. Other limitations come into play in cases where the gap height or clearance is of the same order as the mean free path of the gas molecules. A gradual transition from viscous to inertial flow at high shaft speeds may also have the effect of limiting capacity; this change in flow conditions is liable to occur in externally pressurized as well as in self-acting bearings, though in practice it does not limit the capacity of bearings in the former class. These phenomena, which assume some importance in extreme cases, cannot be gone into here, and the reader is referred to the literature on the subject [9])[10])[11]).

Limits to stability

Mention of the instability phenomena to which contactless bearings are liable takes us into a difficult field which is still far from being fully explored. Because of their great importance we must touch upon these phenomena in passing, but for a fuller study the reader must consult the literature [9])[11]).

In the simplest possible terms, a contactless bearing can be regarded as a damped system possessing mass and springiness. A system of this kind has a natural frequency which, according to a well-known formula, is roughly equal to the square root of its stiffness/mass ratio. A periodic disturbance — consequent on rotor unbalance in the case of a journal bearing, for example — will set the system into forced vibration. In the case of the unbalanced journal bearing, the frequency of the vibration will be equal to the rotor speed in r.p.m. If this happens to be the same as the natural frequency of the system, resonance effects will occur. In general, designers try to obviate such resonances by giving the bearing a natural frequency high above that of any possible disturbing effect, or ensuring that it will pass quickly through the range of critical frequencies.

The system may also get into a state of *self-sustained vibration* which is due, not to any lack of balance, but to the particular way the stabilizing force varies as a function of the layer thickness. In a radial self-acting bearing which is running dead-centre — a state of affairs that is undesirable for this as well as for the previously mentioned reason — the variation in question may give rise to what is known as "half-speed whirl". This is a dangerous and destructive phenomenon in which the shaft axis whirls around the centre-line of the bearing at a speed approaching half the angular velocity of the shaft, its distance from the centre-line gradually increasing. The shaft may thus start chattering in the bearing. Both journal and thrust bearings in the externally pressurized type may be subject to a similar kind of vibration if the gap height or the volume of the "chamber" (see fig. 3) is large. The latter phenomenon, which is comparable with all known forms of relaxation vibration, is termed the "air hammer" (because air hammers are based on that particular kind of vibration).

[9]) See for example W. A. Gross, Gas film lubrication, Wiley, New York 1962, p. 64.

[10]) Gas lubricated bearings, a critical survey, U.S. Dep. Commerce, Office of Technical Services, 1958.

[11]) O. Pinkus and B. Sternlicht, Theory of hydrodynamic lubrication, McGraw Hill, New York 1961.

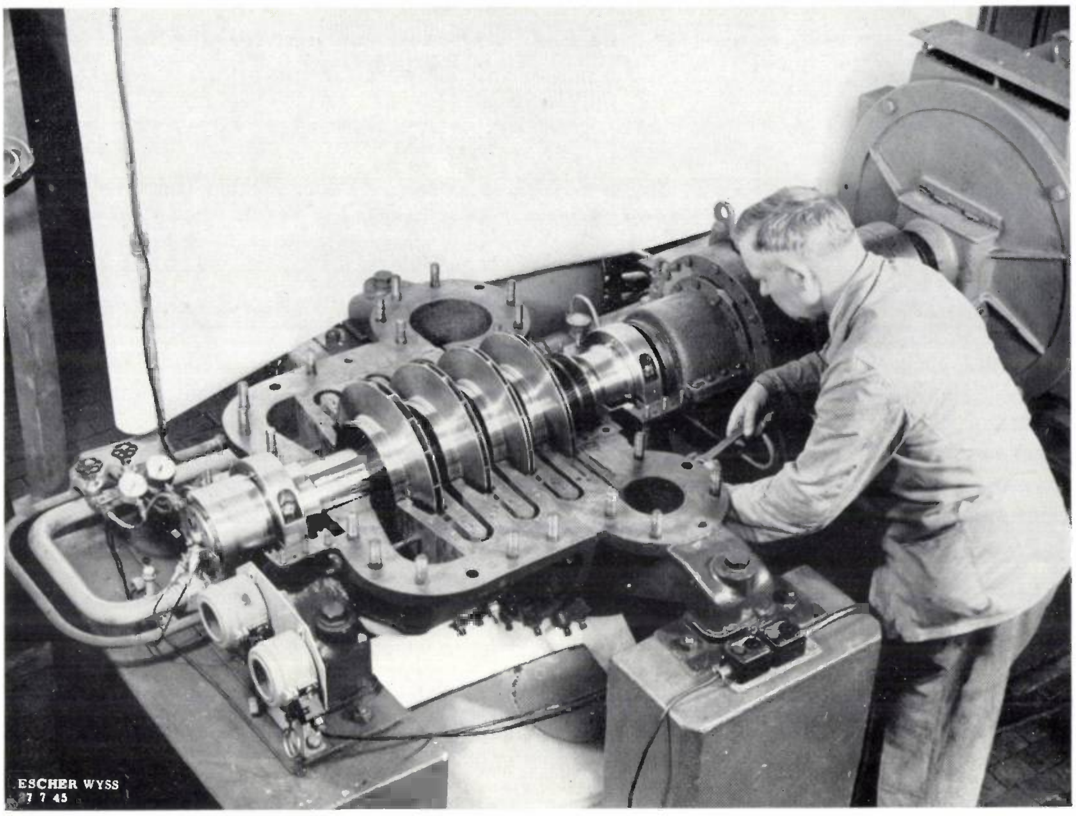


Fig. 10. Experimental four-stage centrifugal compressor made by Escher Wyss, partly dismantled. Gas bearings afford both axial and radial support for the shaft. Pipes and pressure gauges for the incoming gas supply for these bearings may be seen on the left. This compressor runs at higher speeds than normal oil-lubricated types, and is therefore able to produce higher compressions. At the same time the compressed gas is completely free from oil. (Reproduced from E. Loch, Escher Wyss Mitteilungen **33**, no. 1/2/3, p. 118, 1960.)

In cases where one type of bearing may seem on the face of it to be the more suitable, instability phenomena may tip the scale in favour of the other type. If contactless bearings are to be further improved it is of the greatest importance that the forms of instability to which they are susceptible should be properly understood, and much of the research work at present being done in various laboratories does in fact have as its object the better understanding of these phenomena.

Some examples of gas bearing applications

Mention has already been made of certain applications that do not require a great deal of explanation. We shall now go on to consider applications having some special feature or other that must claim our fuller attention.

Spectacular, though still experimental, compressors embodying gas bearings offer a good example of the desirability and feasibility of working at higher speeds than normal, and so developing higher pressures than are possible with oil lubrication [12]. *Fig. 10* shows a four-stage centrifugal compressor, the main shaft of which is supported at left and right in thrust and journal bearings. The bearing gas intake ducts and gauges are clearly visible on the left. The self-acting type of bearing would obviously be less suitable here because the compressors are not kept in continuous operation, and the wear consequent on frequent starting would cause a lot of trouble. A further advantage of using gas bearings in this equipment is that the gas delivered under pressure is entirely free from oil. This is of particular importance when oxygen is being compressed, since the presence of oil would involve an explosion hazard.

Self-acting gas bearings are also employed in CO_2 circulators for nuclear reactors [2]. As is well known, carbon dioxide gas is the coolant used in British atomic power stations like that at Calder Hall. Oil-lubricated bearings were unsuitable for the CO_2 circulators because oil has the tendency to break down under the action of irradiation. This fact gave a considerable stimulus to the development of gas bearings, the first successful application of which aroused a great deal of interest at the time. The con-

[12] E. Loch, Escher Wyss Mitteilungen **33**, No. 1/2/3, p. 118, 1960.

tinuously circulating CO_2 serves at the same time for lubricating the bearings, which together with the entire cooling system are self-contained — a great help in preventing the dissemination of radioactive material. *Fig. 11* shows the circulator and the shaft suspension arrangements, which consist of two journal bearings *a* and *b* and a thrust bearing *c*. The last named is of the spiral grooved type that will be discussed at length in Part III of this article. *Fig. 12* is a photograph of the rotor assembly of the CO_2 circulator, the two bushes for the journal bearings appearing in the foreground.

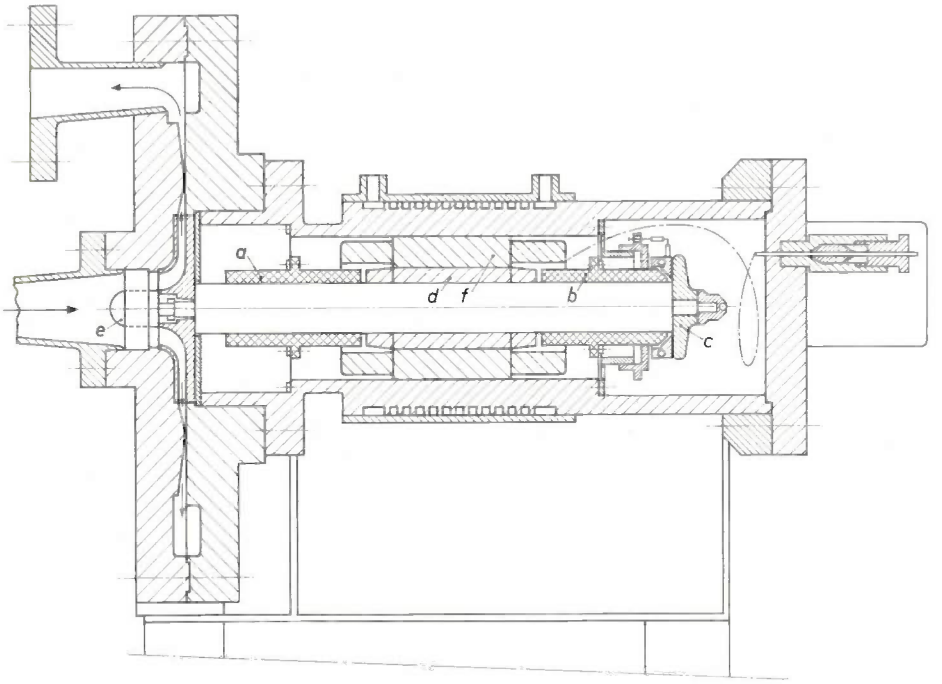


Fig. 11. CO_2 circulator for a nuclear reactor. The carbon dioxide gas used to cool the reactor (or, to put it more accurately, used to recover the heat developed in the reactor for exploitation elsewhere) is forced through the circuit by a radial fan *e* driven by an electric motor (stator *f*, rotor *d*). The arrows show the direction of the circulating gas. Since oil tends to break down chemically under the action of irradiation, gas-lubricated bearings have been incorporated in these circulators, the same CO_2 gas supply serving as lubricating fluid. The shaft is carried in two journal bearings; *a* and *b* are their bushes, and *c* is a spiral groove bearing for taking up the thrust (see Part III of this article). (This diagram and fig. 12 are reproduced from G. W. K. Ford, D. M. Harris and D. Pantall, Proc. Inst. Mech. Engrs. **171**, 93, 1957.)

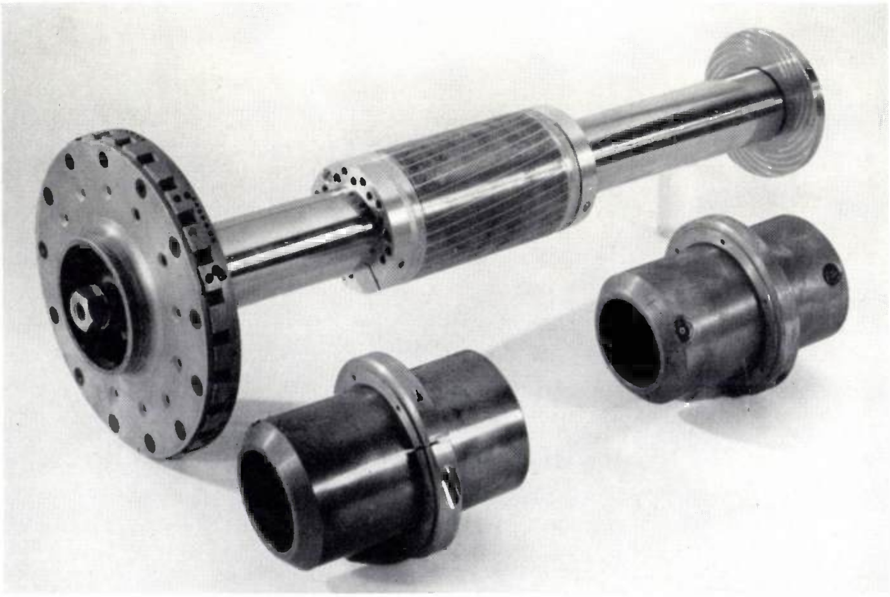


Fig. 12. Rotor assembly of the CO_2 circulator drawn in fig. 11, in which the gas circulated by the machine serves to lubricate its thrust and two journal bearings. In the foreground, the bushes for the journal bearings.

Gas bearings are now to be found in a whole series of smaller devices. Amongst some experimental devices made in our laboratory is a small rotor (*fig. 13*) for which compressed air supplies motive power as well as a supporting lubricating film. Because of the low friction losses, the rotor can be driven at speeds up to 450 000 r.p.m. without overheating. The same idea has been exploited in the newest types of dentists' drills. Turbo-jet engines and gyroscopic compasses also make use of gas bearings in view of the high speeds at which they are required to operate.

An interesting class of gas bearing applications is that in which the main consideration is to keep the friction constant rather than reduce it. It is mainly in connection with sensitive measurements that this demand is encountered. The set-up in *fig. 14* is for measuring, with the aid of an eddy-current dynamometer, the torque exerted by a miniature electric motor. This torque is small compared with the frictional couple that would arise if the rotor of the dynamometer,

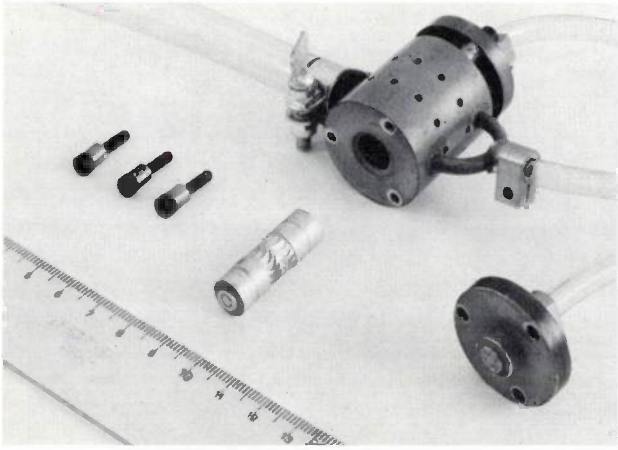


Fig. 13. Small experimental rotor which is driven and cushioned by compressed air. It can attain a speed of about 450 000 r.p.m. Recesses acting as turbine "buckets" have been milled in the rotor, which may be seen in the foreground. The split pipe entering the housing on the right supplies air for lubrication of the journal bearings on either side of the recesses. Thrust bearings are also provided, at both ends of the shaft. The front thrust bearing has been removed; four small orifices through which the air is blown (they have the same function as the restriction in fig. 3) may be seen in the middle.

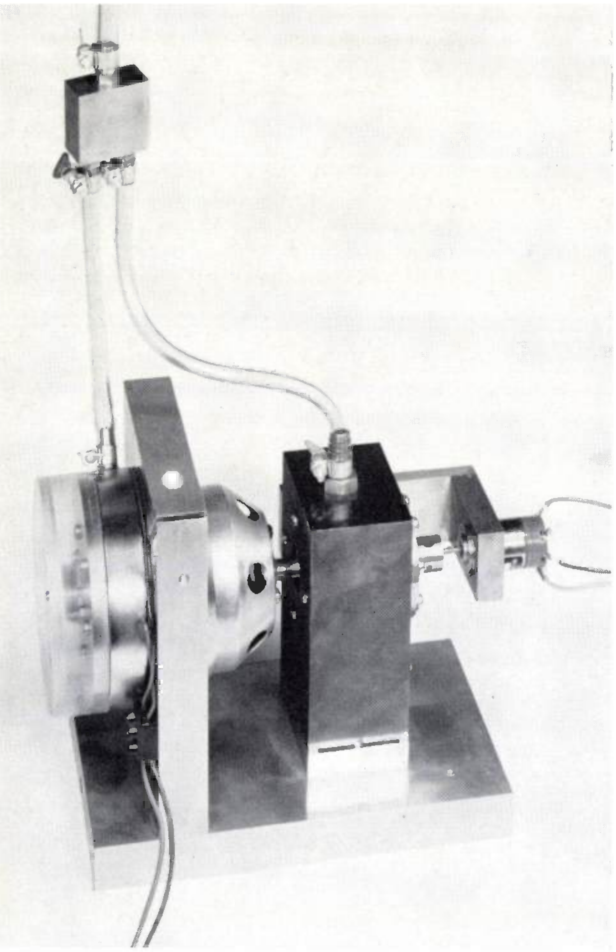


Fig. 14. Arrangement for measuring the torque delivered by a very small electric motor, mounted in the holder on the extreme right. The measuring instrument (left) contains a disc which is coupled to the motor spindle and which undergoes braking due to eddy currents; the braking couple is adjustable. To eliminate inaccuracies due to uneven bearing friction, compressed air is used to provide radial support for the disc, giving an extremely low and constant level of friction.

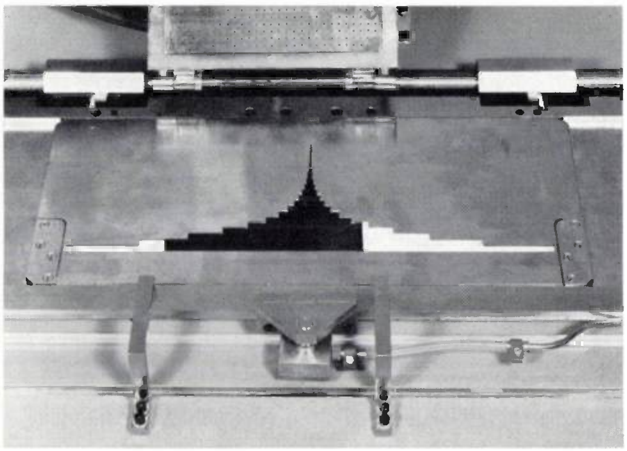


Fig. 15. Sensitometer for determining the sensitivity and other characteristics of photographic material. A masking pattern moves slowly across the instrument, with the result that the photographic material is exposed stepwise to light of known spectral composition and intensity from a source (not visible here) under the masking pattern. By means of a vacuum the material to be tested is held on a hinged grid, which in the photograph is in the raised position. It is necessary that the pattern should move with extreme smoothness and regularity if the desired aim, that of varying the exposure time in exactly determined steps, is to be realized. To ensure this, the carriage to which the masking pattern is fixed runs in tubular air-lubricated bearings that serve as guides (to the left and right of the grid hinges), and rests on a flat air-lubricated pad. The pipe supplying air to the pad can be seen in the foreground.

which gives the braking effect, were supported in ball-bearings, and inevitable small variations in the frictional couple would make the measurements very unreliable [13]). Something rather similar applies to the sensitometer [14]) appearing in *fig. 15.* Mounted on a carriage is a screen or masking pattern which serves to vary, as a function of place, the exposure time of photographic material placed underneath it. The problem was to make the translatory movement of the screen completely smooth. The slight jerks and hesitations involved by rolling over steel balls or the like, though these are practically inperceptible even to the careful observer, would have introduced quite impermissible discontinuities into the blackening pattern on the exposed film. Another advantage in the use of air as a lubricant, and which has already been referred to, is that of not contaminating materials it comes in contact with — in this case photographic film. As a final example of this class of application let us take an apparatus for calibrating acceleration pick-ups that has been developed elsewhere [15]). The pick-up is fixed to a carriage that is

[13]) The torque meter appearing in the photograph is based on damping caused by eddy currents, and was designed by W. Bähler of these Laboratories.

[14]) This instrument was the result of cooperation between our photochemical group and the Research Laboratories designers.

[15]) TNO-Nieuws 18, 154, 1963 (No. 3). In Dutch.

required to vibrate sinusoidally in an exactly defined way. Mechanical and electrodynamic excitation proved to have various drawbacks; either the movement was distorted out of the sinusoidal, or noise and interference were introduced by the ball-bearings etc. In the design just referred to, air was again used for supporting the carriage, which is held between two helical springs. The carriage is brought into free oscillation by displacing it and so compressing one of the springs. It is stated in the article cited that thanks to the very low frictional losses which remain constant within extremely narrow limits, very true sinusoidal oscillations are maintained for minutes at a time.

The last application to be discussed here falls into a quite different class. The device in question is to be found amongst other things in the magnetic-drum memories of computers, e.g. PASCAL. It is essential that there should be accurate transference of signals between the magnetic surface layer of the rotating drum and a fixed magnetic reading or writing head. But this will only be guaranteed if the clearance between head and drum surface is small and not subject to any great variation. In virtue of the air-bearing principle employed in PASCAL, a clearance of 8 μm is maintained between the head and the drum [16]). In the applications discussed earlier, maintenance of a given separation between bearing surfaces was just a consequence of fulfilling the condition for stability essential to contactless bearing operation. Here the constant clearance has become an aim in itself.

Let us examine the air-bearing action exploited in PASCAL in rather greater detail, and with reference to numerical values deriving from an experimental set-up. Fig. 16 shows schematically how the reading or writing head is suspended. A weak spring (of stiffness $c_w = 1$ N/mm) tends to press the head (of mass $m = 7$ grammes) against the rotating drum, but this force is opposed by the air in the more or less wedge-shaped space in between, which acts as a second, very strong spring (of stiffness $c_s = 300$ N/mm). The desired clearance (8 μm) is maintained despite slow changes — temperature fluctuations, for example — but also despite fast ones due to "ripples" on the surface of the rotating drum. Let us suppose that the drum has an angular velocity of ω and that its surface shows a regular sinusoidal deformation, x peaks of amplitude $\hat{h}$ being present around the perimeter. The sinusoidal rippling of the drum surface gives a regularly recurring impulse to the suspended magnetic head, and so sets it oscillating in a manner that is likewise sinusoidal. If its damping is zero the difference between the amplitudes of the two oscillations will be

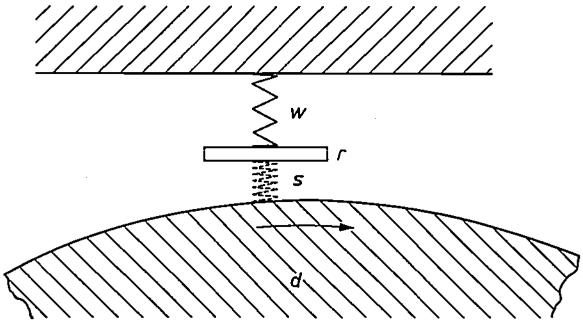


Fig. 16. Schematic drawing to show suspension of a magnetic head r for writing on to or reading off from a drum memory d such as is employed in the PASCAL computer. Spring w, which has a low stiffness value, tends to press the head against the rotating drum. In virtue of the convergence effect the air cushion between head and drum acts as a very strong spring s. Using this principle, it is possible to ensure that an 8 μm clearance will be maintained, with only very small variations, between the head and that part of the rotating drum surface which is passing it.

$$\Delta\hat{y} = \frac{(x\omega)^2/\omega_n^2}{1-(x\omega)^2/\omega_n^2}\,\hat{h}\,,$$

where ω_n is the resonant angular frequency of the magnetic head:

$$\omega_n = \sqrt{\frac{c_s + c_w}{m}}\,.$$

Since in the case under consideration $c_w \ll c_s$, we may write

$$\omega_n \approx \sqrt{\frac{c_s}{m}} \approx 6000 \text{ rad/s}\,.$$

This is in fact quite close to the actual value of ω_n for the air-supported magnetic heads in PASCAL. On this basis, and taking $\omega = 300$, we find that if there are three ripples in the drum perimeter, the oscillating magnetic head will follow the rippling so closely that the clearance between the head and the drum surface will not vary by more than about 3% of the ripple amplitude.

With this example we close the part of the article concerning gas bearings, and turn our attention to bearings of the spiral groove type.

[16]) The suspension system was designed by H. J. Hubers of this laboratory. See also J. H. Wessels, A magnetic wheel store for recording televisions signals, Philips tech. Rev. 22, 1-10, especially p. 6. In the case discussed in that article a constant clearance not exceeding 1 μm has been achieved.

III. SPIRAL GROOVE BEARINGS

Instigation of the work on spiral groove bearings

Our work on spiral groove bearings in the Research Laboratories began in 1960, when we were consulted on the following problem. Would it be possible to make a wear-resistant thrust bearing able to take up to about 1 kg thrust from a shaft rotating at at least 50 000 r.p.m., without incurring a power consumption much in excess of 1.5 watts? The bearing was to support the end of the shaft, and it had to be suitable for operation in a vacuum.

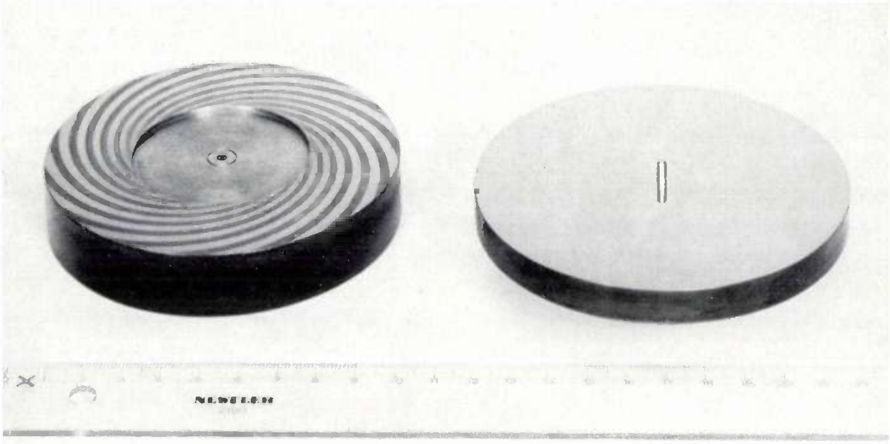


Fig. 17. Spiral groove bearing made in the Research Laboratories, consisting of a smooth and a grooved circular plate. When the plates are assembled and one is turned by hand with respect to the other (it must be turned in the right direction) complete separation occurs at a speed as low as one revolution per second, the air layer between the bearing surfaces then being about 11 μm thick.

To start with, it will be shown that none of the bearings so far discussed was likely to provide a ready solution to the above problem. Bearings involving sliding friction were out of the question, as also were ball-bearings and the like, on account of the wear they are subject to and also because the frictional losses would certainly be excessive. This narrowed the field to contactless bearings. As far as externally pressurized types were concerned, the pump alone would be responsible for more than 2 watts power consumption. The self-acting class seemed to offer good prospects (as did magnetic bearings, but their possibilities in this connection were not investigated in detail).

Formulae for the load-carrying capacity of and the frictional force arising in a self-acting bearing were given in Part I (p. 258). Transformed for the general case of a thrust bearing, formulae (1) and (2) provide a basis for writing down the two conditions that now had to be satisfied:

$$W = \frac{K_1'\eta\omega r^4}{h_2^2} \geqslant 10 \text{ (newton)}, \qquad \text{(1a)}$$

$$P = \frac{K_2'\eta\omega^2 r^4}{h_2} \leqslant 1.5 \text{ (watt)}. \qquad \text{(2a)}$$

It will be recalled that W denotes load-carrying capacity and P frictional losses in the bearing; η is the viscosity, ω the angular velocity, r half the outside diameter, and h_2 the smallest gap height. The bearing coefficients (which are different again from those in equations 1, 2 and 4, 5) are denoted by K_1' and K_2'.

Dividing one condition into the other gives

$$\frac{W}{P} = \frac{K_1'}{K_2'\omega h_2} \geqslant \frac{10}{1.5} \text{ (seconds/metre)},$$

from which it follows that

$$h_2 \leqslant \frac{1.5}{10}\,\frac{1}{\omega}\,\frac{K_1'}{K_2'} \text{ (metres)}.$$

A Michell bearing was the type best fitted to the purpose. Inserting $\omega \approx 5000$ rad/s and optimum values for K_1' and K_2', which calculation reveals to be 0.06 and 0.81 respectively, we obtain

$$h_2 \leqslant 2 \text{ μm}.$$

So small a gap height may well be feasible, in view of the American achievements in connection with radial bearings [5]), but at the time we felt very doubtful about the possibilities of building a bearing with such a small diametral clearance, and the prospect of doing so was not in any case an attractive one.

What else could be done? We were now reduced to looking for a bearing with a different kind of geometry, and having different coefficients, in order to yield a bigger K_1'/K_2' ratio. In physical terms, the problem was essentially one of reducing the coefficient of friction (see formula (3) in Part I).

The spiral groove idea

H. Rinia, of the Research Laboratories, suggested an arrangement in which a smooth circular plate rotated relative to a similar plate in which grooves had been cut (*fig. 17*). Lying oblique to the relative velocity vector of the two plates, the grooves would have a twofold effect. Firstly there would be a convergence effect; looked at in the direction of the velocity vector, the air or other fluid would be pushed out of the grooves into the smaller gap above the ridges. ~~This pump effect can be directed towards or away from the centre of rotation, depending on the direction of rotation and on the grooves.~~ Secondly, provided rotation was in the right sense the fluid would be pumped towards the centre, in consequence of the angle made by the grooves with the velocity vector at any given point, resulting in the plates being pushed apart. Having built the spiral groove bearing illustrated in fig. 17, we found that it sufficed to turn the upper plate by hand (at a speed of about one

[17]) H. Rinia, Verslag gewone Vergad. Afd. Natuurk. Kon. Ned. Akad. Wetensch. (in Dutch), 70, 144, 1961 (No. 10).

revolution per second) in order to built up an air layer about 11 μm thick completely separating the two bearing surfaces.

The reader may wonder just what advantage, in terms of reduced friction (which was what we were after), is provided by the combination of effects just referred to. The situation can be summed up by saying that, whereas pressure is zero in the centre of the Michell bearing, because the relative velocity is zero at that point, the pumping action of the spiral groove bearing results in a pressure distribution that shows a sharp increase along a line drawn from the perimeter to the centre. In the spiral groove bearing, therefore, a bigger area contributes to load-carrying capacity *without this entailing any great increase in frictional losses.* A more quantitative account of the difference between the two types will be given below.

As we later discovered, work had already been done in various quarters on developing the spiral groove bearing. The first reference in the literature is a description of the principle by Gümbel in 1925 [18]); Gümbel did not apply for a patent. Whipple carried out the first calculations in 1949, and some time elapsed before these were allowed to be published. It was on the basis of his calculations that a spiral groove bearing was designed for the CO_2 circulator which was discussed in Part II and which is employed in reactor cooling systems [19]).

It will however be as well to point out that Whipple undertook his calculations with an aim rather different from ours. He was looking into the possibilities of making a gas bearing that would be of simpler construction than the Michell bearing. He does not state explicitly that spiral groove bearings represent a possibility of reducing friction coefficients, and it is not therefore remarkable that he did not devote any further effort to the theoretical and experimental elaboration of this possibility.

Optimization of a spiral groove bearing

Now that a general line of investigation had been opened up, our main concern was to design a bearing with optimum performance. The first experimental version was very far from satisfying the requirements that had been laid down. A brief account will now be given of the calculations that were undertaken with a view to arriving at optimum load-carrying capacity. Frictional losses as well as load-carrying capacity were taken into account in the full calculations [20]).

We started by considering a model consisting of rectilinear grooves and ridges — see *fig. 18.* Assuming steady-state operation of the model in an incompressible working fluid, the pressure distribution therein is described by the Laplace differential equation:

$$\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} = 0 \,. \qquad \qquad (7)$$

No analytical solution can be found for the relevant boundary conditions. The simplest approximation was to consider two linear functions that would satisfy the differential equation, one giving the pressure $p_{\rm d}$ above a ridge and the other giving the pressure $p_{\rm g}$ above a groove.

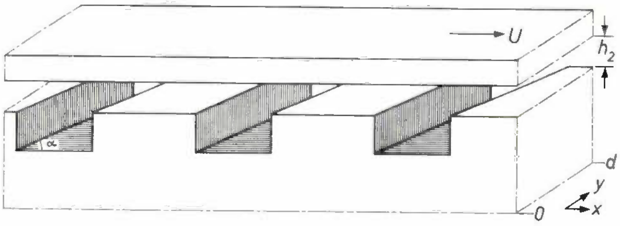


Fig. 18. Rectilinear groove model used in calculations for optimum spiral groove bearing.

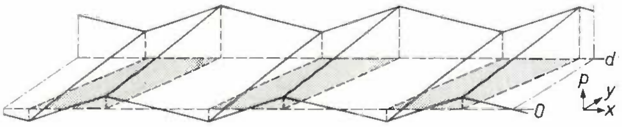


Fig. 19. First approximation to the distribution of pressures in the model appearing in fig. 18. This approximation still does not satisfy all the boundary conditions of the differential equation. By superimposing a correction which has been derived elsewhere, it is possible to obtain a distribution in which pressures are uniform along edges $y = 0$ and $y = d$, as they must necessarily be in reality.

$$p_{\rm d} = p_1 + A_1x + B_1y \,, \qquad \qquad (8)$$

$$p_{\rm g} = p_2 + A_2x + B_2y \,. \qquad \qquad (9)$$

These two functions do not suffice to satisfy the boundary conditions at $y = 0$ and $y = d$; the approximation implies a sawtooth variation in pressure along these coordinates, as in *fig. 19.* The six parameters occurring in the above two functions were evaluated from the six remaining boundary

[18]) L. Gümbel and E. Everling, Reibung und Schmierung im Maschinenbau, Krayn Verlag, Berlin 1925. In particular, see p. 81. To judge from the fact that the spiral groove bearing figures in this book as an exercise problem, Gümbel did not rate its practical importance very high, and it may be that he underestimated the difficulties of exhaustive theoretical treatment.

[19]) See Ford, Harris and Pantall [2]) and also R. T. P. Whipple, The inclined groove bearing, A.E.R.E., T/R 622, revised Oct. 1958, and the same author's contribution to First int. Symp. on gas-lubricated bearings, Washington 1959, pp. 361-382.

[20]) For a more complete account of the solution to this problem, see E. A. Muijderman, De Ingenieur 75, W 35, 1963 (no. 10) and the same author's thesis, "Spiral groove bearings" (University of Delft, 1964). The theory was elaborated in close collaboration with J. A. Haringx of this laboratory.

conditions, deducible from the need for continuity and from the fact that equations (8) and (9) must necessarily yield the same values for pressures along the edge between ridge and groove-wall. We used the parameters thus obtained to calculate Δp, the mean difference of pressure between $y = 0$ and $y = d$. In the case where there is no net flow of working fluid in the y direction:

$$\Delta p = g(\alpha,\delta,\gamma)\,\frac{\eta U d}{h_2^2}\,. \quad . \quad . \quad . \quad . \qquad (10)$$

Here η is the viscosity of the fluid, U the relative velocity of the bearing surfaces, h_2 the gap height above the ridges, and $g(\alpha,\delta,\gamma)$ is a function of the angle α at which the grooves are cut, the ratio δ between the groove depth and h_2, and the ratio γ between the groove width and the ridge width. What we had to do was to choose α, δ and γ in such a way as to get the largest possible $g(\alpha,\delta,\gamma)$. We were able to deduce that γ must in any case be unity. In *fig. 20*, $g(\alpha,\delta,1)$ has been plotted as a function of δ for various values of α. It will be noted that $g(\alpha,\delta,1)$ reaches a peak when $\alpha = 16°$ and $\delta = 0.4$.

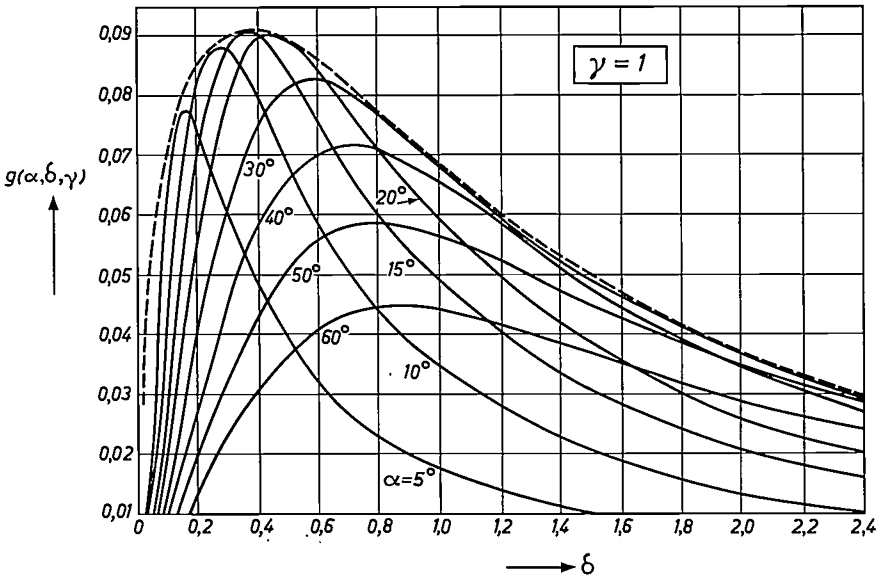


Fig. 20. Function $g(\alpha,\delta,\gamma)$ plotted against δ for $\gamma = 1$ and a set of different α values. An envelope enclosing this family of curves has been drawn as a dashed line. Inspection shows that the highest peak is that of the curve for $\alpha \approx 16°$, this maximum being attained at $\delta = 0.4$.

In order to be able to apply findings from the simple model to a bearing plate in which spiral grooves had been cut, we imagined the latter to be broken up into a large number of small elements, each of which resembled the model. Each element had to contribute as much as possible to the lift afforded by the bearing plate; consequently each spiral groove had to be cut in such a way that at any point along its length, a line drawn tangentially to the groove would make an angle of 16° with the tangential direction; in other words, the groove had to have a logarithmic spiral shape. Also, at any point along the groove, the groove width must equal the ridge width ($\gamma = 1$). The groove depth was fixed by $\delta = 0.4$.

So far we had been working on the sole basis of equations (8) and (9), which only provide an approximation to the pressure distribution. For the manner in which corrections were introduced, reference should be made to the more detailed publication already cited [20]), which also explains how the frictional couple in the actual spiral groove bearing was calculated.

The rear part of *fig. 21* shows the pressure distribution in an optimum spiral groove bearing employed to support a shaft-end. It will be seen that a "pressure hill", whose shape is somewhat reminiscent of an old-fashioned pudding basin, builds up over the whole circular bearing surface. To allow comparison, the pressure distribution over a Michell bearing consisting of six slipper blocks has been drawn into the front part of the figure. A separate "pressure hill" arises on top of each block, and it will be quite obvious that integrating the pressures above the whole set of slipper blocks will yield a much weaker total lift than integrating over the pudding-shaped hill created by the spiral groove bearing. Yet the total area of bearing surface occupied by the hills in the Michell bearing is not much less (about $\frac{3}{4}$) than that occupied by the single hill in the spiral groove bearing, and it is this area (because here the

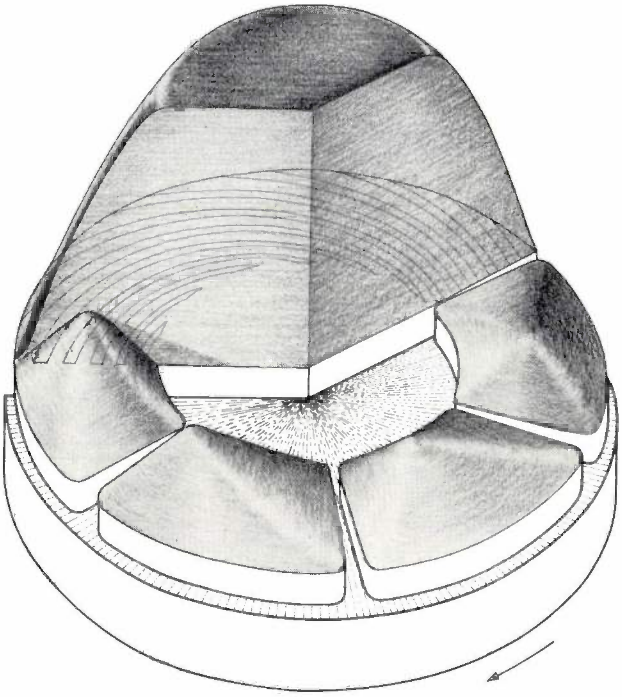


Fig. 21. Three-dimensional sketch showing calculated pressure distributions in an optimum spiral groove bearing (at the rear) and in a Michell bearing of the same size, equipped with a total of six slipper blocks (in front). A smooth upper member, rotating in the sense indicated by the arrow, is to be imagined above the fixed member with its grooves (or slipper blocks). The pressure prevailing at each point on the circular surface of the lower member has been plotted on a vertical scale that is the same for the two types of bearing. The volume of the big "pressure hill" or the total volume of the smaller "hill" is a measure of load-carrying capacity.

layer is the smallest) that determines the frictional losses.

Checking the theory

Measurements to check the theory were done with the apparatus shown schematically in *fig. 22.* The upper plate *a*, the one with the spiral grooves cut in it, was made to rotate at various speeds relative to the smooth plate *b*. A small ball bearing *c* was responsible for keeping the two plates properly centered. The couple transferred to the lower member by friction was measured with a small dial-reading dynamometer. The shaft was suspended both radially and axially with compressed air, in the manner described in Parts I and II of this article. By means of capacitance measurements we were able to determine the height of the gap between the two members of the spiral groove bearing.

Measurements of gap height at a given lift and speed yielded values roughly 12% in excess of those originally calculated. After working out the correction terms and improving the accuracy of the measurements, we managed to reduce the disparity to 1%; ultimately, then, the agreement between theory and experiment could be regarded as highly satisfactory. Theoretical and experimental values of frictional torque were found on average to differ by 6%.

Results

The optimum spiral groove bearing that we designed and made along these lines has a K_1' coefficient greater by a factor of 8.6 than that of a Michell bearing used to support a shaft-end. The K_2' coefficient is also greater, but only by a factor of 1.2. This means that we have managed to make a bearing with a coefficient of friction 7.4 times smaller than that of a Michell bearing with the same radius and clearance. So far there is no other bearing in the self-acting class that has such a low coefficient of friction compared to the spiral groove type. Using the spiral groove bearing in question, the maximum clearance imposed by the problem outlined on p. 266 becomes $h_2 \leqslant 15$ µm. This condition can be fulfilled without great trouble.

We have achieved a further increase in the relatively high load-carrying capacity of the spiral groove

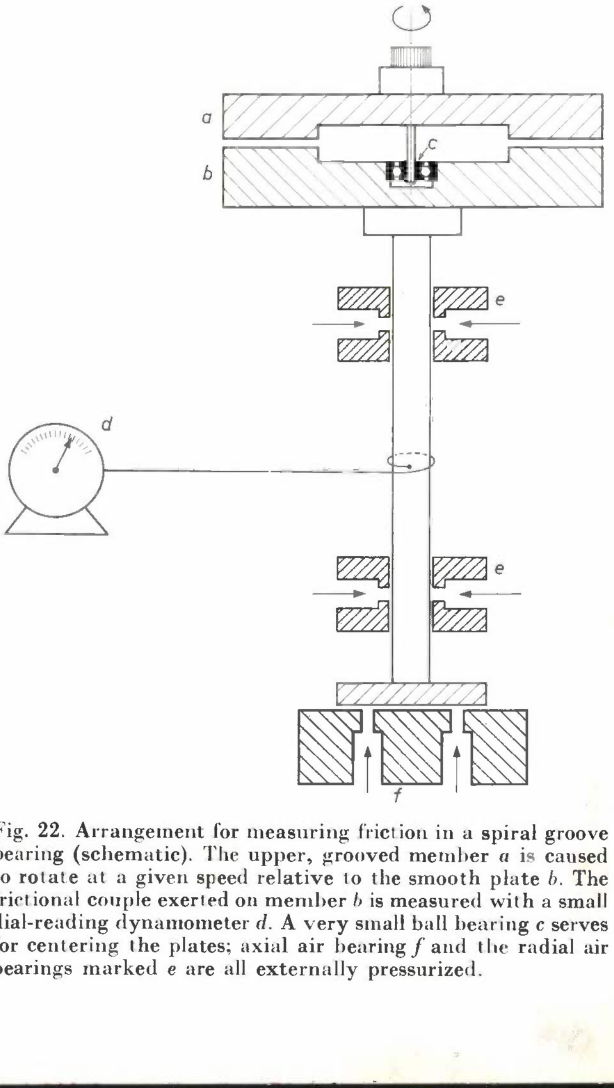
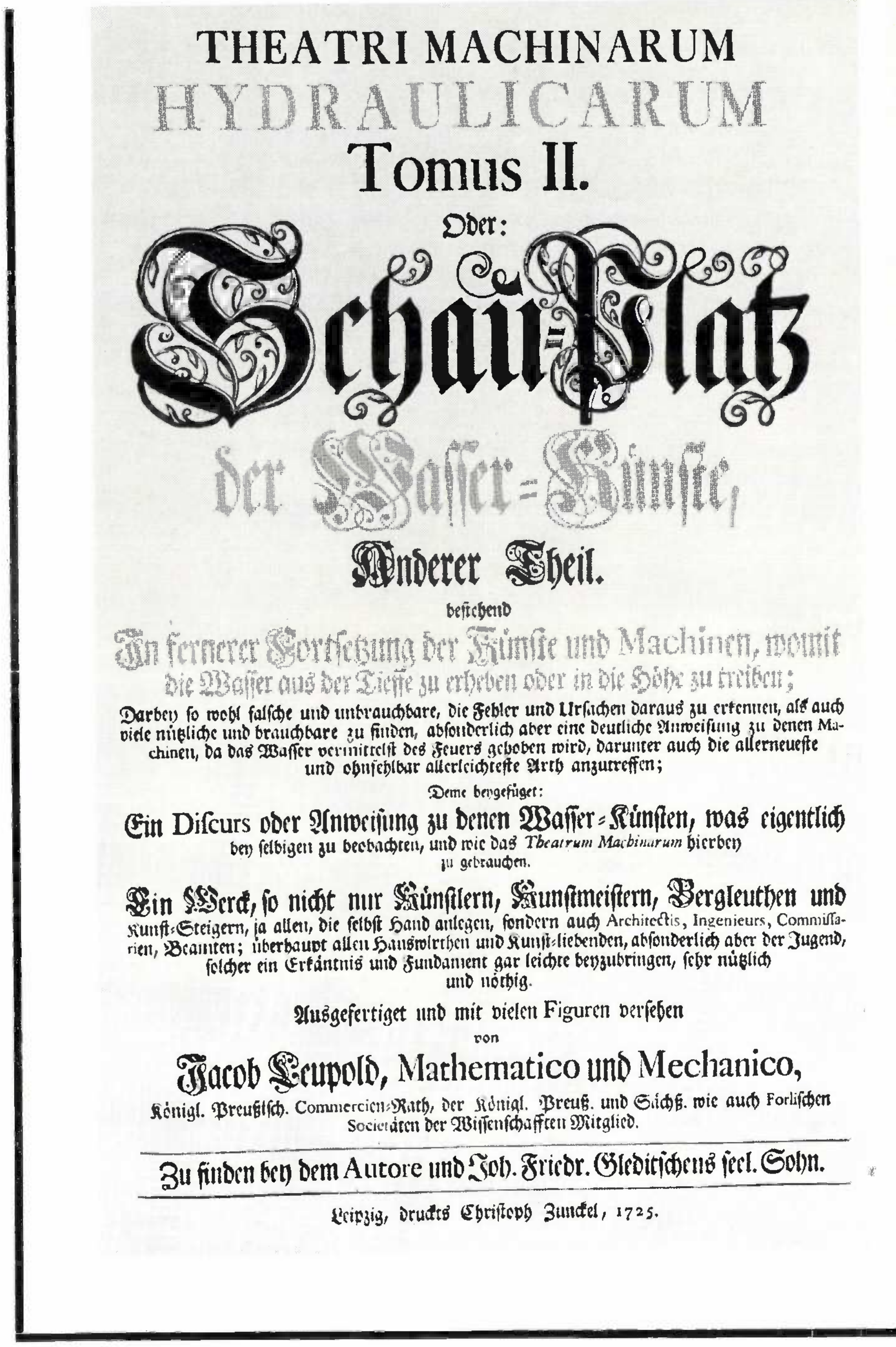


Fig. 22. Arrangement for measuring friction in a spiral groove bearing (schematic). The upper, grooved member *a* is caused to rotate at a given speed relative to the smooth plate *b*. The frictional couple exerted on member *b* is measured with a small dial-reading dynamometer *d*. A very small ball bearing *c* serves for centering the plates; axial air bearing *f* and the radial air bearings marked *e* are all externally pressurized.

bearing (it will be recalled that its K_2' is 8.6 times that of a corresponding Michell bearing) by redesigning the bearing surface in the form of a spherical cap. If a flat thrust bearing be compared with a spherical one having the same radius and operating at the same speed, it can readily be seen that the surface elements of the latter have, on average, a greater velocity relative to the fixed bearing surface. Moreover, the distances from centre involved in integrating over the surface elements are also longer in the case of the spherical bearing. As can be directly inferred from formula (1a), both factors make for greater load-carrying capacity. Indeed, this becomes so high as to amount to about 25 times that of a Michell bearing of the same diameter.

Other favourable features of a spherical as compared with a plane spiral groove bearing are that it is easier to centre and that, to some limited extent, it is able to take up a radial as well as an axial load.

Fig. 23 shows two spherical type spiral groove bearings using oil lubrication which have been made and tested in our laboratory. The diameter of the smallest spherical cap is about 3 mm; the cap and its concave seating are completely separated by an air film at 5000 r.p.m., with the bearing taking up a thrust of 1 kg. Under these conditions the pressure developed at the centre of the bearing is about 25 atm.



THEATRI MACHINARUM
HYDRAULICARUM
Tomus II.
Oder:
Schau=Platz
der Waſſer=Künſte,
Anderer Theil.

beſtehend

In fernerer Fortſetzung der Künſte und Machinen, womit die Waſſer aus der Tieffe zu erheben oder in die Höhe zu treiben;

Darbey ſo wohl falſche und unbrauchbare, die Fehler und Urſachen daraus zu erkennen, als auch viele nützliche und brauchbare zu finden, abſonderlich aber eine deutliche Anweiſung zu denen Machinen, da das Waſſer vermittelſt des Feuers gehoben wird, darunter auch die allerneueſte und ohnfehlbar allerleichteſte Arth anzutreffen;

Deme beygefüget:

Ein Diſcurs oder Anweiſung zu denen Waſſer=Künſten, was eigentlich bey ſelbigen zu beobachten, und wie das *Theatrum Machinarum* hierbey zu gebrauchen.

Ein Werck, ſo nicht nur Künſtlern, Kunſtmeiſtern, Bergleuthen und Kunſt=Steigern, ja allen, die ſelbſt Hand anlegen, ſondern auch Architectis, Ingenieurs, Commiſſarien, Beamten; überhaupt allen Hauswirthen und Kunſt=liebenden, abſonderlich aber der Jugend, ſolcher ein Erkäntnis und Fundament gar leichte beyzubringen, ſehr nützlich und nöthig.

Ausgefertiget und mit vielen Figuren verſehen
von
Jacob Leupold, Mathematico und Mechanico,
Königl. Preußiſch. Commercien=Rath, der Königl. Preuß. und Sächß. wie auch Forliſchen Societäten der Wiſſenſchafften Mitglied.

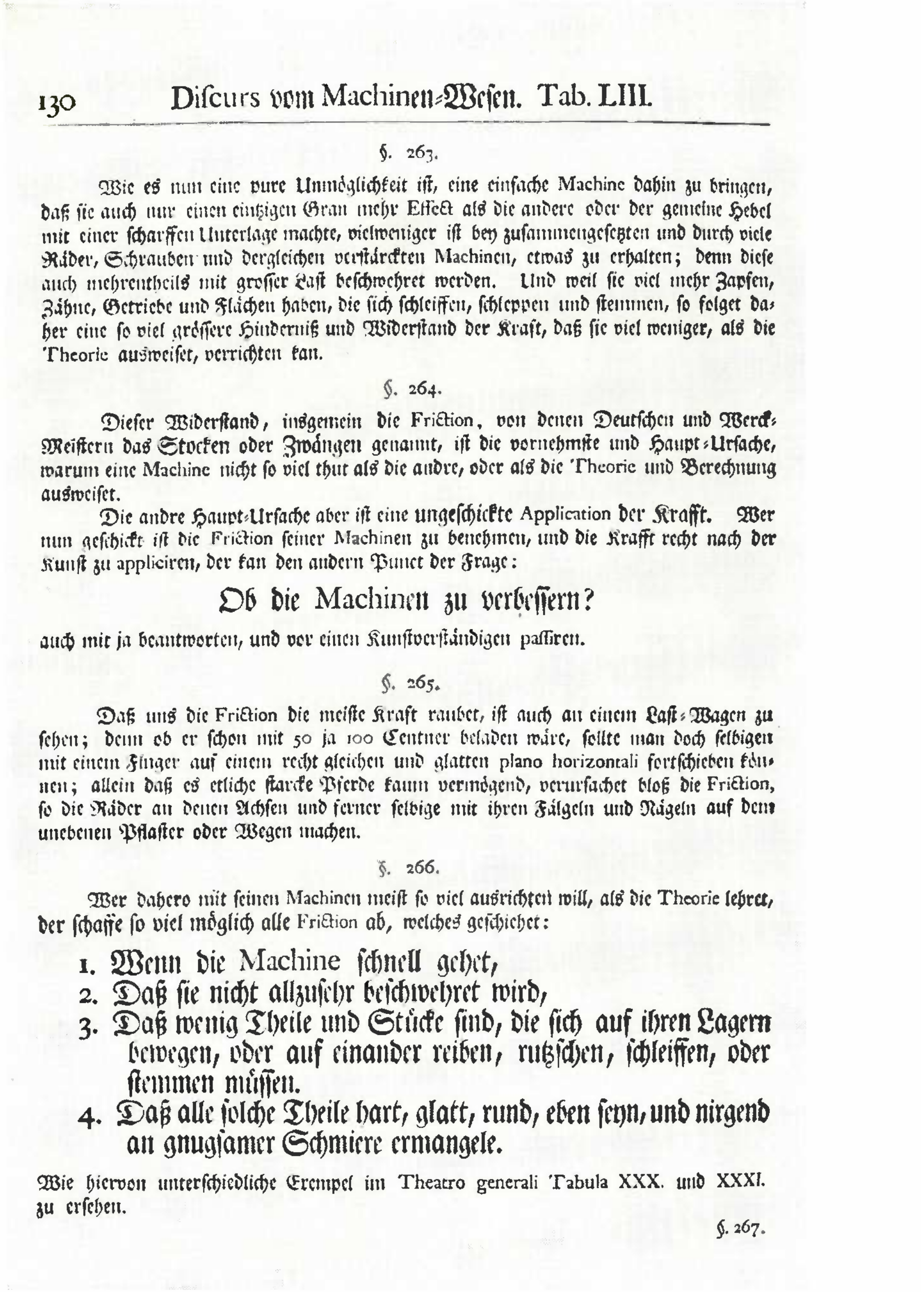
Zu finden bey dem Autore und Joh. Friedr. Gleditſchens ſeel. Sohn.

Leipzig, druckts Chriſtoph Zunckel, 1725.

Applications

The possible applications of spiral groove bearings are not limited to cases where a low coefficient of friction is essential. Another feature of the bearing which is certainly just as important, particularly where small size is required, is the ease with which they can be made. A diameter of 3 mm, such as the smaller of the bearings shown in fig. 23, would be quite impossible for Michell-type bearings.

Spiral groove bearings can also be designed to take up thrust at points along the length of the shaft (instead of at the ends). Leakage takes place in consequence of the shaft passing through the bearing, the pumping action of which no longer serves to maintain a high pressure near the shaft; thus the pudding-shaped pressure hill collapses in the middle, becoming ring-shaped. A remedy for the loss of pressure has been found in the partial and herringbone types of spiral groove bearing (*fig. 24*). In the former type an effort has been made to limit leakage by leaving a large area ungrooved in the middle of the plate. The grooved outer portion, which provides the



130 Diſcurs vom Machinen-Weſen. Tab. LIII.

§. 263.

Wie es nun eine pure Unmöglichkeit iſt, eine einfache Machine dahin zu bringen, daß ſie auch nur einen eintzigen Gran mehr Effect als die andere oder der gemeine Hebel mit einer ſcharffen Unterlage machte, vielweniger iſt bey zuſammengeſetzten und durch viele Räder, Schrauben und dergleichen verſtärckten Machinen, etwas zu erhalten; denn dieſe auch mehrentheils mit groſſer Laſt beſchwehret werden. Und weil ſie viel mehr Zapfen, Zähne, Getriebe und Flächen haben, die ſich ſchleiffen, ſchleppen und ſtemmen, ſo folget daher eine ſo viel gröſſere Hinderniß und Widerſtand der Kraft, daß ſie viel weniger, als die Theorie ausweiſet, verrichten kan.

§. 264.

Dieſer Widerſtand, insgemein die Friction, von denen Deutſchen und Werck-Meiſtern das Stocken oder Zwängen genannt, iſt die vornehmſte und Haupt-Urſache, warum eine Machine nicht ſo viel thut als die andre, oder als die Theorie und Berechnung ausweiſet.

Die andre Haupt-Urſache aber iſt eine ungeſchickte Application der Krafft. Wer nun geſchickt iſt die Friction ſeiner Machinen zu benehmen, und die Krafft recht nach der Kunſt zu appliciren, der kan den andern Punct der Frage:

Ob die Machinen zu verbeſſern?

auch mit ja beantworten, und vor einen Kunſtverſtändigen paſſiren.

§. 265.

Daß uns die Friction die meiſte Kraft raubet, iſt auch an einem Laſt-Wagen zu ſehen; denn ob er ſchon mit 50 ja 100 Centner beladen wäre, ſollte man doch ſelbigen mit einem Finger auf einem recht gleichen und glatten plano horizontali fortſchieben können; allein daß es etliche ſtarcke Pferde kaum vermögend, verurſachet bloß die Friction, ſo die Räder an denen Achſen und ferner ſelbige mit ihren Fälgeln und Nägeln auf dem unebenen Pflaſter oder Wegen machen.

§. 266.

Wer dahero mit ſeinen Machinen meiſt ſo viel ausrichten will, als die Theorie lehret, der ſchaffe ſo viel möglich alle Friction ab, welches geſchiehet:

1. **Wenn die Machine ſchnell gehet,**
2. **Daß ſie nicht allzuſehr beſchwehret wird,**
3. **Daß wenig Theile und Stücke ſind, die ſich auf ihren Lagern bewegen, oder auf einander reiben, rutzſchen, ſchleiffen, oder ſtemmen müſſen.**
4. **Daß alle ſolche Theile hart, glatt, rund, eben ſeyn, und nirgend an gnugſamer Schmiere ermangele.**

Wie hiervon unterſchiedliche Exempel im Theatro generali Tabula XXX. und XXXI. zu erſehen.

§. 267.

In this article in this issue, E. A. Muijderman cites some examples of expedients adopted in antiquity to reduce friction — the use of oil to make a heavily laden sledge run more easily, and the primitive ball-bearing fitted under a turntable. Now, over the long centuries during which Western technology was evolving, a point must sooner or later have been reached at which designers of machines, and for even greater reason those who had to build them, became all too aware of the important part that frictional losses were going to play in their ever more complicated and ambitious constructions. This awareness is clearly reflected in the writings of Jacob Leupold, a Prussian "Mathematico and Mechanico" (i.e. engineer), who published his *Theatrum Machinarum Hydraulicarum* at Leipzig in 1725. The book is largely devoted to the fundamental industrial problems of the time, dealing with the design of water-raising machines such as were required for pumping out mineshafts or supplying towns, estates or fountains. Leupold never neglects an opportunity of emphasizing the importance of reducing friction between the moving parts of machines. Here we reproduce the title-page of the second part of his work, the picturesque wording of which is characteristic of the age, and the page of text in which Leupold formulates his doctrines with forceful clarity. His advice to let the machine run fast (point 1 of § 266) should not be interpreted as a remarkable example of foresight into the operation of self-acting contactless bearings, for which a certain minimum shaft speed is necessary; what he had in mind is that slow and none too regular motion is liable to involve sticking and jamming ("Stocken oder Zwängen"). During these brief intervals of motionlessness, static friction becomes operative in place of dynamic friction.

pumping action, is thus separated by an area of high flow resistance from the central opening for the shaft, through which leakage occurs. Leakage has been cut out altogether in the herringbone bearing, an opposite pumping action being provided by the grooves on the inside. Since the area over which the pressure hill builds up is rather more restricted in these types of bearing, they are not of course ideal from the viewpoint of reducing friction. Even so, the coefficient of friction of a herringbone bearing is still rather smaller than that of a Michell bearing under comparable conditions. A herringbone bearing that has been incorporated in an experimental hot gas engine [21]) in the Research Laboratories may be seen in *fig. 25*.

Another interesting variant is the conical type of spiral groove bearing. Its most notable property is its ability to take up a radial load as well as a thrust; as already stated, the spherical type is also able to do this, but only to a much lower degree. (Theoretical

[21]) The practical development work was undertaken by R. J. Meijer and coworkers of this laboratory.

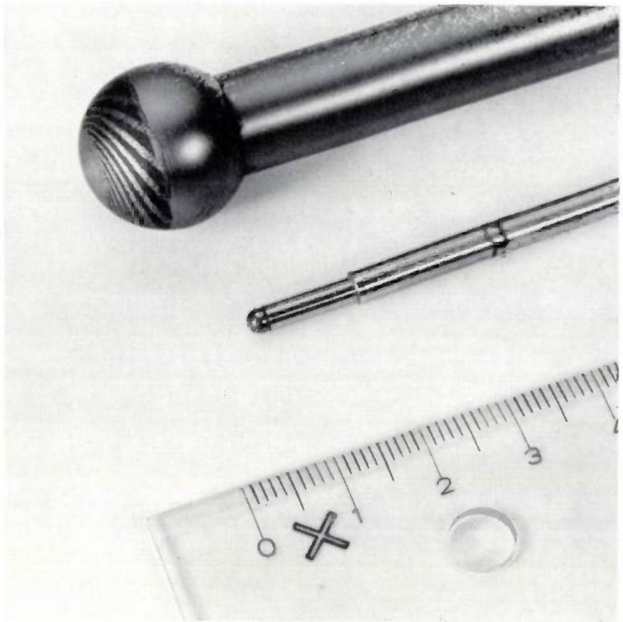


Fig. 23. Spindle-ends for spherical-type spiral groove bearings with diameters of 18 mm and 3 mm. The grooves are cut into hardened and polished steel balls (such as are used in ball-bearings) by an etching process. The surface of the metal is coated with light-sensitive lacquer, and the desired spiral groove pattern is transferred to this layer by a photographic printing method. Subsequent treatment in an etching bath affects only those areas of the layer that have been exposed to light, and only the steel under the exposed areas is attacked by the etchant (which might be $FeCl_3$, for example).

Alternatively, the grooves can be cut into the concave member of the bearing. In this case the seating might be made of aluminium, for which $FeCl_3$ is also a suitable etchant.

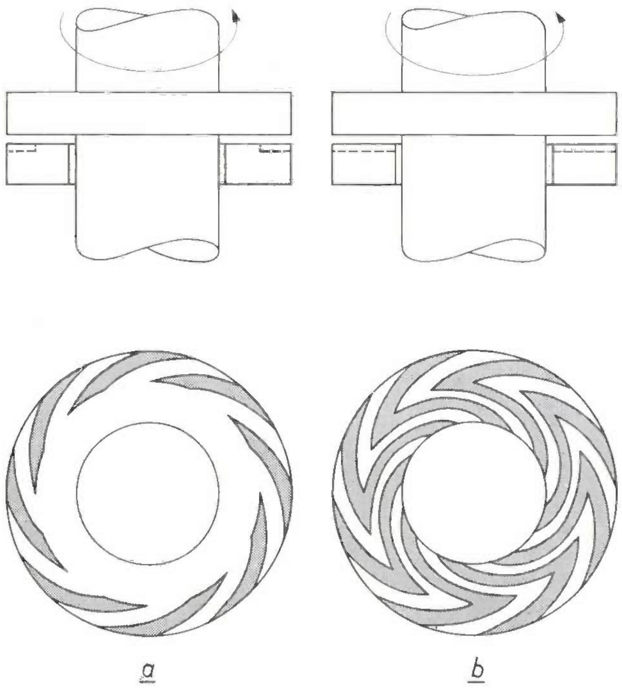


Fig. 24. Two types of spiral groove bearings suitable for taking up axial thrust from a shaft that passes through the bearing. (*a*) Partial spiral groove bearing, in which the grooved area is restricted to the outer margin of the plate. (*b*) Herringbone bearing.

studies have shown that nothing is gained, from the viewpoint of reduced friction, by cutting spiral grooves in a "pure" radial bearing.) *Fig. 26* shows a conical spiral groove bearing that was recently designed in our laboratory for a small electric motor.

Optimum bearing coefficients for all the types discussed above are displayed in *Table I*.

Lubrication with grease

In conclusion, a word may be said about an unexpected finding that has engaged much of our attention over the past few years — the discovery that there is a lot to be gained by lubricating small spiral groove bearings with *grease*.

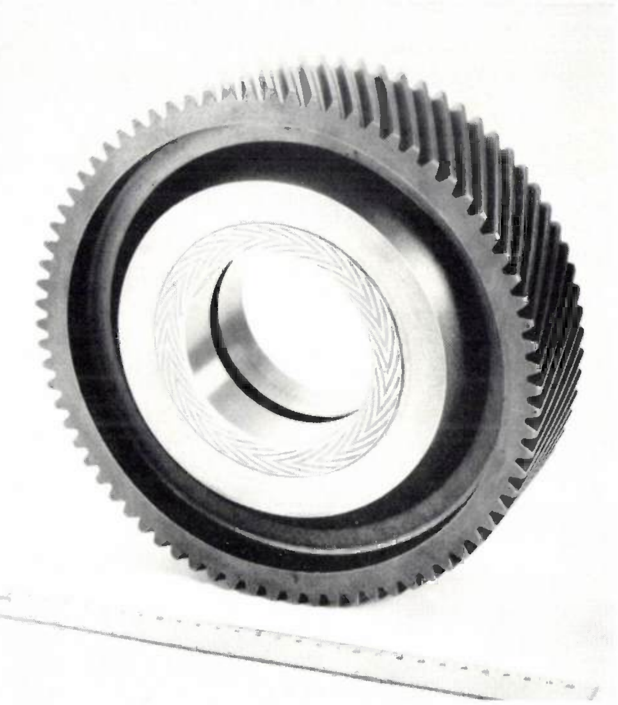
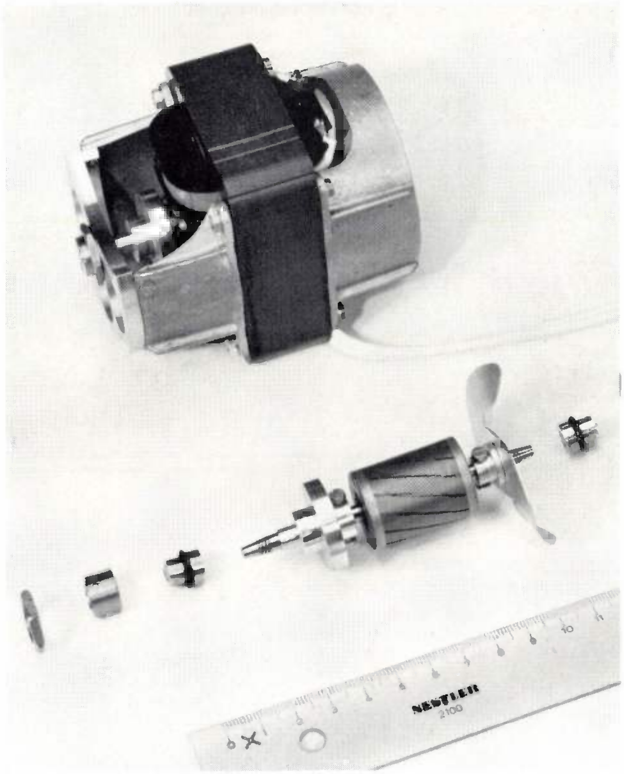


Fig. 25. A herringbone bearing employed in a hot gas engine developed in the Research Laboratories [21].

In bearings with diameters of only a few millimetres, lubrication starts to become a problem in itself. The provision of continuous or drip lubrication becomes disproportionately complicated and expensive. What we have particularly in mind here are small domestic and other appliances which are in operation at irregular intervals, and which must be able to run or to rest in any conceivable position.

If grease is used as a lubricant it will be sucked into the bearing clearance, just as oil is, in virtue of the pumping action of the spiral groove bearing; but it is obvious that grease will be far less liable than oil is to escape from the bearing during periods of rest. We have in fact established that after long periods of immobility in various



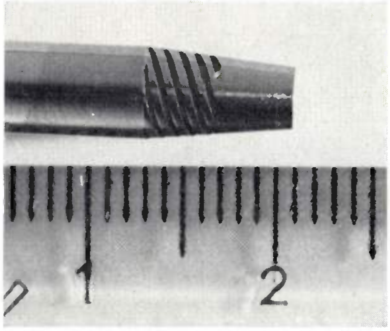
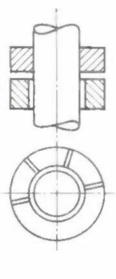
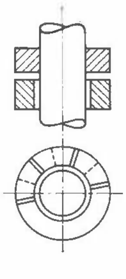
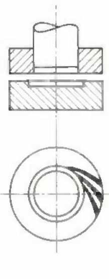
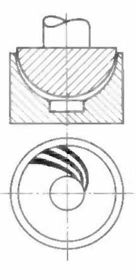
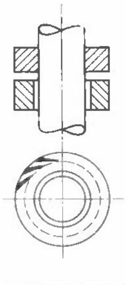
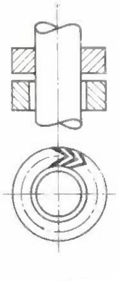


Fig. 26. Experimental type of miniature electric motor embodying a conical spiral groove bearing able to carry appreciable radial loads as well as taking up thrust. In front of it, the rotor of another motor of the same type, dismantled to give an "exploded" view. The photograph above shows one of the grooved spindle-ends. The bearing is grease lubricated.

positions, periods which had ranged from a few months to a year, lubrication conditions on restarting left nothing to be desired, complete separation of the bearing surfaces having occurred at the same relatively slow speeds as had been noted during the first trials. Further investigations are still proceeding.

One point we are interested in, in this connection, is whether the grease retains its good lubricating properties when the bearing has been in operation for a lengthy time; whether its chemical composition alters, for example, or whether it is liable to drain out of the bearing in the course of a long hot spell during the summer. Ad hoc tests that have extended over periods varying from a few months to two years have so far failed to reveal any shortcoming whatsoever in this respect.

Table I. Various forms of thrust bearing, with calculated values of their K_1' and K_2' coefficients appropriate to optimum running conditions. All types are assumed to have the same outside diameter; that of flat bearings is twice the inside diameter. Half-diametral clearance (half the difference between the cup diameter and the ball diameter) is assumed in the case of the spherical cap bearing.

	Michell bearing	step bearing	spiral groove bearings			
			flat type for shaft-end	spherical cap	for mid-shaft support	
					partial	herringbone
						
K_1'	0.042	0.047	0.366	1.11	0.124	0.106
K_2'	22.0	ca. 25	2.9	2.8	8.1	10.2

Though they will still have to undergo the test of time in practical applications, we personally feel that these miniature grease-lubricated contactless bearings have good future prospects.

Summary. In this article types of bearing are discussed in which friction and wear are reduced to very small proportions by the agency of a viscous fluid (lubricant) that serves to maintain a small and more or less constant clearance between the bearing surfaces. A "contactless" bearing of this kind does not necessarily have to be lubricated with a *liquid*, such as oil; a gas is also effective. Particular attention is given to gas-lubricated bearings, which have been a practical possibility for the past 15 years or so. A distinction is made between two classes of contactless bearing, the externally pressurized types and the self-acting types. In a comparative account of the two classes, some indication is given of the applications for which each is most suitable. Examples of gas bearing applications may be found in turbo-jet engines, gyroscopic compasses, nuclear reactors, dentists' drills etc.

On the basis of the self-acting contactless bearing a new type — the spiral groove bearing — has been developed, this work having largely been done in Philips Research Laboratories at Eindhoven. The main advantage of the new type is its very low coefficient of friction. A flat thrust bearing of the spiral groove type, designed to support a shaft-end, has a friction coefficient only about one-seventh that of a Michell bearing under comparable optimum conditions. Besides a low coefficient of friction, a variant in which the grooves are cut into a spherical cap has a load-carrying capacity about 25 times greater than that of a corresponding Michell bearing. Small spiral groove bearings having diameters of a few millimeters, and lubricated with grease, promise to be useful in small domestic and other appliances.

A 4 mm RADAR INSTALLATION

by H. ALLARIES *).

621.396.967.029.65

Radar installations operating at a wavelength of 8 mm are now widely employed in cases where a high degree of detail discrimination is required within a restricted area, e.g. for manoeuvring shipping in narrow and busy water-ways. This article describes an experimental 4 mm radar installation, built in Philips Research Laboratories, which uses various new tubes including a 4 mm magnetron and a 4 mm reflex klystron. The screen images obtained with this installation show that, although the maximum range is no more than a few kilometres, the detail discrimination at short distances is exceptionally good.

Various wavebands are at present in use for radar installations, the choice being determined by the operating range for which the installation is intended. Observations over distances up to several hundreds of kilometres require wavelengths of 25 cm [1]) and longer, distances up to about 100 km are covered with 10 cm waves, and for shorter ranges use is made of 3 cm waves and lower. In general it can be said that shortening the wavelength reduces the range of the installation. On the other hand, it improves the resolution, that is to say the power to discriminate on the radar screen between closely adjacent objects.

For certain special applications, for example where ships have to be manoeuvred in narrow and busy water-ways, this has led to the use of millimetre wavelengths. An article published earlier in this journal described an 8 mm installation [2]). The range at such wavelengths is relatively small, but sufficient for the purpose. The resolution, however, is exceptionally high, making visible on the screen large numbers of details which are of importance, for example, to navigation.

The availability of several new tubes — including a 4 mm magnetron and a 4 mm klystron — led to the construction in our Research Laboratories of a 4 mm radar installation. Part of it can be seen in *fig. 1.* This equipment has been used to investigate the effects of a further shortening of the wavelength on the resolution and range, and also the technical problems which this involves.

General design

To obtain the high estpossible *radial resolution*, the transmitted radar pulses must be extremely short. On the other hand, the magnetron cannot work properly below a certain minimum pulse length. In practice the rule of thumb is used that the number of cycles in a pulse must be equal to the figure of merit Q of the cavity resonator system of the magnetron under optimum loading conditions. For most magnetrons this figure is roughly 200, which, at a wavelength of 4 mm, gives a minimum pulse length of 4 nanoseconds. The pulse length actually chosen is 5 nanoseconds. Theoretically this should result in a radial resolution of 0.75 m, but in practice it will be somewhat less favourable, owing to inevitable pulse distortion.

*) Philips Research Laboratory, Eindhoven.

[1]) A 25 cm radar is described by H. van Lambalgen and J. van der Plas in: The SGR 200 long-range surveillance radar at Schiphol Airport, Philips Telecomm, Rev. 22, 51-62, 1961.

[2]) J. M. G. Seppen and J. Verstraten, An 8 mm-high resolution radar installation, Philips tech. Rev., 21, 92-103, 1959/60.

Though they will still have to undergo the test of time in practical applications, we personally feel that these miniature grease-lubricated contactless bearings have good future prospects.

Summary. In this article types of bearing are discussed in which friction and wear are reduced to very small proportions by the agency of a viscous fluid (lubricant) that serves to maintain a small and more or less constant clearance between the bearing surfaces. A "contactless" bearing of this kind does not necessarily have to be lubricated with a *liquid*, such as oil; a gas is also effective. Particular attention is given to gas-lubricated bearings, which have been a practical possibility for the past 15 years or so. A distinction is made between two classes of contactless bearing, the externally pressurized types and the self-acting types. In a comparative account of the two classes, some indication is given of the applications for which each is most suitable. Examples of gas bearing applications may be found in turbo-jet engines, gyroscopic compasses, nuclear reactors, dentists' drills etc.

On the basis of the self-acting contactless bearing a new type — the spiral groove bearing — has been developed, this work having largely been done in Philips Research Laboratories at Eindhoven. The main advantage of the new type is its very low coefficient of friction. A flat thrust bearing of the spiral groove type, designed to support a shaft-end, has a friction coefficient only about one-seventh that of a Michell bearing under comparable optimum conditions. Besides a low coefficient of friction, a variant in which the grooves are cut into a spherical cap has a load-carrying capacity about 25 times greater than that of a corresponding Michell bearing. Small spiral groove bearings having diameters of a few millimeters, and lubricated with grease, promise to be useful in small domestic and other appliances.

A 4 mm RADAR INSTALLATION

by H. ALLARIES *).

621.396.967.029.65

Radar installations operating at a wavelength of 8 mm are now widely employed in cases where a high degree of detail discrimination is required within a restricted area, e.g. for manoeuvring shipping in narrow and busy water-ways. This article describes an experimental 4 mm radar installation, built in Philips Research Laboratories, which uses various new tubes including a 4 mm magnetron and a 4 mm reflex klystron. The screen images obtained with this installation show that, although the maximum range is no more than a few kilometres, the detail discrimination at short distances is exceptionally good.

Various wavebands are at present in use for radar installations, the choice being determined by the operating range for which the installation is intended. Observations over distances up to several hundreds of kilometres require wavelengths of 25 cm [1]) and longer, distances up to about 100 km are covered with 10 cm waves, and for shorter ranges use is made of 3 cm waves and lower. In general it can be said that shortening the wavelength reduces the range of the installation. On the other hand, it improves the resolution, that is to say the power to discriminate on the radar screen between closely adjacent objects.

For certain special applications, for example where ships have to be manoeuvred in narrow and busy water-ways, this has led to the use of millimetre wavelengths. An article published earlier in this journal described an 8 mm installation [2]). The range at such wavelengths is relatively small, but sufficient for the purpose. The resolution, however, is exceptionally high, making visible on the screen large numbers of details which are of importance, for example, to navigation.

The availability of several new tubes — including a 4 mm magnetron and a 4 mm klystron — led to the construction in our Research Laboratories of a 4 mm radar installation. Part of it can be seen in *fig. 1.* This equipment has been used to investigate the effects of a further shortening of the wavelength on the resolution and range, and also the technical problems which this involves.

General design

To obtain the high estpossible *radial resolution*, the transmitted radar pulses must be extremely short. On the other hand, the magnetron cannot work properly below a certain minimum pulse length. In practice the rule of thumb is used that the number of cycles in a pulse must be equal to the figure of merit Q of the cavity resonator system of the magnetron under optimum loading conditions. For most magnetrons this figure is roughly 200, which, at a wavelength of 4 mm, gives a minimum pulse length of 4 nanoseconds. The pulse length actually chosen is 5 nanoseconds. Theoretically this should result in a radial resolution of 0.75 m, but in practice it will be somewhat less favourable, owing to inevitable pulse distortion.

*) Philips Research Laboratory, Eindhoven.

[1]) A 25 cm radar is described by H. van Lambalgen and J. van der Plas in: The SGR 200 long-range surveillance radar at Schiphol Airport, Philips Telecomm, Rev. 22, 51-62, 1961.

[2]) J. M. G. Seppen and J. Verstraten, An 8 mm-high resolution radar installation, Philips tech. Rev., 21, 92-103, 1959/60.

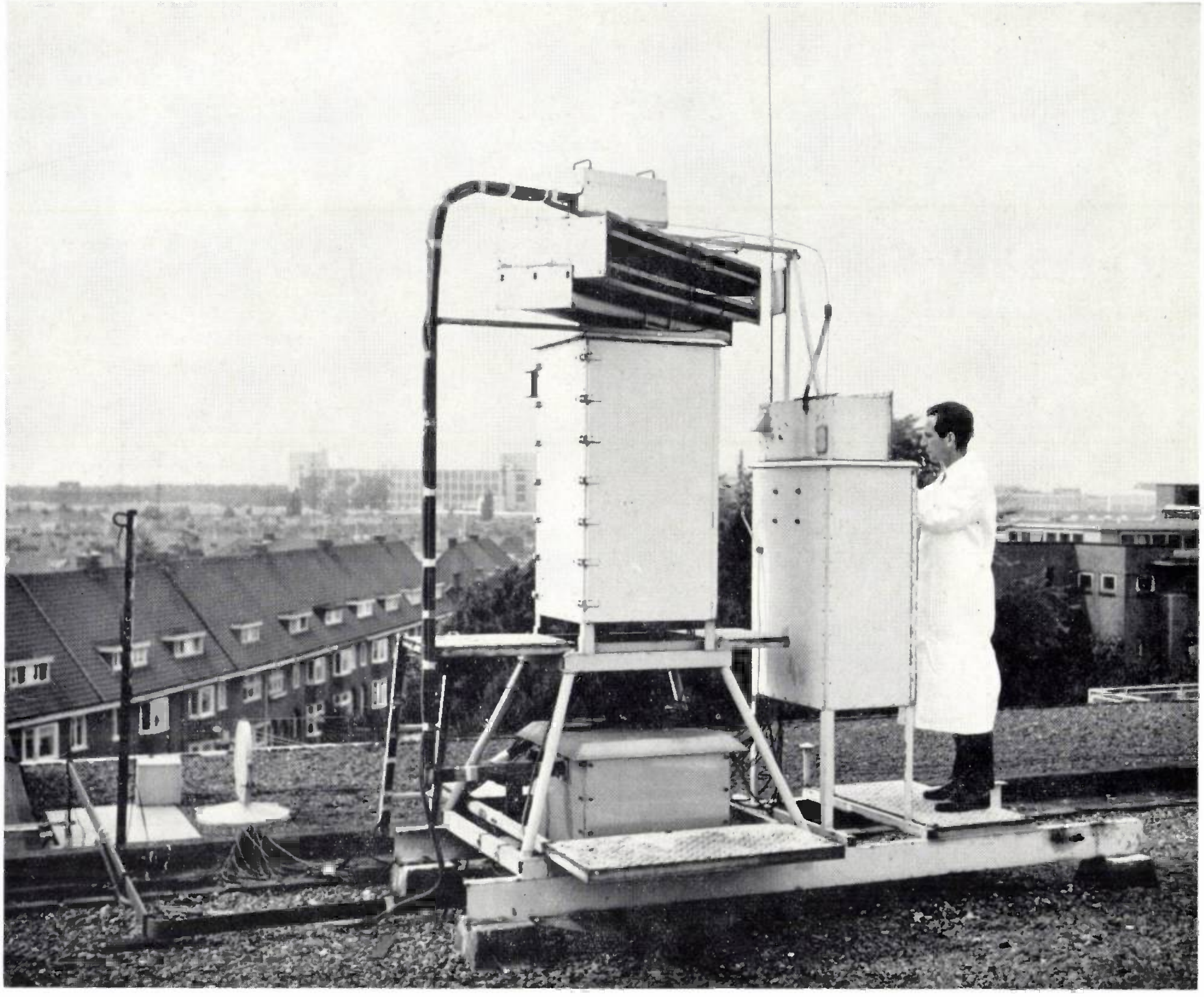


Fig. 1. Experimental 4 mm radar installation: the antenna system and part of the transmitter and receiver, mounted on the roof of the Eindhoven Research Laboratory.

The *tangential resolution* [3]) is determined at every point by the width of the transmitted beam of RF energy. In installations of the type considered here it has to be remembered that the radar beam close to the antenna has roughly constant width, which is that of the antenna itself. At a certain distance away, depending on the aperture angle of the antenna, the beam becomes wider than the antenna by divergence. On the one hand, then, the antenna cannot be made too wide without adversely affecting the tangential resolution at very short ranges; on the other hand a narrow antenna has a large aperture angle, which results in a poorer tangential resolution on longer ranges. In the present case we compromised with an antenna width of 1.5 m. In the horizontal plane the antenna has an aperture angle (beam width) of 0.0027 radians, i.e. 0.15 degrees. From this it can be calculated that the beam only becomes wider than the antenna itself at a distance of 560 m. Up to that point the tangential resolution is virtually constant and equal to 1.5 m.

At wavelengths of 3 cm and longer it is customary to use a common antenna for transmitting and receiving. To prevent the transmitted pulses from passing directly into the receiver, thereby overloading it, the receiver input during pulse transmission is short-circuited by means of a TR switch (duplexer) of the gas-discharge type. The firing (ionization) time of such switches, however, is in the order of several nanoseconds, so that with pulses as short as in millimetre radar they can provide no protection. For this reason, as in the case of the 8 mm radar installation, separate antennas are used for transmitting and receiving. Each antenna has the form of a very short parabolic cylinder, sandwiched between two flat, parallel plates ("cheese antenna", see fig. 1). Both antennas have the above

[3]) See for example S. Silver, Microwave antenna theory and design, M.I.T. Radiation Laboratory Series No. 12, McGraw-Hill, New York 1949, Chapter 1.

mentioned width of 1.5 m. The vertical aperture is 35 mm, giving a beam width in the vertical plane of 6.7°. The directive gain of one antenna is $G_0 = 45$ dB.

If full use is to be made of the theoretical resolution, the cathode ray tube must be capable of displaying sufficient detail. The tube employed has a screen diameter of 40 cm, while the spot — provided the brightness is not excessive — can be reduced to a diameter of 0.5 mm. Therefore in order to obtain the optimum radial resolution of 0.75 m on the screen, the swept area should not exceed 300 m radius. Since the range of the installation is much greater, it is possible to switch over to areas with radii of 1 km and 3 km too; the screen does not then display all available radial details. For the *tangential* resolution it can be calculated that, within a 300 m range, the screen can display more details than the radar can detect, while the reverse is again the case for the other ranges.

A further design parameter, the *repetition frequency* of the radar pulses, can be given both as a minimum and a maximum. The minimum is governed by the consideration that the antenna, while traversing an angle of rotation equal to its beam width, must transmit at least one pulse, otherwise small objects would no longer be detected. The maximum speed of revolution of the antenna was fixed at $w = 120$ r.p.m., corresponding to an angular velocity of 4π rad/s. At the above-mentioned horizontal beam width of $\Theta_\mathrm{h} = 0.0027$ radians we thus arrive at a minimum pulse repetition frequency of $f_{\mathrm{r\ min}} = w/\Theta_\mathrm{h} = 4730$ c/s.

The maximum pulse repetition frequency is governed by two considerations. In the first place it must not be smaller than the time taken by a pulse to travel the theoretical range of the radar installation, there and back. As we shall see, in our case we can calculate for the latter a value of 4.7 km. If, to be on the safe side, we take the range to be 6 km, we then find a maximum pulse repetition frequency of 25 kc/s.

A second upper limit is set to the repetition frequency by the permissible thermal loading of the magnetron, or, to be more exact, by the permissible ratio between average power and peak power. This ratio, (called the duty cycle) which is equal to τf_r (τ being the pulse duration) can be roughly 1/5000 for the magnetron employed, giving for f_r a maximum permissible value of 40 kc/s.

Making f_r equal to the smaller of these two maxima we then obtain the maximum possible number of hits per target, i.e. as bright a display as possible. As will be seen, however, it proved necessary to reduce the pulse repetition frequency to 12.5 kc/s.

In order to turn to full advantage the possibilities offered by using such a short pulse duration (τ = 5 ns), the receiver should have an adequate *bandwidth*. A larger bandwidth improves the pulse discrimination but increases the relative noise, thus reducing the sensitivity obtainable. A good compromise [4]) is to choose the value of the bandwidth B between τ^{-1} and $2\tau^{-1}$. At the very short distances for which the 4 mm radar installation is intended, good pulse discrimination is more important than optimum sensitivity, for which reason we took $B = 2\tau^{-1} =$ 400 Mc/s. Given the intermediate frequencies commonly used for radar and with conventional tubes, it would not be possible to achieve such a bandwidth; in this experimental installation, however, use is made of several travelling-wave tubes (four in cascade), allowing the choice of a 4 Gc/s intermediate frequency. The relative bandwidth is then only 10%.

We shall return to the design of the amplifier when we discuss the receiver. At this stage it is useful to note that a noise figure of $F = 222\times$ was measured on the receiver. The apparent noise power at the input is equal to FkT_0B, where k is Boltzmann's constant and T_0 the standard noise temperature of 290 °K. The minimum detectable signal power S_min is roughly equal to this:

$$S_\mathrm{min} \approx F\,k\,T_0\,B = 3.61 \times 10^{-10}\ \mathrm{W}\,.$$

The range R_max of the radar installation can now be calculated from the familiar formula:

$$R^4_\mathrm{max} = P_0 \frac{A^2 \sigma}{4\,\lambda^2\,S_\mathrm{min}}\,,$$

where P_0 is the peak power of the magnetron, A the effective area of the antenna and σ the reflecting area, normally put at 1 m^2. After inserting this value and writing $P_0 = 20$ kW and $A = 0.042$ m^2, we find for R_max the value 4.7 km.

The foregoing data and some additional data of the installation are presented in *Table I*.

Table I. Data of experimental 4 mm radar installation.

Wavelength	λ	= 4 mm
Antenna width	l_h	= 1.5 m
Antenna height	l_v	= 35 mm
Maximum speed of revolution of antenna	w	= 120 r.p.m.
Pulse length	τ	= 5 ns
Pulse repetition frequency	f_r	= 12.5 kc/s
Peak transmitting power	P_0	= 20 kW
Average transmitting power	$P_0 \tau f_\mathrm{r}$	= 1.3 W
Bandwidth	B	= 400 Mc/s
Centre frequency of IF band	f_m	= 4 Gc/s
Ranges		0.3 km; 1 km; 3 km
Noise figure of receiver	F_tot	= 23.5 dB
Conversion loss of crystal mixer	L_1	= 13.6 dB

[4]) See J. F. Reintjes and G. T. Coate, Principles of radar, McGraw-Hill, New York 1952, p. 392 ff.

We shall now describe at greater length the main parts of the installation, viz: the transmitter, the receiver, the display unit and the various ancillary equipment.

The transmitter

Modulator circuit

The chief problem in designing the transmitter concerned the generation of the very short high-voltage pulses needed for setting the magnetron in oscillation. We shall therefore examine this problem in some detail.

To operate properly, the magnetron employed requires a voltage pulse with an amplitude of 17 kV. If, instead of insisting on a rectangular pulse, we can be satisfied with a triangular pulse with a half-height width of 5 ns, then the rise time will also be 5 ns.

In practice the anode of the magnetron is always at earth potential, and cathode and filament are given a high negative potential by the pulse. The circuit supplying this pulse is called the modulator. The modulator circuit used in this installation, and which will now be discussed, is represented schematically in *fig. 2*.

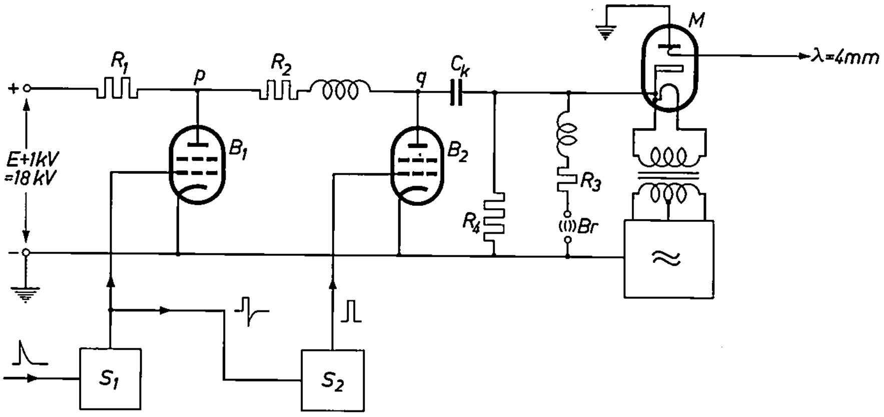


Fig. 2. Simplified circuit of the modulator. M 4 mm magnetron. The circuit must periodically supply to the magnetron cathode a negative voltage pulse of —17 kV, with a sufficiently short rise time to cause the magnetron to generate a wave train of 5 nanoseconds duration. The first part of the pulse is generated with B_1, the second with B_2. These tetrodes receive their triggering pulses from the respective blocking oscillators S_1 and S_2.

In view of the high potential it receives, the magnetron filament has to be fed by means of a transformer. In spite of careful construction, the intrinsic capacitance of the transformer cannot easily be kept below 10 pF. Moreover the magnetron has an input capacitance of 5 pF, and the tube delivering the voltage pulse has an output capacitance of 10 pF. Adding to this roughly 5 pF for the wiring capacitance, we arrive at a total capacitance C of 30 pF, which has to be charged up in 5 ns to —17 kV. For a linearly rising voltage this calls for a continuous charging current $i = C\Delta V/\Delta t$ of 102 A. Compared with this the current of 3 to 4 A consumed by the oscillating magnetron itself is very small, and will therefore from now on be disregarded. The average power $\frac{1}{2}C(\Delta V)^2 f_r = 54$ W, required for charging, presents no particular problems either, but the very high charging current of 102 A would obviously make exceptionally high demands on the pulse-generating tube.

To avoid this difficulty we took as our starting point the consideration that the magnetron does not begin to oscillate before the cathode voltage has dropped to about —14 kV, so that before that time the shape of the voltage pulse is of little significance. It can be seen from *fig. 3* that the pulse is therefore built up in two parts, the first with a rise time of 50 ns and the second with a rise time of 5 ns. The

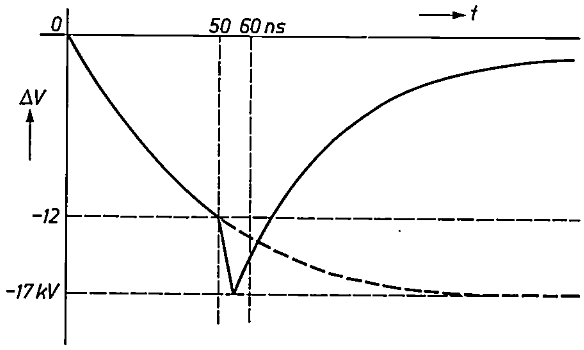


Fig. 3. The voltage pulse ΔV on the cathode of the 4 mm magnetron is built up in two steps. In the first part the voltage is made —12 kV negative in 50 nanoseconds, and in the second it is brought to —17 kV in 5 ns.

Without the measures described in the text, the cathode voltage at the end of the pulse would return to zero following the right-hand part of the solid curve.

second part of the pulse was generated by a tetrode type QEP 20/18, which can deliver a current of 30 A at a somewhat increased heater voltage. For a linearly rising pulse it is possible with a current of this magnitude to charge a capacitance of 30 pF to 5 kV in 5 ns. The first part of the pulse now has to bring the cathode potential from zero to —12 kV while the second part brings the cathode in 5ns to —17 kV.

As can be seen from the diagram in fig. 2, two tubes B_1 and B_2 are used for the modulator, the anodes of which are connected in parallel via a resistor R_2 and coupled to the magnetron cathode via a capacitor C_k. For B_1 a type QB 5/1750 tetrode was chosen, while B_2 is the QEP 20/18 tetrode already mentioned.

When the circuit S_1 (which will presently be discussed, together with S_2) receives a triggering pulse, it applies to the grid of B_1 a pulse of 50 ns duration and 800 V amplitude. As B_1 is made suddenly conductive, the cathode voltage of the magnetron drops in 50 ns to —12 kV. Thereupon circuit S_2 applies to the grid of B_2 a pulse of 5 ns duration and 800 V amplitude. Since R_2 is a wire-wound resistor it has significant self-inductance. For a transient as fast as a pulse of 5 ns the impedance of R_2 is sufficient to ensure the almost complete separation of points p and q, so that during the voltage drop of 5 kV, caused by B_2 becoming conductive, only the abovementioned capacitance of 30 pF needs to be charged up. Although the impedance of R_2 for the first part of the pulse is much lower, here too this resistance still plays a useful role; because of its presence the potential at point p drops somewhat faster than at point q. The result is that the power dissipated in B_1 is kept within permissible bounds.

After the end of the pulses from S_1 and S_2, tubes B_1 and B_2 stop conducting and points p and q have again to be raised via R_1 to the original potential of +18 kV. Because of the various stray capacitances present, this takes place relatively slowly. Without further precautions, the cathode pulse would therefore acquire the long tail indicated in fig. 3. This is prevented by incorporating a spark gap in parallel with the cathode (*Br* in fig. 2). The charging current for the capacitance of the spark gap produces in resistor R_3 a voltage drop which delays the sparkover of the gap until just after the cathode voltage has reached its maximum value. This gives the voltage on the cathode of the magnetron the form shown in *fig. 4*. The leading edge of the pulse is thus built up in two steps over a high impedance, while the discharge of the relevant capacitances takes place via the spark gap and the low impedance of R_3.

The spark gap is situated in a strong air current, with the object of rapidly extinguishing the spark. At the moment of extinction a certain residual charge may be present in the coupling capacitor C_k, which must be drained off through R_4. This resistance must be small enough to allow all charge to leak away before the next pulse arrives. At the same time the potential at points p and q must then again have risen to 18 kV, which means that R_1 cannot exceed a maximum value. That maximum, however, should be chosen as high as possible to ensure that the power supply is not loaded too unevenly. A value of 100 kΩ proved to be a good compromise.

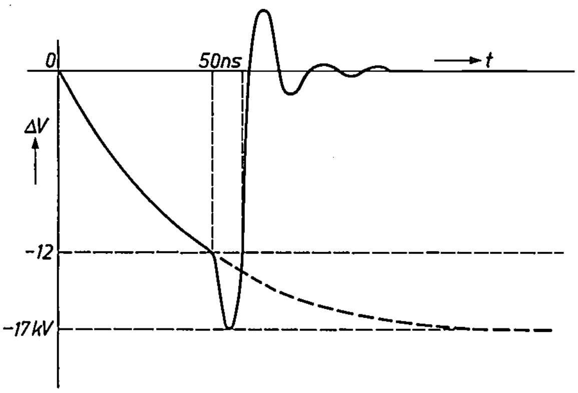


Fig. 4. By means of a spark gap (*Br* in fig. 2), the voltage on the cathode of the magnetron is very rapidly returned to zero at the end of the pulse.

The 50 ns pulse for the grid of B_1 in fig. 3 is generated in the blocking oscillator circuit S_1, a diagram of which is shown in *fig. 5*. In order for tube B_1 to be made strongly conductive in a short time, the grid must be raised in the same time to a high positive potential, resulting in a high grid current, i.e. a low grid impedance. The special feature of the blocking oscillator circuit is therefore that it can supply a large current through a low impedance. For this purpose the four tetrodes type QQE 02/5, operated in parallel in this oscillator, have a high mutual conductance and an amply dimensioned cathode. An examination of this circuit shows that the current tapped from the autotransformer T_1, to which transformer T_2 is connected, is the sum of the anode current, the grid current and the current through the voltage-limiting diode OA 211, i.e. a total of 20A. From the secondary side of T_2 a current of 5A is taken at a voltage of 900 V, approximately 2A flowing through the terminal resistance of 440 ohms and the remainder being available for grid current.

The circuit S_2 in fig. 2, which delivers the pulse for the modulator tube B_2, is a blocking oscillator of the same type as in fig. 5, but now using two QQE 02/5

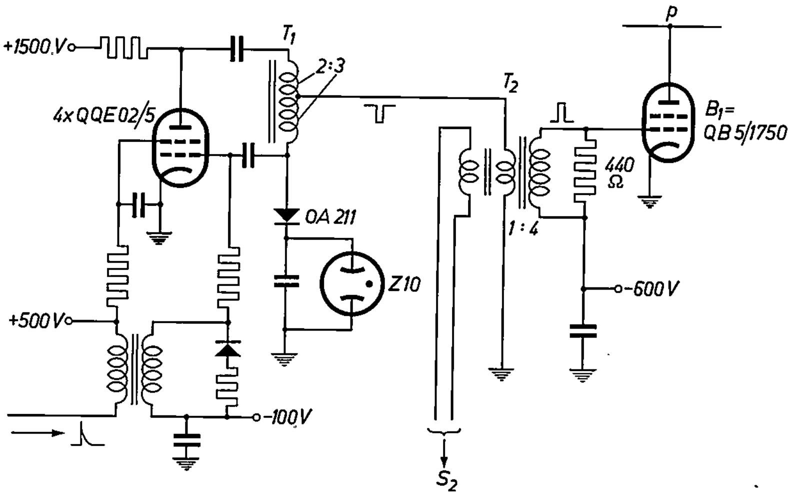


Fig. 5. Circuit diagram of the blocking oscillator S_1, which supplies the driving pulse to the grid of B_1 in fig. 2.

tetrodes in parallel. The circuit diagram can be seen in *fig. 6*. The pulse that starts this oscillator is obtained from an auxiliary winding of transformer T_2 in fig. 5. The second blocking oscillator is required to start about 50 ns after the first; the necessary delay is provided by the grid resistor R_g together with coupling capacitor C_g. The length of the delay can further be regulated slightly with the potentiometer P.

Since the pulse to be delivered by S_2 has to be much shorter than that from S_1, additional measures are taken to limit the length of the pulse from the blocking oscillator. In the first place the autotransformer T_3 is rated for saturation. Secondly an extra stage containing four QQE 02/5 tetrodes is employed. As a result of the high negative bias of this stage, it can only be triggered by the peaks of the pulses from the oscillator.

The cores used for transformers T_3, T_4 and T_5 are ferroxcube cores with a diameter of 10 mm. The high voltages occurring in the circuits described gave rise initially to some flashover between the base pins of the tubes. This was overcome by applying silicon grease between the pins.

Overload protection

As mentioned when discussing the design, the theoretically permissible modulator pulse repetition frequency of 25 kc/s was not feasible and had to be reduced to 12.5 kc/s. The reason for this is that, although the magnetron does not begin to oscillate in the desired mode at a wavelength of 4 mm until the voltage is 14 kV, it starts oscillating in another mode at a lower frequency earlier, owing to the rather slow initial rise of the anode voltage in our pulse (fig. 4). The energy thereby generated cannot leave the magnetron and causes extra heating of the cathode.

To prevent damage to the cathode an H-bend is fitted directly to the output waveguide of the magnetron, i.e. a piece of rectangular waveguide bent over its short side. As can be seen in *fig. 7*, the wall of the bend is slotted on the outside, forming a kind of grille. While the 4 mm energy follows the waveguide, the light radiated from the hot cathode passes through the slots on to a photoconductor. The latter is incorporated in a circuit which automatically reduces the heater current of the magnetron if the light intensity, i.e. the operating temperature of the cathode, exceeds the permissible maximum.

The 4 mm energy is finally conducted to the rotating antenna by means of a rotary joint of conventional construction.

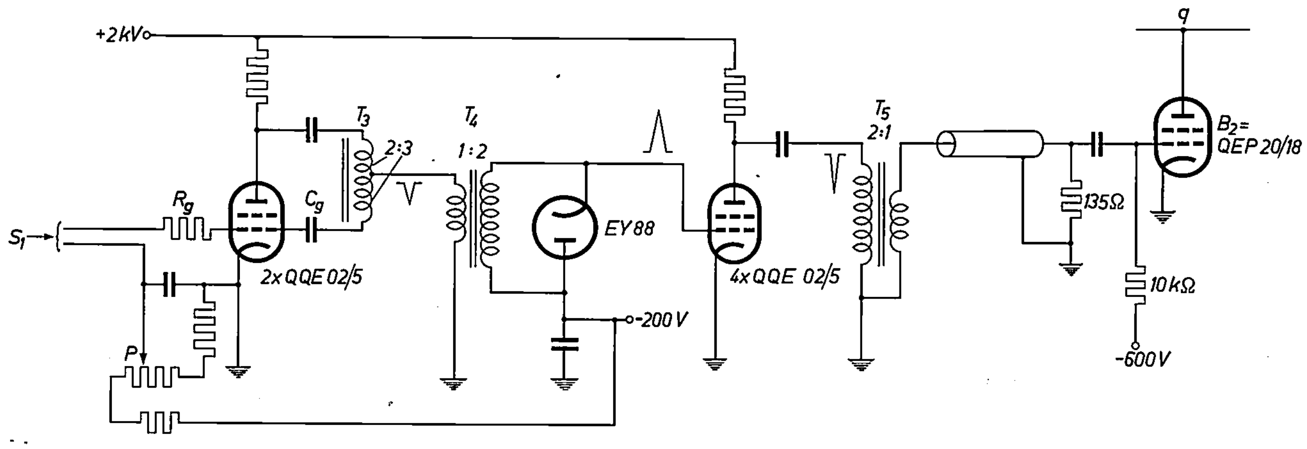


Fig. 6. Circuit diagram of blocking oscillator S_2, with follower stage which supplies the pulse to the grid of B_2 in fig. 2.

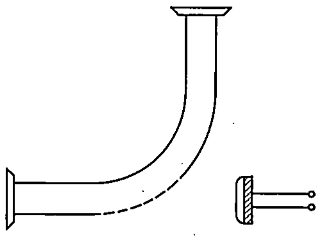


Fig. 7. *H*-bend in the output waveguide of the magnetron, slotted on the outside. The light radiated from the cathode passes through the slots to a photocell, and is used for automatically reducing the heater current to prevent thermal overloading of the cathode.

The receiver

The parts of the receiver are shown in *fig. 8*. The crystal mixer *K* is mounted directly above the receiving antenna *B*, from which it receives its signals via a rectangular wave guide and a rotary joint *C*. In the local oscillator a reflex klystron generates RF energy with a frequency of 71 Gc/s; the intermediate frequency is therefore 4 Gs/s (wavelength 7.5 cm). The local oscillator signal is mixed by means of a directional coupler.

The large bandwidth of 400 Mc/s required for the image quality does not of course present any problems at the signal frequency of 75 Gc/s, the relative bandwidth in this case being only 0.5%. Nor are there any difficulties in the IF amplifier, where the relative bandwidth is no more than 10%; we mentioned at the outset that the relatively high intermediate frequency of 4 Gc/s was chosen for this very reason. Because of this high intermediate frequency, however, a new problem arises in the crystal mixer: the difference of 4 Gc/s between the signal and the local oscillator frequency would require from the mixer device a relative bandwidth of more than 5% in order to be able to supply to the crystal the received signal together with the LO signal. This problem was solved by departing from the tuning procedure commonly used in radar installations for longer wavelengths. In the normal procedure the plunger in the waveguide which incorporates the crystal mixer is adjusted so as to produce the highest possible crystal current during the operation of the local oscillator; in other words the system is tuned to the local oscillator frequency. Making use of the fact that the available reflex klystron delivers sufficient local oscillator power, we decided to tune to the signal frequency. In spite of the mismatch the local oscillator then still delivers a current of 0.5 mA in the crystal mixer.

Another advantage of using such a high intermediate frequency is that a substantial part of the local oscillator's contribution to the noise is eliminated. The local oscillator, too, generates a spectrum of noise frequencies, and the width of that spectrum is determined by the quality of the output cavity of the klystron. In radar operating on conventional wavelengths the mutual beat frequencies of the noise spectrum lie partly in the IF amplifier band, but at our high intermediate frequency they lie outside it.

The crystal with its cat's whisker is fitted in a mount which has the same inside dimensions as a 4 mm waveguide. To obtain a low intrinsic capacitance and a large mixing effect, the cat's whisker must be extremely thin at the point where it makes

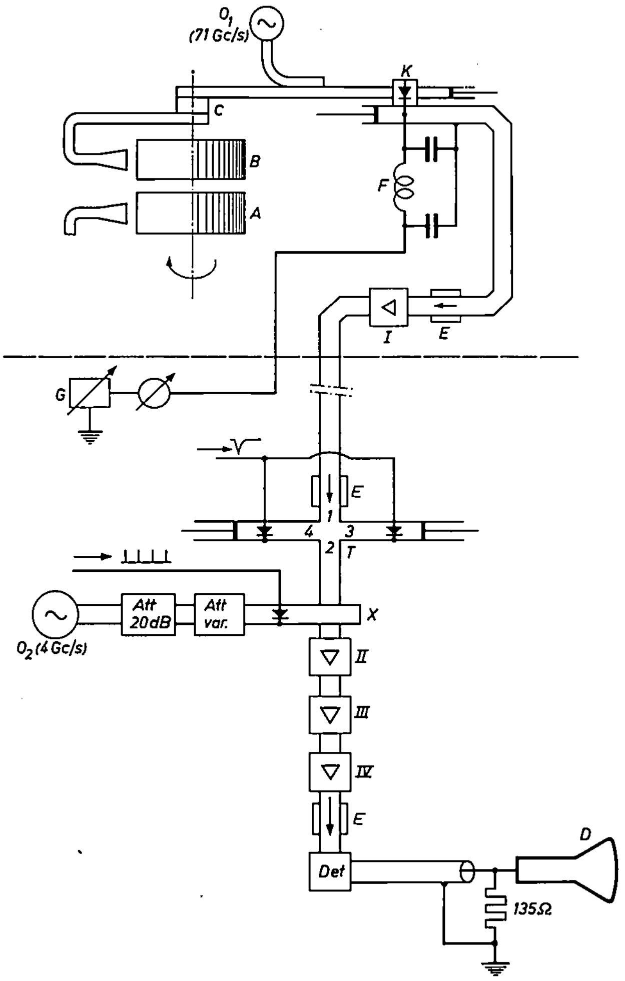


Fig. 8. Schematic survey of the 4 mm radar receiving system. *A* transmitting antenna. *B* receiving antenna. *C* rotary joint. O_1 local oscillator. *K* crystal mixer, in which the signal band of 400 Mc/s width is shifted from the signal frequency 75 Gc/s to the intermediate frequency 4 Gc/s (7.5 cm wavelength). *F* filter. *G* DC voltage source. *E* directional isolators. *I*, *II*, *III*, *IV* amplifiying stages using travelling-wave tubes. *T* crystal switch, providing the swept gain. *X* cross-guide coupler. O_2 auxiliary oscillator, one purpose of which is to provide range marker rings on the screen. *Det* detector (disc-seal triode type EC 157). *D* cathode ray tube.

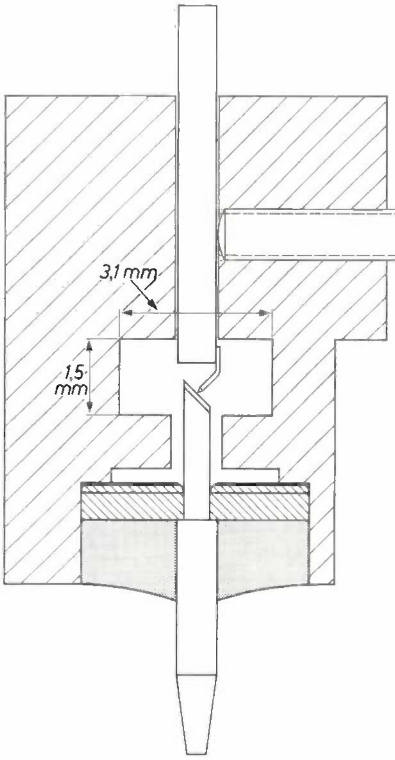


Fig. 9. Mount of the 4 mm mixing crystal, specially designed to minimize the length of the cat's whisker, this keeping the inductance as low as possible.

contact with the crystal wafer. A thin cat's whisker, however, implies a high series inductance which, in crystal mounts of conventional construction, is further increased by the fact that the cat's whisker has a loop in it, so that the point makes spring contact with the crystal.

This drawback has been overcome by the construction illustrated in *fig. 9*. A germanium wafer is mounted at an angle of 45° on the output pin. The cat's whisker is kept very short and has a diameter of only 20 μm; the tip of the whisker is pointed electrolytically. The whisker is welded to a thick pin, which largely bridges the height of the waveguide. When the pin is rotated the whisker comes in contact with the crystal, the contact pressure being derived from a slight torsion in the thin whisker. Because of its higher electron mobility — and hence its lower resistance — germanium was chosen for the crystal wafer rather than silicon.

In spite of the above-mentioned precautions the conversion loss is high. This loss, which was found to be 14 dB, occurs partly at the crystal output. Although the capacitance here is kept as low as possible, in this case it is not compensated in the usual manner by means of a tuned circuit at the input of the IF stage.

With a view to keeping the bandwidth as large as possible, the 4 mm waveguide containing the crystal mount is coupled directly to a 7.5 cm waveguide for the IF output signal. As can be seen in *fig. 10*, the IF signal is conducted via a cylindrical pin coaxially through a hole in the common wall. The pin is at the same time a component of a crossbar coupling, the purpose of which is to obtain a good match between the crystal output and the 7.5 cm waveguide. The crossbar is not metallically connected to the wall of the waveguide but is led out capacitively, making it possible to measure the crystal current at this point. The test line contains, in addition to a filter

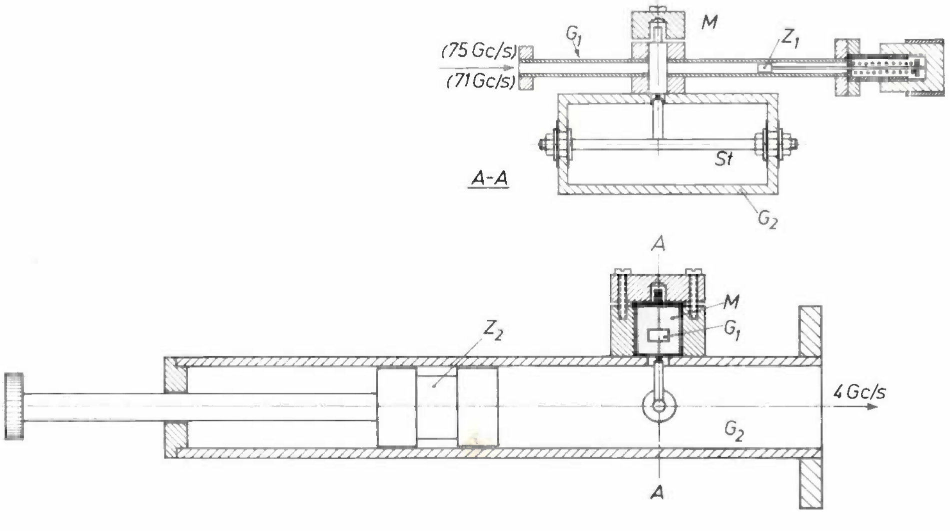


Fig. 10. Mechanical details of the mixer stage, showing the crystal mount M, the coupling to the 4 mm waveguide G_1, which receives the radar signal (75 Gc/s) and the local oscillator signal (71 Gc/s), and the coupling to the 7.5 cm waveguide G_2 for the IF signal. The latter coupling is a crossbar type St. Both waveguides contain a tuning plunger (Z_1, Z_2); the plunger for the 4 mm guide, the adjustment of which is extremely critical, has practically no backlash.

for protecting the crystal from voltage surges, a variable biasing device for adjusting the crystal to the optimum operating point.

The IF signal passes along the 7.5 cm waveguide to the first travelling-wave tube (*I*), still fitted in the antenna pedestal. This guide contains a directional isolator, which prevents very strong radar echoes from causing repeated reflections, which would appear on the screen as a series of dots. The first ampliflying tube is set to give a low noise figure for a moderate gain (18.5 dB). The former was measured at 9 dB (or 8×).

The signal from the first ampliflying tube is conducted through a 12 metre-long waveguide to the installation in the operating room. This waveguide has a low and wide cross-section with a view to avoiding excessive dispersion, i.e. undue transit-time differences between the components of the very broad frequency spectrum, which would give rise to pulse distortion. The broader the guide, the lower these relative differences are.

In front of the second amplifying stage the waveguide contains another directional isolator, specially intended for suppressing reflections in the variable attenuator which follows next in the waveguide. The operation of this component can only be described here very roughly. It contains a magic *T* (see fig. 8), arms *3* and *4* of which contain diodes and shorting plungers [5]). Immediately before the pulse is sent out the diodes are made conductive, as a result of which incoming signals are almost completely reflected and absorbed in the directional isolator. During the moment of transmission the receiver is therefore nonreceptive. After this the diode current decreases exponentially, so that the signal transmission from arm *1* to *2* gradually increases. This time-dependent gain control, which appears as a localized gain control on the PPI tube, makes the brightness of the image more uniform in the radial direction.

The signal next passes a cross-guide coupler. By means of this coupler, which gives 20 dB attenuation, 4 Gc/s pulses can be added to produce range marker rings on the PPI tube. The 4 Gc/s signal can also serve for setting the subsequent amplifier stages *II*, *III* and *IV*, each of which contains a travelling-wave tube. Since stage *II* makes a much smaller contribution to the total noise than stage *I*, it is set to give a somewhat higher gain (25 dB). The measured noise figure of this stage was 11.1 dB (i.e. 13×). Stage *III* has an even higher gain, of 35 dB. Its relative noise contribution is negligible and was therefore not measured. The last stage serves as a power amplifier and can deliver 10 W. The total gain of the four stages in cascade is between 90 and 100 dB.

The noise figure of a multi-stage receiver can be calculated using the familiar formula:

$$F = F_1 + (F_2 - 1)\,G_1^{-1} + (F_3 - 1)\,(G_1G_2)^{-1} + \dots + (F_k - 1)\,(G_1G_2 \dots G_{k-1})^{-1},$$

where F_i and G_i are respectively the noise figure and the gain of stage i of the receiver.

Considering the crystal mixer as the first stage of the receiver, it is better to introduce: L_1 = conversion loss = G_1^{-1}, and n_1 = noise temperature [6]) = F_1G_1. We can then rewrite the formula as:

$$F = L_1 \Big[n_1 + F_2 - 1 + (F_3 - 1)\,G_2^{-1} + \dots \Big].$$

The measured conversion loss of the crystal was 23 (i.e. about 14 dB) and the measured noise temperature 2.5. For determining the noise figures of the various amplifier stages, use was made of a gas-discharge noise source for the 4 mm band [7]). Substitution in the above formula of the separately measured values for the individual stages gives the earlier mentioned total noise figure of 222× (i.e. 23.5 dB).

When adjusting the operating point of the various stages it was taken into account that the gain of a travelling-wave tube is not constant but decreases as the strength of the input signal incresase. This relation can be influenced by a suitable choice of the voltage on the helix: see for example the schematic curves *1* and *2* in *fig. 11.* In the last two stages the helix voltage is chosen so that the gain only begins to decrease at very large signal strengths. The total gain therefore follows a curve as shown schematically by curve *3* in fig. 11. For the largest signals the curve of the gain as a function of input power is flat, which again contributes to the uniformity of the image brightness and makes a limiter for the CRT superfluous.

Detection of the output signal from the last amplifying stage is effected by a disc-seal triode type EC 157, the grid-cathode space of which is incorporated in the 7.5 cm waveguide as illustrated in *fig. 12.* The waveguide profile is shaped in such a way as to match the grid-cathode impedance to that of the

[5]) For an explanation of the operation of the magic T, see: G. L. Ragan, Microwave transmission circuits, M.I.T. Radiation Laboratory Series No. 9, McGraw-Hill, New York 1948, p. 706. For diode switching in waveguides: R. V. Garver, J. A. Rosado and E. F. Turner, Theory of the germanium diode microwave switch, I.R.E. Transactions on microwave theory and techniques **8**, 108-111, 1960. Applications of the magic T are discussed in an article by J. P. M. Gieles, Applications of microwave triodes, Philips tech. Rev. **22**, 16-28, 1960/61.

[6]) See for example H. C. Torrey and C. A. Whitmer, Crystal rectifiers, M.I.T. Radiation Laboratory Series No. 15, McGraw-Hill, New York 1948, pp. 213 and 227.

[7]) See P. A. H. Hart, Standard noise sources, Philips tech. Rev. **23**, 293-309, 1961/62.

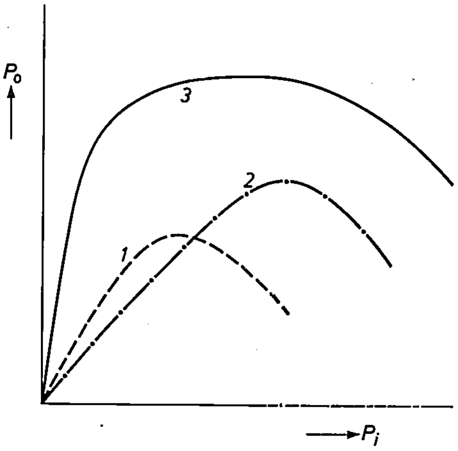


Fig. 11. Output signal power P_0 of a travelling-wave tube as a function of signal input power P_i, represented schematically. At large input signals the gain decreases, finally becoming less than unity. By giving the four tubes in the receiver different operating points (e.g. curve *1* for tube *III* and curve *2* for tube *IV*) a total gain characteristic is obtained as represented schematically by curve *3*. The curve flattens for strong signals.

waveguide while maintaining the required bandwidth of 400 Mc/s.

The choice of an EC 157 as detector was due to the fact that an experimental CRT was available which could be completely cut off by a negative bias as low as 6 V. This made it possible to dispense completely with a video amplifier, which, after the detector, would have needed a bandwidth of 200 Mc/s. This meant, however, that the detector had to be capable of delivering through its output impedance, formed by the 135 ohm coaxial cable through which the video signal is applied to the CRT, a current of at least 6/135 A, i.e. about 50 mA; a crystal detector cannot readily be made to sypply such a current.

The grid bias for the EC 157 is so chosen to make the tube work as an anodebend detector. The video signal appears across an anode resistance of 1000 ohms and is supplied to the coaxial cable via a coupling capacitance of 470 nF. Incorporated in the supply line to the control grid of the CRT, by a method familiar in television engineering [8]), is a small inductance designed to obtain the required bandwidth of 200 Mc/s in the transmission of the video signal. In order to be able to display the video signal on an oscilloscope, the terminating impedance for the coaxial cable is not fitted to the CRT but in a removable plug on the front panel of the PPI console.

No measures are taken for DC restoration [9]) because most of the time there is no signal and therefore the charge displacement on the coupling capacitor of 470 nF is negligible.

Display unit and ancillary equipment

The main components of the display unit, apart from the CRT, are the sawtooth generator, which energizes the CRT deflection coil, and the servo system that synchronizes the deflection coil with the antenna. These devices follow familiar principles of construction and will not be described here in detail. It is, however, useful to note the considerable extent to which the currents and voltages to be delivered by the sawtooth generator are increased at the short ranges at which this radar installation operates. The echo from an object at a distance of 300 m away reaches the receiver 2 microseconds

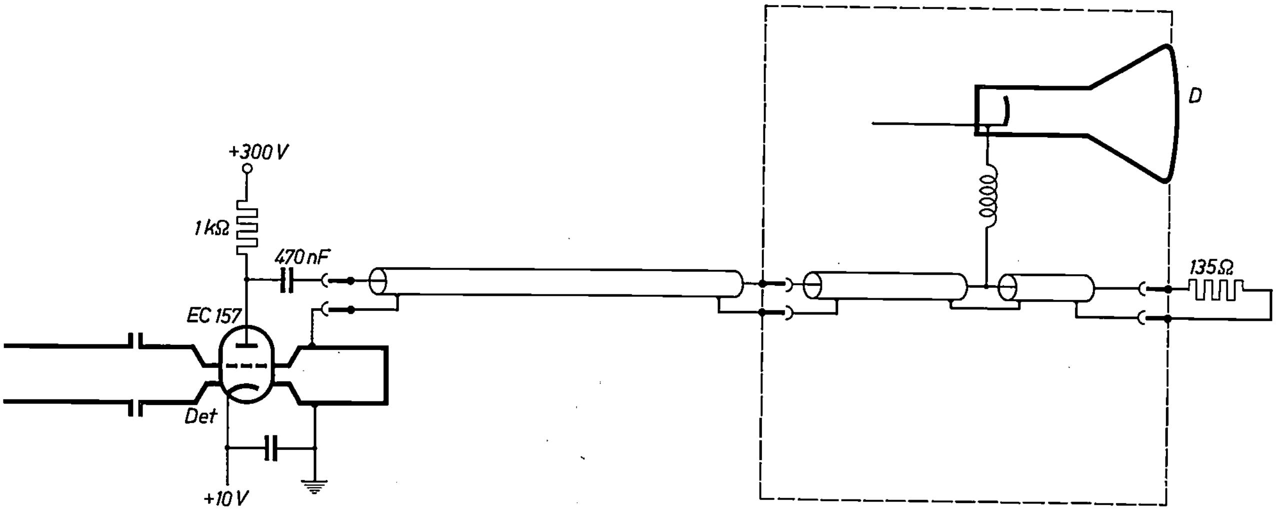


Fig. 12. The intermediate frequency detector is an EC 157 disc-seal triode, whose grid-cathode space is incorporated in the 4 Gc/s waveguide. Particulars of the connection to the cathode ray tube *D* are mentioned in the text.

[8]) See G. E. Valley and H. Wallman, Vacuum tube amplifiers, M.I.T. Radiation Laboratory Series No. 18, McGraw-Hill, New York 1948, Chapter 2.

[9]) See F. Kerkhof and W. Werner, Television, Philips Tech. Library, Centrex, Eindhoven 1952, Part I, p. 88.

after transmission of the radar pulse. Where the range is set to 300 m this means that the spot has to move from the centre of the tube to the periphery in 2 μs. With a view to keeping the voltage across the deflection coil within reasonable limits we gave the coil relatively few windings, but this necessitated a relatively high deflection current. From the inductance value obtained, $L = 800$ μH, and taking into account the number of ampere-turns required for the maximum deflection of the spot, the amplitude of the energizing current is easily calculated to be 1.5 A and the amplitude of the voltage 600 V. Since the output tube in the sawtooth generator in the conductive state must maintain a certain reserve of anode voltage, a supply voltage of 800 V was needed. The required current can be obtained with two E 130 L tubes in parallel.

The equipment further comprises, in addition to the power pack, all circuits needed for generating the various pulse and triggering signals. These signals are obtained by frequency division from the output signal of a crystal-controlled 1500 kc/s generator. The highest frequencies in the sequence are those used for producing the range marker rings on the PPI tube, while the lowest is the pulse repetition frequency (12.5 kc/s). Since, after a number of fre-

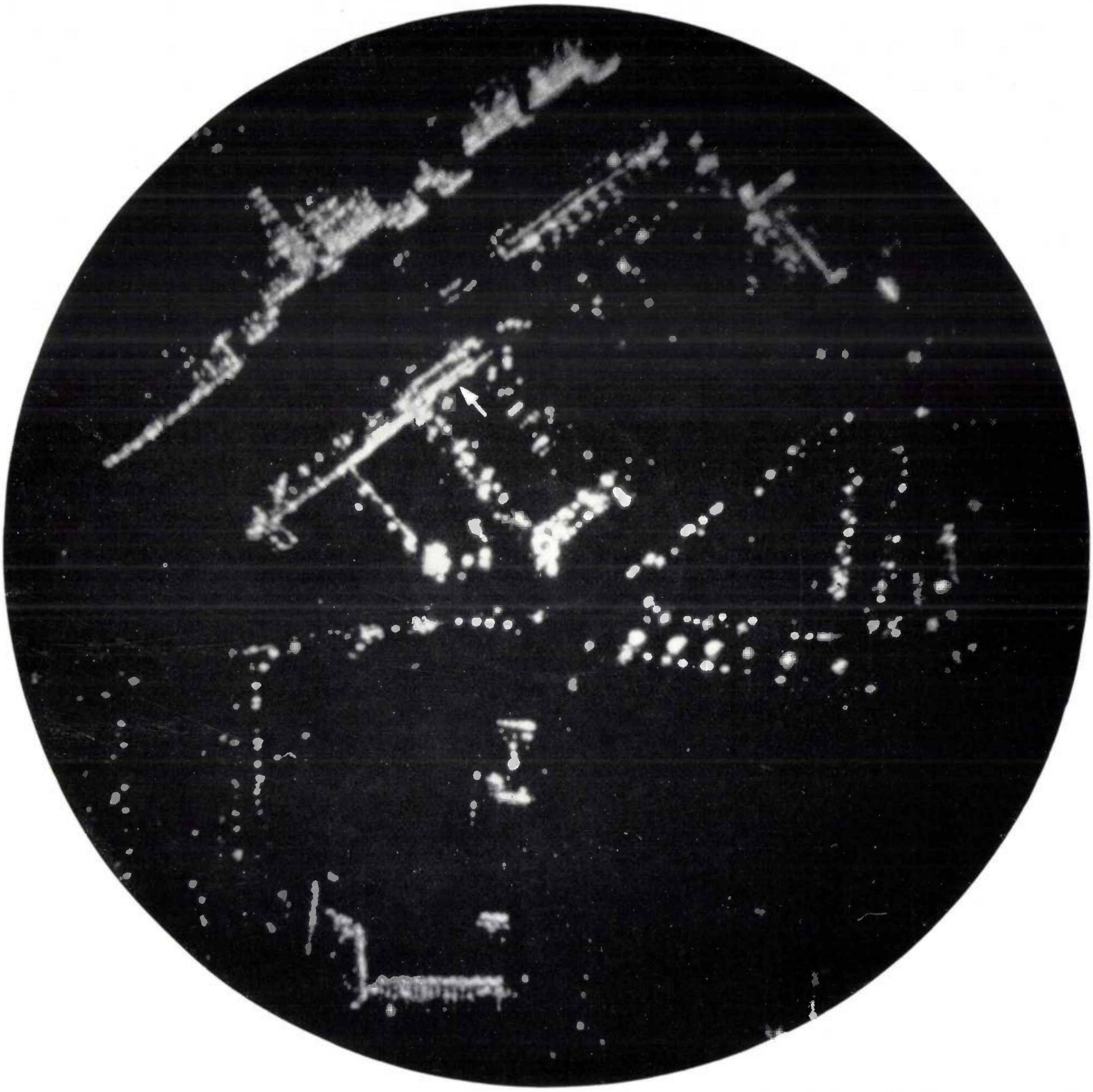


Fig. 13. The radar display covering the part of the town and the Philips factories immediately surrounding the Eindhoven Research Laboratories. The range was set here at 300 m.

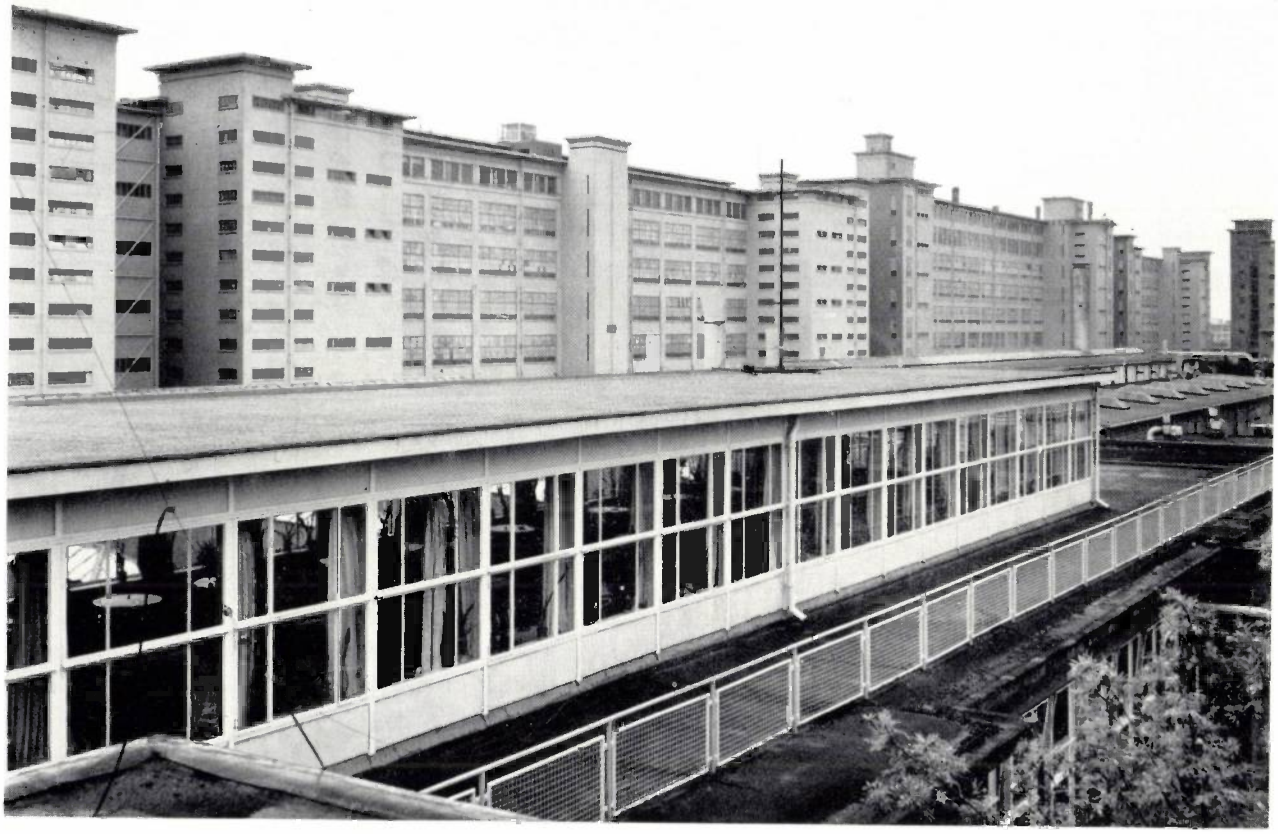


Fig. 14. Glass wall of a canteen on the roof of the Research Laboratories and the roof balustrade 3 m in front of it, each separately visible in the radar display in fig. 13 (marked by arrow). The tall factory buildings in the background can be seen top left on the radar screen.

quency divisions, the leading edges of the pulses may show some jitter, the leading edges of the pulses for the lowest frequencies are directly coupled with those of the oscillations of the crystal oscillator.

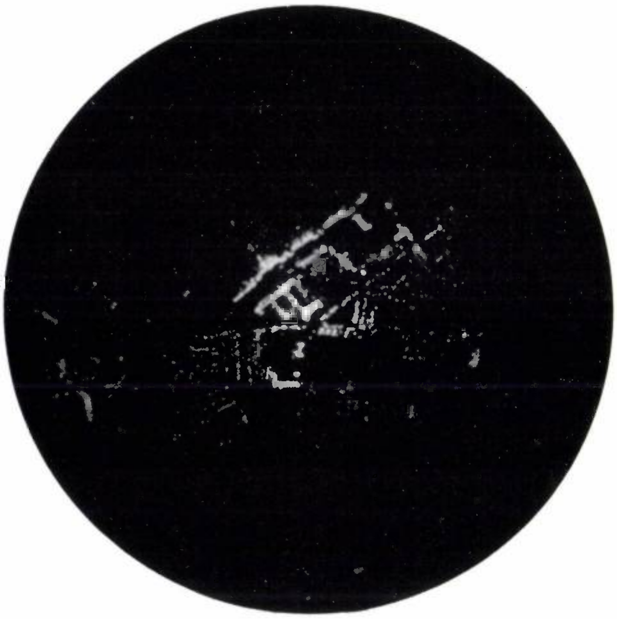


Fig. 15. As fig. 13, but with the range set at 1000 m. Beyond 500 m the amount of detail reproduced rapidly decreases.

Performance and discussion

A radar display obtained with this installation when the range is set to 300 m is shown in *fig. 13.* At the point marked by an arrow it can be seen how the front wall of a building is separated in the display quite distinctly from a balustrade about 3 m in front of it. The actual situation on the site is shown in *fig. 14.* The rear wall of the building also appears in the display. This is bound up with the fact that the front wall consists to a large extent of glass and thus passes part of the 4 mm energy.

Fig. 15 shows the radar display of the same objects when the range is set to 1000 m. As can be seen, not many echoes are received beyond this range. The effective maximum range of the installation is therefore roughly 1 km, which is a lot shorter than the theoretical value of 4.7 km. There are several possible explanations for this.

a) The microwave energy passes at the transmitting side through about 1.5 m of waveguide, and at the receiving side through about 1 m. To minimize the losses the guide used for this purpose is the 5 mm type (WG 25) which is just narrow enough not to transmit energy in the wrong mode (TE_{02} mode). Further, red copper is used instead of brass for the waveguide. The damping of the two sections of waveguide together is still 3.5 dB, however. Added to this are the losses of the two rotary joints, together amounting to 2.4 dB, so that the total damping in the microwave part is about 6 dB. It is possible to reduce the damping by shortening the waveguide connections, which might be done in the present installation for example by shifting the point of rotation of the antennas.

b) The influence of atmospheric losses, which occur during the propagation of energy in free space, is apparently relatively low. The pictures of the displays were taken in fine weather, and, as reported in the literature [10]), the absorption of 4 mm waves in oxygen is 0.2 dB/km and for uncondensed water vapour 0.05 dB/km at a degree of humidity of 44%. This adds up to 0.25 dB/km, which, for radar, where the energy travels twice the distance to the object, means 0.5 dB per km range. Compared with the damping mentioned under (a) this is not very significant.

In rain or mist the situation would be quite different. Under these conditions the following figures apply (including the radar multiplication factor of 2):

In mist	
of 0.23 g/m³ (400 m visibility):	1.6 dB/km
of 2.3 g/m³ (100 m visibility):	10 dB/km
In rain	
of 0.25 mm per hr. (drizzle):	0.4 dB/km
of 1 mm per hr. (light rain):	1.6 dB/km
of 16 mm per hr. (heavy rain):	16 dB/km

To these figures we must add the 0.5 dB/km for absorption in oxygen and water vapour.

In mist, but particularly in rain, the loss in question is not so much due to absorption as to scattering. Part of the scattered energy will arrive in the receiving antenna and cause interference comparable to noise, blurring the echoes which are already attenuated by absorption. The effects of this scattering can be overcome to a large extent by using antennas that polarize the waves circularly instead of linearly.

c) The range formula, which we discussed when dealing with the design, is presumably limited in its application to 4 mm waves. There is reason to assume that the reflection of such short waves involves more losses than in the case of longer microwaves. As mentioned in connection with fig. 13, the 4 mm waves can readily pass through certain objects, such as glass walls. On the other hand, with waves as short as this there is a real chance of actual reflection from parts of an object, the reflected beam then being *directed*, and in general *not* back to the receiving antenna. All in all, then, the formula referred to gives only a first order approximation of the true range.

When dealing with the receiver we stated that the first travelling-wave tube has a noise figure of 9 dB (8×). Because of the high conversion loss of the crystal mixer (14 dB, see page 281) the noise figure of the first travelling wave tube has a fairly marked influence on the total noise [11]). A considerable improvement can therefore be achieved by using a tube with lower inherent noise. Admittedly this does not have very much influence on the range, which is of course inversely proportional to the fourth root of the minimum detectable signal — the latter being directly proportional to the total noise — but it does mean that with less noise more echoes are visible within the swept area. In the experiment described here the main emphasis was placed on achieving the best possible resolution. For this purpose the less favourable noise figure of the first travelling wave tube was not felt to be a disadvantage.

[10]) See C. W. Tolbert and A. W. Straiton, Attenuation and fluctuation of millimeter radio waves, IRE Nat. Conv. Rec. 5, Part 1, pp. 12-18, 1957.

[11]) This can be derived directly from the calculation given in small print on page 282: F_2 is the noise figure of the first tube; the term $n_1 = F_1G_1$ beside F_2 is relatively small owing to the high conversion loss G_1^{-1}.

Summary. To investigate the extent to which the resolution of a short-range radar can be improved by using wavelengths shorter than the conventional 8 mm, tests were made with an experimental 4 mm installation. A high radial resolution depends on the transmitted pulses being as short as possible. The pulse length of 5 ns adopted is very close to the minimum required for the proper operation of the magnetron. To achieve the very short rise time for the magnetron pulses a modulator was built that delivers a voltage pulse in two steps. For the purpose of amplifying the received echo pulses without distortion the IF amplifier in the receiver should have a bandwidth of 400 Mc/s. The choice of a very high intermediate frequency, i.e. 4 Gc/s, allowed the relative bandwidth of the IF amplifier to be limited to 10%. Amplification at 4 Gc/s was made possible by using in the receiver four travelling-wave tubes connected in cascade. The crystal mixer before the amplifier is contained in a specially designed mount. The detector for the 4 Gc/s signal is a disc-seal triode type EC 157, which can deliver more video energy than conventional crystal detectors. Because of this and the use of an experimental CRT of exceptional sensitivity, a video amplifier between detector and CRT could be dispensed with.

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.

3256: S. Duinker: Search for a complete set of basic elements for the synthesis of non-linear electrical systems (Network theory, pp. 221-250, Pergamon Press, Oxford 1963).

3257: K. M. Adams: A review of the synthesis of linear three-terminal networks composed of two kinds of elements (as **3256**, pp. 65-73).

3258: G. Blasse: The structures of $ZnLiVO_4$ and $ZnLiNbO_4$ (J. inorg. nucl. Chem. **25**, 136-137, 1963, No. 1).

3259: D. J. van Ooijen and J. D. Fast: Electrical resistance of hydrogen-charged iron wires (Acta metallurgica **11**, 211-216, 1963, No. 3).

3260: A. J. Th. Mollet and L. E. Vrenken: Fülldruckmessungen an Inhalten von 0,2 cm³ und grösser im Druckbereich von 20 bis 100 Torr (Vakuum-Technik **12**, 6-12, 1963, No. 1). (Filling-pressure measurements on volumes of 0.2 cm³ upwards in the pressure range from 20 to 100 torr; in German.)

3261: A. J. Bosman and E. E. Havinga: Temperature dependence of dielectric constants of cubic ionic compounds (Phys. Rev. **129**, 1593-1600, 1963, No. 4).

3262: C. McD. Hargreaves: Multiple-beam 'transmission-like' Fizeau fringes in the reflexion interference system (Nature **197**, 890-892, 1963, No. 4870).

3263: D. Polder and A. Baelde: Theory of noise of transistor-like devices (Solid-state electronics **6**, 103-110, 1963, No. 2).

3264: W. F. Druyvesteyn and D. J. van Ooijen: Influence of neutron irradiation at 78 °K on the critical field of superconducting lead (Physics Letters, Amsterdam, 4, 170-172, 1963, No. 3).

3265: C. J. M. Rooymans: Structure of the high pressure phase of CdS, CdSe and InSb (Physics Letters, Amsterdam, 4, 186-187, 1963, No. 3).

3266: G. Blasse: The crystal structure of some compounds of the type $LiMe^{3+}Me^{4+}O_4$ and $LiMe^{2+}Me^{5+}O_4$ (J. inorg. nucl. Chem. **25**, 230-231, 1963, No. 2).

3267: J. S. C. Wessels: Separation of the two photochemical systems of photosynthesis by digitonin fragmentation of spinach chloroplasts (Proc. Roy. Soc., London, B **157**, 345-355, 1963, No. 968).

3268: C. W. Berghout: The uniaxial magnetic anisotropy of the F.C.C. cobalt precipitate in copper (Phys. Chem. Solids **24**, 507-516, 1963, No. 4).

3269: P. Massini: Aminotriazolylalanine: a metabolic product of aminotriazole from plants (Acta bot. neerl. **12**, 64-72, 1963, No. 1).

3270: E. W. Gorter: Interactions in various magnetic compounds (J. appl. Phys. **34**, 1253-1259, 1963, No. 4^{II}).

3271: C. J. M. Rooymans: A phase transformation in the wurtzite and zinc blende lattice under pressure (J. inorg. nucl. Chem. **25**, 253-255, 1963, No. 3).

3272: C. Wansdronk: Threshold mechanism and masking by noise (J. Acoust. Soc. Amer. **35**, 726-728, 1963, No. 5).

3273: J. D. Wasscher, W. Albers and C. Haas: Simple evaluation of the maximum thermo-electric figure of merit, with application to mixed crystals $SnS_{1-x}Se_x$ (Solid-state electronics **6**, 261-264, 1963, No. 3).

3274: J. C. Courvoisier, W. Haidinger, P. J. W. Jochems and L. J. Tummers: Evaporation-condensation method for making germanium layers for transistor purposes (Solid-state electronics **6**, 265-270, 1963, No. 3).

3275: G. Blasse: The structure of some new mixed metal oxides containing lithium (J. inorg. nucl. Chem. **25**, 743-744, 1963, No. 6).

R 459: 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).

R 460: J. A. W. van der Does de Bye: Transient recombination in silicon carbide (Philips Res. Repts **17**, 419-430, 1962, No. 5).

R 461: L. J. van de Polder: Simple representation of the behaviour of a diode in converter circuits (Philips Res. Repts **17**, 431-450, 1962, No. 5).

R 462: U. Enz, J. F. Fast, S. van Houten and J. Smit: Magnetism of EuS, EuSe and EuTe (Philips Res. Repts **17**, 451-463, 1962, No. 5).

R 463: B. Okkerse: Anomalous transmission of X-rays in germanium. Part I: The imaginary part of the atomic scattering factor (Philips Res. Repts **17**, 464-478, 1962, No. 5).

R 464: F. van der Maesen, C. A. A. J. Greebe and H. J. C. A. Nunnink: Steady-state injections in semiconductors (Philips Res. Repts **17**, 479-512, 1962, No. 5).

R 465: R. Bleekrode, J. Dieleman and H. J. Vegter: Electron-spin resonance of Fe in GaAs (Philips Res. Repts **17**, 513-517, 1962, No. 5).

R 466: A. J. W. Duijvestijn: Electronic computation of squared rectangles (Philips Res. Repts **17**, 523-613, 1962, No. 6):
Thesis Eindhoven, 1962.

R 467: J. H. N. van Vucht: Equilibrium pressures in the system Th_2Al-hydrogen (Philips Res. Repts **18**, 1-20, 1963, No. 1).

R 468: J. H. N. van Vucht: X-ray diffraction of Th_2Al containing hydrogen (Philips Res. Repts **18**, 21-34, 1963, No. 1).

R 469: J. H. N. van Vucht: Neutron diffraction and proton magnetic resonance of deuterium and hydrogen solutions in Th_2Al (Philips Res. Repts **18**, 35-52, 1963, No. 1).

R 470: J. H. N. van Vucht: The problem of order or disorder of hydrogen in $Th_8Al_4H_8$ (Philips Res. Repts **18**, 53-60, 1963, No. 1).

R 471: H. de Lang: Ferromagnetic domain structure as a microscopic object (Philips Res. Repts **18**, 61-64, 1963, No. 1).

R 472: C. A. A. J. Greebe and F. van der Maesen: Graphical approach to small injections in semiconductors (Philips Res. Repts **18**, 65-70, 1963, No. 1).

R 473: J. J. van Loef: Influence of hydrogen on electric and magnetic properties of a nickel-palladium alloy (Philips Res. Repts **18**, 71-74, 1963, No. 1).

R 474: W. Steinmaier: Thermodynamical approach to the growth rate of epitaxial silicon from $SiCl_4$ (Philips Res. Repts **18**, 75-81, 1963, No. 1).

R 475: B. Okkerse and P. Penning: Anomalous transmission of X-rays in germanium. Part II: Effects due to elastic deformation (Philips Res. Repts **18**, 82-94, 1963, No. 1).
Continued from **R 463**.

R 476: M. T. Vlaardingerbroek and K. R. U. Weimer: On wave propagation in beam-plasma systems in a finite magnetic field (Philips Res. Repts **18**, 95-108, 1963, No. 2).

R 477: M. J. Hogendijk: Random dense packing of spheres with a discrete distribution of the radii (Philips Res. Repts **18**, 109-126, 1963, No. 2).

R 478: F. H. Stieltjes and G. Diemer: Two-carrier photoconductance (Philips Res. Repts **18**, 127-146, 1963, No. 2).

R 479: W. Elenbaas: Rate of evaporation and heat dissipation of a heated filament in a gaseous atmosphere (Philips Res. Repts **18**, 147-160, 1963, No. 2).

A 60: A. Klopfer: An ionization gauge for measurement of ultra-high vacua (1961 Trans. 8th Nat. Vacuum Symp. + 2nd Int. Congress on Vacuum Sci. and Technol., Washington D.C., Vol. 1, pp. 439-442, Pergamon Press, Oxford 1962).

A 61: R. Groth, E. Kauer and P. C. van der Linden: Optische Effekte freier Träger in SnO_2-Schichten (Z. Naturf. **17a**, 789-793, 1962, No. 9). (Optical effects of free carriers in SnO_2 layers; in German.)

A 62: H. G. Reik: Theoretische Untersuchungen zum Problem der heissen Elektronen in Halbleitern (Festkörperprobleme I, pp. 89-121, Vieweg, Brunswick 1962). (Theoretical studies on the problem of hot electrons in semiconductors; in German.)

A 63: H. G. Grimmeiss and R. Memming: Origins of the photovoltaic effect in vapor-deposited CdS layers (J. appl. Phys. **33**, 3596-3597, 1962, No. 12).

A 64: H. Hörster and E. Kauer: Untersuchung der Kinetik von Gas-Feststoff-Reaktionen mittels elektrischer Messungen (Z. Elektrochemie, Ber. Bunsenges. physik. Chemie **66**, 667-671, 1962, No. 8/9). (Study of the kinetics of gas-solid reactions, using electrical measurements; in German.)

A 65: N. Hansen: Chemisorption und Widerstandsänderungen an aufgedampften Zirkoniumfilmen (Z. Elektrochemie, Ber. Bunsenges. physik. Chemie **66**, 726-731, 1962, No. 8/9). (Chemisorption and resistance changes in vacuum-deposited zirconium films; in German.)

A 66: A. Rabenau: Siliziumnitrid, ein keramisches Material für hohe Temperaturen (Ber. Dtsch. Keram. Ges. **40**, 6-12, 1963, No. 1). (Silicon nitride, a ceramic material for high temperatures; in German.)

A 67: E. Andrich and T. van der Sterre: Aufbau und Eigenschaften von PTC-Widerständen (Elektron. Rdsch. **17**, 63-64, 1963, No. 2). (Structure and properties of PTC resistors; in German.)

A 68: E. Schwartz: Die Änderung der Leistung in gestörten linearen Gleichstromnetzwerken (Arch. elektr. Übertr. **17**, 198-202, 1963, No. 4). (Power changes in perturbed linear DC networks; in German.)

A 69: P. Gerthsen: Ein neues Verfahren zur Kristallzüchtung von hochschmelzenden Stoffen (Z. angew. Phys. **15**, 301-307, 1963, No. 4). (A new procedure for growing crystals from substances with a high melting point; in German.)

A 70: G. Schiefer: Kleine Ferritantennen für Meterwellen (Arch. elektr. Übertr. **17**, 289-294, 1963, No. 6). (Small ferrite aerials for metre wavelengths; in German.)
See also Philips tech. Rev. **24**, 332-336, 1962/63 (No. 10).

A 71: P. Gerthsen and K. H. Härdtl: Eine Methode zum direkten Nachweis von Leitfähigkeitsinhomogenitäten an Korngrenzen (Z. Naturf. **18a**, 423-424, 1963, No. 3). (A direct method for demonstrating conductivity inhomogeneities at grain boundaries; in German.)

VOLUME 25, 1963/64, No. 11/12 pp. 289-358 Published 7th January 1965

Philips Technical Review

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

SOLID STATE MASERS AND THEIR USE IN SATELLITE COMMUNICATION SYSTEMS

by J. C. WALLING *) and F. W. SMITH *).

621.375.9:538.56

The design of a maser for practical applications calls for a combination of electronics and solid state physics. The article below discusses this combination. A full description is also given of the travelling wave maser which was built at the Mullard Research Laboratories for use in experiments with the satellites Telstar and Relay.

Introduction

It is now well known that the solid state maser amplifier provides the most sensitive available means of amplifying microwave radiation. It is potentially capable of permitting the detection of signals in the microwave region having energies of only a few quanta. The device has therefore excited considerable interest among radio astronomers and more recently among communications engineers. Indeed the spectacular success of the recent satellite communication projects Telstar and Relay was due, in no small part, to the use of solid state maser amplifiers in the receiving stations.

It is the purpose of this article to discuss the design, performance and applications of travelling wave maser (TWM) amplifiers with particular reference to the maser recently developed at Mullard Research Laboratories for use at the Communication Satellite Earth Station of the British General Post Office at Goonhilly Down, Cornwall. A preliminary account of this device has already been given elsewhere [1]) **).

Satellite communications

The first experiments in communication via an artificial earth satellite were carried out in 1960 with equipment designed and constructed by Bell Telephone Laboratories and using the 30 m diameter metallized balloon Echo I [2]) which was launched from Cape Kennedy into an almost circular orbit in August 1960 at an altitude of about 1600 km. Echo acts as a passive reflector of signals from a ground transmitter, and the scattered signal is received at other ground stations.

The path loss over the complete circuit may be shown to be

$$L = \frac{(4\pi)^3\, d_1^{\,2} d_2^{\,2}}{G_1 G_2\, \lambda^2 \sigma},$$

where d_1 and d_2 are the path lengths from the two ground stations to the satellite, G_1 and G_2 are the aerial gains over an isotropic radiator appropriate to the two stations, λ is the wavelength and σ is the scattering cross section of the satellite; see *fig. 1.*

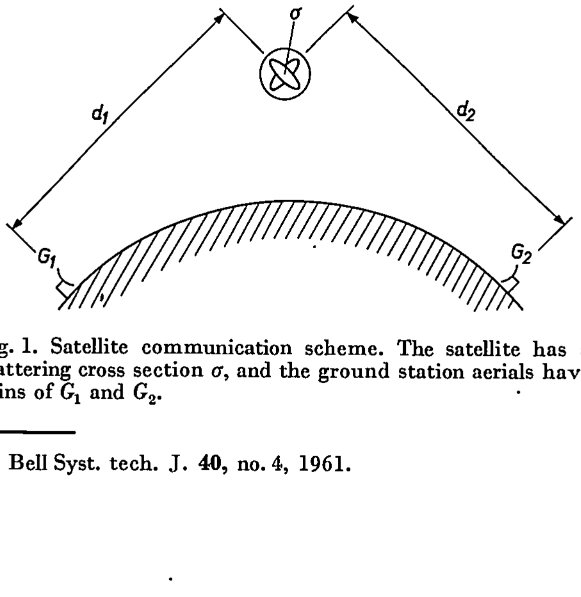


Fig. 1. Satellite communication scheme. The satellite has a scattering cross section σ, and the ground station aerials have gains of G_1 and G_2.

*) Mullard Research Laboratories, Redhill (Surrey), England.

**) *Note of the editor:* In view of the fact that the Gaussian system is in common use in the current literature on masers, we have refrained from converting to the rationalized Giorgi system. For such a conversion see Appendix III.

[1]) J. C. Walling and F. W. Smith, Brit. Commun. and Elec., Aug. 1962.

[2]) Bell Syst. tech. J. **40**, no. 4, 1961.

From this expression the magnitude of the received signal can be calculated. Assuming that the transmitter and receiver aerials have gains of about 45 dB (a practical figure), then with the satellite (scattering cross section ~730 m^2) in the most favourable position with respect to two stations 4800 km apart, we find that the path loss at 2400 Mc/s is about 180 dB and therefore with a 10 kW transmitter the received carrier power is 10^{-14} watt; signals of this order may easily be swamped by noise. For example, a good low noise receiver with 1 Mc/s bandwidth and a noise factor of 1.06 dB (equivalent noise temperature [3]) of 50 °K) has an equivalent input noise power of 0.7×10^{-15} watt and could thus, for this case, only realize a carrier-to-noise power ratio of about 10 dB.

Initial experiments were carried out with receivers of much narrower bandwidth than this to achieve satisfactory signal-to-noise ratios, and although the experiments were successful the use of a narrow bandwidth always severely limits the amount of information which can be put over the channel.

In a world-wide communication system it is desirable to use satellites at greater altitudes than the 1600 km at which Echo I is orbiting because of the need to obtain longer periods of mutual visibility between ground stations. However this invariably implies an even greater path attenuation and even smaller received signals.

The second generation of communication satellites, Telstar, developed by the Bell Laboratories and launched in 1963 [4]), and Relay, developed by R.C.A. for N.A.S.A. and launched in 1963 [5]), employs active repeaters in the satellite. The signal from one ground station is received in the satellite, frequency converted, amplified and re-radiated at a power level of about 2 watts (Telstar I and II) and 10 watts (Relay I). By introducing gain in the system in this way it is possible to work at greater satellite altitudes with longer periods of mutual visibility between ground stations and also to increase the bandwidth of the system. Furthermore, use of these greater altitudes is desirable in order that the active satellites avoid the van Allen radiation belts. More sophisticated aerials with gains above 55 dB have been built for these experiments and signals received from Telstar I at slant ranges (e.g. d_1 in fig. 1) of between 4000 and 10 000 kilometres were in the range between 10^{-12} and 10^{-13} watt. This follows from the ratio of power received at the aerial P_A to that radiated substantially isotropically by the satellite P_S which is:

$$\frac{P_A}{P_S} = \frac{G\lambda^2}{16\pi^2 d^2} \quad (G \text{ being the aerial gain}).$$

The bandwidth of the system using an active repeater in the satellite may be increased to about 25 Mc/s since the noise power introduced by a receiver with such a bandwidth and an equivalent noise temperature of 50 °K is now 1.75×10^{-14} watt and even at ranges up to 10 000 kilometres the carrier-to-noise power ratio is greater than 10 dB. Sophisticated modulation techniques such as wideband frequency modulation with FM feedback in the receiver are currently used to improve further the overall signal-to-noise ratio [5]).

From the foregoing it is apparent that receivers with noise temperatures of the order of 50 °K or less are necessary if wide band satellite communication systems are to be successfully realized with practical satellite transmitter powers which, in the present state of the art, are limited to a few watts. Clearly then it is necessary to pay careful attention to the reduction of all sources of noise in the receiving system and it is thus desirable that the first stage amplifier should have the lowest possible noise temperature; this amplifier should then be a maser.

In the receiving system at Goonhilly Down the maser amplifier is mounted in a cabin at the back of a steerable, 26 metre diameter, parabolic antenna illustrated in *fig. 2. Figure 3* is a photograph of the maser in its operating position in the cabin.

The combination of this antenna and the maser amplifier results in a receiving system having the necessary high directivity and very low noise temperature.

Basic principles of maser operation

Before describing a particular device we shall first make a brief re-examination of the basis of maser operation.

The fundamental principle of maser operation is indicated in the name [6]): microwave amplification

[3]) When we speak of the equivalent noise temperature T of an amplifier we mean that the noise added to the signal by the amplifier is equivalent to that which would be generated by a matched resistive termination at temperature T radiating into the input waveguide of the amplifier. However the noise factor F of the amplifier is, according to the standard definition:

$$F = \frac{\text{input signal}}{\text{input noise}} \Big/ \frac{\text{output signal}}{\text{output noise}}$$

when the input is generated by a matched source at a temperature T_s = 290 °K. Hence:

$$F = 1 + \frac{T}{290}\,.$$

[4]) "The Telstar experiment", Bell Syst. tech. J. 42, no. 4, 1963.

[5]) Programme and Conference Digest, I.E.E. Conference on Satellite Communications, Session 8, Nov. 1962.

[6]) J. P. Gordon, H. J. Zeiger and C. H. Townes, Phys. Rev. **99**, 1264, 1955.

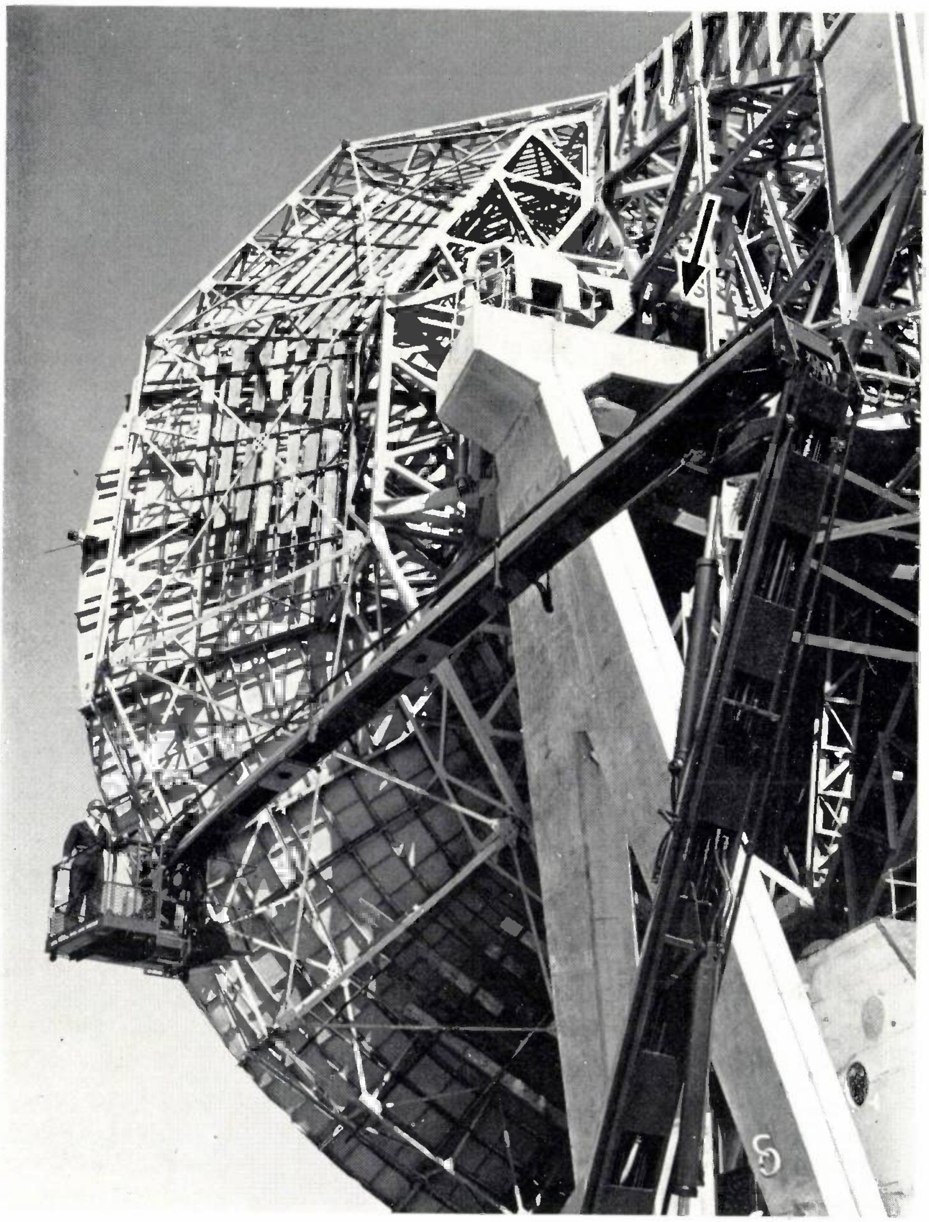


Fig. 2. The 26 meter diameter antenna at the Communication Satellite Earth Station at Goonhilly Down, Cornwall. This antenna is used for receiving signals from the Telstar and Relay satellites. The arrow indicates the cabin in which the maser is mounted. (Photo published by permission of the British G.P.O.)

by stimulated emission of radiation. The radiation is emitted from particles which make a transition from the upper to the lower of two quantum energy states. In the Bloembergen solid state maser the fundamental phenomenon which gives rise to the relevant energy states is paramagnetic resonance. We will first discuss paramagnetic resonance for the simple case of an assembly of free electrons and the stimulated emission which may occur in such a system.

Each electron in our assembly possesses a spin $h/4\pi$ and a magnetic moment of one Bohr magneton:

$$\beta = \frac{eh}{4\pi mc}, \quad \ldots\ldots \quad (1)$$

where e = the electronic charge, h = Planck's constant, m = the electronic mass, and c = the free space velocity of light. If the system is subjected to the action of a static magnetic field H_0, each electron can occupy one of two energy states often referred to as energy levels, the energy difference between the states being

$$\Delta E = 2\,\beta H_0. \quad \ldots\ldots \quad (2)$$

In thermal equilibrium the relative populations n_1 and n_2 of the two states are determined by Boltzmann statistics, i.e.:

$$\frac{n_2}{n_1} = \exp\left(-\frac{\Delta E}{kT}\right). \quad \ldots\ldots \quad (3)$$

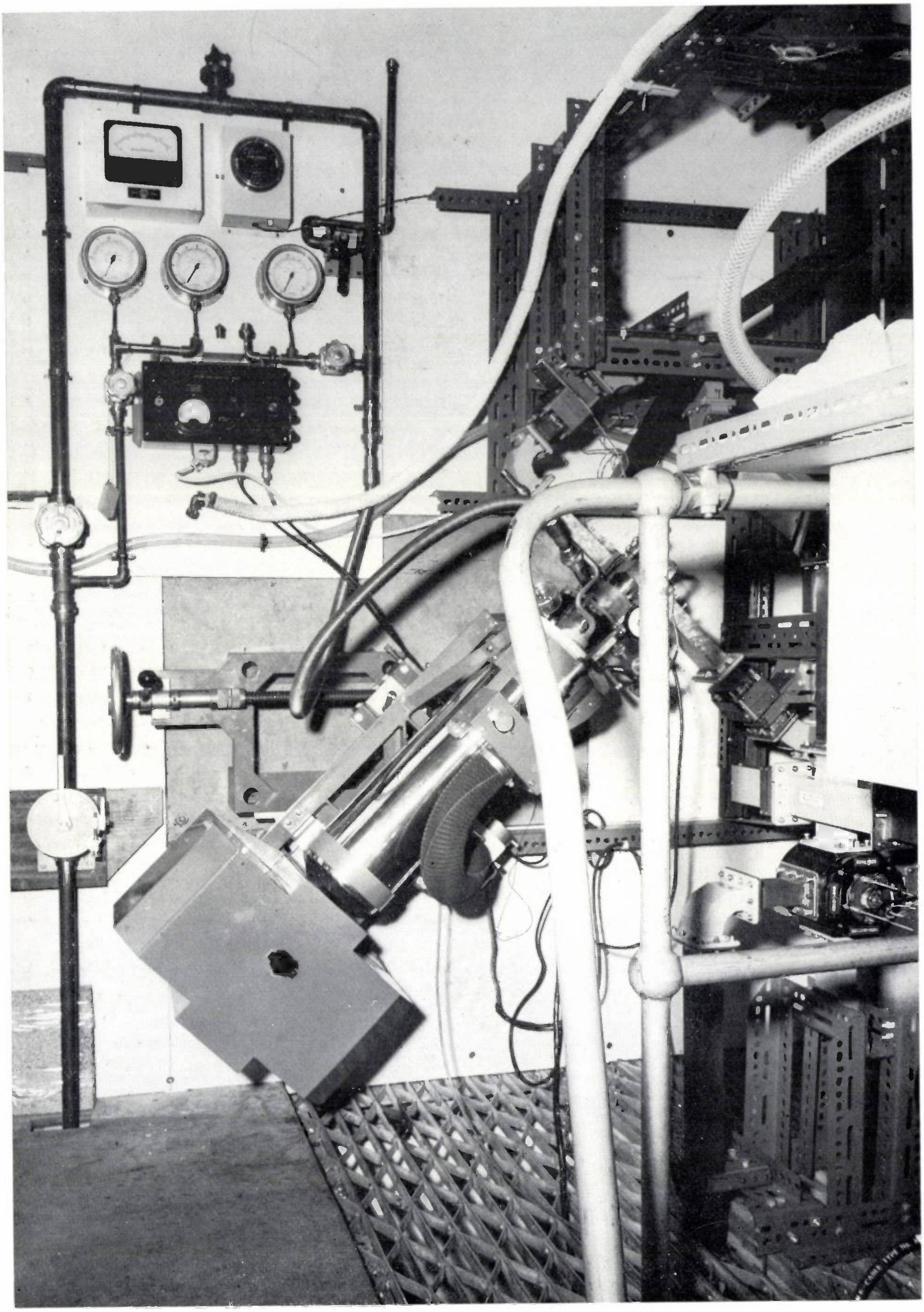


Fig. 3. Travelling wave maser amplifier in its operating position on the parabolic antenna shown in fig. 2. (Photo published by permission of the British G.P.O.)

where n_2 corresponds to the higher energy state. It was shown by Einstein [7]) that if electromagnetic radiation of frequency f given by

$$hf = \Delta E \quad . \ . \ . \ . \ . \ . \ . \ . \ . \quad (4)$$

is incident upon the system, transitions are *stimulated* from level *1* to level *2* and from level *2* to level *1*, and that upward and downward transitions are equally probable. From equations (2) and (4):

$$f = \frac{eH_0}{2\pi mc} \quad . \ . \ . \ . \ . \ . \ . \quad (5)$$

This type of resonant interaction between a system of spins and a radiation field is known as paramagnetic resonance.

Certain features of this interaction can be understood in classical terms. The assembly of electrons when subject to the action of a magnetic field **H** will have a net magnetic moment **M** and a net angular momentum $1/\gamma\mathbf{M}$. The gyromagnetic ratio γ is equal to e/mc for free electrons. In general $\gamma = ge/2mc$, g being the spectroscopic splitting factor (equal to 2 for free electrons). The equation of motion for the angular momentum of the system is:

$$\frac{1}{\gamma}\frac{\mathrm{d}\mathbf{M}}{\mathrm{d}t} = \mathbf{M}\times\mathbf{H}. \quad . \ . \ . \ . \ . \quad (6)$$

If we suppose that the magnetic field **H** has a large static component $H_z = H_0$ and small time dependent components H_x and H_y varying as $\mathrm{e}^{\mathrm{j}\omega t}$, then M_z is time independent and for M_x and M_y we have:

$$\frac{\mathrm{d}M_x}{\mathrm{d}t} = \mathrm{j}\omega M_x = \gamma M_y H_0 - \gamma M_z H_y\,,$$

$$\frac{\mathrm{d}M_y}{\mathrm{d}t} = \mathrm{j}\omega M_y = \gamma M_z H_x - \gamma M_x H_0\,.$$

Writing $\omega_0 = \gamma H_0$ and resolving the (x, y) magnetization and field into left and right circularly polarized components:

$$M_x + \mathrm{j}M_y = M_+ \quad \text{and} \quad M_x - \mathrm{j}M_y = M_-,$$

$$H_x + \mathrm{j}H_y = H_+ \quad \text{and} \quad H_x - \mathrm{j}H_y = H_-,$$

we obtain:

$$M_+ = \frac{\gamma M_z}{\omega + \omega_0} H_+\,,$$

$$M_- = \frac{-\gamma M_z}{\omega - \omega_0} H_-\,.$$

These equations indicate that the interaction is of a resonant character, circularly polarized fields of positive sense strongly exciting circularly polarized magnetization of the same sense at $\omega = \omega_0$, whilst for the opposite direction of applied field ($\omega = -\omega_0$) a resonant interaction is observed for the opposite sense of polarization. In this classical situation the magnetization executes a precessional motion about the direction of the applied static field with frequency ω_0.

This classical analysis demonstrates that the phenomenon of paramagnetic resonance is due to an interaction between an RF magnetic field and the magnetic moment of the spin system, and that this interaction will only occur if the RF magnetic field is of the correct frequency and contains a circularly polarized component of appropriate sense.

The two energy levels of our earlier discussion might be regarded in the classical case as corresponding to the magnetic moments of individual electrons being aligned parallel or antiparallel to the applied field.

We have seen that in the two-level system transitions are stimulated between the two levels with equal probability in either direction. Downward transitions take place with the emission of radiation and upward transitions are accompanied by absorption. The radiation emitted in the downward transitions is coherent and in phase with the original stimulating field. However, although the probabilities of upward and downward transitions for each electron are equal, the number of upward transitions exceeds the number of downward transitions in a system initially in thermal equilibrium because of the Boltzmann distribution (eq. 3) between the levels, and thus the net effect is that of absorption of energy from the incident RF field.

The maser is based on the following idea: If by some means the distribution of electrons between the levels can be changed so that the number in the upper level exceeds that in the lower, the dominant paramagnetic resonance process would be that of stimulated emission rather than absorption, and since the emitted radiation is in phase with the stimulating radiation, amplification of the original signal would be observed. For readily available magnetic fields paramagnetic resonance occurs at microwave frequencies (2.8 Gc/s at 1000 Oe according to eq. (5)) and thus this phenomenon offers the possibility of microwave amplification.

Although there are means (reviewed by Wittke [8])) of inverting the populations of a solid state two-level system, the inversion is transient and continuous amplification cannot be obtained. It was

[7]) A. Einstein, Phys. Z. **18**, 121, 1917.

[8]) J. P. Wittke, Proc. I.R.E. **45**, 291, 1957.

not until 1956 when Bloembergen [9]) proposed the three-level maser that a method of obtaining continuous amplification of an RF signal by a spin system became available.

Paramagnetic ions in crystals

In the three-level maser we are concerned, not with an assembly of electrons, but with paramagnetic ions in a crystal lattice.

The magnetic properties of paramagnetic ions may be rather complicated because we are concerned both with the spin and orbital motion of the unpaired electrons. Both spin and orbit have angular momentum and an associated magnetic moment. If the spin and orbital angular momenta are characterised by quantum numbers S and L respectively then the corresponding angular momenta and magnetic moments are:

$$\text{spin:}\quad \frac{h}{2\pi}\sqrt{S(S+1)} \quad \text{and} \quad -\frac{g_S\,\beta h}{2\pi}\sqrt{S(S+1)},$$

$$\text{orbit:}\quad \frac{h}{2\pi}\sqrt{L(L+1)} \quad \text{and} \quad -\frac{g_L\,\beta h}{2\pi}\sqrt{L(L+1)},$$

β being the Bohr magneton (1) and g_S and g_L the spin and orbital spectroscopic splitting factors. However since $g_S = 2$ and $g_L = 1$ there is no simple relation between the *resultant* magnetic moment and angular momentum. The latter is characterised by a quantum number J which is the vector sum of L and S and has magnitude $(h/2\pi)\sqrt{J(J+1)}$. The resultant magnetic moment is $-(g_J\beta h/2\pi)\sqrt{J(J+1)}$, g_J being the Landé splitting factor. When acted upon by a magnetic field such an ion can occupy one of $(2J+1)$ equally spaced energy levels.

Detailed treatment of the effect of the electric field produced by neighbouring ions of the crystal lattice on the paramagnetic ion lies beyond the scope of this article; for further discussion of this subject the reader is referred to the review article by Bleaney and Stevens [10]). We may note however that the effect of this crystal field depends on its symmetry and on the relative magnitudes of the field and the spin-orbit coupling in the paramagnetic ion.

In most maser materials the effect of the crystal field is to cause large energy differences (corresponding to frequencies in the optical region of the spectrum) between states of different orbital motion and only the lowest of these levels is occupied at ordinary temperatures (cf. equation 3). For example the ground state of the Cr^{3+} ion (the active ion in ruby — to date the most widely used maser material) is split by the action of a field of cubic symmetry into a low lying orbital singlet and two very much higher triplets (*fig. 4a*). Ions in the lowest state have no net orbital moment and the paramagnetism of this material is due solely to the electron spin. There are three unpaired electrons in this ion and thus four spin states all of which have the same energy (fourfold spin degeneracy). As a result of deviations of the crystal field from cubic symmetry and some residual spin orbit coupling, the fourfold degenerate lowest level is split into two doubly degenerate levels; this latter energy splitting is called the ground state zero field splitting and generally corresponds to frequencies in the microwave range (in ruby it is 11.46 Gc/s; fig. 4*b*).

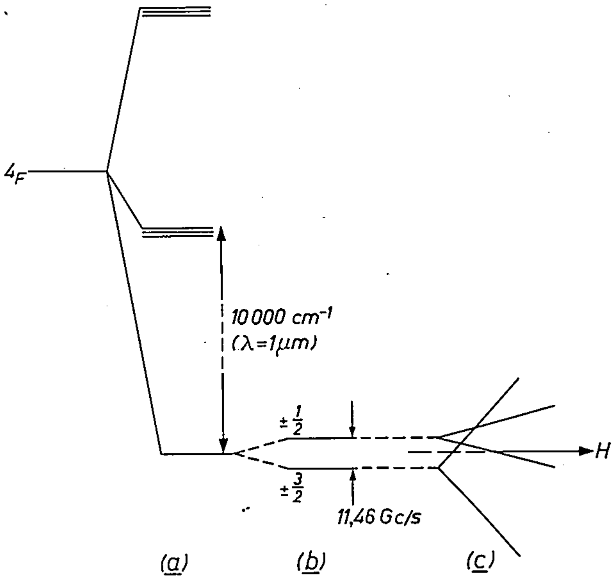


Fig. 4. Energy level diagram. The splitting of the ground state of the free chromium (Cr^{3+}) ion by a cubic crystal field with trigonal distortion.
a) Splitting of orbital levels by a crystal field of cubic symmetry.
b) Partial lifting of four-fold spin degeneracy of lowest orbital level by combined action of axial field and spin orbit coupling (zero field splitting).
c) Zeeman splitting of ground state levels in a magnetic field *H*.

In the presence of a magnetic field each of the doubly degenerate ground state levels is further split and it is with these Zeeman levels (so called because this type of level splitting gives rise to the splitting of optical spectral lines first observed by Zeeman in 1896) of the ground state that we are concerned in the maser (fig. 4*c*).

The behaviour of the Zeeman levels depends on the magnitude and direction of the applied static magnetic field with respect to the crystal field. In the case in which the crystal is axially symmetrical

[9]) N. Bloembergen, Phys. Rev. **104**, 324, 1956.

[10]) B. Bleaney and K. W. H. Stevens, Rep. Prog. Phys. **16**, 108, 1953 — An introductory article on paramagnetic resonance by J. S. van Wieringen appeared in Philips tech. Rev. **19**, 301, 1957/58.

and the static magnetic field is parallel to this axis, the ground state Zeeman level splitting is linearly dependent on the applied field and the corresponding spin states may be individually characterised by magnetic quantum numbers m ($mh/2\pi$ being the component of angular momentum along the magnetic field). Such states are called pure states. For an ion with three unpaired spins (such as Cr^{3+}) m can take the values $+\frac{3}{2}, +\frac{1}{2}, -\frac{1}{2}, -\frac{3}{2}$.

Only transitions between levels corresponding to $\Delta m = \pm 1$ (or 0) are allowed. However, in the general case in which the static magnetic field is not parallel to the axis of symmetry the energy differences between states do not vary linearly with the applied magnetic field. This is shown in *fig. 5*, which is the energy level diagram for the ground state levels of the chromium ion in ruby when the angle between the magnetic field and the axis of symmetry of the crystal field is 90°. Furthermore these energy states are not pure states but linear combinations of pure states; thus they can no longer be characterized by a single magnetic quantum number — such states are called mixed states. As a consequence of this mixing of states, transitions between any pair of ground state levels are allowed although all transitions are not of equal probability.

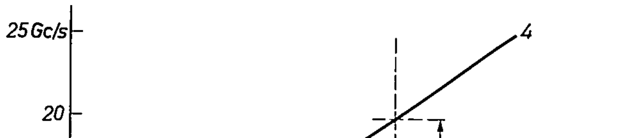


Fig. 5. Ground state Zeeman levels in ruby (Cr^{3+} in Al_2O_3) as functions of the magnetic field H. The magnetic field is at 90° to the three-fold axis of symmetry of the crystal field. The arrows indicate the operating pump transition *1-4* (30 150 Mc/s) and the signal transition *1-2* (4170 Mc/s) in a field of 3280 Oe.

The three level maser

We have now seen broadly what determines the ground state Zeeman levels of paramagnetic ions in a crystal and in 1956 Bloembergen proposed a simple means of producing a continuously inverted population distribution between a pair of these levels which provides the basis of the continuous wave solid state maser as we know it today.

Let us consider a system in which we have three unequally spaced Zeeman levels (*fig. 6*). We will suppose that we have *two* impressed signals which have frequencies f_{12} and f_{13} and which stimulate transitions between levels *1* and *2* and *1* and *3* with probabilities W_{12} and W_{13} respectively ($W_{12} = W_{21}$, $W_{13} = W_{31}$). In addition to the transitions stimulated by these external signals there are also spontaneous downward transitions and the thermal transitions which are responsible for the energy exchange between the spin system and the lattice and therefore determine the rate of approach to thermal equilibrium of a perturbed spin system — the "spin lattice relaxation". These latter transitions are radiationless and their probabilities are different in the two directions; these probabilities are w_{12}, w_{21}, w_{13}, w_{31}, w_{23} and w_{32}. The spontaneous transitions are very improbable at microwave frequencies and may be neglected in this discussion. If the three levels have populations n_1, n_2, n_3 then we may write:

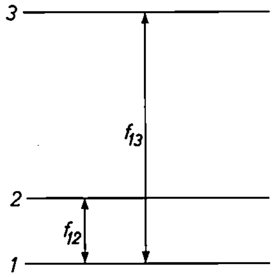


Fig. 6. Three energy level system. f_{12} signal frequency. f_{13} pump frequency.

$$\left.\begin{aligned}\frac{dn_1}{dt} &= -n_1(W_{13} + w_{13} + W_{12} + w_{12}) + \\ &\quad + n_2(W_{21} + w_{21}) + n_3(W_{31} + w_{31}), \\ \frac{dn_2}{dt} &= -n_2(W_{21} + w_{21} + w_{23}) + \\ &\quad + n_1(W_{12} + w_{12}) + n_3(w_{32}).\end{aligned}\right\} \quad . \quad (7)$$

If we have no impressed signals the thermal equilibrium populations of the levels are determined by

Boltzmann's law (equation 3) and it follows from the Principle of Detailed Balancing [11]) that:

$$\left.\begin{aligned} \frac{w_{13}}{w_{31}} &= \left(\frac{n_3}{n_1}\right)_0 = \exp\left(-\frac{nf_{13}}{kT}\right) \\ \frac{w_{12}}{w_{21}} &= \left(\frac{n_2}{n_1}\right)_0 = \exp\left(-\frac{hf_{12}}{kT}\right) \\ \frac{w_{23}}{w_{32}} &= \left(\frac{n_3}{n_2}\right)_0 = \exp\left(-\frac{hf_{23}}{kT}\right) \end{aligned}\right\}, \quad . \; . \quad (8)$$

the subscript 0 signifying the magnitude in thermal equilibrium.

If we now make W_{13} very large, i.e., we make the amplitude of the signal at the frequency f_{13} very large, the transition at f_{13} becomes saturated, that is, the populations of levels *1* and *3* become equal and we find from (5) and (6) at equilibrium ($\mathrm{d}n_1/\mathrm{d}t = \mathrm{d}n_2/\mathrm{d}t = 0$):

$$\frac{n_2}{n_1} = \frac{w_{23}\exp\,(hf_{23}/kT) + W_{12} + w_{12}}{W_{12} + w_{12}\exp\,(hf_{12}/kT) + w_{23}}. \quad . \quad (9)$$

If our signal at f_{12} is very small we may write $W_{12} = 0$ and hence

$$\frac{n_2 - n_1}{N} = \frac{w_{23}\,[\exp(hf_{23}/kT) - 1] - w_{12}\,[\exp(hf_{12}/kT) - 1]}{w_{23}[2 + \exp(hf_{23}/kT)] + w_{12}[2\exp(hf_{12}/kT) + 1]} \quad . \; . \; . \quad (10)$$

in which $N = n_1 + n_2 + n_3 = 2n_1 + n_2$ is the total number of active ions involved. This equation expresses the fact that in a three level system the population of level *2* can be made to exceed that of level *1* by applying a large amplitude signal of frequency f_{13} (called the "pump") if

$$w_{23}\,[\exp(hf_{23}/kT) - 1] > w_{12}\,[\exp(hf_{12}/kT) - 1]. \quad (11)$$

If both hf_{23} and $hf_{12} \ll kT$ this condition becomes:

$$w_{23}f_{23} > w_{12}f_{12}\, . \; . \; . \; . \; . \; . \quad (12)$$

In the absence of precise information concerning the relative magnitudes of the thermal transition probabilities w_{12}, w_{23} it is usually assumed that they are equal, which may not be too far from correct provided we are dealing with frequencies of the same order of magnitude. Making this assumption we see that a condition for maser action in a system of this kind is that the pump frequency ($f_{13} = f_{12} + f_{23}$) should be greater than twice the signal frequency f_{12}.

Three level maser action is illustrated by the diagram of *fig. 7* in which population is plotted against energy. The populations of levels *1* and *3* which result from applying the pump signal are indicated by the heavy lines. The lower the ambient temperature the larger the population difference ($n_2 - n_1$) becomes and thus masers are commonly operated at very low temperatures.

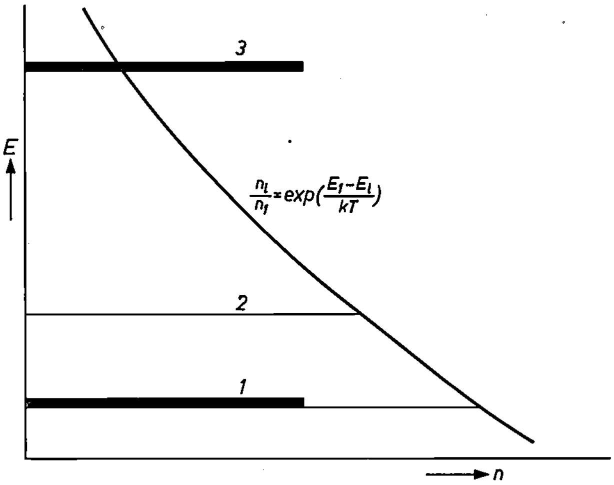


Fig. 7. Population inversion in a three-level system. Population is plotted horizontally for the three energy levels. The exponential curve indicates the Boltzmann distribution for thermal equilibrium at a given temperature. The heavy lines indicate the populations of levels *1* and *3* when the pump signal is applied. The relative population of level *2* is determined by the thermal transition probabilities w_{12}, w_{23}. For $w_{23}f_{23} > w_{12}f_{12}$ an inversion of population between levels *1* and *2* obtains. (In the opposite case, $w_{23}f_{23} < w_{12}f_{12}$, the populations of levels *2* and *3* would be inverted.)

The majority of solid state masers have used crystals in which the active ion is Cr^{3+} (e.g. ruby, chromicyanide, emerald, rutile) and here as we have seen we have four Zeeman levels in the ground state. Three-level maser action can still be obtained by choosing three of the four available levels although the rate equations are somewhat complicated, due to the greater number of relaxation processes which can now take place. Occasionally double pumping schemes in which all four energy levels take part can be used to advantage. In the case where the energy levels are completely symmetrical — which occurs in ruby, for example, when the angle between the static applied magnetic field and the symmetry axis is $\cos^{-1}(1/\sqrt{3})$ i.e. 54°44′ — it is possible to employ "push pull" pumping, the signal transition being between levels *2* and *3* and pump transitions between levels *1* and *3* and *2* and *4*, a pump of one frequency saturating both the *1-3* and *2-4* transitions [12]) (*fig. 8*).

[11]) According to this principle every process of transformation or exchange of energy which occurs in a system in thermodynamic equilibrium is invariably accompanied by an analogous reverse process and the two processes occur with equal frequency.

[12]) G. Makhov *et al.*, Phys. Rev. **109**, 1399, 1958.

It is convenient to relate the inverted population difference (Δn) to the thermal equilibrium population difference (Δn_0) through

$$\Delta n = - I\, \Delta n_0 \,,$$

I being known as the inversion produced by the pumping scheme.

Transition probabilities

The probability W_{ij} that an applied RF magnetic field will stimulate transitions between a given pair of Zeeman levels i and j is a vital parameter in maser design. It can be calculated by means of perturbation theory but we will content ourselves by quoting the result of this calculation which is:

$$W_{ij} = W_{ji} = \frac{1}{4}\, \gamma^2 \varphi(f)\, |(\mathbf{A} + \mathrm{j}\mathbf{B})\, (\mathbf{H} + \mathrm{j}\mathbf{H}')|^2, \quad (13)$$

and by briefly discussing this expression and the occurring symbols. γ is the gyromagnetic ratio already introduced:

$$\gamma = \frac{2\pi g \beta}{h} \,. \quad . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

$\varphi(f)$ is a normalized line shape function:

$$\int_0^\infty \varphi(f) \mathrm{d}f = 1$$

which expresses the fact that because a given ion spends only a limited time in a particular energy state before exchanging its energy with another ion (or the lattice), the energy levels have a finite width and resonant interaction is obtained at signal frequencies not only equal to f_0 but also around f_0. When interaction between neighbouring ions (dipole-dipole or spin-spin) is the dominant process, $\varphi(f)$ is a "Lorentzian line shape function":

$$\varphi(f) = \frac{2/T_2}{4\pi^2\,(f - f_0)^2 + 1/T_2{}^2} \,. \quad . \quad . \quad (15)$$

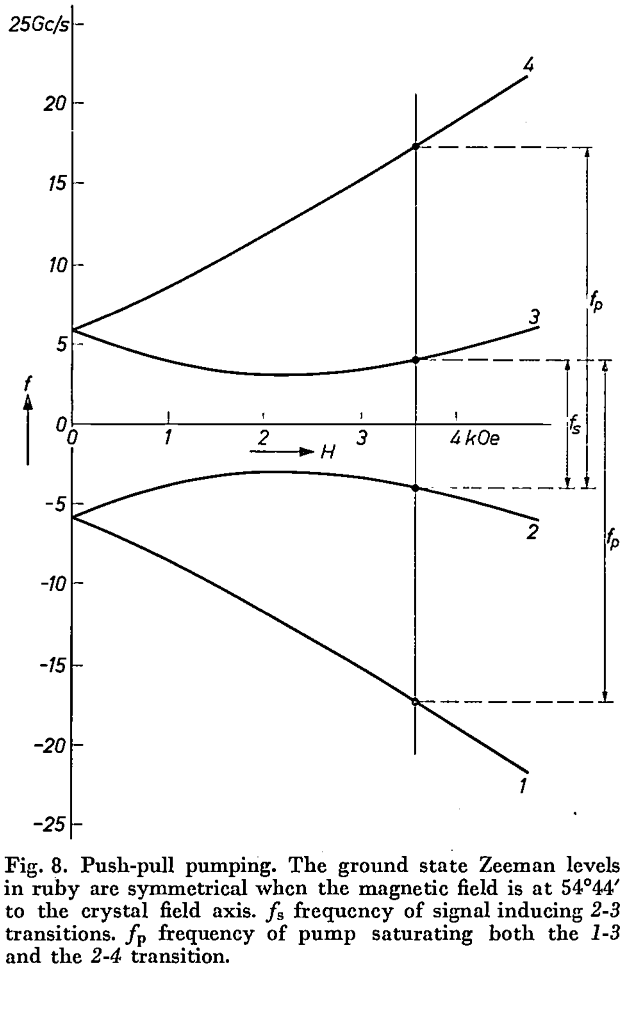


Fig. 8. Push-pull pumping. The ground state Zeeman levels in ruby are symmetrical when the magnetic field is at 54°44′ to the crystal field axis. f_s frequency of signal inducing *2-3* transitions. f_p frequency of pump saturating both the *1-3* and the *2-4* transition.

T_2 is the relaxation time characterizing the interaction and is typically of the order of 10^{-9} s. T_2 determines the paramagnetic resonance linewidth $1/\pi T_2$. The RF magnetic field is represented by the term $\mathbf{H} + \mathrm{j}\mathbf{H}'$ (that is, the actual RF magnetic field is the real part of $(\mathbf{H} + \mathrm{j}\mathbf{H}')\, \exp\, (2\pi \mathrm{j} f t)$). $\mathbf{H}$ and $\mathbf{H}'$ are two mutually orthogonal magnetic field vectors.

$\mathbf{A}$ and $\mathbf{B}$ are two mutually perpendicular vectors (termed matrix element vectors) whose magnitude and direction can be calculated [13]) from a knowledge of the electron spin resonance spectrum of the ion in the particular crystal environment under consideration. For transitions between pure states, the magnitudes of $\mathbf{A}$ and $\mathbf{B}$ are equal and in general of the order of unity for allowed transitions (and zero for forbidden transitions). Thus for pure states eq. (13) expresses the fact that maximum transition probability is obtained when the RF field is circularly polarized and of the appropriate sense which is in agreement with the earlier classical analysis.

In general, however, we are not concerned with pure states and for a given transition A and B are not equal and depend both on the magnitude of the static magnetic field and the angle a between this field and the axis of symmetry of the crystal field. As an illustration of this the variation of A and B for two transitions in ruby is plotted in *fig. 9* from the data of Chang and Siegman [13]).

By analogy with the classical model we may regard the crystal field as producing an anisotropic constraint on the precession of the magnetization and thus maximum interaction occurs in this case not with circularly polarized fields but with elliptically polarized fields of the appropriate ellipticity (A/B) and sense.

[13]) W. S. C. Chang and A. E. Siegman, Stanford Technical Report 156-2, 1958 (essential data from this report are quoted by J. Weber, Rev. mod. Phys. **31**, 681, 1959).

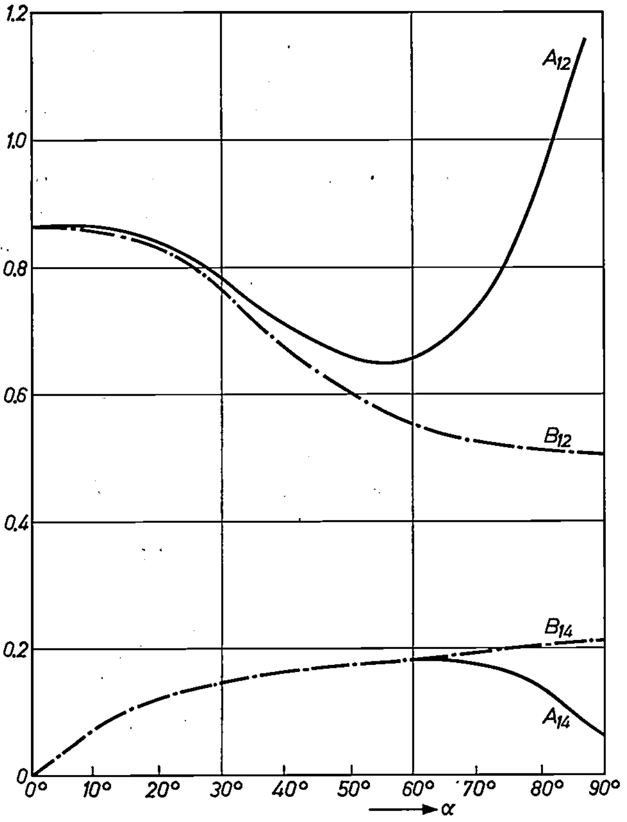


Fig. 9. The matrix element vectors A and B as a function of α, the angle between magnetic field and crystal field axis, for the *1-2* and the *1-4* transition in ruby in a magnetic field of 3000 Oe.

Choice of material for a maser

We can now state the conditions which a good maser material should satisfy. These are as follows:

(1) The active ions should occupy equivalent sites in the crystal; otherwise in general at a given orientation of the crystal only ions on a certain site will take part in the interaction with the RF field.

(2) The active ions should not have a nuclear magnetic moment. This would cause hyperfine splitting of energy levels and effectively decrease the number of ions available for maser action at a particular frequency.

(3) The interaction between the spin system and the lattice should not be too strong if we are to saturate the pump transition fairly readily — this is another reason for operating at very low temperatures.

(4) The material should be mechanically robust and should withstand repeated cooling to very low temperatures. For instance, chromium doped potassium cobalticyanide ($K_3Co(CN)_6$) which is otherwise a very satisfactory material fails in this respect.

(5) The material should be available as large single crystals.

Few, if any, materials fulfil all these requirements: synthetic ruby (chromium-doped aluminium oxide) comes very close and is widely used, emerald looks a promising material but is at present difficult to synthesize as large single crystals, chromium- and iron-doped rutile are also frequently used but present difficulties at high frequencies as a result of the very high dielectric constant which necessitates the use of very small microwave structures in the maser. A new material which is extremely interesting is chromium doped (Cr^{3+}) zinc tungstate [14]).

Maser devices; the travelling wave maser

Having seen how stimulated emission from a paramagnetic crystal can occur at microwave frequencies, we must now consider how it can be used in an amplifier. The strength of the interaction between the spin system and the RF magnetic field is, as we have seen, determined by the square of the amplitude of this field within the crystal (eq. 13) and it is therefore desirable, for a given signal power input to the maser, that this amplitude should be made as large as possible.

The simplest way of obtaining a large RF field within the crystal is to place it in a resonant cavity and this technique has been widely used [15]). Such a cavity maser can be constructed as a reflection device (*fig. 10*) in which the input and output signals

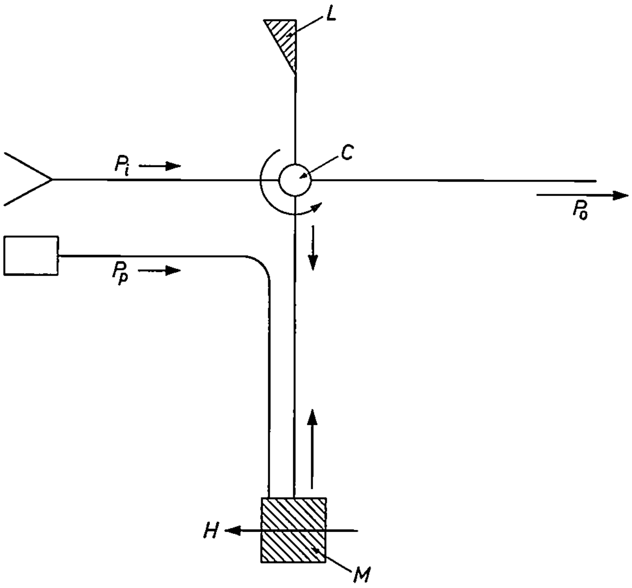


Fig. 10. Cavity maser circuit. P_p pump power. P_i signal input power. P_o signal output power. M maser cavity filled with active material. H static magnetic field. C circulator. L matched load.

[14]) L. G. van Uitert and S. Preziosi, J. appl. Phys. **33**, 2908, 1962.

[15]) H. E. D. Scovil, G. Feher and H. Seidel, Phys. Rev. **105**, 762, 1957; A. L. McWhorter and J. W. Meyer, Phys. Rev. **109**, 312, 1958.

are separated by means of a circulator [16]) or as a transmission device with distinct input and output waveguides. In either case the device is regenerative, that is to say the energy which is emitted by an element of the maser material increases the RF field in the cavity and thus the rate of emission of energy from the element of material itself increases. As a consequence of this positive feedback the product of the bandwidth and the square root of the numerical gain for a single cavity device is constant, and the gain is very sensitive to variations in load impedance and hence there is a tendency to oscillation at high gains. Although it is true that for multiple cavity devices larger bandwidths can be obtained at a given gain, the positive feedback is always present and so too is the possibility of self oscillation.

Another means of obtaining a large RF field is to use a slow wave structure in which the velocity of propagation of electromagnetic energy is substantially less than the free space velocity. Field concentration occurs in such a structure as a consequence of the simple law relating the stored energy per unit length (W_S) to the power P in a wave travelling in the structure which is

$$P = v_g W_s , \quad \ldots \ldots \ldots \quad (16)$$

v_g being the group velocity. W_s is determined by the square of the RF fields integrated over the cross section of the structure and thus if, at a given power level, W_s is made large by velocity reduction so also is the RF field.

A two port device employing this latter principle is known as a travelling wave maser (TWM) and the first working TWM was described by DeGrasse *et al* [17]).

The situation in the TWM is quite different from that in the cavity device in so far as the energy emitted by an element of maser material leads to an increase in the RF field travelling in the structure and thus increases the rate of emission of energy from succeeding elements of material but does not react on the original element. Regeneration and oscillation in a TWM can then only occur as a result of reflections in the external circuitry.

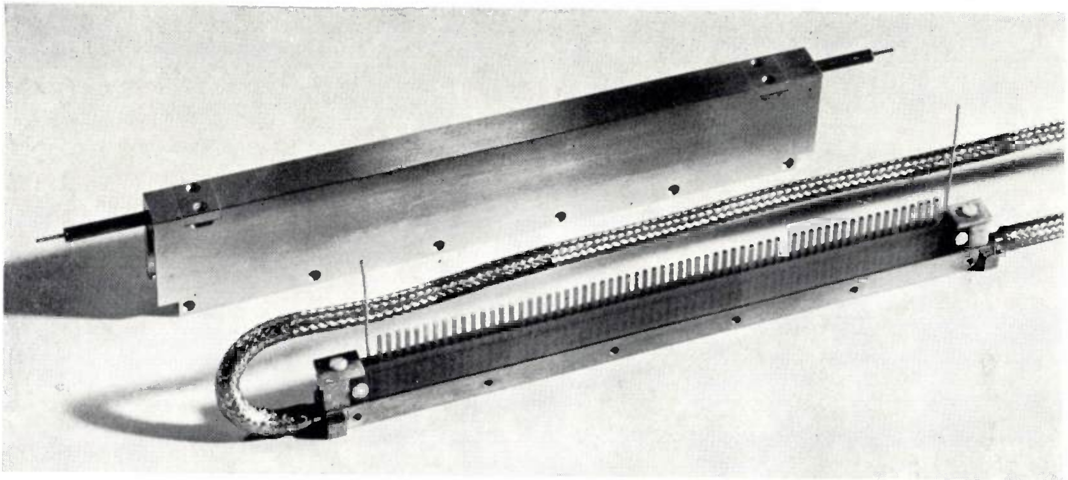


Fig. 11. Photograph of comb slow wave structure as used in a TWM.

The TWM thus has some important advantages over the cavity maser in that it is non-regenerative and has therefore no intrinsic tendency to oscillate. Furthermore, the bandwith is determined mainly by the paramagnetic resonance linewidth of the sample, and decreases only slowly with increasing gain.

Moreover, as will be shown, the device can be constructed to be completely non-reciprocal and hence the need for ancillary non-reciprocal elements essential for cavity maser operation is removed. In addition, operation can be obtained over a very wide centre frequency range — some hundreds of megacycles in a single structure — the centre frequency being simply shifted by adjusting the magnetic field and the pump frequency. These features make the TWM the only really practical embodiment of the maser principle. A typical maser slow wave structure loaded with active material is shown in *fig. 11*.

[16]) A. G. Fox *et al.*, Bell Syst. tech. J. **34**, 5, 1955.

[17]) R. W. DeGrasse, E. O. Schulz-Du Bois and H. E. Scovil, Bell Syst. tech. J. **38**, 305, 1959.

Theory of the travelling wave maser

The small signal gain of the TWM can readily be expressed in terms of two quality factors, Q_0, the intrinsic quality factor of the propagating structure, determined by ohmic and dielectric power losses per unit length P_0, and Q_m, the magnetic quality factor of the maser material which is determined by the power P_m emitted per unit length of the maser material:

$$\frac{1}{Q_\mathrm{m}} = \frac{P_\mathrm{m}}{2\pi f W_\mathrm{s}}\ , \quad \frac{1}{Q_0} = \frac{P_0}{2\pi f W_\mathrm{s}}. \quad . \quad . \quad (17)$$

Let us consider an element of the TWM of length dz (*fig. 12*). Energy is emitted by the maser material in this element but, since only the forward travelling increments of field add in phase, virtually all the energy emitted by the element travels in the forward direction. Energy absorption also takes place in the element due to ohmic and dielectric losses and thus the change dP in power level P in the element is:

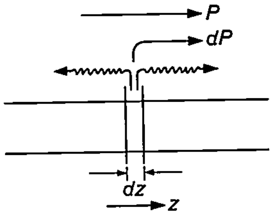


Fig. 12. Emission from an element of a TWM.

$$\mathrm{d}P = (P_\mathrm{m} - P_0)\mathrm{d}z.$$

Using (16) and (17):

$$\mathrm{d}P = 2\pi f W_\mathrm{s}\left(\frac{1}{Q_\mathrm{m}} - \frac{1}{Q_0}\right)\mathrm{d}z$$

$$= P\frac{2\pi f}{v_\mathrm{g}}\left(\frac{1}{Q_\mathrm{m}} - \frac{1}{Q_0}\right)\mathrm{d}z,$$

whence the net gain (in dB) of a maser of length L is:

$$G_\mathrm{n} = 20\pi(\log_{10} \mathrm{e})\frac{fL}{v_\mathrm{g}}\left(\frac{1}{Q_\mathrm{m}} - \frac{1}{Q_0}\right)$$

$$= \frac{27.3 fL}{v_\mathrm{g}}\left(\frac{1}{Q_\mathrm{m}} - \frac{1}{Q_0}\right). \quad . \quad . \quad . \quad . \quad (18)$$

To obtain a high gain within a reasonable length L, Q_m must be made small, Q_0 large and v_g as small as is practicable. The first term of G_n is the "electronic gain" G, i.e. the gain which would be obtained if there were no structure losses.

Both v_g and Q_m are determined by the way in which we combine our propagating structure and active material and this is the main problem in designing a TWM.

It is shown in the Appendix I that Q_m as defined above is given by

$$\frac{1}{Q_\mathrm{m}} = \Delta n h \gamma^2 \frac{\varphi(f)\eta}{2(1+R)}[(A + B)^2 + R(A - B)^2 +$$

$$+ 2\cos 2\Theta (A^2 - B^2)R^{\frac{1}{2}}\ ], \quad (19)$$

in which A and B are the signal transition matrix element vectors, previously defined; Δn is the inverted population density difference of the signal levels; $\varphi(f)$ is the line shape function, in many cases given by eq. (15); Θ is the angle between the larger of the two matrix element vectors and the major axis of the elliptically polarized RF magnetic field. R is a measure of the degree of ellipticity of the field polarization — for circular polarization $R = 0$ or $R = \infty$ depending on the sense of polarization, whilst for linear polarization $R = 1$ (cf. eq. 33). The factor η, the "filling factor", is the fraction of the total magnetic energy stored in the maser material (eq. 38).

The gain of the travelling wave maser is a function of frequency and by combining eq. (18) and (19) the instantaneous bandwidth of the device may readily be calculated. We find that for a Lorentzian line (i.e. $\varphi(f)$ given by eq. (15)) the bandwidth to half power points is:

$$B = \frac{1}{\pi T_2}\sqrt{\frac{3}{G_0 - 3,}}, \quad . \quad . \quad . \quad . \quad (20)$$

G_0 being the uniform field peak electronic gain (in dB) of the TWM and $1/\pi T_2$ the paramagnetic resonance line width (cf. eq. (15)). This relation is plotted for $T_2 = 4.9 \times 10^{-9}$ s in *fig. 13* [18]).

The bandwidth of the TWM may be increased at the expense of gain if we arrange that different sections of the device are acted upon by different static

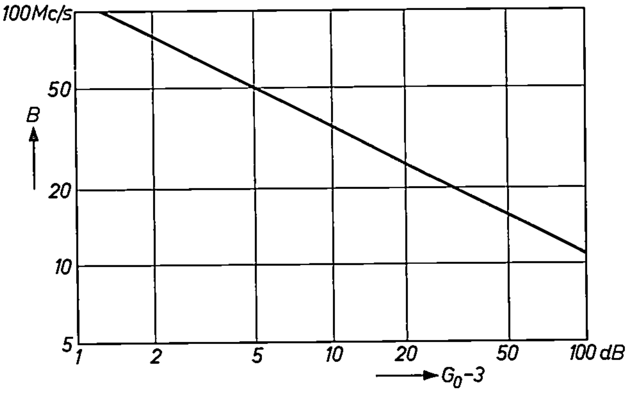


Fig. 13. Bandwidth B as a function of peak electronic gain G_0 for a TWM in a uniform magnetic field, with $T_2 = 4.9 \times 10^{-9}$ s.

[18]) The gain-bandwidth characteristics of the single cavity maser and the *TWM* have been compared by DeGrasse *et al.* [17]).

magnetic fields, thus spreading the centre resonant frequencies of these sections.

This is the most powerful technique for increasing the bandwith of the TWM and we shall consider three types of static magnetic field variation along the length of the structure. These are a single step occurring half way along the structure and linear and sinusoidal variations.

For these cases, assuming the line shape function $\varphi(f)$ of equation 15, the electronic gain G is given as a function of frequency by the following expressions:

1) single step:

$$G = \frac{G_0}{2}\left[\frac{1}{1+(x+b)^2} + \frac{1}{1+(x-b)^2}\right]. \quad (21)$$

2) sinusoidal variation:

$$G = \frac{G_0}{[(1-x^2+b^2)^2+4x^2]^{\frac{1}{4}}} \cos\left[\tfrac{1}{2}\tan^{-1}\frac{2x}{1-x^2+b^2}\right]. \quad (22)$$

3) linear variation:

$$G = \frac{G_0}{2b}\cdot\tan^{-1}\frac{2b}{1+x^2-b^2}\ . \quad . \quad . \quad (23)$$

In these expressions we have written:

$$\left.\begin{aligned} x &= 2\pi T_2(f-f_0) \\ b &= \pi T_2 \Delta f \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad (24)$$

G_0 is the peak electronic gain as in eq. (20). The factor Δf is the total spread in local centre resonant frequencies. As the centre resonant frequency is a function of magnetic field H (to be obtained e.g. from fig. 5), Δf is directly given by ΔH, the difference between the highest and the lowest field acting on the maser crystal. (Similar expressions to the above have been given by Ostermayer [19].)

It follows from these expressions that both single step and sinusoidal field variations produce gain frequency characteristics with a minimum at the centre frequency when the field staggering exceeds certain limits, e.g. *fig. 14*. For a single step this occurs when

$$b^2 > \tfrac{1}{3},$$

and for a sinusoidal variation when

$$b^2 > 2.$$

There is usually no practical advantage in increasing the bandwidth beyond the point at which the centre gain dip exceeds 3 dB.

In *fig. 15* the ratio of peak electronic gain G_c to the uniform field peak electronic gain G_0 is plotted against the field-staggering parameter, $\pi T_2 \Delta f$, for the three types of field variation. In *fig. 16* the 3 dB bandwidth resulting from the different types of field variation is plotted against peak electronic gain G_c for a TWM having a uniform field peak electronic gain G_0 of 50 dB. Clearly the field staggering parameter Δf corresponding to each point of fig. 16 can be derived from fig. 15.

An alternative technique to field staggering for increasing the bandwidth of a system employing a TWM as the first stage amplifier is to compensate the rather sharply peaked gain frequency characteristic of the uniform field maser by a suitable passive network in a subsequent stage of amplification — typically one of the intermediate frequency amplifiers. This technique which is known as gain equalization [20], may be regarded to a first approximation as

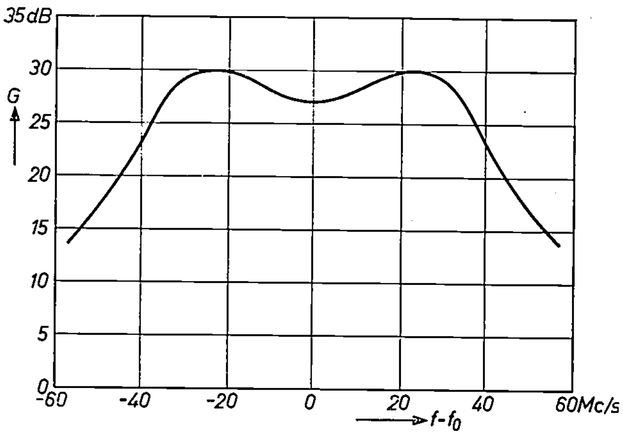


Fig. 14. Electronic gain G as a function of frequency in a TWM with a single step in the magnetic field. The curve represents eq. (21) for G_0 = 50 dB and b = 1.0.

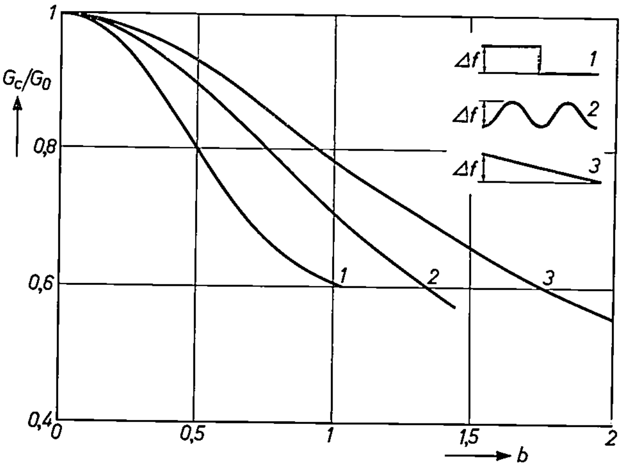


Fig. 15. Peak electronic gain G_c (relative to the uniform field peak electronic gain G_0) as a function of field-staggering parameter $b = \pi T_2 \Delta f$, for three types of field variation:
1. single step,
2. sinusoidal variation,
3. linear variation.

[19] F. W. Ostermayer, 2nd quarterly report "Solid State Maser Research" B.T.L., 1960.

[20] W. J. Tabor and J. T. Sibilia, Bell Syst. tech. J. **42**, 1863, 1963.

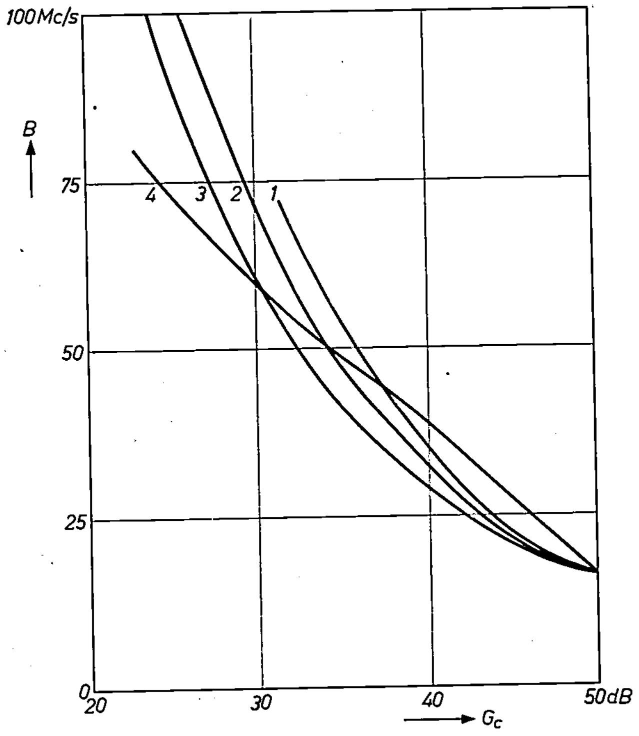


Fig. 16. 3 dB bandwidth B as a function of centre frequency electronic gain G_c for a uniform field peak electronic gain $G_0 =$ 50 dB and $T_2 = 4.9 \times 10^{-9}$ s for three types of field variation and for gain equalization.
1 field variation: single step,
2 field variation: sinusoidal,
3 field variation: linear,
4 gain equalization (curve refers to *system* gain). (Curve *1* is not drawn in the region where the centre gain dip exceeds 3 dB.)

removing the peak of the uniform field gain characteristic (*fig. 17*). The 3 dB bandwidth of the system may then readily be calculated and is plotted in fig. 16 against effective system gain.

The noise temperature of the travelling wave maser

The most important characteristic of the TWM is its very low noise temperature. Calculation of the maser noise temperature is made in the Appendix and we find that the equivalent noise temperature is given to a good approximation by:

$$T_n = \frac{1}{(1-a)}\left[aT_a + \frac{T_0}{I}\right]. \quad . \quad . \quad . \quad (25)$$

In this equation a is the absorption coefficient of the input lead to the maser and T_a the mean temperature of this lead; T_0 is the ambient temperature of the active part of the maser and I is the inversion. Typically $T_0 = 4.2$ °K, $I \approx 3$, $a \approx 0.045$ (0.2 dB loss) and $T_a \approx 100$ °K, whence we find:

$$T_n = \frac{1}{0.955}\ [4.5 + 1.4] = 6.2\ ^\circ\mathrm{K} \quad . \quad (26)$$

Evidently if the low noise potentialities of the maser are to be fully realized, great attention must be given to reduction of loss in the input leads.

Design of a travelling wave maser

In designing a TWM the first consideration must be the choice of active material. To date, for masers operating with signal frequencies in the 1 - 10 Gc/s range, no better material than synthetic ruby has come to light. The optimum concentration of chromium ions in the ruby depends on the operating temperature proposed for the device, and if this is in the liquid helium range it is found that the largest value of the product $\Delta n_0 I T_2$ (hence the lowest value of Q_m) can be obtained at a chromium concentration of about 0.05 Cr^{3+} ions per 100 aluminium ions (0.05% ruby) — at higher concentrations multiple spin-spin interactions (cross relaxation) rapidly tend to reduce the inversion which may be obtained. In choosing the orientation of the ruby crystal with respect to the applied static magnetic field we look for a large transition probability at the signal frequency and freedom from low-order cross-relaxation effects which even at low concentrations can seriously affect maser operation [21]). For example, it is found that if a simple integral relation exists between the pump and signal frequencies, e.g.,

$$nf_s = mf_p\ ,$$

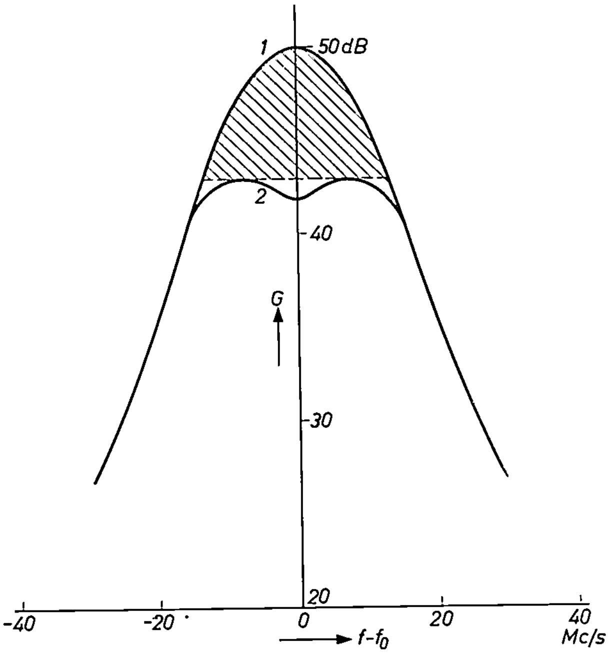


Fig. 17. Gain equalization.
1 Uniform field gain as a function of frequency — no equalization.
2 Effective gain as a function of frequency after equalization. The shaded area is the excess gain removed by equalization.

[21]) S. A. Ahern, P. A. Gould and J. C. Walling, J. Electronics and Control 9, 477, 1960.

there is a tendency for the pump frequency to saturate the signal transition, the tendency being particularly marked for low values of m and n. It is found that in ruby the best operation (for signal frequencies below 7 Gc/s) is obtained with the magnetic field at right angles to the three fold symmetry axis of the ruby and in this case the ground state Zeeman levels are as indicated in figure 5.

We can now use eq. (19) to ascertain what value of Q_m might be expected. For simplicity let us assume that the RF magnetic field is circularly polarized ($R = 0$) and has a frequency f_0. Eq. (19) then becomes:

$$\frac{1}{Q_\mathrm{m}} = \Delta n h \gamma^2 \eta T_2 (A + B)^2 . \quad . \quad . \quad (27)$$

Δn may be evaluated in terms of the chromium ion population density N, the signal frequency f, the ambient temperature T and the inversion I as follows.

We have:

$$N = n_1 + n_2 + n_3 + n_4,$$

and in thermal equilibrium (subscript 0):

$$N = n_{1_0}\left[1 + \exp\left(-\frac{hf_{12}}{kT}\right) + \exp\left(-\frac{hf_{13}}{kT}\right) + \exp\left(-\frac{hf_{14}}{kT}\right)\right]$$
$$= n_{1_0}\, c\,.$$

Therefore the equilibrium population difference between levels *1* and *2* is given by

$$(n_2 - n_1)_0 = \Delta n_0 = \frac{N}{c}\left[\exp\left(-\frac{hf_{12}}{kT}\right) - 1\right]$$

and hence the inverted population difference

$$\Delta n = -I\Delta n_0 = -\frac{IN}{c}\left[\exp\left(-\frac{hf_{12}}{kT}\right) - 1\right] \approx \frac{NIhf_{12}}{kTc}.$$

Thus:

$$\frac{1}{Q_\mathrm{m}} = \frac{NIh^2 f_\mathrm{s}}{ckT}\,\gamma^2 \eta T_2 (A + B)^2 .$$

Let us consider the particular case of a maser using ruby as the active material and operating in the 90° orientation (fig. 5) at a signal frequency of 4170 Mc/s, the signal transition being between levels *1* and *2*. In this case we find using the tables of Chang and Siegman [13]) that the matrix element vectors are $A = 1.16$, $B = 0.56$. For 0.05% ruby $N = 2.35 \times 10^{19}$ cm^{-3}, $T_2 = 4.9 \times 10^{-9}$ s, and by pumping between levels *1* and *4* an inversion of about 2.5 (possibly better) may be obtained. The value of γ^2 is 3.09×10^{14} gauss^{-2} s^{-2}, $h = 6.62 \times 10^{-27}$ erg s and $k = 1.38 \times 10^{-16}$ erg deg^{-1}. Therefore for this case:

$$\frac{1}{Q_\mathrm{m}} = 3.54 \times 10^{-1}\,\frac{\eta}{cT}.$$

For an operating temperature of 1.5 °K we find that

$$c = 2.85\,,$$

and if we assume that $\eta = 0.20$:

$$\frac{1}{Q_\mathrm{m}} = 1.66 \times 10^{-2}\,.$$

We may note that for waves propagating in the backward direction the sense of polarization is opposite to that for waves propagating in the forward direction. Thus for backward travelling waves we must write $R = \infty$ in eq. (19), so that the ratio of Q_m in the forward and backward directions is:

$$\left(\frac{A - B}{A + B}\right)^2 \approx \frac{1}{8}.$$

The device is thus, to some extent, intrinsically non reciprocal.

According to eq. (18) the gain in dB is proportional to the length of the structure. Various practical considerations such as the manufacture of the slow wave structure, the obtaining of large uniform single crystals of ruby and the provision of magnetic fields of high uniformity over a large volume tend to limit the length of the device. We will consider the case of a TWM operating under the above conditions and having a length of 10 cm. We find using eq. (18) that the forward electronic gain is:

$$G_0 = 0.63\, r \text{ dB},$$

where r is the ratio between the free space velocity of light and the structure group velocity v_g, and is called the slowing factor.

Such a maser using a structure having a slowing factor of 100 will therefore have an electronic gain of 63 dB — the net gain of course will be less than this by the structure insertion loss. In a practical structure the polarization of the RF field may deviate from circularity and the complete form of eq. (19) must be used.

We see then that successful operation of masers of this type depends on the use of a propagating structure having a substantial slowing factor. Slow wave structures are familiar through their application in travelling wave tubes and in linear accelerators. In these devices, however, we are concerned with reducing the *phase* velocity of the propagating wave whereas in the TWM we wish to reduce the group velocity and furthermore (with the construction of a non-reciprocal device in mind) to provide regions in the structure where the RF magnetic fields are substantially circularly polarised. Structures consisting essentially of arrays of parallel conductors have been found suitable and that most widely used is the comb structure (fig. 11) which consists of an array of conductors approximately a quarter wavelength long at the signal frequency and short circuited at one end. (In this structure the RF magnetic field associated with the propagation lies in the plane normal to the extension of the conductors and is substantially circularly polarized, the sense of polarization being opposite on the two sides of the array of conductors. This means that by placing the active material on one side of the array it is possible to utilize the non reciprocal characteristics of the ruby already discussed. The device may be made to be completely non reciprocal if ferrite material, dimensioned so that ferrimagnetic resonance is obtained at the signal frequency for the same static magnetic field as is required by the ruby, is placed on the other side of the structure. In this way a resonant interaction between the magnetization of the ferrite and the backward travelling circularly polarized waves in the structure is obtained, thus backward travelling waves experience large attenuation whilst the forward travelling waves do not interact with the ferrite to any significant extent. If sufficient backward loss is provided for the loop gain to be less than unity when the structure is terminated by a short circuit, the device is completely stable. Furthermore, since the ferrite elements are at a very low temperature, any loss which they might introduce in the forward travelling wave does not significantly affect the maser noise temperature (Appendix II).

A 4170 Mc/s packaged travelling wave maser

As an example of a practical packaged TWM we will now describe briefly the amplifier designed and built at the Mullard Research Laboratories and used at the General Post Office satellite communication ground station at Goonhilly Down in Cornwall for experiments with the Telstar and Relay communication satellites.

Maser structure

The active material in this maser is 0.05% ruby used with the applied static magnetic field normal to the three-fold axis of symmetry of the crystal as described in the last section. The signal transition is that between levels *1* and *2* at 4170 Mc/s and a population inversion of about 2.7 is obtained by pumping between levels *1* and *4* at 30 150 Mc/s. These frequencies correspond to a magnetic field of 3280 Oe (fig. 5).

The ruby sample was cut from a large synthetic single crystal in the form of a rod, the *c*-axis being inclined at an angle of 60° to the axis of the rod. The *c*-axis is therefore at an angle of 60° to the direction of propagation in the slow wave structure, in the plane normal to the array of parallel conductors.

The isolator material used in the maser described here is yttrium iron garnet (*YIG*) in the form of flat discs which are supported in a polycrystalline alumina slab. By suitable choice of the thickness of the discs (determining the shape demagnetizing factors) it is arranged that ferrimagnetic resonance occurs in the discs at the same static magnetic field as is required by the ruby. The required demagnetizing factor is calculated from the Kittel resonance equation [22]) which for flat discs mounted away from conducting walls may be written:

$$\omega = \gamma \left[H_0 - \frac{4\pi M_s}{2} (3N_z - 1) \right],$$

where H_0 is the applied magnetic field, N_z is the demagnetizing factor in the direction of H_0, and $4\pi\ M_s$ is the saturation magnetization of the garnet at the operating temperature. Yttrium iron garnet is one of the few ferrimagnetic materials which are suitable for this application as it still has a relatively narrow ferrimagnetic resonance line at liquid helium temperatures.

The slow wave structure is a comb (cf. fig. 11) having the dimensions shown in *fig. 18b* and provides a slowing factor of 110. The assembly is shown in a cut-away view in fig. 18*a*.

The input signal is launched on to the comb structure through a coaxial line the centre conductor of which continues as the conductor *E* of *fig. 19*. This conductor terminates in a short length of coaxial line *X* the length of which may be adjusted by means of the short circuiting plunger *P*. Impedance matching is achieved by adjusting the position of this plunger and varying the proximity of the conductor *E* to the first conductor of the comb proper. A similar arrangement is provided at the output end.

An advantage of slow wave structures of the iterated conductor type is that they can be introduced in a magnetic equipotential plane of a rectangular waveguide propagating, say, an H_{01} mode, without substantially perturbing this propagation. The structure of fig. 18 will therefore propagate certain rectangular waveguide modes at frequencies above that corresponding to a free space wavelength of approximately 3 cm. In the present maser the pump energy propagates in such a waveguide mode and is introduced into the end of the slow wave structure

[22]) C. Kittel, Phys. Rev. **73**, 155, 1948.

by a direct connection from rectangular waveguide, the broad face of the waveguide being parallel to the plane of the comb conductors. A pump power of about 30 mW is required to saturate the *1* → *4* transition.

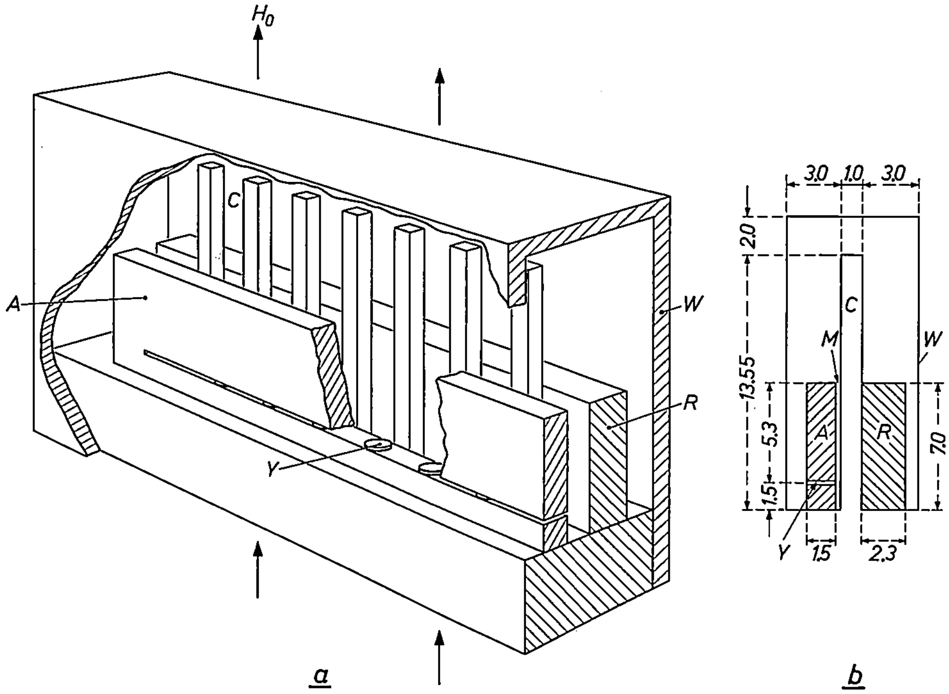


Fig. 18. Cut-away view (*a*) and cross section (*b*) of the 4170 Mc/s TWM. The dimensions are in millimeters. The slow wave comb structure *C* (pitch 2 mm) permits the propagation of waves with a group velocity of 1/110 times the free space velocity of light, and with an RF magnetic field circularly polarized in a plane perpendicular to the extension of the comb conductors. The sense of polarization in a travelling wave is opposite on the two sides of the comb, and opposite for forward and backward waves. In the applied static magnetic field H_0 of 3280 Oe the Cr^{3+} ions in the single crystal ruby slab *R* (Al_2O_3 with 0.05% Cr^{3+}) are resonant at the signal frequency (4170 Mc/s, *1-2* transition) and at the pump frequency (30 150 Mc/s, *1-4* transition). At the signal frequency they amplify forward waves. The yttrium iron garnet discs *Y*, held in place by the polycrystalline alumina slab *A*, and also resonant in H_0 at the signal frequency, attenuate backward waves. A sheet of 0.025 mm Melinex *M* is used to adjust the transmission characteristics of the slow wave structure. Pump power (30 150 Mc/s) is transmitted by the waveguide *W*.

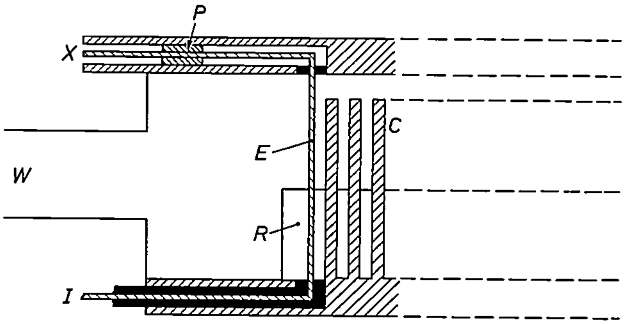


Fig. 19. Matching system from coaxial line to comb. *I* coaxial input line. *E* conductor, extension of the central conductor of both the input (*I*) and the terminating (*X*) coaxial line. *P* short circuiting plunger. *C* comb structure. *R* ruby slab. *W* pump waveguide. Matching is achieved by adjusting position of *P* and distance of *E* from comb.

The cryostat

The maser structure is housed in the tail of a stainless steel double dewar vessel (*fig. 20*), the magnetic field of 3280 Oe being provided by a permanent magnet as shown in fig. 3. Centre frequency adjustment is obtained by varying the current through two coils attached to the pole faces of the permanent magnet. Connections from the top of the cryostat to the coaxial leads on the maser comb structure are made through low thermal conductivity coaxial lines which terminate at the top of the cryostat in waveguide-coaxial transitions (*fig. 21*). The pump energy is introduced via a low thermal conductivity waveguide (WG 22).

To exclude air from the system (air is of course solid at liquid helium temperatures) it is necessary that all these leads should be vacuum tight. The input and output waveguides are sealed with thin terylene sheet windows and the pump waveguide with a mica window, all the seals being in the room temperature part of the apparatus. A series of carbon resistors attached at various levels to the connecting lead structure permits monitoring of the liquid helium level in the dewar vessel.

The whole maser package is mounted on a cradle which allows it to be tipped through 45° from the vertical once it has been charged with liquid nitrogen and helium. This is its operating position when the aerial on which it is mounted is directed towards the horizon, and as the aerial moves to the zenith the maser moves through 90° to a position 45° the other side of the vertical.

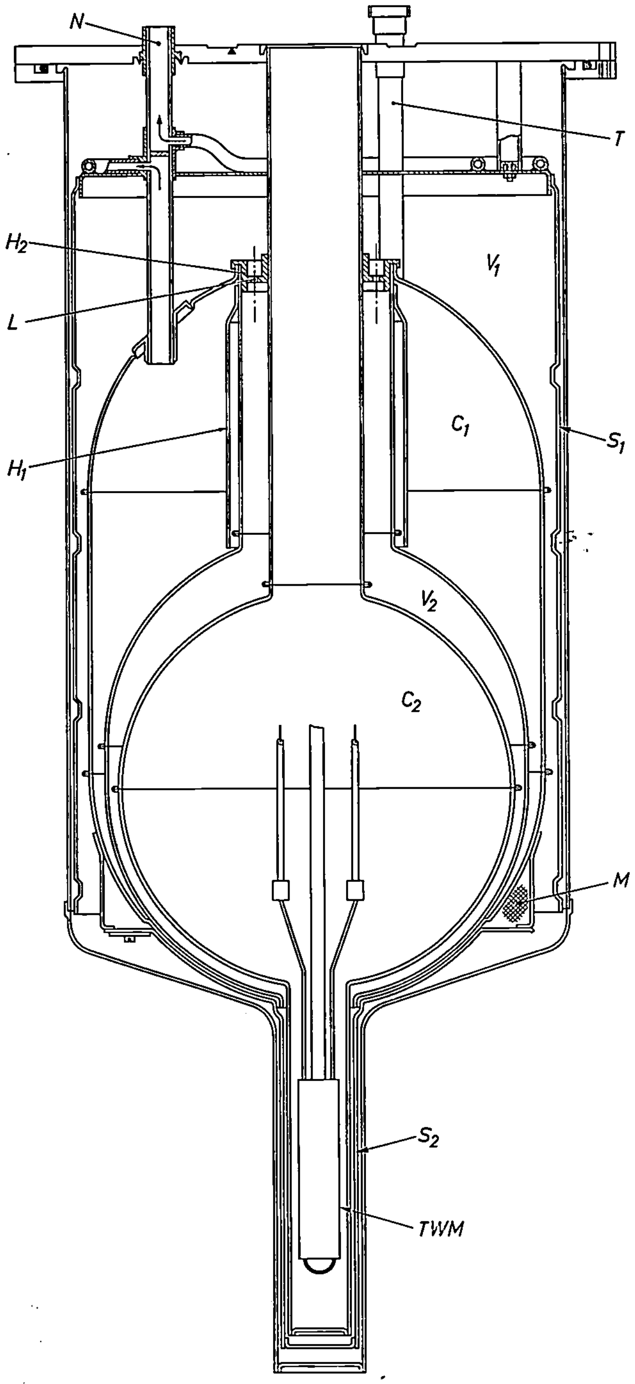


Fig. 20. Sectional drawing of double dewar vessel. Length 100 cm, max. diameter 33 cm.
C_1 liquid nitrogen container. C_2 liquid helium container. V_1, V_2 vacuum spaces. N nitrogen vapour outlet. S_1 copper radiation shield cooled at its top by nitrogen vapour. S_2 copper radiation shield cooled at its top by liquid nitrogen. H_1, H_2, copper heat conductors keeping a fixed point on the neck of the helium container at liquid nitrogen temperature. L holes linking vacuum spaces. T filling and venting tubes. M molecular sieve (getter material to absorb residual gases). TWM travelling wave maser.

The complete package is shown in figure 3 mounted on the aerial structure in its operating position, with a dewar vessel which allows an operating time per filling of liquid helium of about 8 hours. The maser is operated at a temperature of 1.5 °K, the pressure over the boiling liquid being reduced by means of a pump situated lower on the aerial structure. *Fig 22* shows a later version of the maser with a much larger dewar vessel which gives an operating time per filling of about 2 days.

The maser has been operated in a homogeneous magnetic field and also with a suitable stepped magnetic field to obtain greater bandwidth at the expense of gain. The performance figures of the maser under such conditions are:

Operation in a homogeneous magnetic field:

Electronic gain	52.5	dB
Bandwidth to 3 dB points	16	Mc/s
Total structure forward loss	11	dB
Total structure backward loss	70	dB
Net forward gain	41.5	dB
Noise temperature	15±4	°K

Operation in stepped magnetic field ($\Delta H \approx 5$ Oe):

Net forward gain	30	dB
Bandwidth to 3 dB points	28	Mc/s

These performance figures are in reasonable agreement with the design predictions outlined previously.

Systems considerations and conclusions

The packaged travelling wave maser which is described above is one of the first such devices to be regularly employed in a systems application, and its operation has proved to be reliable and consistent. The principal operational difficulty is the need for repeated liquid helium transfers on site, a difficulty which is mitigated by the use of very large storage capacity maser dewar vessels, and which would be eliminated by the use of a closed cycle helium temperature refrigeration system.

A further operational difficulty encountered was a drift in the centre frequency of the maser due to changes in the field provided by the permanent magnet. These were the result of ambient temperature variations. This effect can be entirely eliminated by the employment of persistent current superconducting electromagnets housed in the dewar

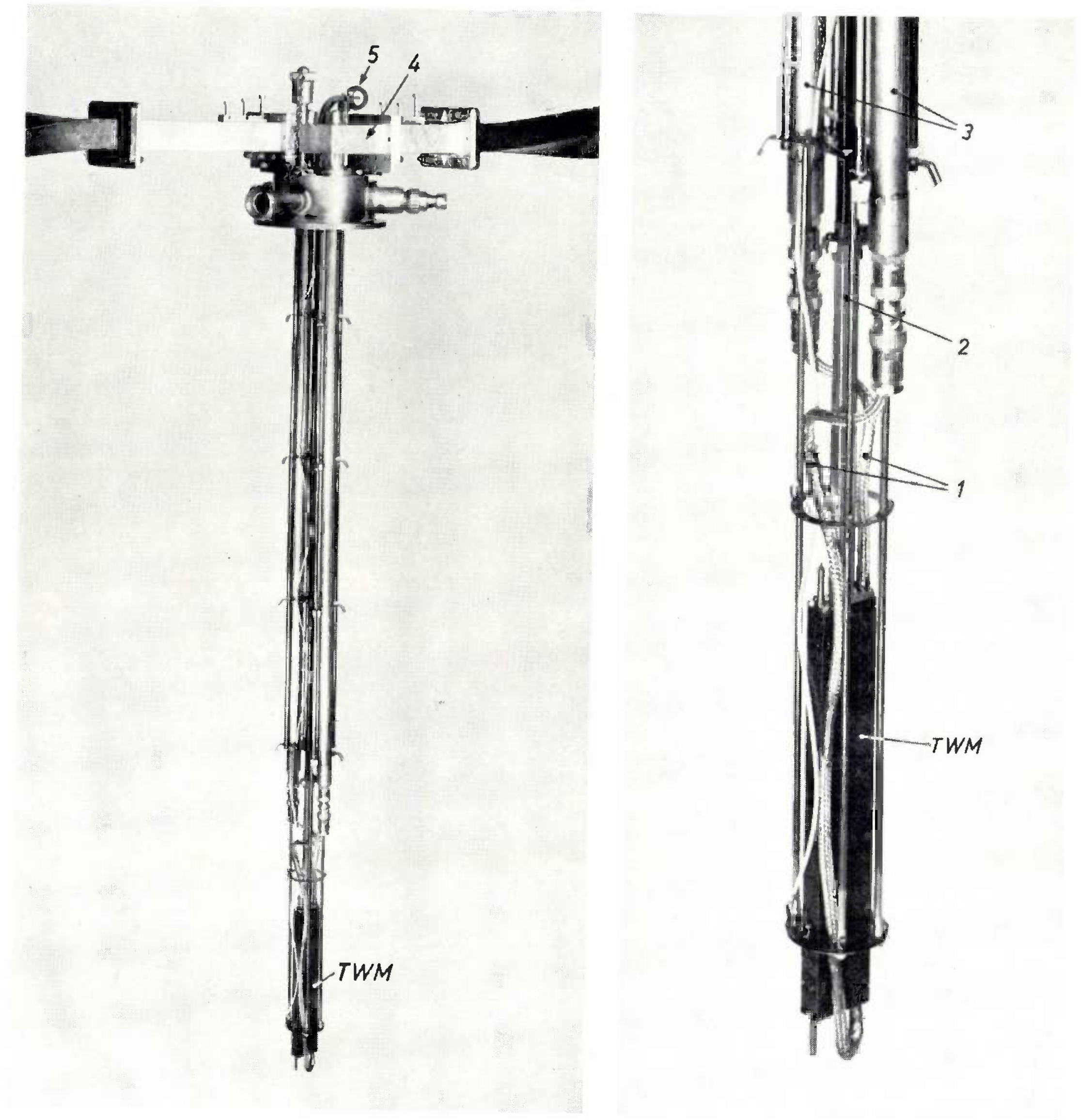


Fig. 21. Cryostat head and maser connection leads (enlarged view of lower end at right). *TWM* travelling wave maser. *1* coaxial leads (cf. fig. 19). *2* low thermal conductivity pump waveguide. *3* low thermal conductivity coaxial lines. *4* one of the two waveguide-coaxial transitions. *5* connection of pump waveguide.

vessel, such as have recently been developed at Mullard Research Laboratories, based on the work of P. P. Cioffi [23]).

We may note that the noise performance of the maser described may be further improved by careful attention to the reduction of radio frequency loss in the input lead (eq. 25 and 26). This loss is the major contributor to the noise temperature of the maser amplifier, and the use of a waveguide feed down the neck of the dewar vessel would substantially reduce this. A noise temperature for the maser of 3-4 °K, as measured at its input terminals, is entirely practical. However, the spectacular decrease in noise temperature of the first amplifier of a receiver system which is offered by the maser compared with more conventional amplifiers (a good travelling wave tube at this frequency has a noise temperature of 900 °K), may not be fully utilized at present because of the noise contributions of other parts of the system. These sources of noise arise as a result of loss in other necessary circuit components which precede the maser such as filters, diplexer, feed horns etc. The antenna itself also may make an appreciable contribution to the overall system noise temperature as it will have side lobes of its radiation pattern which accept noise

[23]) P. P. Cioffi, J. appl. Phys. **33**, 875, 1962.

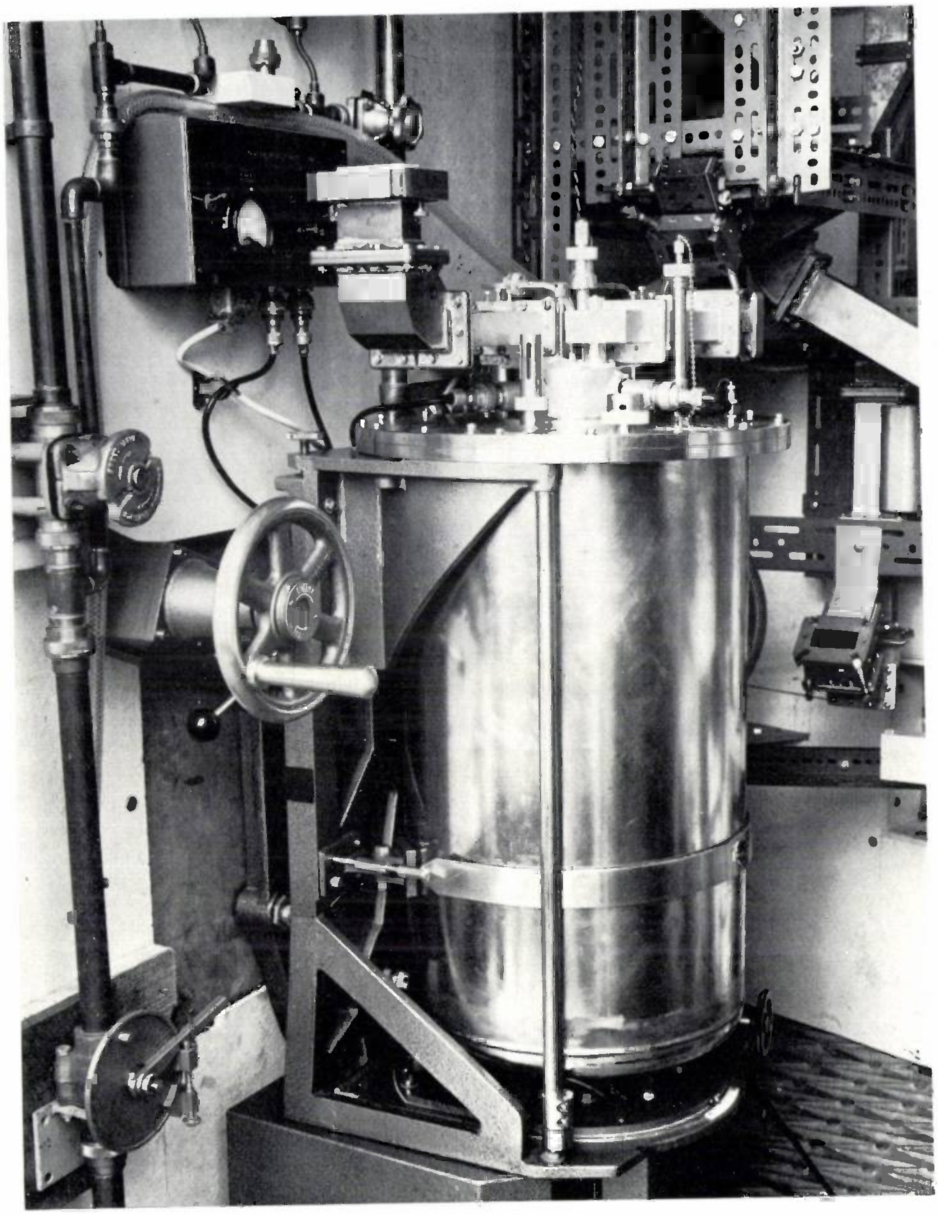


Fig. 22. Travelling wave maser with large storage capacity dewar vessel. (Photo published by permission of the British G.P.O.)

radiated from the ground. It is to be expected that these other noise contributions will be substantially reduced as further attention is directed towards their elimination, and that systems with very low overall noise temperature which can fully utilize the potentialities of the maser noise performance will soon be realized.

Appendix I

Calculation of Q_m for a TWM

The quality factor Q_m of the maser material has been defined as:

$$Q_m = \frac{2\pi f_s W_s}{P_m}. \qquad (28)$$

P_m in this equation is the power emitted per unit length of the maser material and may be written:

$$P_m = \Delta n \, V_c \, h f_s \, W_{ij}, \qquad (29)$$

where Δn is the population density difference between the two levels i and j between which the signal transition takes place. W_{ij} is the transition probability defined (eq. 13) and V_c is the volume of crystal per unit length of the structure.

Δn may be written in terms of the equilibrium population difference Δn_0 and the inversion I through:

$$\Delta n = -I \Delta n_0.$$

In general the RF magnetic field amplitude and polarization will vary within the active crystal and thus if a correct value is to be obtained for P_m, the transition probability W_{ij} should be integrated over the volume of the crystal.

The power emitted from an element of material as indicated in *fig. 23* is:

$$\mathrm{d}P_\mathrm{m} = \Delta n\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z\, hf_\mathrm{s}\, W_{ij}$$

$$= \Delta n\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z\, hf_\mathrm{s} \frac{\gamma^2}{4}\, \varphi(f)\, |(\mathbf{A} + \mathrm{j}\mathbf{B}) \cdot (\mathbf{H} + \mathrm{j}\mathbf{H}')|^2. \quad (30)$$

Let us now suppose that the plane of polarization of the RF magnetic field and the plane defined by the matrix element vectors **A** and **B** are coincident. (This is a case frequently met in practice but if this condition is not satisfied then the vectors **A** and **B** must be replaced, in what follows, by their projections on the plane of polarization of the RF field.) If the angle between the vectors **A** and **H** (the larger field component) is Θ and if we resolve the elliptically polarized field into two circularly polarized components H_+ and H_- of opposite sense such that:

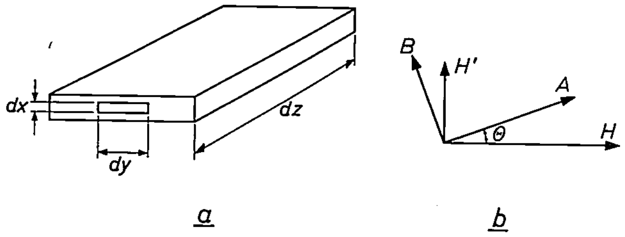


Fig. 23. *a*) Volume element of the maser material. *b*) H and H' are the magneticfield vectors. **A** and **B** are the matrix element vectors.

$$H = H_+ + H_-, \qquad H' = H_+ - H_-,$$

then equation (30) may be rewritten:

$$\mathrm{d}P_\mathrm{m} = \Delta n\, hf_\mathrm{s} \frac{\gamma^2}{4}\, \varphi(f)\, [H_+^2(A + B)^2 + H_-^2(A - B)^2 +$$

$$+ 2H_+H_- \cos 2\Theta\, (A^2 - B^2)]\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z. \quad . \quad . \quad . \quad . \quad . \quad (31)$$

Integrating over unit length of the crystal (volume V_c):

$$P_\mathrm{m} = \Delta n\, hf_\mathrm{s} \frac{\gamma^2}{4}\, \varphi(f)\, [(A+B)^2 \int_{V_c} H_+^2 \mathrm{d}v + (A - B)^2 \int_{V_c} H_-^2 \mathrm{d}v +$$

$$+ 2 \cos 2\Theta\, (A^2 - B^2) \int_{V_c} H_+H_-\, \mathrm{d}v]\, . \quad . \quad . \quad . \quad (32)$$

The integrals in this expression can only be evaluated if we have an exact knowledge of the field distribution, which in most TWM's we do not have, although in the cases of uniformly loaded rectangular waveguides or strip lines eq. (32) can be integrated directly. If we define a structure reciprocity factor R as:

$$R = \int_{V_c} H_-^2\, \mathrm{d}v / \int_{V_c} H_+^2\, \mathrm{d}v, \quad . \quad . \quad . \quad . \quad . \quad (33)$$

then (32) may be written approximately as:

$$P_\mathrm{m} = \Delta n\, hf_\mathrm{s} \frac{\gamma^2}{4}\, \varphi(f) \int_{V_c} H_+^2\, \mathrm{d}v\, [(A + B)^2 + \dot{R}(A - B)^2 +$$

$$+ 2 \cos 2\Theta\, (A^2 - B^2) R^{\frac{1}{2}}]. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (34)$$

The mean magnetic energy stored in the maser material per unit length is (in c.g.s. units):

$$M_\mathrm{m} = \frac{\mu}{8\pi}(1 + R) \int_{V_c} H_+^2\, \mathrm{d}v = \frac{\mu}{16\pi} \int_{V_c} \mathbf{H} \cdot \mathbf{H}^*\, \mathrm{d}v,$$

whence:

$$\int_{V_c} H_+^2\, \mathrm{d}v = \frac{1}{2(1 + R)} \int_{V_c} \mathbf{H} \cdot \mathbf{H}^*\, \mathrm{d}v. \quad . \quad . \quad . \quad . \quad (35)$$

The mean magnetic energy stored in the whole structure per unit length is

$$\frac{W_\mathrm{s}}{2} = \frac{\mu}{16\pi} \int_V \mathbf{H} \cdot \mathbf{H}^*\, \mathrm{d}v. \quad . \quad . \quad . \quad . \quad . \quad (36)$$

From (34), (35) and (36) we find eq. (19):

$$\frac{1}{Q_\mathrm{m}} = \Delta n\, h \frac{\gamma^2}{1 + R} \frac{\varphi(f)}{2}\, \eta\, [(A + B)^2 + R(A - B)^2 +$$

$$+ 2 \cos 2\Theta\, (A^2 - B^2) R^{\frac{1}{2}}], \qquad (37)$$

in which

$$\eta = \int_{V_c} \mathbf{H}\, \mathbf{H}^*\, \mathrm{d}v / \int_V \mathbf{H}\, \mathbf{H}^*\, \mathrm{d}v \quad . \quad . \quad . \quad . \quad . \quad (38)$$

is the filling factor.

Appendix II

Calculation of the noise temperature of the TWM

The equivalent noise temperature of a maser amplifier is its most important characteristic. To calculate the noise temperature we recall a basic result of radiation theory that a body at a temperature T emits incoherent radiation (noise), the power in a single mode and in a frequency band of width Δf being:

$$P_\mathrm{n} = \frac{a\, hf \Delta f}{\exp\left(\frac{hf}{kT}\right) - 1} = a\Phi(T)\Delta f \quad . \quad . \quad . \quad . \quad (39)$$

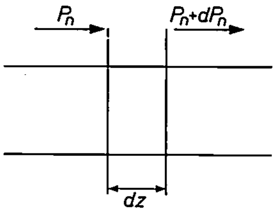
$$= a\, kT \Delta f \quad \text{for } \frac{hf}{kT} \ll 1. \quad . \quad . \quad . \quad . \quad . \quad (40)$$

a is the emission coefficient of the body which is equal to its absorption coefficient.

Ditchfield [24]) has shown that this result may be applied to bodies containing energy levels the populations of which are inverted. Considering an element of the TWM and a narrow frequency band around the signal centre frequency, we may then write (cf. *fig. 24*) the increment of noise power as:

$$\mathrm{d}P_\mathrm{n} = \underset{\text{structure loss}}{-a_\mathrm{L} P_\mathrm{n}\, \mathrm{d}z} + \underset{\text{thermal noise from structure}}{a_\mathrm{L} \Phi(T_0) \Delta f\, \mathrm{d}z} + \underset{\text{gain}}{g P_\mathrm{n} \mathrm{d}z} - \underset{\text{thermal noise from maser material; abs. coeff.} = -g}{g \Phi(T_\mathrm{m}) \Delta f\, \mathrm{d}z}.$$

Fig. 24. Increase in noise power P_n in an element of length dz of the travelling wave maser.



Integrating this expression we find that the output noise power is given by

$$P_{\mathrm{n}_0} = G' P_{\mathrm{n}_\mathrm{i}} + \frac{[a_\mathrm{L} \Phi(T_0) - g \Phi(T_\mathrm{m})] \Delta f}{g - a_\mathrm{L}} (G' - 1), \quad (41)$$

in which we have written

$$G' = \mathrm{e}^{(g - a_\mathrm{L})L} = \text{net numerical gain of the maser.}$$

The noise power at the input terminals of the active part of the maser is made up of two components, the source noise $\Phi(T_\mathrm{s})\, \Delta f$ and the thermal noise contributed by loss in the leads to the active part of the maser. Thus:

$$P_{\mathrm{n}_\mathrm{i}} = [(1 - a)\, \Phi(T_\mathrm{s}) + a\Phi(T_a)] \Delta f, \quad . \quad . \quad . \quad . \quad (42)$$

where a is the loss factor of the input leads and T_a their mean temperature. The loss factor of the output leads can be taken

[24]) C. R. Ditchfield, Solid State Electronics 4, 171, 1962.

into account similarly but noise due to this cause is usually negligible. Assuming $G' \gg 1$ and $|g| \gg |\alpha_L|$ and using the linear approximation for $\Phi(T)$, $\Phi(T) = kT$, we have:

$$P_{n_0} = G'(P_{n_i} + k|T_m|\Delta f) = G'\,[(1-a)kT_s + akT_a + k|T_m|]\Delta f.$$

The net gain of the maser including the input leads is $G'(1-a)$. Thus the noise figure (cf [3]) of the TWM including the input leads is:

$$F = \frac{1}{G'(1-a)}\,\frac{P_{n_0}}{kT_s\Delta f} = 1 + \frac{1}{T_s}\,\frac{aT_a + |T_m|}{1-a}, \qquad (43)$$

whence the equivalent noise temperature of the maser is:

$$T_n = \frac{aT_a + |T_m|}{1-a}. \quad \dots\dots \quad (44)$$

The spin temperature T_m can be related to the ambient temperature T_0 and the inversion I. The thermal equilibrium population difference Δn_0 is:

$$\Delta n_0 = n_{2_0} - n_{1_0} = n_{1_0}\left[\exp\left(-\frac{hf_{12}}{kT_0}\right) - 1\right]$$

$$\approx -n_{1_0}\frac{hf_{12}}{kT_0} \text{ if } \frac{hf_{12}}{kT_0} \ll 1. \quad \dots\dots \quad (45)$$

The population difference during operation of the maser is, in the same way,

$$\Delta n \approx -n_1\frac{hf_{12}}{kT_m} \quad \text{if} \quad \frac{hf_{12}}{|kT_m|} \ll 1 \quad \dots\dots \quad (46)$$

The actual population of level *1* is not appreciably altered by the pumping (i.e. $\Delta n \ll n_{1_0}$), so that we may put $n_1 \approx n_{1_0}$. Then, by definition of the inversion I and (45) and (46):

$$I = -\frac{\Delta n}{\Delta n_0} \approx -\frac{T_0}{T_m},$$

or, for $T_m < 0$:

$$|T_m| \approx \frac{T_0}{I},$$

whence from (44) we obtain eq. (25):

$$T_n = \frac{1}{(1-a)}\left[aT_a + \frac{T_0}{I}\right].$$

Appendix III (Note of the editor)

In the rationalized Giorgi system the "magnetic dipole moment" and the "magnetic area moment" (electromagnetic moment) have different dimensions (Wbm and Am^2 respectively), which calls for special attention when using these and related concepts such as "magnetization" and "gyromagnetic ratio". Conversion to the rationalized Giorgi system and a notation in accordance with (expected) recommendations of the International Electrotechnical Committee and the International Organization for Standardization is obtained as follows:

1) Read everywhere "magnetic induction **B**" instead of "magnetic field **H**" and 10^{-4} Wb/m^2 instead of "1 Oe".
2) Drop the *c* in the expressions for β, γ and the resonant frequency on pages 291 and 293 (the unit for β becomes Am^2; the resonance equation reads $\omega = \gamma B$).
3) Multiply the expressions (19), (27), the one following (27), and (37) for $1/Q_m$ by $\mu_0/4\pi$ (this results after matching the expressions for the stored energy to the Giorgi system in Appendix I). In the calculation on p. 29 substitute 1 cm^{-3} = 10^6 m^{-3}, 1 gauss = 10^{-4} Wb/m^2, 1 erg = 10^{-7} joule.
4) Read the Kittel resonance equation (p. 304) as $\omega = \gamma[B_0 - \mu_0 M_s(3N_z - 1)/2]$, in which M_s is the saturation magnetization.

Summary. The signals received from communication satellites are of the order of 10^{-13} watt, even with highly directional aerials at the ground stations. Signals of this level dictate the use of ground station receivers with very low noise temperatures, and thus the solid state travelling wave maser, which has the lowest available noise temperature, is used as the first amplifier in such systems.

The operation of a solid state maser depends on the resonant interaction of radiofrequency signals with paramagnetic ions occupying discrete energy levels. This interaction is discussed for the simple case of free electrons in a static magnetic field and for paramagnetic ions in crystals. To obtain amplification from such systems a population inversion is required between a pair of levels resonant at the signal frequency. The method, proposed by Bloembergen, of obtaining a continuous population inversion ("pumping" in a three-level system) is discussed.

The factors affecting the design of travelling wave masers and the choice of suitable materials are reviewed. The method of increasing the instantaneous bandwidth of the maser, by applying a static field which varies along the length of the structure, is discussed. The design and construction at Mullard Research Laboratories of a maser operating at 4170 Mc/s for use in the Goonhilly Down satellite communication ground station of the British General Post Office is outlined. This maser, consisting of a slow wave comb structure containing as the active material a ruby single crystal at 1.5 °K in a magnetic field of 3280 Oe and pumped at a frequency of 30 150 Mc/s, has a noise temperature of less than 20 °K.

"COMPATIBLE" SINGLE-SIDEBAND MODULATION

by T. J. van KESSEL *).

621.376.24

As broadcasting transmitters are continuously increasing in power and number, it is desirable to reduce their bandwidth in the frequency channels allocated to them. One means to this end is single-sideband modulation. If signals modulated in this way can be received by sets currently in use, then the single-sideband modulation is said to be "compatible" with the conventional double-sideband system, and can therefore be introduced gradually. This article describes various methods of single-sideband modulation, with special reference to a new method, called the "squaring system".

In radio broadcasting an audio frequency signal, i.e. music or speech, is transmitted by means of a radio-frequency carrier wave. For this purpose the carrier is modulated by the audio signal, which is then recovered in the receiver by demodulation.

The most commonly used method of modulation is *amplitude modulation*, in which the amplitude of the carrier wave is varied in accordance with the waveform of the audio signal to be transmitted. Let this audio signal be a simple cosine oscillation with angular frequency p. The modulated RF signal will then have the following form as a function of time:

$$v(t) = (1 + a \cos pt) \cos \omega t, \text{ where } a \leqq 1 . \quad (1)$$

In this expression a is the modulation depth and ω is the angular frequency of the RF oscillation. The modulation depth a must be less than unity, since the demodulator (detector) in the receiver cannot distinguish a "negative" amplitude from a "positive" one.

The amplitude-modulated signal can be regarded as the sum of three unmodulated signals. This can be seen by writing (1) in the following form:

$$v(t) = \cos \omega t + \frac{a}{2} \cos(\omega + p)t + \frac{a}{2} \cos (\omega - p)t . \quad . \; . \; . \; (1a)$$

The last two terms in the right-hand side of (1a) represent signals called the upper and lower *sidebands.* They lie symmetrically at a frequency distance $\pm p$ from the unmodulated carrier. It is evident, then, that the amplitude-modulated signal takes up an unnecessary amount of space in the frequency channel available, since all the information to be transmitted is already present in one of the sidebands. In view of the overcrowding of the frequency bands, attempts have understandably been made to find a method of reducing the bandwidth of the transmitter. An early proposal was to transmit only one of the sidebands. Without taking other special measures, however, this method is not practicable. To make this clear, we shall first consider what happens when one of the sidebands is omitted. This will make it easier to understand the methods that can in fact be used.

In the following we assume throughout that the transmitted signal can be received by a conventional receiver, in which the demodulator responds to the instantaneous amplitude of the RF signal. This implies that the receiver is also capable — without conversion — of handling the normal amplitude-modulated signal with two sidebands. The new form of modulation is then *compatible* with normal amplitude modulation, and if necessary can gradually be introduced for broadcasting transmitters now in existence or yet to be built.

In *fig. 1* the first term on the right-hand side of (1a), i.e. the constant *carrier*, $\cos \omega t$, is represented in the conventional way by a vector OA rotating at an angular velocity ω. The second term is taken as the single sideband to be transmitted, and this is represented by the vector AC, which, with respect to OA, has the relative angular velocity p. The instantaneous value of the amplitude of the RF signal is then equal to the length of the vector OC.

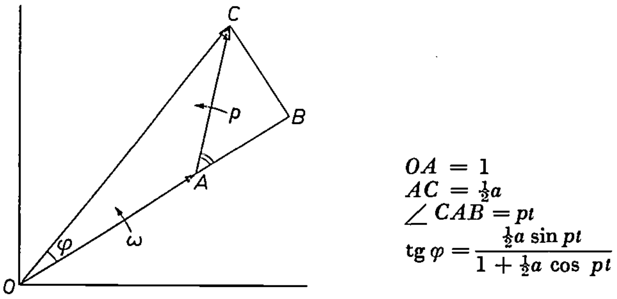

C
B
A
O
p
ω
φ
$OA = 1$
$AC = \frac{1}{2}a$
$\angle CAB = pt$
$\mathrm{tg}\,\varphi = \dfrac{\frac{1}{2}a \sin pt}{1 + \frac{1}{2}a \cos pt}$

Fig. 1. Vector diagram of an amplitude-modulated RF signal with only one side component (sideband). The carrier is represented by the vector OA, rotating with angular velocity ω. With respect to OA the side component AC has the relative angular velocity p, the latter being the modulation frequency. The sum OC of both vectors represents the RF signal. This shows an amplitude modulation which is not purely sinusoidal, and also phase modulation.

*) Philips Research Laboratories, Eindhoven.

The relation

$$(OC)^2 = (OB)^2 + (BC)^2 = \\ = \left(1 + \frac{a}{2}\cos pt\right)^2 + \left(\frac{a}{2}\sin pt\right)^2,$$

reduces, after elementary mathematical treatment, to:

$$OC = \sqrt{1 + \frac{a^2}{4} + a\cos pt} \quad . \; . \; . \quad (2)$$

Using the binominal theorem, a square root form of this kind can be expanded to form the series:

$$\sqrt{1+x} = 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16} - \frac{5x^4}{128} + \dots .$$

Applying this to expression (2), and taking only the first terms that interest us, we find after some manipulation:

$$OC = 1 + \frac{a^2}{16} + \frac{a}{2}\left(1 - \frac{a^2}{32}\right)\cos pt - \frac{a^2}{16}\cos 2pt + \\ + \frac{a^3}{64}\cos 3pt \dots \quad . \; . \; . \; . \quad (3)$$

We can see from this that by simply omitting one sideband the amplitude waveform, or "envelope", of the radio-frequency signal also contains higher harmonics of the desired audio frequency. After demodulation, the audio signal will therefore show harmonic distortion, particularly in the case of deep modulation. We may therefore conclude that the single-sideband method, as remarked, is not practicable without corrective measures.

From fig. 1 we can also read a second effect which occurs when only one sideband is transmitted. The vector OC, which corresponds to the total RF signal, then no longer rotates at a constant angular velocity ω, but periodically lags and leads with respect to OA, with an angle φ given by:

$$\tan\varphi = \frac{\frac{1}{2}a\sin pt}{1 + \frac{1}{2}a\cos pt}. \quad . \; . \; . \quad (4)$$

In other words, the single-sideband signal shows *phase modulation*. In itself this is not serious, because a normal amplitude or envelope detector does not respond to phase modulation. We shall see presently that the signals that can suitably be used for single-sideband modulation also show phase modulation.

If it were possible to arrange the single-sideband modulation without giving rise to the above-mentioned distortion due to higher harmonics in the envelope of the RF signal, two advantages would be obtained. One of them, the reduction of the transmitter bandwidth, has already been mentioned. With the constantly increasing demand for transmitters in the frequency bands available for broadcast purposes, single-sideband modulation would make it possible to use more transmitters or to reduce the interference between them.

A second advantage of a narrower frequency channel is that it reduces the noise and similar interference received. In general the interference energy received is proportional to the bandwidth of the receiver. The general introduction of single-sideband modulation could thus lead to an improvement of the signal-to-noise ratio by a factor of 2. Given the same radiant power, the transmitter could therefore cover a larger area.

We shall now consider three methods by which the distortion can be reduced in the uncorrected single-sideband signal given by expressions (2) and (3). The third method to be discussed, that of the "squaring" of this signal, will receive special attention. This method was devised and developed in Philips Research Laboratories in Eindhoven [1]).

Use of negative feedback

An obvious method of reducing the distortion in the envelope of the SSB (single-sideband) signal is represented schematically in *fig. 2*. It is analogous with the method of reducing distortion by means of *negative feedback* in low-frequency systems.

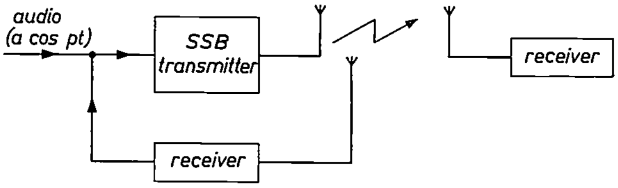


Fig. 2. Application of the negative feedback principle to single-sideband modulation. The amplitude modulation of the transmitted RF signal, not initially free from distortion, is added, after detection, in opposite phase to the modulating audio signal.

The audio signal is supplied to a transmitter with normal amplitude modulation, where one of the sidebands is removed by a filter. The signal which this transmitter sends out therefore shows the envelope distortion discussed above. This distorted signal is received and detected at the same location, and the audio signal thus obtained is then added in opposite phase to the original modulating signal. The result is reduced distortion in the transmitted RF signal.

We shall not discuss this method here in detail, but comment briefly on the result. Without going

[1]) 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. Review, Part A, no. 71, 12-19, 1962.

into mathematics, it can be seen that the signal sent out will extend further into the frequency spectrum on the side of the side-band transmitted. Evidently, in order to remove the second harmonic in the detected signal, the RF signal must contain a compensating side component with the frequency $\omega + 2p$, etc. The result is a broadening of the transmitted frequency spectrum on one side, which partly cancels the bandwidth limitation aimed at. It will be shown in the following, however, that this need not in practice be such a serious drawback. The main disadvantage of this method is much rather a technical problem, for it is in fact impossible to apply sufficient negative feedback. Just as with other systems employing negative feedback, phase shifts — due in particular here to the SSB filter — very soon give rise to instability. The system then goes into self-oscillation ("motor-boating"). For technical application, then, some other procedure has to be found.

Combination of phase and amplitude modulation

We have seen that the removal of a sideband causes phase modulation in the RF signal. Kahn [2]) has described a method of obtaining compatible single-sideband modulation which turns this phase modulation to good use. Since the envelope must not ultimately be distorted, the RF signal sent out will in principle be of the form:

$$v(t) = (1 + a \cos pt) \cos [\omega t + \Omega(pt)]. \quad . \quad . \quad (5)$$

The question here is what condition the phase modulation $\Omega(pt)$ must satisfy in order for the signal to be a single sideband one. In eq. (5) ω again denotes the carrier frequency, and $a \cos pt$ is again the audio signal.

Calculations show that the required $\Omega(pt)$ can be fairly well approximated by starting from the "normal" phase modulation (4) and then increasing the phase deviation by a factor of 1.4. This is done by first multiplying the frequency by 7 and then dividing the resultant frequency by 5. Such transformations of a carrier-wave frequency, in which the modulating frequency does not change, are standard practice in radio engineering.

A block diagram of Kahn's system is given in *fig. 3*. A carrier with a frequency having $\frac{5}{7}$ of the desired value is modulated in amplitude with the audio signal, after which a sideband is suppressed. A limiter cuts off the amplitude modulation from the SSB signal thus obtained, leaving a carrier which is modulated purely in phase. Next, this carrier is multiplied in frequency by the factor $\frac{7}{5}$ in the manner described, after which it is modulated in amplitude again with the audio signal. The signal produced in this way is a very reasonable approximation to the ideal SSB signal with an undistorted envelope. Further particulars of this system, which is already used by some broadcasting stations in the United States, will be found in the article mentioned below [2]).

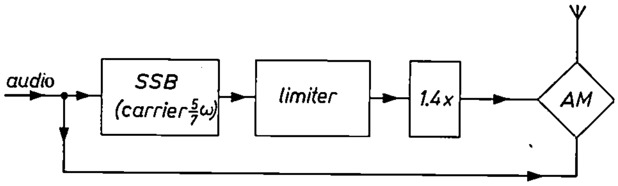


Fig. 3. Schematic diagram of a single-sideband transmitter combining phase and amplitude modulation. The amplitude modulation is removed from the original SSB signal by a limiter. The phase modulation is then increased by a factor 1.4, after which the RF signal is again modulated in amplitude.

Squaring method

At the beginning of this article we recalled that a normal amplitude-modulated signal consists of three equidistant components: the carrier in the centre, and on either side a component having a frequency difference equal to the modulating audio frequency p. The sum of these components is an amplitude-modulated radio-frequency signal without envelope distortion.

The method we shall now discuss for obtaining an SSB signal with an undistorted envelope is based on the remarkable fact that three equally spaced components can be arranged in such a way that one of the *outer* components acts as the carrier wave.

To make the new procedure clear we shall first deal separately with the conditions in which an RF signal with three components of frequencies ω, $\omega + p$ and $\omega + 2p$ has an undistorted envelope with the modulation frequency p.

In general we may write:

$$v(t) = A \cos \omega t + B \cos(\omega + p)t + C \cos(\omega + 2p)t, \quad . \quad . \quad . \quad (6)$$

and we choose A, B and C so that the envelope of this signal has the form:

$$m(t) = P + Q \cos pt. \quad . \quad . \quad . \quad . \quad . \quad (7)$$

If we analyse this problem, we find that the desired result can be achieved in two ways, given respectively by:

$$B = P, A = C = Q/2 \quad . \quad . \quad . \quad . \quad . \quad (8)$$

and

$$B = Q, \; 4AC = B^2, \; A + C = P. \quad . \quad . \quad (9)$$

The first case, given by (8) can be recognized as a normal amplitude-modulated signal with two side-

[2]) L. R. Kahn, Compatible single sideband, Proc. Inst. Radio Engrs. 49, 1503-1527, 1961.

bands, the B component in (6) functioning as the carrier. This case, then, offers nothing new.

The second case given by (9) does, however, present an interesting new possibility. As Q denotes the depth of modulation, and because $B=Q$, we can regard the B component as a sideband. Since A and C are symmetrical in (9), we can now take one of these components as the carrier. The other then becomes a corrective side component, at the frequency distance $2p$, which together with the B component yields an undistorted envelope. It is this possibility that underlies the third method, or squaring method, to be discussed below.

The fact that the signal postulated by (6) and (7) in fact leads to the possibilities given by (8) and (9) can be understood as follows. If, using the basic method represented in fig. 1, we calculate the square of the envelope of the signal defined by (6), we find:

$$[m(t)]^2 = (A - C)^2 + B^2 + 2B(A + C) \cos pt + 4AC \cos^2 pt .$$

In view of (7), this must correspond to:

$$[m(t)]^2 = P^2 + 2PQ \cos pt + Q^2 \cos^2 pt .$$

We may therefore conclude:

$$\left.\begin{aligned} &(A - C)^2 + B^2 = P^2 , \\ &B(A + C) = PQ , \\ &4AC = Q^2 , \end{aligned}\right\} \quad . \; . \; . \; . \; . \quad (7a)$$

We may consider this as three equations containing the two unknown quantities P and Q. If they are to have a solution, then the three equations must be interdependent. It is clear that this interdependence, which provides us with a relation between A, B and C, is given by: $P^2Q^2 = (PQ)^2$. We may therefore write:

$$[B(A + C)]^2 = 4AC\,[(A - C)^2 + B^2] .$$

This yields:

$$B^2(A - C)^2 = 4AC(A - C)^2 .$$

From this it follows that $A = C$, or $B^2 = 4AC$. Together with the three equations of (7a) this gives us precisely the cases (8) and (9).

The addition of a corrective side component at twice the frequency distance from the carrier has, in principle at least, the disadvantage of cancelling the bandwidth limiting aimed at. There is one circumstance, however, which in fact removes this drawback. From eq. (3) it can be seen that the distortion of the uncorrected SSB signal is serious only when the modulation is very deep. Experience has shown that, in the transmission of music or speech, very deep modulation occurs only at low audio frequencies, lower for example than 1000 to 2000 c/s. It is therefore sufficient to add a corrective side component pertaining to these frequencies, and this remains within the frequency spectrum also transmitted with uncorrected single-sideband modulation. (Similar considerations apply to the other methods mentioned for obtaining "compatible" single-sideband modulation.)

An advantage of the method using the corrective side component is the ease with which the required signal can be obtained. To illustrate this, taking a simple case, we write in (6): $A = 1$ and $B = 2b$. In conjunction with eq. (9) this gives: $C = b^2$, $P = 1 + b^2$ and $Q = 2b$. The required RF signal then has the form:

$$v(t) = \cos \omega t + 2b \cos(\omega + p)t + b^2 \cos(\omega + 2p)t . \quad . \; . \; . \quad (6a)$$

This signal is obtained by starting from a normal SSB signal with $\omega/2$ as carrier frequency, and given by:

$$v(t) = \cos \frac{\omega}{2}\, t + b \cos\left(\frac{\omega}{2} + p\right) t . \quad . \quad (10)$$

By *squaring* this signal and then selecting the RF component in the proximity of the frequency ω, we find the required signal (6a). The squaring process is effected by passing the signal (10) through an amplifier stage which has a non-linear, preferably square-law characteristic and which is followed by a network tuned to the frequency ω. This is a standard procedure in radio engineering.

This "squaring" method would indeed be a very simple solution for compatible single-sideband modulation — henceforth called CSSB — if it were not for the fact that modulation with two or more audio frequencies gives rise to intermodulation distortion in the envelope [1]).

Let the signal before squaring have the form:

$$v_1(t) = \cos \frac{\omega}{2}\, t + b_1 \cos\left(\frac{\omega}{2} + p_1\right) t + b_2 \cos\left(\frac{\omega}{2} + p_2\right) t . \quad . \; . \; . \quad (11)$$

After squaring we then find as the envelope:

$$\begin{aligned} m_2(t) = {} & 1 + b_1^2 + b_2^2 + 2b_1 \cos p_1 t + \\ & + 2b_2 \cos p_2 t + 2b_1 b_2 \cos (p_2 - p_1)t . \quad (12) \end{aligned}$$

To make the squaring method really practicable for CSSB it is necessary to suppress the last term in (12), which we call the product or intermodulation term. We shall now examine the method by which this is done.

Eliminating the intermodulation from the amplitude of the squared signal

The interfering intermodulation is removed from the envelope of the squared signal by removing the amplitude modulation of the signal in an amplitude limiter and remodulating the remainder with the undistorted audio signal. We shall presently see how this is done by considering a block diagram of the circuit. First, however, some fundamental points will be dealt with.

The operations referred to do not change the phase modulation of the RF signal, but they do change the spectral composition. Obviously, the spectral composition after squaring only fits the envelope given by eq. (12), i.e. the envelope with the intermodulation term. The disappearance of the intermodulation term is due to the addition of two new side components at the frequency distance $p_2 - p_1$ on either side of the carrier. This is represented in *fig. 4b*, and fig. 4*a* gives the uncorrected spectral composition appertaining to eq. (12). (The new side components together produce an amplitude modulation which exactly cancels out the intermodulation term.)

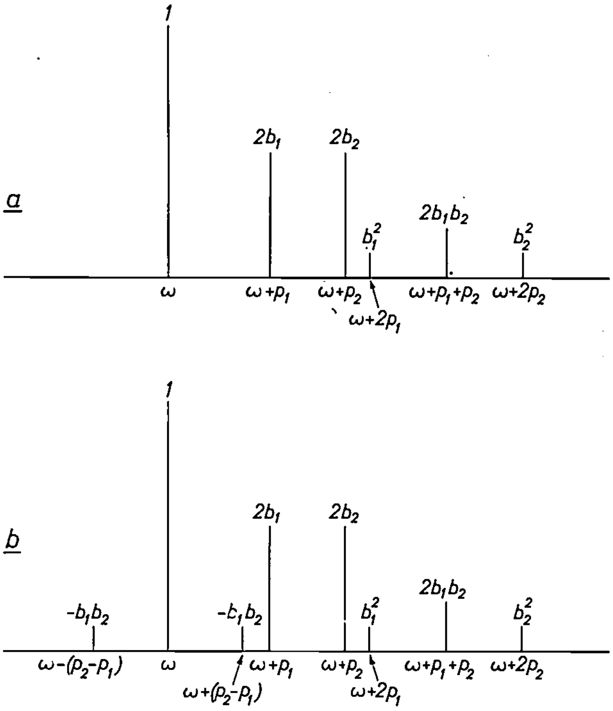


Fig. 4. *a*). Frequency spectrum of an SSB signal, modulated by two audio frequencies p_1 and p_2, and obtained by the squaring method. The spectrum contains an intermodulation component $\omega + p_1 + p_2$, which coincides with the difference tone $p_2 - p_1$ occurring in the signal envelope given by eq. (12). This envelope, then, does not contain the sum frequency $p_2 + p_1$.
b). When the interfering intermodulation term has been removed from the amplitude modulation by a limiter and remodulation, the frequency spectrum is seen to be somewhat wider. The two components not yet found in (*a*) thus cancel out the nterfering difference frequency in the envelope.

The removal of the intermodulation thus has the effect of broadening the radiated frequency spectrum on the side where the sideband is suppressed (not transmitted). This broadening is not, however, of the *difference frequencies* of the low-frequency tones p_2 and p_1, both of which are already fairly low. For, as we have seen, it is only in their case that modulation depths occur that might lead to serious distortion.

We are now sufficiently advanced to be able to discuss the block diagram of a complete transmission system for CSSB, shown in *fig. 5*. On the left, the audio-frequency speech or music signal is conducted to an SSB modulator, which is also supplied with the carrier wave having the frequency $\omega/2$. The circuit of this modulator is designed so that it only delivers the sideband occurring in eq. (10). Before squaring takes place it is therefore necessary to add a carrier. This is done in the block marked with a plus sign. The purpose of the block marked "expandor" will be touched upon presently.

The squaring and amplitude-limiting operations are carried out in the next two stages in the block diagram. At the output of the limiter stage, then, a signal modulated purely in phase is presented to the transmitter, where the signal is remodulated with the undistorted audio signal. The method by which this audio signal is obtained calls for some further explanation.

The simplest procedure would be to modulate the transmitter with the original audio signal used in the SSB modulator. This is ruled out, however, because of the phase shift which the SSB filter in this modulator produced in the audio frequency. The original audio signal therefore no longer "matches" the SSB signal. For this reason the remodulation must be done with an audio signal which is derived from the SSB signal.

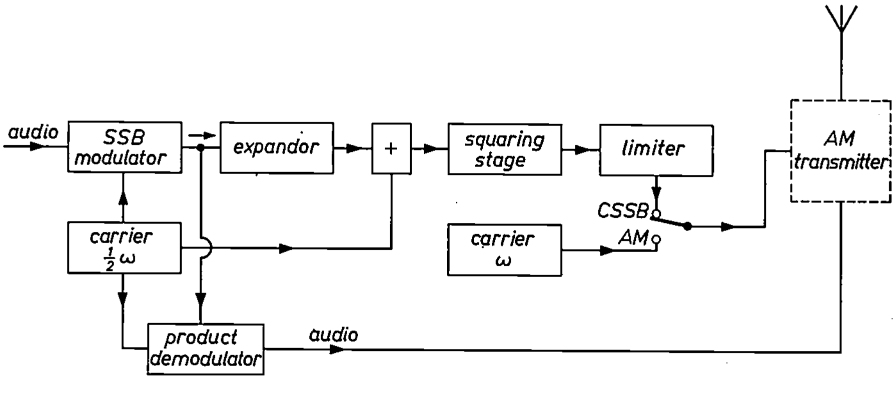


Fig. 5. Complete block diagram of a transmitter for "compatible" single-sideband modulation by the squaring method developed at Philips. The frequency of the primary SSB signal is doubled in this process. The non-linear distortion of the modulation depth in the squaring stage is removed by the preceding "expandor". The AM transmitter is supplied with a signal modulated only in phase, which is then remodulated in amplitude. This ensures that the ultimate amplitude modulation is free from distortion.

In the block marked "product demodulator" in fig. 5 the SSB signal, i.e. the sideband of the signal given by eq. (10), is mixed with the carrier having the frequency $\omega/2$. The result of this operation is a signal with audio frequency p, and it is this that is finally used for the amplitude modulation of the transmitter.

Introduction of a non-linear element

We now have to consider the network referred to in fig. 5 as "expandor". Its function is to compensate for a change which squaring causes in the depth of modulation.

We have already seen that the signal given by (6) or (6a) has the modulation depth Q/P given by (7). In view of the relations in (9) it follows that the modulation depth of the signal (6a) is:

$$\frac{Q}{P}=\frac{2b}{1+b^2}. \quad \ldots\ldots \quad (13)$$

Since the modulation depth is b before squaring, we see that this operation almost doubles the modulation depth for small values of b. For large values of b this is no longer the case. Generally speaking, eq. (13) gives a non-linear relation between the modulation depth before and after squaring.

The audio signal used for modulating the transmitter is derived from the RF signal before squaring. In general, then, this audio signal is proportional to b, and therefore does not exactly fit the required modulation depth given by (13).

This drawback can be partly overcome if we approximate (13) as a function of b by a linear relation in which the modulation depth is somewhat smaller than $2b$. This method, which has previously been dealt with in detail [1]), will not be discussed here. A better method is to remove the non-linearity. This is done in the block diagram shown in fig. 5 by amplifying the transmitted sideband non-linearly before squaring. In the "expandor" the large amplitudes of the SSB signal are therefore given more amplification than the low ones. In conjunction with the non-linear function (13) this then produces a modulation depth which is proportional to that of the original audio signal.

Of course, the audio signal could be subjected to non-linear attenuation or compression before being used for remodulation. The "compressor" would then have to follow behind the product demodulator. The construction of such a non-linear audio amplifier is, in view of the need to minimize distortion, more difficult than one for a radio-frequency signal.

The limiter in fig. 5, the function of which is to suppress the amplitude modulation of the SSB signal, generally works well only up to modulation depths of about 95%, as the instantaneous value of the amplitude should not be too small. The signal delivered by the squaring stage should not therefore be modulated up to 100% in amplitude. In the AM transmitter it is of course possible to remodulate to 100%, and this is basically the procedure adopted. As the fully modulated signal no longer completely matches the phase modulation present, it will show some slight change in spectral composition, but the consequent distortion upon reception is very low compared with the distortion already present as a rule in ordinary receivers. We shall return to the latter point presently.

To complete the description of the transmitting system shown in fig. 5, a further interesting feature should be noted. Since the AM transmitter can be 100% modulated with a non-distorted audio signal, this transmitter can easily be switched over to normal amplitude modulation. To do so it is only necessary to replace the phase-modulated RF signal from the limiter by an unmodulated carrier with the frequency ω. This is indicated in the figure by the switch with two positions: CSSB and AM. On the transmitter side too, then, the method described is completely compatible with the method of amplitude modulation hitherto employed:

Fig. 6 shows a switching unit containing all the elements of fig. 5, except the normal AM transmitter. The relatively small dimensions are sufficient indication that experiments with, or the introduction of this system of CSSB need present no technical problems.

Tests on the squaring system

In cooperation with engineers of the Netherlands G.P.O., tests have been carried out on a 10 kW medium-wave transmitter, operating with signals from a CSSB modulator in accordance with the method described. The results showed that the linear, non-linear and intermodulation distortion did not differ significantly from the corresponding distortion in normal amplitude modulation of the same transmitter.

At modulation depths of 50%, 80% and 90%, sideband suppression was better than 40 dB, 35 dB and 30 dB respectively. From this it can be concluded that the aerial signal fully complies with the requirements which CSSB may be expected to meet.

With a view to the possible introduction of CSSB modulation, it is especially important to have some idea of the reception results in a standard AM receiver. A survey of these results is presented below.

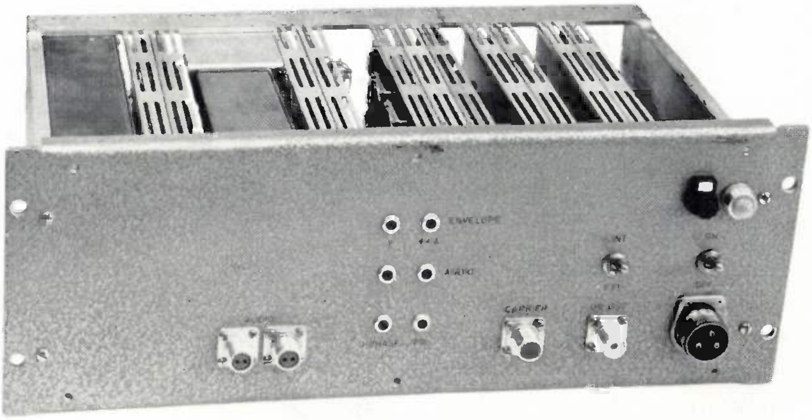


Fig. 6. Example of a switching unit containing all elements required for adapting a standard AM transmitter to CSSB based on the squaring method. Together with this transmitter the unit shown corresponds to the block diagram in fig. 5. The relatively small dimensions make it technically a simple matter to experiment with, or introduce, this system of C.S.S.B.

The quality is governed primarily by the bandpass characteristic of the IF filters in the receiver. These should be lined up so as to avoid any disturbance of the ratio between the three components of the CSSB signal. Since, as explained, the correctivesideband is of interest only at relatively low audio frequencies, this requirement is generally satisfied.

For a conventional AM receiver, *fig. 7* gives a plot of the linear and non-linear distortion as a function of the audio frequency p, both for AM and CSSB, at 50% modulation depth. As the new modulation method makes it possible, because of the smaller bandwidth, to obtain better reproduction of the high audio frequencies by slight detuning, test results for CSSB are also given with the receiver detuned 1 kc/s and 2 kc/s. The result, as can be seen from the figure, is less linear distortion without any troublesome increase in non-linear distortion.

The curves for the second harmonic, $2p$, which may be taken as representative of the non-linear distortion, indicate that at audio frequencies above about 1000 c/s the non-linear distortion is greater with CSSB than with AM. At this point we repeat that a modulation depth of 50% seldom occurs at these frequencies. The increase of distortion can easily be explained by the attenuation which the corrective side component suffers in the IF filter. When the receiver is detuned, the ratio between the components of the detected signal changes. This too gives rise to increased non-linear distortion at the higher audio-frequencies.

Above about 3000 c/s, however, it can be seen that the non-linear distortion decreases again. This is bound up with the fact that at these frequencies the main side component also suffers attenuation. From eq. (6a) it can be deduced that the corrective side component should then decrease quadratically in order to produce again an undistorted RF signal. This more or less quadratic attenuation occurs in the IF filter.

The energy distribution of speech or music in the audio frequency spectrum is such that tones above 2 kc/s seldom account for more than 15% of the total modulation. This means that, as far as these tones are concerned, the corrective component is in fact unimportant, and the uncorrected SSB signal is already sufficiently free from distortion. For this reason, detuning of the receiver is quite permissible with CSSB, and in practice it results in a noticeably better reproduction of the

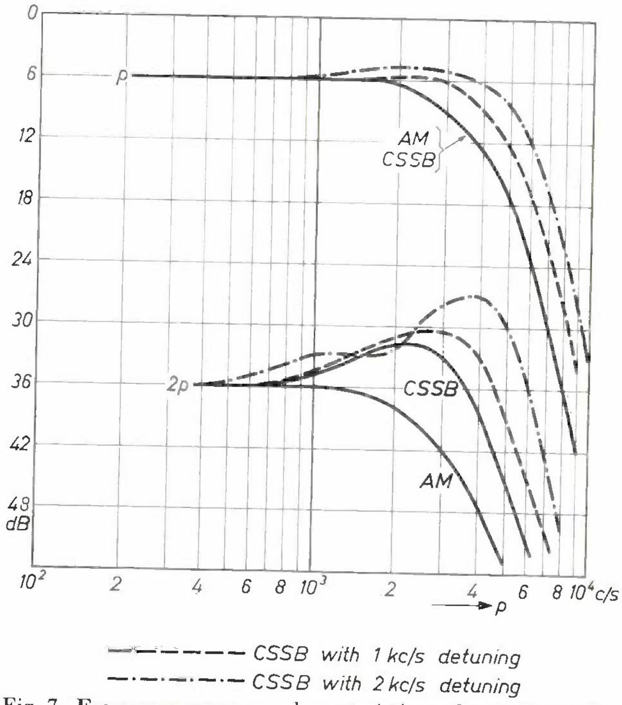


Fig. 7. Frequency response characteristic and ~~non~~-linear distortion in normal AM and CSSB reception with a conventional AM receiver at 50% modulation depth. The strength of the second harmonic ($2p$) may be considered representative of the non-linear distortion. By detuning the receiver 1 to 2 kc/s for CSSB reception, the reproduction of the high audio frequencies is improved without appreciably increasing the non-linear distortion. When interpreting the curves it should be remembered that, with speech and music, a modulation depth of 50% hardly ever occurs at frequencies above 1 kc/s. At smaller depths of modulation the distortion decreases quadratically.

high tones without any significant increase of nonlinear distortion. This improvement of reception quality, together with the other advantages mentioned in the foregoing, is a particularly strong argument in favour of the gradual introduction of CSSB.

Interference between transmitters

Apart from the reception quality when a conventional type of receiver is used, an important factor in the judgement of CSSB is the interference between transmitters. To determine this interference, tests have been carried out on the lines illustrated in *fig. 8*.

A standard programme, consisting of filtered noise having an energy distribution corresponding to that of music, is transmitted via the first transmitter in conjunction with a receiver tuned to it. Connected to the output of the receiver is a psophometer, which measures the strength of the programme. (A psophometer is a voltmeter preceded by a filter which simulates the sensitivity of the human ear).

Next, the programme is switched over to an interfering transmitter, which is likewise connected to the input of the receiver. When a certain difference exists between the carrier frequencies of the two transmitters, the output of the interfering transmitter is adjusted until the psophometer gives the same reading again.

The results were plotted to produce the contours shown in *fig. 9*, which represent the relative strength (interference ratio) of the transmitters as a function of the frequency difference. If the meter is adjusted in this test not to the same but to a smaller deflection for the interfering transmitter, the contours can be moved correspondingly upwards. The tests related to three different situations:

1) Both transmitters with normal amplitude modulation (solic curve).
2) Both transmitters with CSSB modulation and the sideband on the same side of the carrier (dashed curve).
3) Both transmitters with CSSB modulation and the sideband at different sides of the carrier (dot-dash curve).

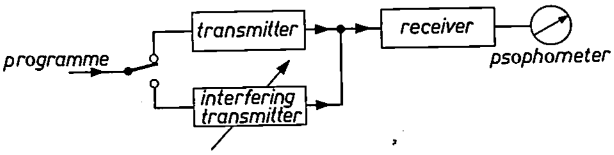


Fig. 8. Test arrangement for determining the interference between two transmitters, both for normal AM and for CSSB The transmitter under test and the interfering transmitter are modulated in turn with a standard programme. The relative strength (interference ratio) of the transmitters is measured when the interfering transmitter is heard just as strongly as the other. A filter in the psophometer simulates for this purpose the sensitivity characteristic of the human ear.

Whether, in the event of CSSB being introduced, case (2) or perhaps even case (3) will be encountered depends upon the arrangements agreed upon, and cannot yet be predicted. Both cases therefore had to be studied in these tests. When CSSB was applied, the receiver was detuned 2 kc/s, since this may be considered as the normal mode of operation.

The form of the curves can best be explained by distinguishing three separate regions of frequency difference between the transmitters.

a) If the frequency difference is so slight that the interference tone of the carrier waves falls in the

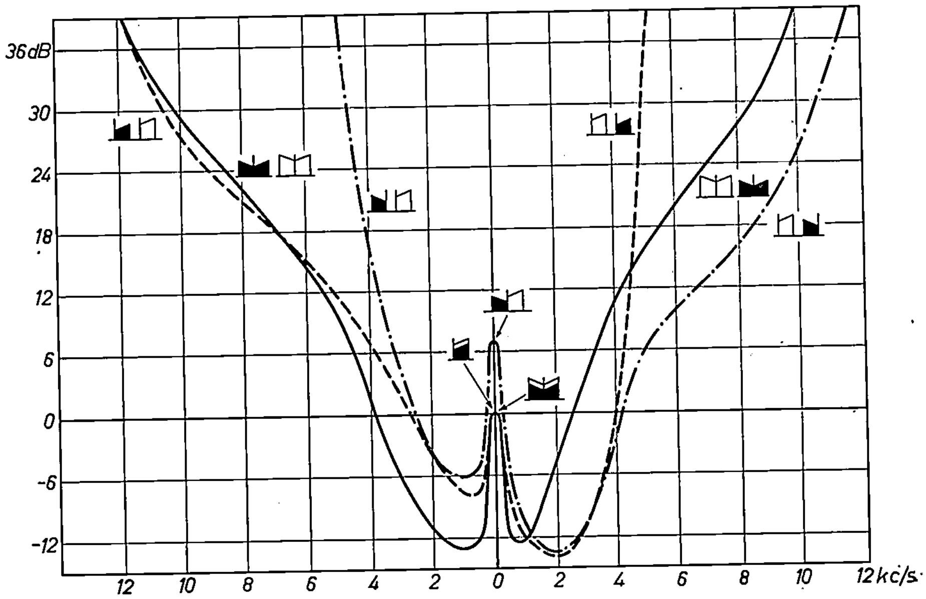


Fig. 9. Results of measurements with the set-up shown in fig. 8. The relative strength (interference ratio) of the interfering and the transmitter under test, for reception at equal strength, is plotted as a function of the difference in the carrier frequencies of both transmitters. The three contours relate to the interference between two transmitters both with AM or both with CSSB In the latter case, two further situations are distinguished: the transmitted sidebands are on the same side or on different sides of the carrier. The sidebands are represented by the black and white symbols (the white ones representing the transmitter under test, the black ones the interfering transmitter). In CSSB reception (sideband on the right) the receiver is detuned 2 kc/s on the right to improve the reproduction of the high audio frequencies (see fig. 7). This accounts for the asymmetry in the relevant contours. As compared with AM the interference with CSSB is substantially lower if the sideband transmitted by the interfering station is on the opposite side from that transmitted by the desired station.

region of low aural sensitivity, i.e. below about 100 c/s, the interference is largely caused by the modulation of the interfering transmitter. In the case of AM (case 1) both transmitters are as it were interchangeable under these conditions, and the interfering transmitter will have to be just as powerful as that under test if the interference is to produce the same deflection on the meter as the standard programme. In case (2), i.e. CSSB with the sideband of both transmitters on the same side, the situation is still roughly the same. The situation is different in case (3), i.e. CSSB with the sideband of the transmitter under test on a different side from that of the interfering transmitter: owing to the detuning of the receiver, the interfering sideband is then amplified less than the other. As can be seen in the figure, one thus gains 7 dB.

b) Where the frequency difference is greater, falling in the region of maximum aural sensitivity, i.e. between about 100 c/s and 4000 c/s, the interference tone of the carriers is the most troublesome effect. Since this does not depend on the method of modulation, the CSSB system will not — on the average — offer any further gain.

The difference tone decreases in strength when the interfering carrier, with increasing frequency difference, is attenuated in the IF filter in the receiver. In this region, then, the contours give a picture of the filter characteristic with that of the psophometer added to it.

c. At frequency differences above 6 kc/s the interference is mainly due to one of the sidebands of the interfering signal penetrating into the spectrum of the desired signal. It is clear that on average CSSB now offers an improvement, due to the absence of a sideband. In evaluating the advantage gained, one should take into account that the remaining sideband is 6 dB stronger than with AM.

Concluding, it may be said that a change-over from AM to CSSB can considerably reduce interference between stations. Moreover, detuning the receivers from the carrier position can lead to an improvement of reproduction quality.

Summary. The ever-increasing power and number of broadcasting transmitters makes it desirable to reduce their bandwidth in the frequency channels allocated to them. Single-sideband modulation is one means to this end. If signals modulated in this way can be received by sets currently in use, the single-sideband modulation is said to be "compatible" with the conventional double-sideband system, and can be introduced gradually. The article deals in particular with the "squaring method" developed by Philips for obtaining the SSB signal required for this purpose. The underlying idea is that amplitude modulation sufficiently free from distortion can be achieved by adding a corrective side component, lying within the transmitted sideband, to a non-corrected SSB signal. The circuitry needed to adapt a standard AM transmitter for SSB modulation by this method can be accommodated in a simple switching unit.

Tests show that compatible SSB modulation offers better reproduction of the high audio frequencies withou tany troublesome increase in non-linear distortion. Greater freedom from interference is also possible, enabling the transmitter to cover a wider area.

A 5000 : 1 SCALE MODEL OF THE MAGNETIC RECORDING PROCESS

by D. L. A. TJADEN *) and J. LEYTEN *).

621.318.24.087

Although magnetic recording is widely used for all kinds of purposes, little is yet known about the actual process by which signals are recorded in the thin magnetic coating of the tape, Attempts are being made to gain a better understanding of this process with the aid of a model of the tape and recording head scaled up 5000 : 1 and in which the distribution of the magnetization over the thickness of the coating is measured.

In the past 15 to 20 years magnetic recording has developed into an indispensable tool, both for recording sound and video signals and for the registration of machine data (measuring, regulating, control and computer signals). This development was based on the improvements in the properties of the magnetic tape and in those of the recording and playback heads. If one asks, however, to what these improvements are due, one has to admit, rather disappointedly perhaps, that they were obtained for the most part by trial and error. In our opinion the reason for this is that fundamental understanding of the factors that govern the recording process is still imperfect. This applies particularly to the

*) Philips Research Laboratories, Eindhoven.

region of low aural sensitivity, i.e. below about 100 c/s, the interference is largely caused by the modulation of the interfering transmitter. In the case of AM (case 1) both transmitters are as it were interchangeable under these conditions, and the interfering transmitter will have to be just as powerful as that under test if the interference is to produce the same deflection on the meter as the standard programme. In case (2), i.e. CSSB with the sideband of both transmitters on the same side, the situation is still roughly the same. The situation is different in case (3), i.e. CSSB with the sideband of the transmitter under test on a different side from that of the interfering transmitter: owing to the detuning of the receiver, the interfering sideband is then amplified less than the other. As can be seen in the figure, one thus gains 7 dB.

b) Where the frequency difference is greater, falling in the region of maximum aural sensitivity, i.e. between about 100 c/s and 4000 c/s, the interference tone of the carriers is the most troublesome effect. Since this does not depend on the method of modulation, the CSSB system will not — on the average — offer any further gain.

The difference tone decreases in strength when the interfering carrier, with increasing frequency difference, is attenuated in the IF filter in the receiver. In this region, then, the contours give a picture of the filter characteristic with that of the psophometer added to it.

c. At frequency differences above 6 kc/s the interference is mainly due to one of the sidebands of the interfering signal penetrating into the spectrum of the desired signal. It is clear that on average CSSB now offers an improvement, due to the absence of a sideband. In evaluating the advantage gained, one should take into account that the remaining sideband is 6 dB stronger than with AM.

Concluding, it may be said that a change-over from AM to CSSB can considerably reduce interference between stations. Moreover, detuning the receivers from the carrier position can lead to an improvement of reproduction quality.

Summary. The ever-increasing power and number of broadcasting transmitters makes it desirable to reduce their bandwidth in the frequency channels allocated to them. Single-sideband modulation is one means to this end. If signals modulated in this way can be received by sets currently in use, the single-sideband modulation is said to be "compatible" with the conventional double-sideband system, and can be introduced gradually. The article deals in particular with the "squaring method" developed by Philips for obtaining the SSB signal required for this purpose. The underlying idea is that amplitude modulation sufficiently free from distortion can be achieved by adding a corrective side component, lying within the transmitted sideband, to a non-corrected SSB signal. The circuitry needed to adapt a standard AM transmitter for SSB modulation by this method can be accommodated in a simple switching unit.

Tests show that compatible SSB modulation offers better reproduction of the high audio frequencies withou tany troublesome increase in non-linear distortion. Greater freedom from interference is also possible, enabling the transmitter to cover a wider area.

A 5000 : 1 SCALE MODEL OF THE MAGNETIC RECORDING PROCESS

by D. L. A. TJADEN *) and J. LEYTEN *).

621.318.24.087

Although magnetic recording is widely used for all kinds of purposes, little is yet known about the actual process by which signals are recorded in the thin magnetic coating of the tape, Attempts are being made to gain a better understanding of this process with the aid of a model of the tape and recording head scaled up 5000 : 1 and in which the distribution of the magnetization over the thickness of the coating is measured.

In the past 15 to 20 years magnetic recording has developed into an indispensable tool, both for recording sound and video signals and for the registration of machine data (measuring, regulating, control and computer signals). This development was based on the improvements in the properties of the magnetic tape and in those of the recording and playback heads. If one asks, however, to what these improvements are due, one has to admit, rather disappointedly perhaps, that they were obtained for the most part by trial and error. In our opinion the reason for this is that fundamental understanding of the factors that govern the recording process is still imperfect. This applies particularly to the

*) Philips Research Laboratories, Eindhoven.

recording process at very short wavelengths, that is to say at high signal frequencies and/or low tape speeds. But even where the recording of longer wavelengths is concerned, it is difficult to predict exactly the behaviour of a tape from its magnetic properties.

How is it that the theory is so far behind the practice? It should be possible in principle to describe the recording process by passing a piece of tape over the recording head and ascertaining how the form of the magnetization curve of the tape material — at points at varying depths in the tape — varies. The situation is complicated, however, because as the tape passes the head the magnetizing field changes not only in strength but also in direction (turning through a total of 180°, see *fig. 1*), and also by the fact that the magnetization at any given instant is determined to a considerable extent by demagnetizing fields. In view of these complications, we have in fact to rely on experimental data to improve our knowledge of the behaviour of a magnetic tape, especially data on the way in which the resultant magnetization vector changes in magnitude and direction over the thickness of the magnetic coating after the recording of a simple signal. Hitherto, such data have been almost completely lacking, and there is little hope of obtaining them from direct measurements on the 0.01 mm thick coatings of normal magnetic tapes.

To overcome this difficulty, a large-scale model of the magnetic recording process has been constructed at the Philips Research Laboratories, Eindhoven; the essential dimensions have been scaled up by a factor of 5000, the recording head being given a gap length of 2 cm and the magnetic coating of the "tape" a thickness of 5 cm. By appropriately choosing the speed at which the tape travels and the frequency of the signal current applied to the recording head, a recorded wavelength is obtained which is in the same ratio to gap length and layer thickness as in the normal magnetic recording process. Theoretical considerations show that, given this condition, and provided the magnetizing field-strengths have the same magnitude, the spatial variation of the magnetization vector in the layer may be expected to be a fairly faithful copy of that in the normal process. In the large model this magnetization vector can be measured directly and accurately on samples removed from the layer after a recording.

It is reasonable to assume that the magnetization pattern is not essentially affected by the other dimensions in the recording process; this is fortunate, for scaling up all dimensions by a factor of 5000 would have led to an impossible construction. *Fig. 2* shows the form of the recording head and tape as originally adopted. The tape is only 60 cm

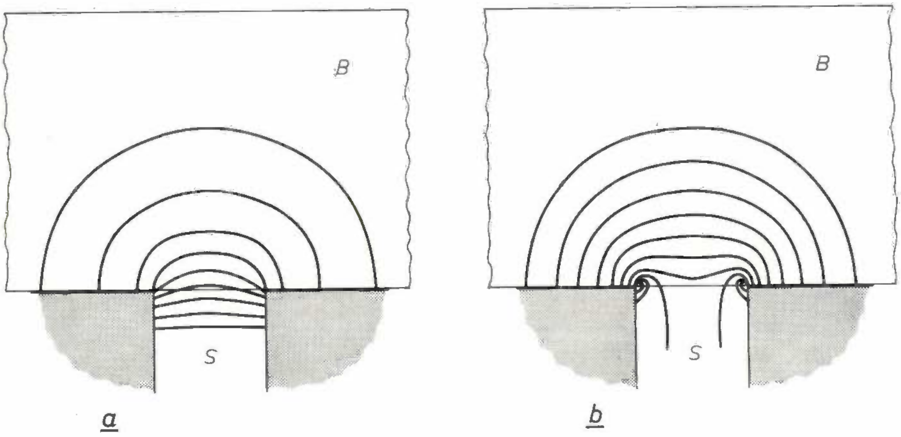


Fig. 1. Qualitative picture of the magnetic field near the gap S of a magnetic recording head. B is the coating of magnetic material on the tape.
a) Conventional pattern of lines of force.
b) Lines of equal field strength (H = constant).

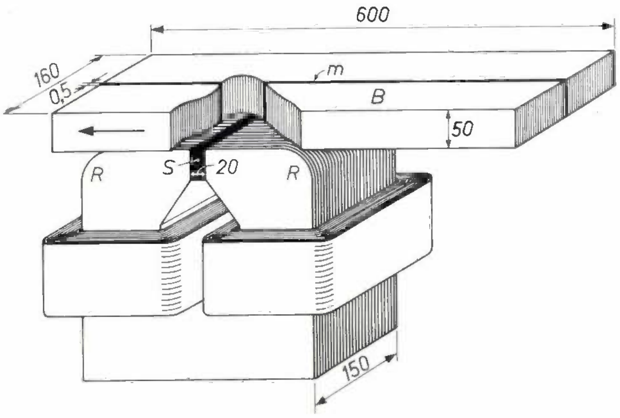


Fig. 2. Dimensional drawing of the large-scale model of recording head and tape. The tape B is composed of three longitudinal strips; the measurements are performed on samples taken from the middle strip m. Part of the tape is here cut away to disclose the gap S in the recoi ding head. Dimensions in mm.

long and consists of a dispersion in a suitable material of grains of γFe_2O_3 — exactly the same grains as used on a normal magnetic tape. The dispersion is contained in a tray of non-ferromagnetic material and consists of three longitudinal strips. It is from the middle strip (*m* in fig. 2), which is only 0.5 mm thick, that the samples are removed for measurement, and this strip is therefore renewed for each experiment.

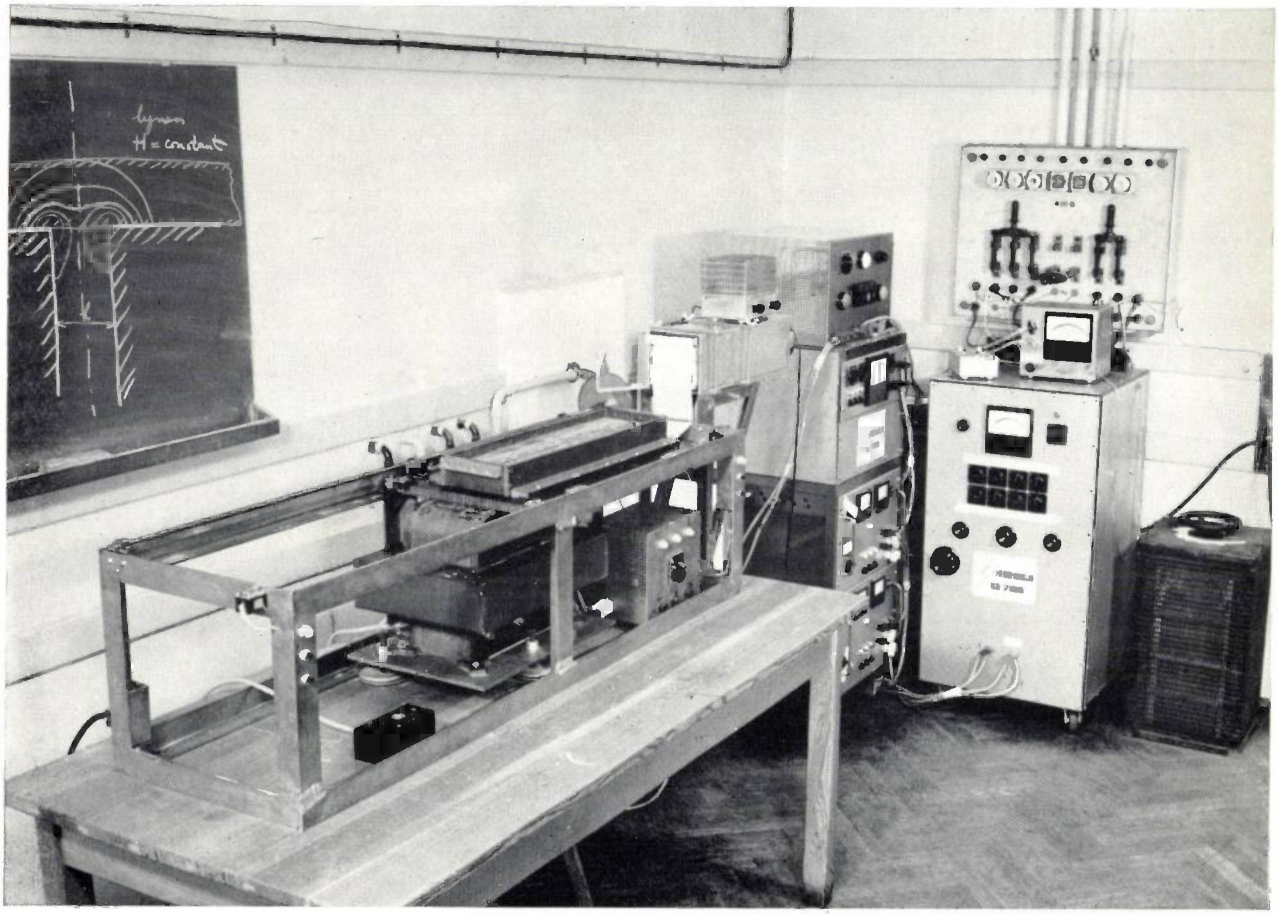


Fig. 3. Complete arrangement of the large-scale model of the magnetic recording process (first version). On the table can be seen the large recording head, and above this the tray with magnetic layer, carried on a roller track. In the background, from left to right: the generator for the signal current of 0.1 c/s (maximum 10 A), surmounted by a cabinet with relay switchgear; then the generator for the bias current of 50 c/s (maximum 60 A); and finally a rheostat for controlling a DC signal (maximum 70 A).

Although in our model the grains are a factor of 5000 too small in relation to the wavelength, this does not — provided we disregard the noise — cause any significant errors. In fact, if the grains were to be scaled up — supposing it were possible — fundamental difficulties would then really arise. In the first place a grain enlarged 5000 times would contain innumerable Weiss domains, whereas most of the real grains contain no more than one Weiss domain (i.e. no Bloch walls). The coercivity of the large grains would therefore be considerably less than that of the small ones. Secondly, the samples, which we want to be relatively small with a view to good spatial definition, would contain so few grains that the measurement results would inevitably show a very wide spread. Although this spread is real, that is to say it would also be ascertainable in the real tape if one were to consider an equivalently narrow strip, in practice it is averaged out by the enormous width of the track in relation to the individual grain.

Construction of the model

The complete experimental set-up in its original form is shown in *fig. 3.* On the table can be seen the large recording head, and above this a track with rollers on which the tray containing the dispersion is moved over the head on a carriage. In this arrangement the carriage is moved by means of a screw shaft driven by an electric motor. The speed of travel can be varied slightly. The tape speed we normally use is 0.4 cm per second. In *fig. 4* the head and the carriage (with the tray removed) are shown more clearly. It can also be seen that the carriage draws along behind it a strip of recording paper. On this strip is traced a record of the signal applied to the head.

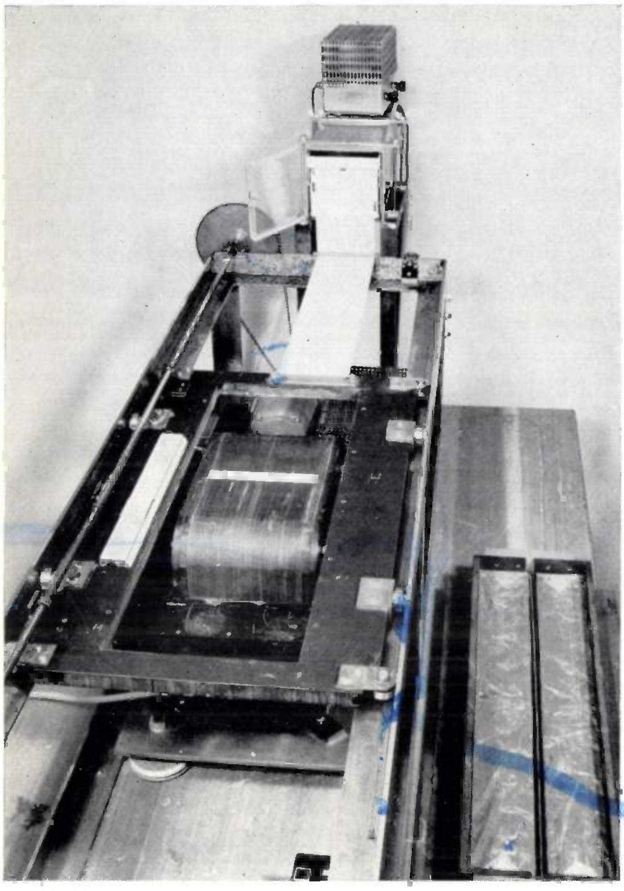


Fig. 4. Recording head and roller track with carriage. The tray containing the magnetic layer has been removed and placed on the table on the right. The recording-head gap, which, like conventional heads, is filled with a non-ferromagnetic material, has been painted white to make it show up. The carriage draws along behind it a paper strip on which a diagram of the signal current applied to the head is recorded.

The signal current has a fixed frequency of 0.1 c/s. The recorded wavelength is then 4 cm, which corresponds to a wavelength of 8 μm in normal recording. The latter wavelength is obtained, for example, at a signal frequency of 12 000 c/s and a tape speed of 9½ cm/sec. For generating the signal current a somewhat unconventional generator is required, not because of the very low frequency (this is easily produced with an *RC* generator), but because of the very high current needed. In normal recording the largest magnetizing signal field-strengths ever required (approx. 10^4 A/m) can be generated at the greatest depth in the tape with about 0.3 ampere turns. In the model the distances from the material to be magnetized to the air gap of the recording head have been enlarged by a factor of 5000, which means

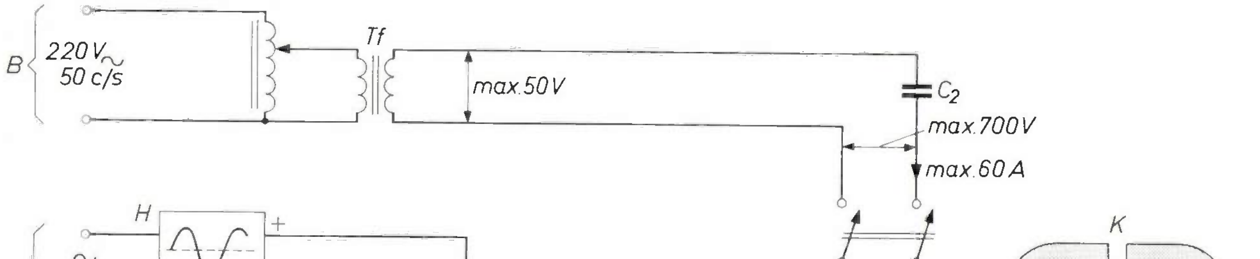


Fig. 5. Simplified diagram of the circuitry for supplying an AC signal (*Si*), a DC signal (*G*) and a bias current (*B*). *K* recording head. The AC signal of 0.1 c/s is delivered by two stabilized power supply circuits, *H* and *H'*, connected in push-pull. The output voltage of each of these circuits is controlled, with the aid of a small electric motor *M*, and sinusoidally varied 6 times per minute between 5 and 15 V, while *H* and *H'* operate in anti-phase. At the moments when both are delivering the same voltage (10 V) the current through the recording head *K* is zero; in the interim periods the current through *K* is alternately positive and negative. L_1-C_1-C_1' is a filter that keeps the 50 c/s bias current outside *H* and *H'*.

The bias current is supplied from the mains via a variable transformer and an isolating transformer T_f. Since the capacitor C_2 with the coil of the recording head constitutes a circuit tuned to the mains frequency, the output voltage of T_f need not be more than 50 V.

that for the same purpose we need roughly 2000 ampere turns. The recording head designed by us contains 175 turns, so that signal currents up to about 10 A should be possible. The principle of the generator used to produce these signal currents is illustrated in *fig. 5* and explained in the caption.

The same figure shows how we generate the "high-frequency" bias current which is always superimposed on the signal in normal recording, and which it must be possible to supply to the recording head in our case too. For convenience we chose for the bias current the frequency of the mains, 50 c/s. (In itself a frequency of a few c/s would be high enough, even at our maximum tape speed, to make the recorded wavelength of the bias current smaller than the diameter of the samples measured.) Bias currents up to 60 A should be obtainable. The voltage on the coil around the recording head (inductance 35 mH) is then about 700 V, a value low enough to avoid insulation difficulties. The power dissipated in the recording head, due to eddy currents, etc., is about 2.5 kW.

The recording of pulses can also be studied in the model. For this purpose the recording head is supplied with a direct current of about 70 A maximum, and the steep leading edge of a pulse is simulated by switching the current on and off.

All the equipment used for generating and controlling the various currents applied to the recording head can be seen in fig. 3 in the background.

The first measurements were on a "tape" consisting of randomly orientated grains dispersed in vaseline. In practice the grains in the tapes used for magnetic recording are usually needle-shaped. They have a single preferred direction of magnetization, and in the manufacturing process this is arranged to lie mainly in the long axis of the tape. We were able subsequently to devise a method of making our tape in such a way as to simulate this preferential orientation, a plastic material with dispersed grains being calendered so as to align the grains. This material is now used for the middle strip from which the samples are taken. The space in the tray on either side of the exchangeable middle strip is also filled with a large number of narrow, tightly packed plastic strips fabricated in the same way.

In the first experiments it was soon apparent that the pole pieces of the recording head, which were relatively far too small in length, caused interfering edge effects: the field concentration at the edges (R in fig. 2) also affected the resultant magnetization in the tape after passing the gap. It is scarcely feasible to scale up the pole pieces to 5000 : 1 (this would make them several tens of metres long); the edge effects can, however, be avoided by making the pole pieces long enough so that the piece of tape used, does not pass their ends. We were able to satisfy this condition in a subsequent version of the model in which the tape speed can also be varied between much wider limits (*fig. 6*).

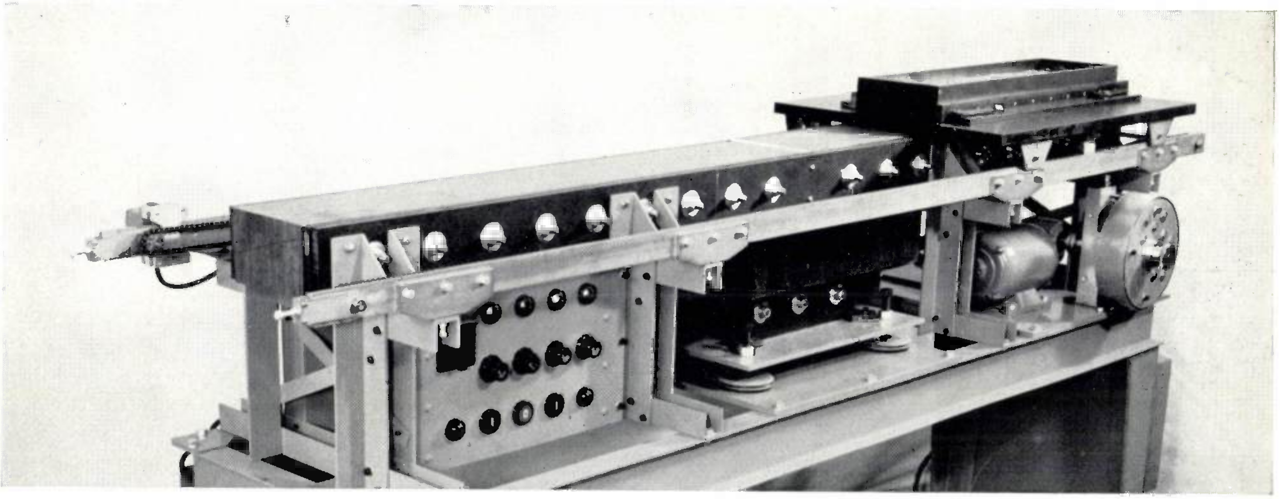


Fig. 6. Model of recording head and tape in its present form. The pole pieces of the head have been lengthened so that their ends are not passed by the tray containing the tape material. The tray in this version is drawn along the track by two chains, which proved to give a smoother motion at high speeds than a screw shaft.

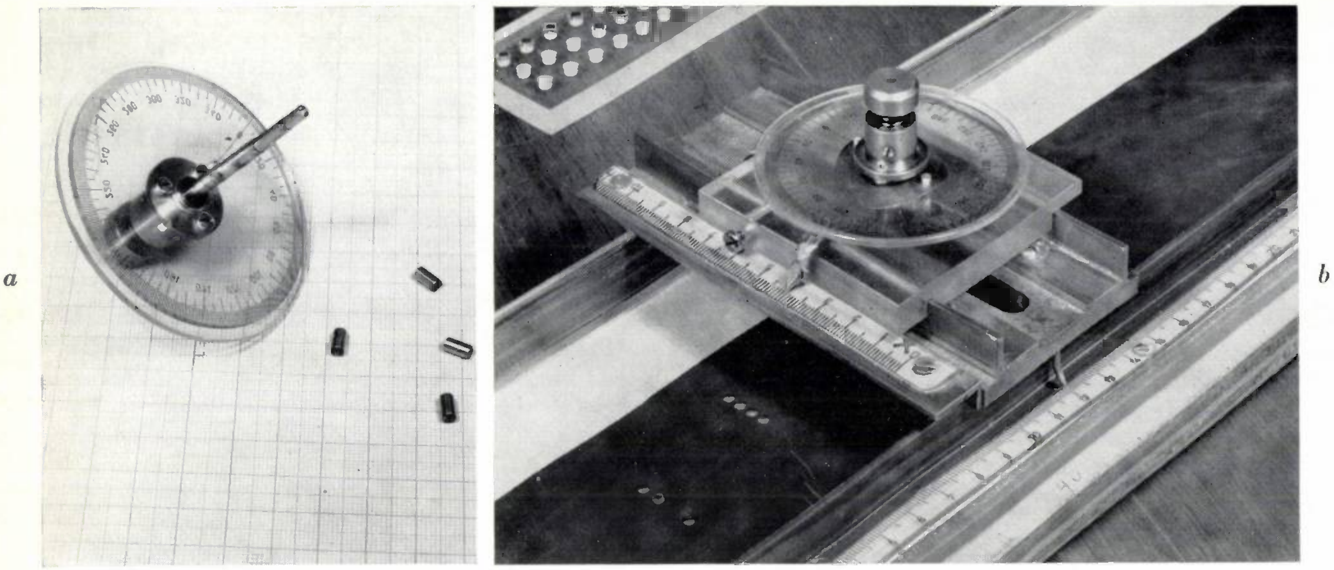


Fig. 7. Device for punching out disc-shaped samples from the 0.5 mm thick middle strip of the tape.

a) Holder with graduated circle; over the end of the spindle is fitted an interchangeable hollow punch which collects the sample.

b) Roller track and carriage with which the holder can be placed above any desired point of the tape. A sample can only be punched out when the zero pointer of the graduated circle on the holder is parallel to the long axis of the track (and hence of the tape). Some samples have already been punched out of the tape.

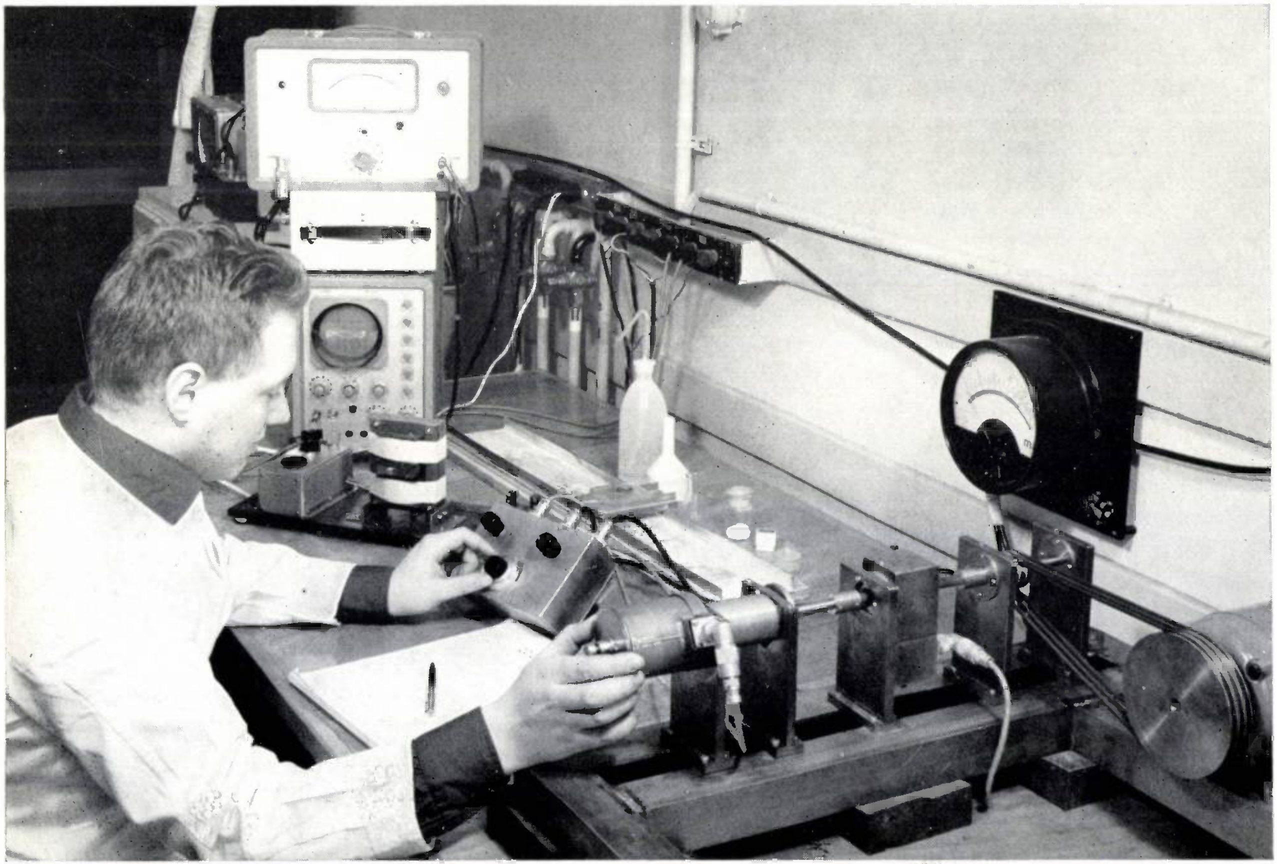


Fig. 8. Equipment for measuring the magnitude and direction of the magnetization of the individual samples. In the right foreground can be seen the magnetometer proper, in which the holder with graduated circle and sample is placed. For further particulars, see fig. 9. The alternating voltage generated by the magnetized sample is compared with a signal generated by a "reference magnet". By rotating the sample holder the phases of both alternating voltages are equalized, and their amplitudes are equalized by regulating a calibrated attenuator.

To eliminate approximately the influence of the demagnetization factor of the sample, each sample after measurement is momentarily magnetized to saturation (this is done by the electromagnet that can be seen in front of the oscilloscope) and the magnitude of the magnetic moment of the remanence is determined in the same way with the magnetometer. This being done, the hollow punch with the sample is removed from the holder and a new measurement started.

Measurement of the samples

The samples — discs of 3 mm diameter — are punched out of the 0.5 mm thick centre strip of our "tape" by means of a device shown in *fig. 7*. The holder in which the sample is collected carries a graduated circle which, in the punching operation, is kept with its zero angle always exactly in the long axis of the tape by means of two guide pins. The location of each sample can be measured on the centimetre scales to within an accuracy of $\frac{1}{2}$ mm. The magnitude and direction of the magnetic moment of each sample are then measured by means of an instrument which is illustrated in *figs. 8* and *9*. It is based on the same principle as the rotating sample magnetometer. The phase and magnitude of the alternating voltage induced in the coil give the direction and magnitude of the magnetic moment of the sample. In our arrangement the roles have merely been reversed: the sample, introduced together with the holder into the instrument, remains stationary, while the two pole pieces rotate around it at 175 r.p.s. For convenience, the coil and the other part of the magnetic circuit are also kept stationary. The small air gap that we have to leave between this part and the pole pieces (*l* in fig. 9*a*) does practically no harm. The advantage of this arrangement is that it takes very little time and trouble to change from one sample to the next — an important consideration in view of the large series of measurements needed for each experiment.

The actual measurement is done by a null method. The magnitude and phase of the induced e.m.f. are compared on an oscilloscope with those of a reference signal, generated in a small coil by a magnet on the shaft of the rotating pole pieces (fig. 9*b*). By turning the holder with the sample through a given angle, the phases of both signals can be equalized, while at the same time the amplitudes are equalized with a variable attenuator. The direction of the magnetic moment of the sample can then be read from the scale around the holder with an accuracy of about $\frac{1}{2}$°, and the magnitude of the magnetic moment can be read from the attenuator.

A drawback of the measuring arrangement described is its sensitivity to constant stray fields. Good mu-metal screening is fitted to counteract such interference.

The magnetic moments of the samples punched from the tape do not give directly the magnetization in situ in which we are interested. After the sample is punched out its magnetic state is subject to a demagnetizing field different from that in the tape. Experiments have shown that the effects of this field can for all practical purposes be eliminated — together with the effect of any inhomogeneity in the dispersion — by determining in addition the magnetic moment of the remanence of each sample after *saturation*, and by dividing the measured moment of the recorded magnetization by the value so found.

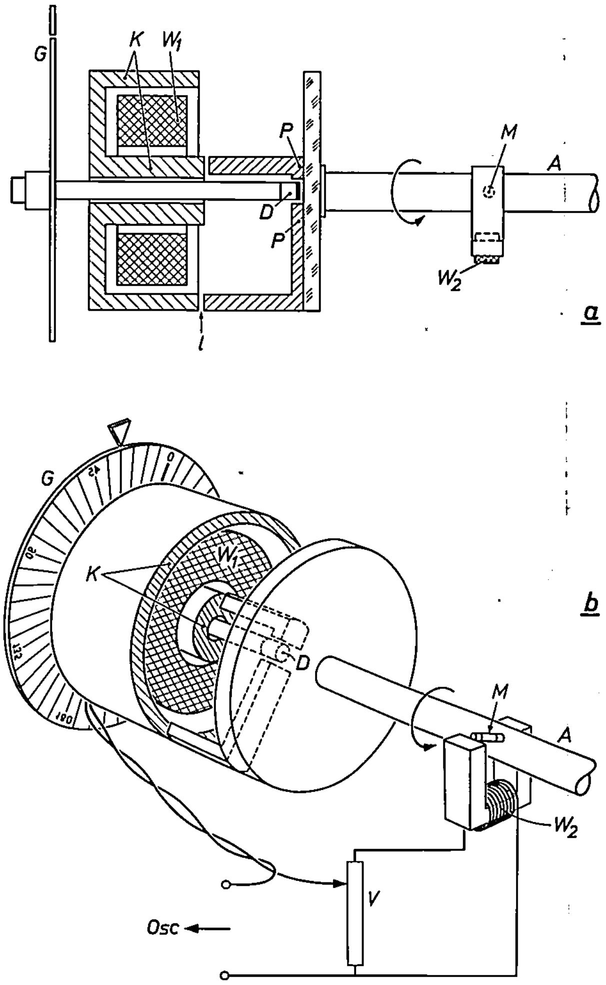


Fig. 9. *a*) Principle of the magnetometer for measuring the samples. *D* hollow punch with magnetized sample, at the end of the holder with graduated circle *G*. The holder remains stationary, but can be turned by hand to any required position. The ferrite core *K* with coil W_1 is also stationary, but the pole pieces *P* of this magnetic circuit, which are separated from the core *K* by a narrow air gap *l*, are fixed to the shaft *A* and rotate at 175 r.p.s. around the sample. In the perspective drawing (*b*) it can be seen that the shaft also carries a small magnet *M* which induces in the coil W_2 an alternating voltage, which serves as a reference signal. The alternating voltage induced by the sample in W_1 is compared with that induced in W_2, both in phase and amplitude, by rotating the holder and regulating the attenuator *V*; this can be monitored on an oscilloscope (*Osc*).

Some results of measurements

Finally, we shall mention some of the results of measurements carried out in the first experiments on our model.

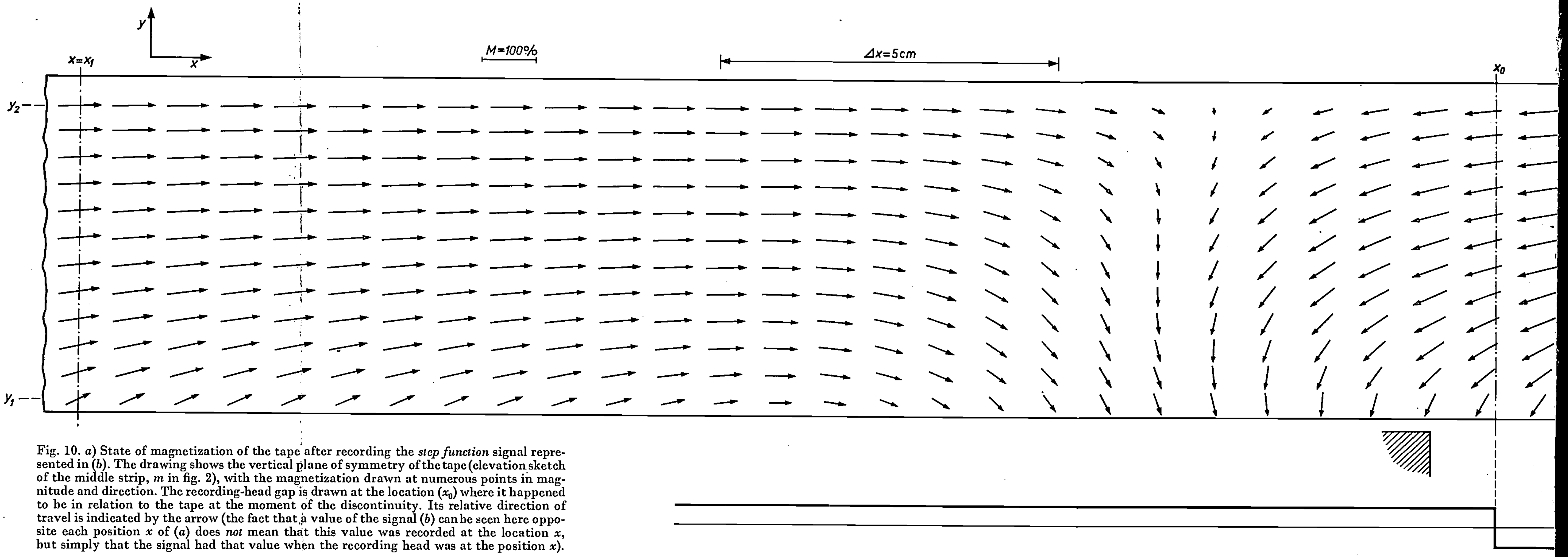


Fig. 10. *a*) State of magnetization of the tape after recording the *step function* signal represented in (*b*). The drawing shows the vertical plane of symmetry of the tape (elevation sketch of the middle strip, *m* in fig. 2), with the magnetization drawn at numerous points in magnitude and direction. The recording-head gap is drawn at the location (x_0) where it happened to be in relation to the tape at the moment of the discontinuity. Its relative direction of travel is indicated by the arrow (the fact that a value of the signal (*b*) can be seen here opposite each position x of (*a*) does *not* mean that this value was recorded at the location x, but simply that the signal had that value when the recording head was at the position x).

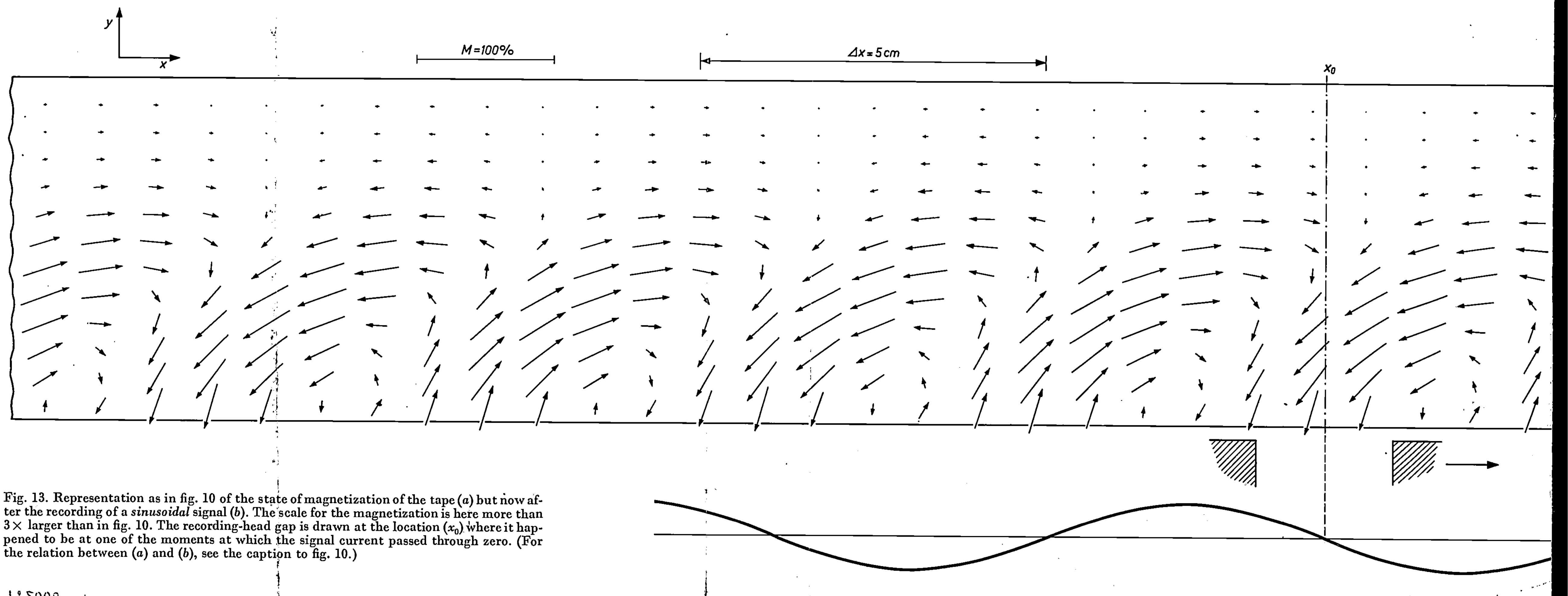
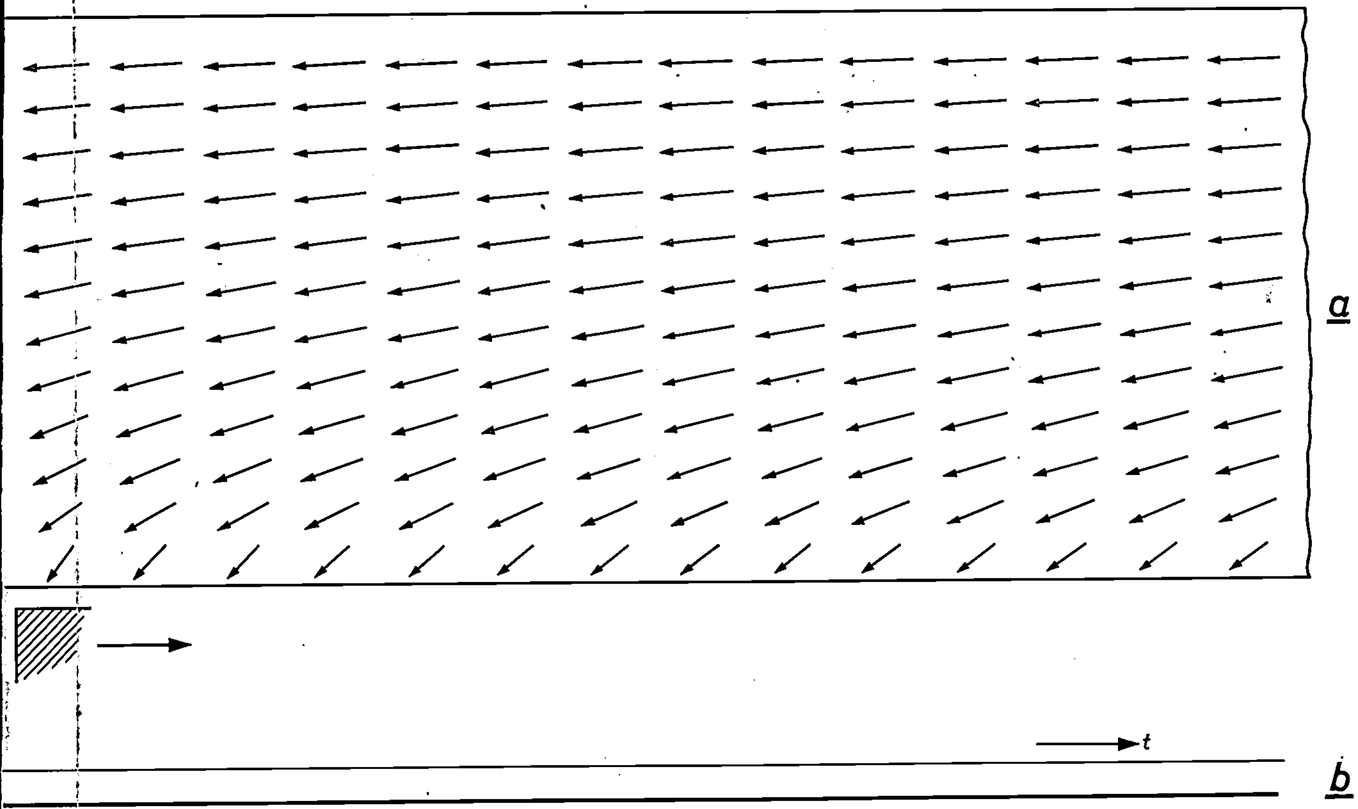
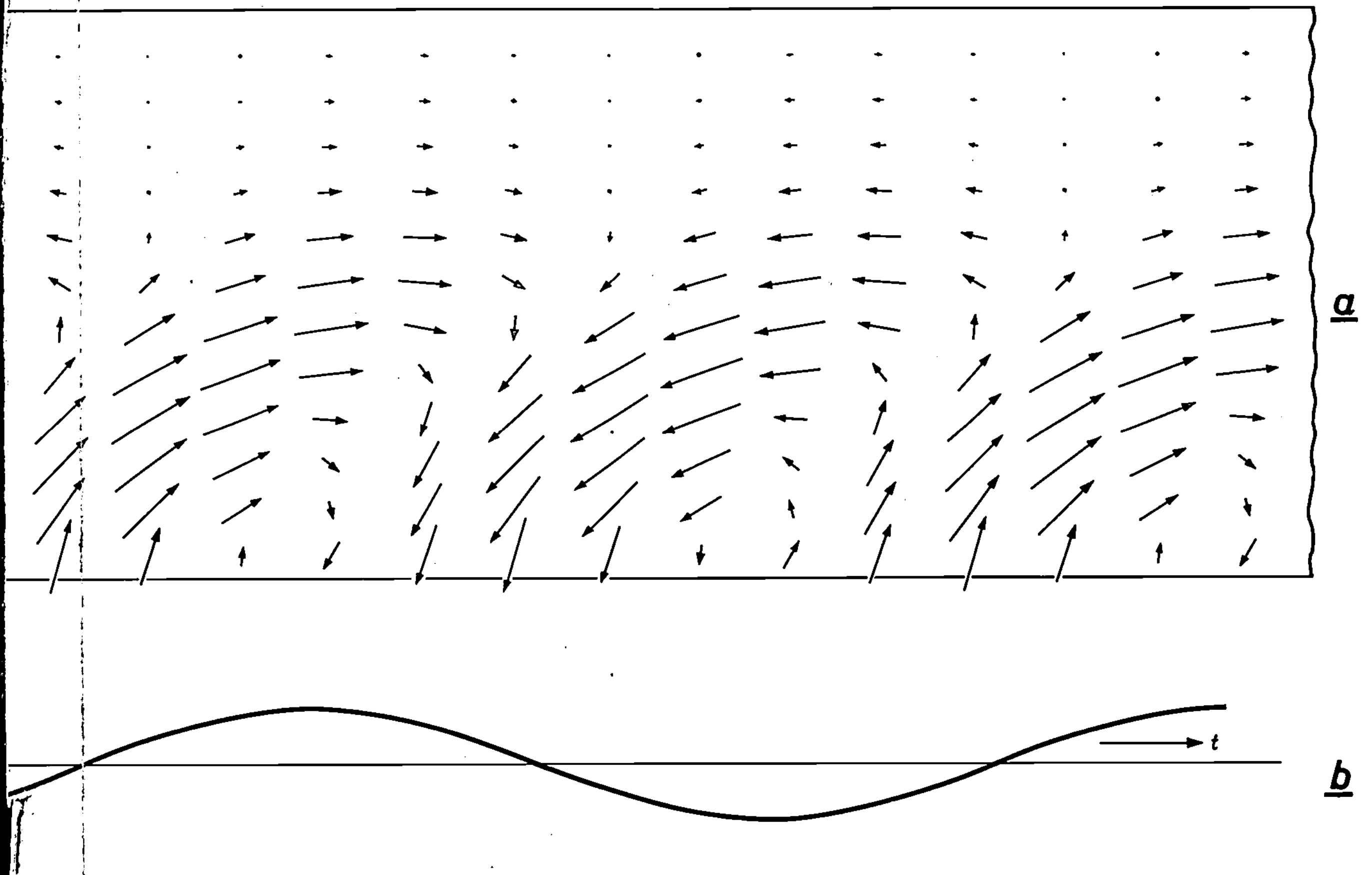


Fig. 13. Representation as in fig. 10 of the state of magnetization of the tape (*a*) but now after the recording of a *sinusoidal* signal (*b*). The scale for the magnetization is here more than 3× larger than in fig. 10. The recording-head gap is drawn at the location (x_0) where it happened to be at one of the moments at which the signal current passed through zero. (For the relation between (*a*) and (*b*), see the caption to fig. 10.)

1:5000


a
t
b


a
t
b

The strip shown in *fig. 10a* represents a section of the vertical plane of symmetry of the magnetic layer. The magnetization, produced by recording a particular signal, has been indicated in this cross-section by drawing to scale at numerous points the magnetization vector with the measured orientation. In this case the signal was a *step function*, obtained by switching a direct current at a certain moment from positive to negative (fig. 10*b*). The recording was made without bias current.

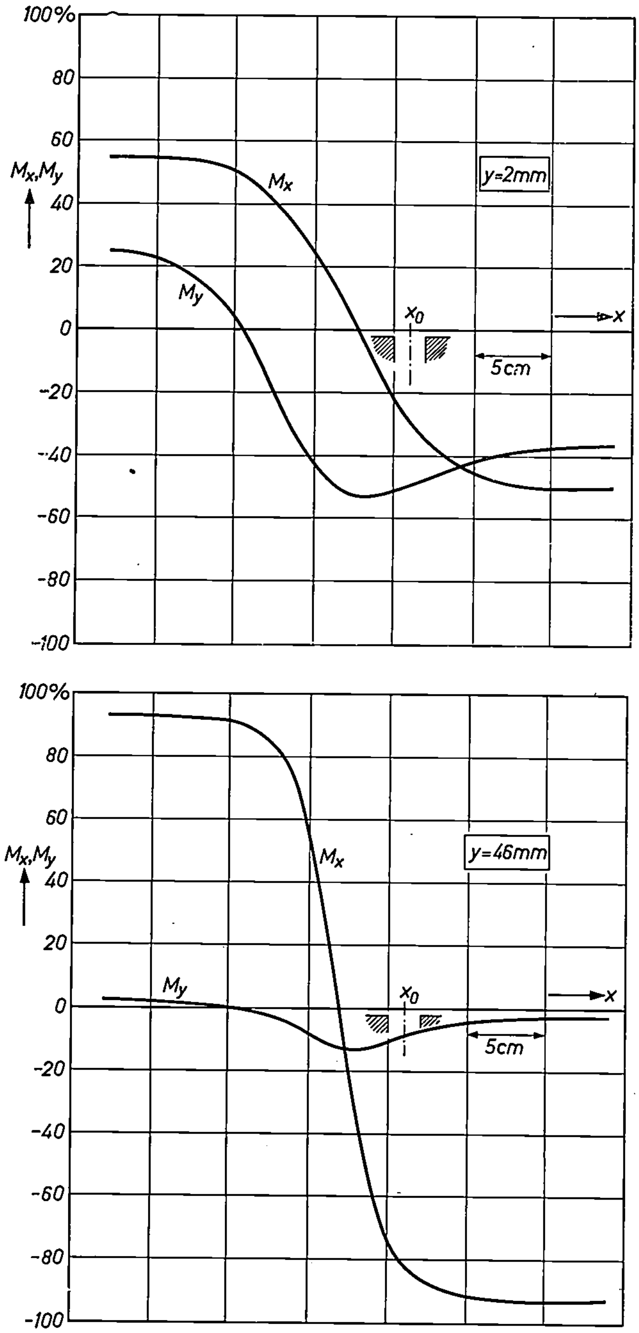


Fig. 11. The components M_x and M_y of the magnetization, measured in the recording represented in fig. 10, are plotted here versus the location x at a given depth y in the tape: in *a*) at a depth y_1 = 2 mm, and *b*) at a depth y_2 = 46 mm. These depths are indicated in fig. 10*a*.

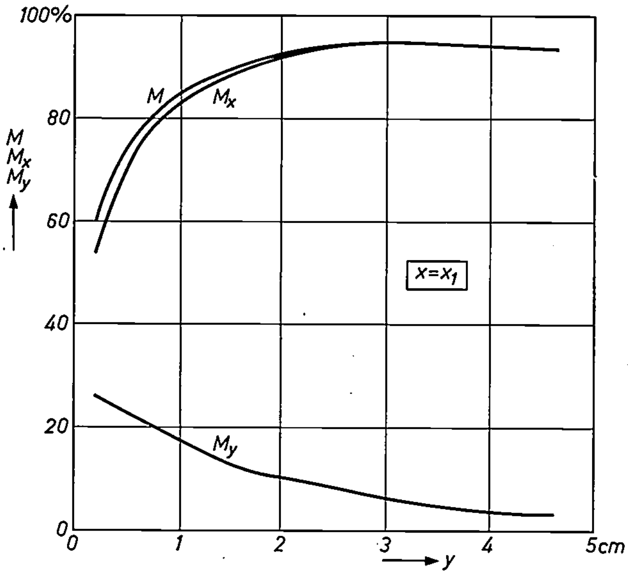


Fig. 12. Magnitude M of the magnetization, and components M_x and M_y, measured in the recording represented in fig. 10, plotted as a function of the depth y in the tape at a given location x_1. The cross-section x_1 is indicated in fig. 10*a*.

Owing to the finite diameter of the samples, the curves begin at y = 2 mm.

To obtain a more quantitative presentation of the measuring results, the longitudinal and vertical components of the magnetization, M_x and M_y, can be plotted as a function of location x at a given depth y in the tape, and also as a function of the depth y at a given location x in the tape. A plot of the former kind is given in *fig. 11* for the results in fig. 10, in respect of two different depths, y_1 = 2 mm and y_2 = 46 mm. Upon reversal of the direction of the current, the component M_x at the depth y_2 is seen to change direction much more abruptly than at the depth y_1; in other words the step function is recorded more sharply deep inside the tape than near the surface. The behaviour of the component M_y contributes to this effect. Remarkably enough, M_y does not reverse simultaneously with M_x but quite a bit "earlier", i.e. at a greater distance x behind the gap, which is moving relative to the tape. As a consequence of this, the step function is less sharply defined at the surface, but deep inside the tape the definition is not affected because there M_y is very small.

In *fig. 12* the values M_x and M_y are plotted in the other manner mentioned, i.e. as a function of the depth y at a given location x on the tape. The location chosen here is x_1 = 21 cm in front of the cross-section (x_0) that was just passing the gap when the current was reversed. The tape material at x_1 knows nothing of the discontinuity about to occur, and has thus "seen" a pure direct current. It follows from the curves — as is also apparent in fig. 10 — that the magnetization near the surface of the tape then makes a fairly large angle with the long axis of the tape. The absolute value M of the magnitization, which is also plotted in fig. 12, is everywhere lower than the remanence that would follow from the given field strength. This is presumably a consequence of the rotation of the field during the recording process. A similar result has recently been found by other investigators working along entirely different lines[1]).

Fig. 13a presents in the same manner as in fig. 10 the magnetic state of the layer after the recording of a sinusoidal signal. The wave-form of this signal is shown in fig. 13*b*. The reasons for the much more complicated picture obtained in this case will not be dealt with here.

In order, however, to make clearer the kind of theoretical problems we hope to bring nearer to a solution with the model described, we shall discuss as a final example the result of recording a simple direct current, this time however together with a bias current. For this purpose the non-oriented tape material was used. In *fig. 14* the magnetization M together with the longitudinal and vertical components M_x and M_y are again plotted as a function of depth y in the layer, for a bias current with an amplitude of 20 A and a direct current of 1 A. The result differs considerably from that found without bias current: up to a certain depth y the longitudinal component M_x now increases more or less proportionately with y. The total magnetization is substantially lower nearer the surface than somewhat deeper in the layer.

From the point of view of playback a state of magnetization of this nature is far more unfavourable than homogeneous magnetization of the whole magnetic layer. It can be calculated that in the latter case the output signal obtained in a long-wave recording would be about 3 dB stronger (assuming that the current used in both recordings is such as to result in the same non-linear distortion from the curved B-H line); at short waves the difference would be even more marked.

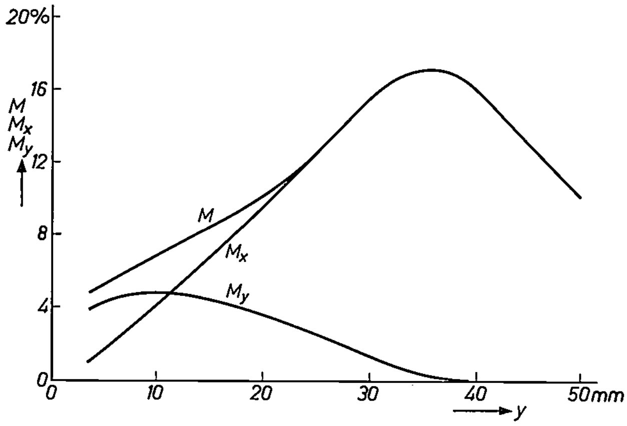


Fig. 14. Measured magnetization M and components M_x and M_y, as a function of the depth y, for a signal consisting of a direct current (1 A) with a superimposed bias current (20 A).

The occurence of a magnetization state as shown in fig. 14 can to some extent be understood as follows [2]). According to the present theory [3]), the actual recording takes place in a ring-shaped zone situated just beyond the centre of the air gap in the recording head; the amplitude of the bias field in this zone is roughly equal to the coercivity of the tape material (see *fig. 15*). This means that every part of the tape emerges from the process with a magnetization M proportional to the signal field-strength to which it was subjected at the instant of passing the "recording zone". If there were no tape present, the field strengths H_x and H_y in this zone would be given to a first approximation by

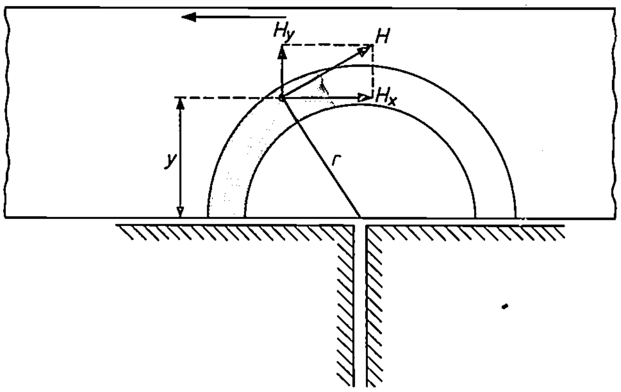


Fig. 15. The actual recording process takes place in a zone around the gap (the "recording zone", shown shaded), where the amplitude of the field H generated by the bias current is roughly equal to the coercivity of the tape material.

$$H_x = \frac{niy}{\pi r^2},$$

$$H_y = -\frac{ni\sqrt{r^2 - y^2}}{\pi r^2},$$

where n is the number of turns, i the signal current and r the radius (partly determined by the bias current) of the recording zone. This field distribution would then cause M_x and M_y to vary as a function of y in the manner shown in *fig. 16*. It can be seen that the curves in fig. 14 agree very nicely with this

[1]) H. S. Templeton and G. Bate, Proc. Intermag. Conf. Washington 1963, p. 7-4.

[2]) D. L. A. Tjaden, Proc. 3rd Int. Congress on Acoustics, Stuttgart 1959, Part II, page 758.

[3]) W. K. Westmijze, Philips Res. Repts. 8, 250, 1953.

picture up to a certain extent (or rather to a certain depth), except that the vertical component M_y is substantially reduced. This effect, which is responsible for the above-mentioned loss in reproduction, is attributable to the demagnetizing effect of the part of the tape which, with a certain magnetization, has already left the recording zone.

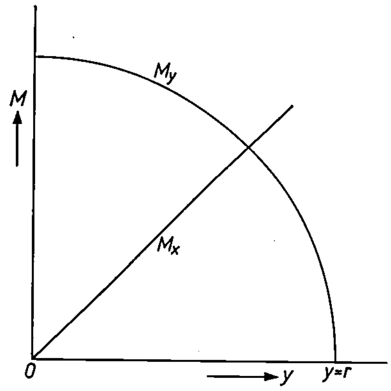


Fig. 16. Theoretical curve of M_x and M_y, derived from fig. 15, as a function of y, for a recording of the same signal as used for fig. 14.

This demagnetizing influence is one of the effects that will be extensively investigated. The considerable size of the model makes it possible to apply other techniques than simply the measurement of samples after a recording. It is intended, for example, to carry out measurements in which, during the recording, the middle strip will carry a number of Hall generators, which will make it possible to observe and record on an oscilloscope the local fields and their fluctuations, during the recording of both DC and AC signals. A report on these investigations will be published in due course.

In conclusion, it should be mentioned that the equipment was devised in cooperation with G. W. van Oosterhout of this laboratory. The construction was largely the work of J. R. P. N. Crüts. For the oriented tape material at present in use we are indebted to M. L. van Splunder of Philips Plastics Laboratory and to the cooperation of the Plastics Research Institute (T.N.O.) of Delft.

Summary. In magnetic recording the magnetic coating of the tape is magnetized over its entire thickness as it passes the recording head. Little is yet known about the manner in which the magnitude and direction of this magnetization vary as a function of depth in the tape. To obtain data on this subject, a model of the tape and recording head has been built in Philips Laboratories at Eindhoven in which all essential dimensions (air gap, tape thickness, recorded wavelength) are scaled up by a factor of 5000. Provided the number of ampere turns of the recording head is scaled up by the same factor — which makes it necessary to build a somewhat unconventional generator — the state of the magnetization produced in the tape (disregarding fluctuation effects) is a true copy of that in normal recording. The "tape" contains an interchangeable middle strip, 5 cm wide and ½ mm thick, from which numerous small samples are punched out after a signal has been recorded. The magnitude and direction of the magnetization of these samples are measured by means of a special magnetometer.

Some results of the measurements hitherto carried out are discussed and their theoretical significance commented upon.

ERRATUM

In the article "New forms of bearing: the gas and the spiral groove bearing", Philips tech. Rev. 25, 1963/64 (No. 10), a sentence from a proof has erroneously remained, viz, the sentence on p. 266, right-hand column, lines 11-8 from the bottom: "This pump effect . . . on the grooves". This sentence should be deleted.

SOME SIMPLE ACTIVE FILTERS FOR LOW FREQUENCIES

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

621.372.542.2

The growing interest shown during recent years in electrical measurements at very low frequencies has led to a demand for special low-frequency filters. Coils, which are commonly used in filters for higher frequencies, have certain disadvantages at low frequencies. In this article the authors describe some circuits in which the use of coils is avoided.

Introduction

In radio engineering and in other branches of electrotechnology it is the practice to give circuits specified frequency-dependent characteristics by using capacitors and coils. These two circuit elements have properties that are to a certain extent opposed to one another. When an alternating voltage is applied, the current flowing in a capacitor leads the voltage in phase, whereas in a coil the current lags in phase. Again, in a capacitor the reactance decreases with rising frequency, whereas in a coil it increases.

Among the passive electric networks composed of combinations of capacitors and coils, the electric filters, which are used for separating signals with certain frequencies from signals with other frequencies, form an important subgroup. For this purpose the use of capacitors *and* coils is not strictly necessary. A filter can be built using only one of these elements in combination with resistors. By combining coils and capacitors, however, it is easy to produce a much sharper separation between wanted and unwanted signals.

One of the simplest and most familiar passive filter networks is the parallel resonant circuit (here referred to as an *LC* circuit) which can be considered as formed from the parallel arrangement of a capacitor, a coil (both loss-free) and a resistor (see *fig. 1*). In such a circuit the modulus of the impedance is maximum at the resonant frequency

$$\omega_0 = 1/\sqrt{LC}\,. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

For any given value ω of the frequency, the impedance can be written in the form:

$$Z(\omega) = \frac{R}{1 + \mathrm{j}\beta Q}, \quad . \quad . \quad . \quad . \quad (2)$$

where β is an abbreviated notation for $\omega/\omega_0 - \omega_0/\omega$. If $|\omega - \omega_0|$ is small compared with ω_0, we can write β in the form:

$$\beta = \frac{2(\omega - \omega_0)}{\omega_0}\,.$$

In this case, then, β is twice the "relative detuning".

In equation (2) Q is the figure of merit, here given by:

$$Q = \frac{R}{\omega_0 L} = \omega_0 CR\,. \quad . \quad . \quad . \quad . \quad (3)$$

The modulus of the impedance Z, plotted as a function of frequency, has the form of the familiar resonance curve (fig. 1). The curve is narrower, i.e. the *LC* circuit more selective, the higher the value of Q. The difference between the two frequencies at which $|Z|$ is a factor of $\sqrt{2}$ smaller than the maximum value is normally referred to as the bandwidth.

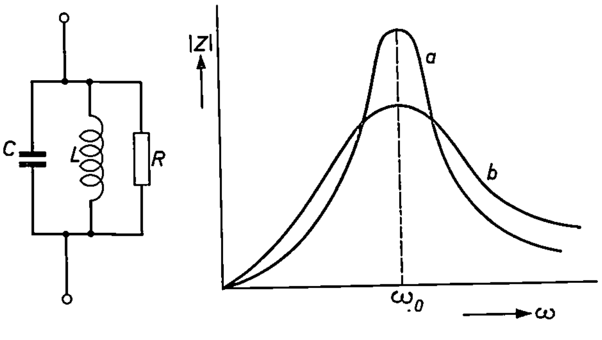


Fig. 1. Left: resonant circuit consisting of a coil, an inductance, a capacitance and a resistance in parallel *LC* circuit. Right: frequency characteristic (resonance curve) for two *LC* circuits. The resonant frequency is ω_0. The network to which curve *a* relates has a higher figure of merit Q.

Filters for very low frequencies

In recent years a growing interest has been shown in electrical measurements at very low frequencies. In the fields of physics and chemistry, for instance,

*) Philips Research Laboratories, Eindhoven.

measurements are frequently carried out with intermittent light; examples are measurements of photoconductivity and of light adsorption in liquids and crystals. The periodic interruption of the light makes it possible to use AC amplifiers for this purpose, which are easier to build and to operate than DC amplifiers. Measurements at very low frequencies are also important in various electro-medical applications, particularly in cardiography, encephalography and myography.

In some of these cases, filters are required which reduce interference to the minimum at the low frequencies involved. Sometimes, too, filters are needed for separating signals of different very low frequencies. This is notably the case in encephalography, where it is occasionally necessary to perform a harmonic analysis of the cerebral voltages, the presence or absence of voltages of certain frequencies having a bearing on the diagnosis of epilepsy. To give some idea of the resolving power required in such a case, it may be mentioned that filters are sometimes needed which have a bandwidth of 1 c/s at a centre frequency of some tens of cycles per second.

To build filters for lower frequencies it is necessary, as can be deduced from eq. (1), to use capacitors of higher capacitance and/or coils of higher inductance. As regards the capacitors, this does not as a rule present insuperable difficulties; the capacitance of a capacitor contained in a given volume can be increased by enlarging the surface area of the plates and by making the dielectric thinner. The construction of coils having a very high inductance involves considerable, and often intractable problems. Coils of this kind always contain a core, which may consist of iron alloy lamellae or of ferroxcube. This makes high-inductance coils relatively bulky and heavy, as also does the larger number of turns required to produce a higher inductance.

Apart from the drawbacks of large bulk and/or weight, there are various other disadvantages attached to the use of high-inductance coils. As a rule they are expensive, because of the large amount of material used, and it takes much more time to produce them. Such a coil, once made, is not very "flexible", that is to say its properties are difficult to modify. In most cases, indeed, it is not practicable to change the inductance value. This, together with the unavoidable variation during manufacture, can make it difficult to produce a coil with an accurately specified inductance. Another drawback is that the properties of coils of this kind tend to vary more markedly with temperature than those of most other circuit elements. Finally, problems can also arise from the necessity of screening the coils against stray magnetic fields. At low frequencies ferromagnetic cans have to be used for this purpose, possibly combined with one or more cans made of a material possessing good electrical conductivity. This again puts up the weight and the price.

The above-mentioned objections to the use of large coils have resulted in the development of circuits which contain no coils, but have the same properties as circuits that do. A familiar example is the feedback amplifier incorporating in the feedback loop a twin-T filter composed of capacitors and resistors [1]. The characteristics of a network of this kind correspond to those of an *LC* circuit. Alignment, however, is somewhat laborious, particularly where several amplifiers tuned to different frequencies are to be connected in cascade.

The object of this article is to draw attention to a type of filter with which the same or even better results can often be achieved, and which is simpler to align than the above-mentioned feedback amplifier. This type of filter has in fact long been known; here, however, after giving a brief outline of its principles, we shall consider some of the less familiar aspects. In particular it will be shown that various measures can be taken to achieve considerable precision with filters of this type and some lesser known applications will be touched upon [2].

Principle of the circuit

Our starting point is a resistance-coupled amplifier stage, as represented by the diagram in *fig. 2.*

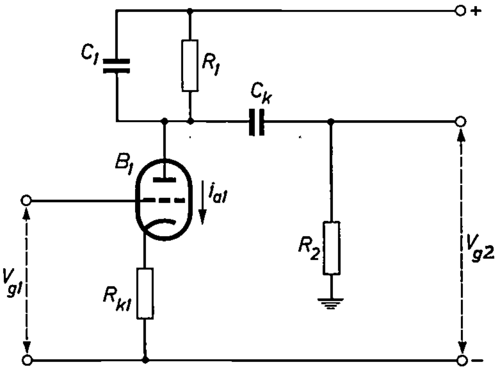


Fig. 2. Amplifying stage with resistance coupling.

[1] A filter of this kind is described by G. Klein and J.J. Zaalberg van Zelst in: A low-frequency oscillator with very low distortion under non-linear loading, Philips tech. Rev. **25**, 22-30, 1963/64 (No. 1), p. 27.

[2] A special application is described by K. Teer, Audibility of phase errors, Philips tech. Rev. **25**, 176-178, 1963/64 (No. 6/7).

Connected in parallel with the resistor R_1 in the anode circuit of the triode B_1 is a capacitor C_1, and the output voltage V_{g2} [3]) of the anode is taken off via a capacitance-resistance coupling C_k-R_2. It can readily be shown that the relation between the anode alternating current i_{a1} of B_1 and the voltage V_{g2} is:

$$\frac{V_{g2}}{i_{a1}} = \frac{-j\omega R_1R_2C_k}{1+j\omega(R_1C_1+R_1C_k+R_2C_k)-\omega^2R_1R_2C_1C_k}. \quad . \; . \; . \quad (4)$$

The same relation between V_{g2} and i_{a1} would have been found if we had incorporated in the anode network of B_1 an LC circuit consisting of a parallel arrangement of a resistance R', an inductance L' and a capacitance C' (see *fig. 3*). These three elements would then have to have the following magnitudes:

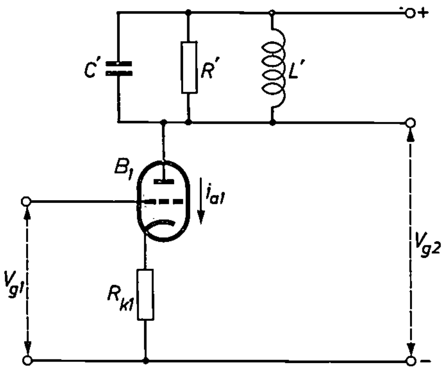


Fig. 3. With appropriate values of L', C' and R' this circuit is equivalent to fig. 2. The figure of merit is in that case very low.

$$R' = \frac{R_1R_2C_k}{R_1C_1+R_1C_k+R_2C_k}, \quad . \; . \quad (5)$$

$$L' = R_1R_2C_k, \quad . \; . \; . \; . \; . \; . \; . \; . \quad (6)$$

$$C' = C_1. \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (7)$$

The resonant frequency of this equivalent circuit would thus be:

$$\omega_0 = \frac{1}{\sqrt{L'C'}} = \frac{1}{\sqrt{R_1R_2C_1C_k}}, \quad . \; . \; . \quad (8)$$

and the figure of merit:

$$Q = \frac{R'}{\omega_0L'} = \frac{\sqrt{R_1R_2C_1C_k}}{R_1C_1+R_1C_k+R_2C_k}. \quad . \quad (9)$$

The quotient of V_{g2} and i_{a1} can then be written in the form:

$$\frac{V_{g2}}{i_{a1}} = \frac{-R'}{1+j\beta Q}. \quad . \; . \; . \; . \quad (10)$$

If the resistance R_{k1} in the cathode lead of B_1 is large compared with the reciprocal of the transconductance, and if, moreover, the amplification factor of B_1 is very high, then $i_{a1} = V_{g1}/R_{k1}$, and eq. (10) becomes:

$$V_{g2} = \frac{-R'}{R_{k1}(1+j\beta Q)}\,V_{g1}. \quad . \; . \quad (11)$$

This analogy with LC circuits is also found with normal amplifying stages. As a rule C_1 is then made up of the anode capacitance of the valve and the wiring capacitance. By way of illustration it will be useful here to insert some conventional values in the expressions (5) to (9), i.e.

$R_1 = 10\,\text{k}\Omega, R_2 = 1\,\text{M}\Omega, C_1 = 10\,\text{pF}$ and $C_k = 0.1\,\mu\text{F}$.

We then find:

$R' = 9.99\ \text{k}\Omega$, $L' = 10^3\,\text{H}$, $C' = 10\,\text{pF}$, $\omega_0 = 10^4$ rad/s and $Q = 10^{-3}$.

The value of Q found here is exceptionally small compared with that of conventional resonant circuits (10 to 100). Further, it is already apparent that the circuit in fig. 2 behaves like an LC circuit containing a coil with a very high value of L. Because of the low value of C_1, however, ω_0 is not particularly small. If we make C_1 and C_k much larger, e.g. both 1 μF, then we find from (5) to (9):

$R' = 9.98\ \text{k}\Omega$, $L' = 10^4$ H, $C' = 1\ \mu$F, $\omega_0 = 10$ rad/s and $Q = 9.8 \times 10^{-2}$.

The resonant frequency is now extremely low, but Q is still very small.

We may now ask whether a higher Q might be obtained by choosing different resistance and capacitance values. This is indeed found to be the case; for Q it is possible to achieve a maximum value of $\frac{1}{2}$, although usually a somewhat lower value is accepted, the reason being that the signals are more strongly attenuated the nearer Q approaches the value $\frac{1}{2}$.

To demonstrate the latter point, we combine (5) and (9) to form the following equation:

$$Q^2 + \frac{R'^2}{R_1R_2} = \frac{\dfrac{R_1(C_1+C_k)}{R_2C_k}}{\left\{\dfrac{R_1(C_1+C_k)}{R_2C_k}+1\right\}^2}. \quad . \; . \quad (12)$$

The right-hand side has an absolute maximum of $\frac{1}{4}$ for

$$\frac{R_1(C_1+C_k)}{R_2C_k} = 1. \quad . \; . \; . \quad (13)$$

[3]) The notation V_{g2} is used here because later in the article this voltage will be applied to the grid of another valve (see e.g. fig. 4).

At a very low value of R', or at very high values of R_1 and/or R_2, Q can therefore approach $\frac{1}{2}$. Now R_1 and R_2 cannot be given unlimitedly high values without affecting the operation of the two valves, these resistances being included in the anode and grid circuits respectively. If Q is to approach $\frac{1}{2}$, therefore, R' will have to be small, which means that the signal transmission by the filter will be poor.

Since, moreover, a Q of $\frac{1}{2}$ is still very low, it would not be possible in this way to design a filter that satisfied practical selectivity requirements, if it were not for the fact that the figure of merit can be considerably increased by the use of feedback. This resembles the effect of feedback in an oscillator. Here too, the effective Q of a circuit is increased by feedback until, at an almost infinitely high Q, oscillation occurs.

Circuit using feedback

Feedback can be obtained by applying to the input of the filter a current which is proportional to the output voltage. A suitable circuit for this purpose is shown in *fig. 4*. The output voltage here is obtained by applying the voltage V_{g2} to the grid of a triode B_2 connected as a cathode follower. The feedback occurs through the resistance R_t shunted between the cathodes. The way the circuit works can be understood in simple terms by assuming that the two triodes function as ideal cathode followers. (To approximate to this ideal, the resistances R_{k1} and R_{k2} should be extremely high, and so too should the amplification factors of the valves.) In this case the cathodes carry alternating voltages V_{k1} and V_{k2} which are equal to the respective alternating grid-voltages V_{g1} and V_{g2}. Calculating V_{g2} ($= V_{k2}$) as a function of V_{g1} ($= V_{k1}$), we then find:

$$V_{g2} = -V_{g1}\frac{R_t + R_{k1}}{R_{k1}(R_t - R')}\,\frac{R'}{1 + j\beta\dfrac{R_t}{R_t - R'}Q}\,. \qquad \ldots \quad (14)$$

The anode current of B_1 equals the sum of the currents in R_{k1} and R_t, hence:

$$i_{a1} = \frac{V_{k1}}{R_{k1}} + \frac{V_{k1} - V_{k2}}{R_t}.$$

Using (10) we write:

$$V_{g2} = \frac{-R'}{1 + j\beta Q}\left(\frac{V_{k1}}{R_{k1}} + \frac{V_{k1} - V_{k2}}{R_t}\right),$$

which leads, after some manipulation, to equation (14).

Since the situation is entirely analogous to that in fig. 2, except that eq. (14) now applies instead of eq. (11), we see that the characteristics of the circuit again correspond to those of an LC circuit, but now with a figure of merit

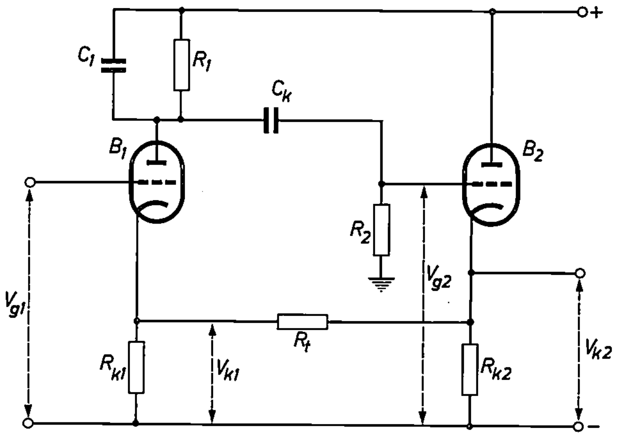


Fig. 4. The figure of merit can be increased by using feedback. In the circuit shown here this is done by introducing the triode B_2 as a cathode follower and a resistance R_t between the cathodes of the two valves.

$$Q' = \frac{R_t}{R_t - R'}Q\,.$$

Thus, by the use of feedback, the figure of merit is increased by a factor $R_t/(R_t - R')$. Raising this factor to 100 or more results in values of Q' comparable with those of LC circuits designed for higher frequencies.

The resonant frequency is not altered by the application of the feedback; the inductance L', however, is increased by a factor $(R_t + R_{k1})/R_t$ and the capacitance C' reduced by the same factor. Instead of the resistance R' from (11) we now have in (14) the resistance $R'(R_t + R_{k1})/(R_t - R')$. The use of feedback thus increases the signal transmission by a factor $(R_t + R_{k1})/(R_t - R')$.

Circuit design; stability

Deciding on the optimum circuit values for a filter under given conditions is usually a complicated problem owing to the large number of variables involved. Moreover the choice of the values of resistances and capacitances is restricted by various practical considerations.

We have seen that the high value of Q required for good selectivity can be achieved by means of feedback. The extent to which this can be done, however, is limited by the fact that a filter is always required to possess a certain *stability*, that is to say its characteristics are required to have a certain constancy under variations that may occur in the components used. In this respect the valves are the greatest danger. If a highly stable circuit is required, they set a limit to the extent to which feedback can be employed.

This can be understood by considering that equation (14), applicable to fig. 4, represents only an approximation since it takes no account of the fact that a cathode follower has an internal resistance which can be said to be roughly equivalent to the reciprocal of the transconductance of the relevant valve. If we take this into account, we find that instead of the resistance R_t shunted between the cathodes of the two valves, we should introduce in the equations a resistance $R_t' = R_t + 1/S_1 + 1/S_2$, where S_1 and S_2 are the transconductances of B_1 and B_2 respectively. Now, for a stable circuit with strong feedback it is most important that the value of R_t' should be highly constant. For if the factor $R_t'/(R_t' - R')$, by which the figure of merit is multiplied, is to be made sufficiently large, R_t' should not differ significantly from R'. A small percentage change in R_t' therefore has a considerable influence on the factor in question, and therefore also on the figure of merit ultimately obtained. Slight variations in the transconductance of the valves can thus cause substantial variations in the Q of the filter.

To minimize this effect it is necessary to use for R_t a highly stable resistance which is large compared with $1/S_1 + 1/S_2$. This means that the circuit should be designed so as to give R' a high value also, and this, according to (5), can only be done by using large values for R_1 and R_2. We have already noted that there are limits to the values we can give to these resistances, and this in fact means a limitation of the degree of feedback that can be adopted when rigorous stability requirements are imposed. In most cases the circuit will therefore be so designed as to obtain the highest possible figure of merit even without feedback, i.e. a Q as close to $\frac{1}{2}$ as a reasonable signal transmission permits. If we choose R_1, R_2, C_1 and C_k so as to satisfy (13), then in accordance with (9) and (5) we have:

$$Q = \tfrac{1}{2}\sqrt{\frac{C_1}{C_1 + C_k}} \quad \text{and} \quad R' = \tfrac{1}{2}R_1 .$$

If, for example, we now take $C_k = \frac{1}{4}C_1$ (to satisfy (13) we must then make $R_2 = 5R_1$), then $Q = 1/\sqrt{5} \approx 0.45$. Since the maximum value of R' is equal to R_1 — i.e. when R_2C_k is large compared with $R_1(C_1 + C_k)$ — the signal transmission at $C_k = \frac{1}{4}C_1$ and $R_2 = 5R_1$ is still 50% of that of the maximum that can be achieved at a given value of R_1. This can be regarded as a reasonable compromise.

In some cases, to ensure that the filter is satisfactorily stable in operation, R_1 and R_2 will have to be higher than is desirable from the point of view of the valves. When the relevant requirements are not too divergent, both can sometimes be met by connecting the anode of B_1 and/or the grid of B_2 to tappings on R_1 and R_2 respectively (see *fig. 5*). It can easily be seen that the filter action is now governed by the total values of R_1 and R_2, while only parts of these resistances are incorporated in the anode circuit of B_1 and the grid circuit of B_2.

Two other limitations of the magnitude of the resistances may be mentioned, the first being the fact that very high resistances possessing high stability (metal film types) are difficult to produce. The other point is that parallel with R_2 is the input capacitance of the valve B_2. If R_2 is high or the frequency for which the relevant filter was designed is not very low (e.g. a few hundred c/s) it may well be that the impedance of the input capacitance of B_2 is not to be disregarded, and since a capacitance of this kind is always a rather unstable quantity, it can have an adverse effect on the stability of the filter. The effect is more pronounced the higher the value of R_2, and this too sets a limit to the permissible value of R_2. A means of reducing the influence of the unstable input capacitance of B_2 is to connect a sufficiently large and stable capacitor in parallel with R_2.

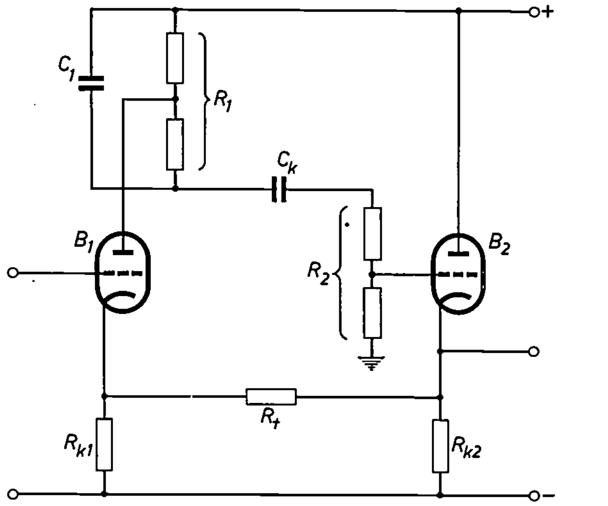


Fig. 5. The resistances in the anode circuit of B_1 and in the grid circuit of B_2 should sometimes be lower than the resistances R_1 and R_2 that govern the filter action. It may then be an improvement to connect the anode of B_1 and the grid of B_2 to tappings on R_1 and R_2.

In this case the filter proper is formed by two resistors and three capacitors (*fig. 6*). From the equation giving the relation between V_{g2} and i_{a1} we find that for this circuit also we can draw an equivalent circuit as represented in fig. 3. Without going any further into this subject, it will be useful to give the equations which then take the place of (5) to (9):

$$R' = \frac{R_1R_2C_k}{R_1(C_1 + C_k) + R_2(C_2 + C_k)}, \quad \ldots \quad (5')$$

$$L' = R_1R_2C_k, \quad \ldots \quad (6')$$

$$C' = \frac{1}{C_k}(C_1C_2 + C_1C_k + C_2C_k), \quad \ldots \quad (7')$$

$$\omega_0 = \frac{1}{\sqrt{R_1R_2(C_1C_2 + C_1C_k + C_2C_k)}}, \quad \ldots \quad (8')$$

$$Q = \frac{\sqrt{R_1R_2(C_1C_2 + C_1C_k + C_2C_k)}}{R_1(C_1 + C_k) + R_2(C_2 + C_k)}. \quad \ldots \quad (9')$$

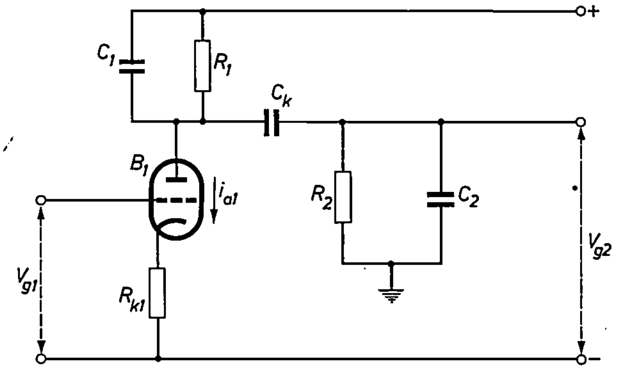


Fig. 6. When the frequency of the signals is not very low and /o the resistance R_2 is very high, the capacitance C_2 may have to be taken into account.

In this case too, as may be deduced from these equations, the maximum value of Q is equal to $\frac{1}{2}$. In this respect, then, the introduction of C_2 makes no difference.

Circuits of high stability

As has been shown, the principal objection to high stability of a filter designed along the lines described above, arises from the variations that may occur in the transconductance of the valves. Substantial improvements can be achieved in this respect by employing circuits in which the feedback is brought about in a different manner. For example, a marked improvement can be obtained by using for the feedback a separate valve, on the principle illustrated in the diagram in *fig. 7*. Valves B_1 and B_2 have the same functions here as in fig. 4. The anode circuit of B_2 now contains a resistance R_{a2}. The alternating anode voltage of B_2 is applied to the grid of B_3, whose anode is connected to the anode of B_1. (The DC biasing of the valves will not be dealt with here.) When the resistances R_{k2} and R_{k3} are very large compared with the reciprocal of the transconductance of the relevant valves, the alternating anode current of B_3 is:

$$i_{a3} = -V_{g2}\,\frac{R_{a2}}{R_{k2}\,R_{k3}}\,.$$

This current, together with the alternating current from the anode of B_1, is fed to the input of the actual filter. Here again, then, the strength of the feedback is determined primarily by a number of resistances, i.e. R_{a2}, R_{k2} and R_{k3}. The smaller the reciprocal values of the transconductances of B_2 and B_3 compared with R_{k2} and R_{k3}, the more accurately is this approximation satisfied, and since the latter resistances can readily be made higher than R_t in fig. 4, variations in the transconductance of the valves can easily be made to have less effect on the strength of the feedback in the circuit of fig. 7 than in that of fig. 4; the Q of the circuit in fig. 7 can therefore be more stable.

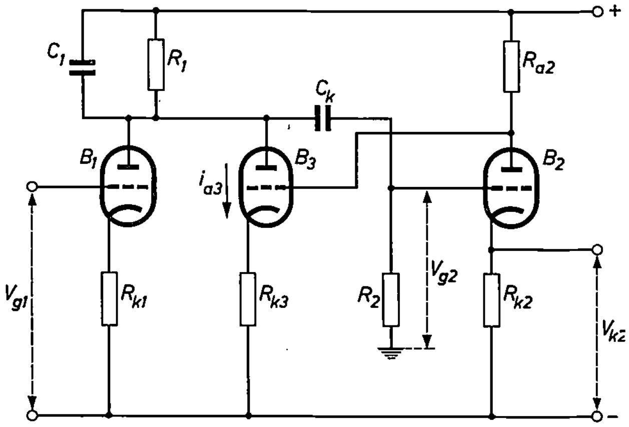


Fig. 7. The stability of the filter can be improved by using a separate valve B_3 for the feedback.

An even greater improvement can be made by using yet another valve, denoted B_4 in *fig. 8*. The operation of this circuit can again be understood by assuming that B_2 functions as an ideal cathode follower, and B_3 likewise. The alternating voltages on both the grid and cathode of B_3 are then equal to $-V_{g2}R_{a2}/R_{k2}$. If we now choose the values of R_2 and R_3 such that $R_2 : R_3 = R_{k2} : R_{a2}$, then the point

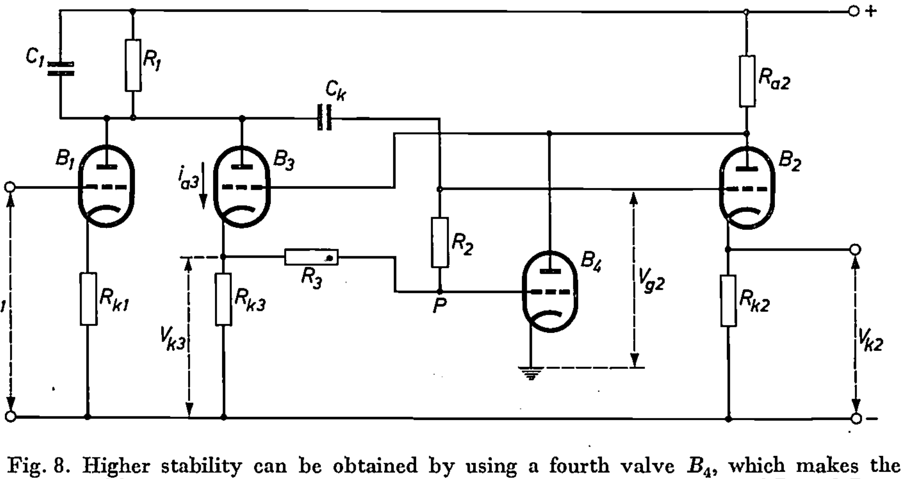


Fig. 8. Higher stability can be obtained by using a fourth valve B_4, which makes the amount of feedback less dependent on variations in the transconductances of B_2 and B_3.

P will carry no alternating voltage with respect to earth, and therefore no alternating anode-current will flow in B_4. The operation of the circuit in fig. 8 is then identical with that in fig. 7. (The only difference is that fig. 7 contains the resistor R_{k3} for alternating current in the cathode lead of B_3, whereas fig. 8 contains R_{k3} and R_3 in parallel.) Any change in this situation produces an alternating voltage at point P. This gives rise in B_4 to an alternating current which has a corrective effect, owing to the anode of B_4 being connected with that of B_2. In this way, then, the voltages and currents always adjust themselves so that there is virtually no alternating voltage at point P, and it can easily be seen that in this case the ratio of the alternating current i_{a3} and the alternating voltage V_{g2} is governed solely by the resistances R_2, R_3 and R_{k3}. Owing to the corrective effect of B_4, variations in the transconductance of the valves have no effect on this ratio, and therefore the feedback is extremely stable, making it possible to give the figure of merit of the filter an exceptionally high and stable value.

Obtaining a flatter frequency characteristic

As we have shown, the frequency characteristic of a filter as drawn in figs. 4 to 8 corresponds to that of an LC circuit. In many cases a characteristic will be required that is flatter at the top and has steeper sides, as can be obtained for example with the aid of two or more non-coupled LC circuits whose resonant frequencies do not coincide (staggered tuning). A frequency characteristic having an exceptionally flat and broad top can be obtained by this method, using three tuned circuits, the figure of merit of one of them being half that of the two others, and the latter two being tuned to specific frequencies which are higher and lower than the resonant frequency of the first one. This principle can also be applied by using, instead of LC circuits, active filters on the principle described above; the alignment of the networks, however, is rather laborious as it is with passive filters.

Another method of obtaining a flat frequency characteristic with passive LC circuits is to couple circuits which are all tuned to the same frequency. This principle can also be applied to the active filters described here, the "coupling" being achieved by introducing negative feedback between pairs of filters.

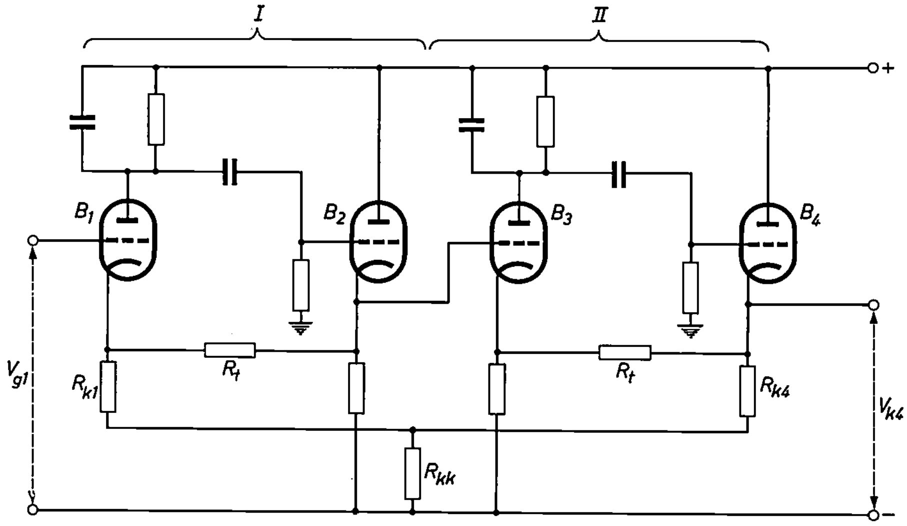


Fig. 9. Combination of two active filters, I and II, with negative feedback obtained by means of R_{kk}. The frequency characteristic achieved with this circuit corresponds to that of two coupled LC circuits.

An example of such a circuit is shown in *fig. 9*. The two filters are denoted by I and II. The negative feedback takes place via a common resistance R_{kk} introduced in the cathode lead of the first valve (B_1) in the first filter and of the second valve (B_4) in the second filter [4]).

Making some simplifying assumptions, i.e. by treating B_2 and B_4 as ideal cathode followers and assuming that R_{kk} is small compared with R_{k1} and R_{k4}, it is

[4]) Instead of the resistances R_{k1}, R_{k4} and R_{kk}, a delta connection of resistances might have been used. A drawback, however, is the very high value that would then be needed for the feedback resistor, which would be difficult to make with sufficient stability. In this respect the small resistance R_{kk} presents no difficulties.

easily computed that the relation between the output voltage V_{k4} and the input voltage V_{g1} is given by the expression:

$$V_{k4} = \frac{A V_{g1}}{(1+j\beta Q') + A p}. \quad . \quad . \quad (15)$$

Here A is the gain of the circuit without negative feedback at the resonant frequency, i.e. at $R_{kk} = 0$, and $\beta = 0$, while

$$p = \frac{R_{kk}}{R_{k4}} \frac{R_t}{R_{k1}+R_t}.$$

This quantity is thus a measure of the strength of the feedback. If we now take $Ap = 1$, say, then (15) becomes:

$$V_{k4} = \frac{A V_{g1}}{(1+j\beta Q')^2 + 1}, \quad . \quad . \quad . \quad (16)$$

and the relation between the moduli of V_{k4} and V_{g1} is then given by

$$|V_{k4}| = \frac{A}{\sqrt{4 + \beta^4 Q'^4}}\ |V_{g1}|. \quad . \quad (17)$$

The behaviour of $|V_{k4}|/|V_{g1}|$ as a function of β is in this case identical with that in a band-pass filter consisting of two critically coupled LC circuits. The characteristic is flatter and has steeper sides than that belonging to a single *LC* circuit.

By using combinations of such filters with different "couplings" (Ap) it is possible to meet widely varying requirements with regard to the frequency characteristic. A filter as in fig. 4 can, for example, be connected in cascade with a filter as in fig. 9; with suitable dimensioning, a frequency characteristic can then be obtained corresponding to that of three *LC* circuits with staggered tuning. The conformity of the equations makes it possible to treat the design of such configurations mathematically in the same way as the design of filters using *LC* circuits.

Filters *I* and *II* in fig. 9 are circuited on the principle represented in fig. 4. Obviously, to achieve greater stability one can also use the circuits of fig. 7 or fig. 8 as "building bricks".

Low-pass filters

The filters described above are of the band-pass type, that is to say the frequency band which is transmitted by the filter is limited on both the low and high frequency sides. For certain experiments, however, *low-pass* filters are wanted, i.e. filters which pass all signals components with frequencies below a certain critical frequency. A filter of this type is approximated in the first instance by a single resistor and a single capacitor (see *fig. 10*). One can think of a low-pass filter as being derived from the band-pass type by making the capacitance C_k infinitely large, i.e. by short-circuiting it, thus making the resonant frequency zero. (The resistors R_1 and R_2 in fig. 2 are now in parallel for alternating current and in fig. 10 are replaced by a single resistance R.)

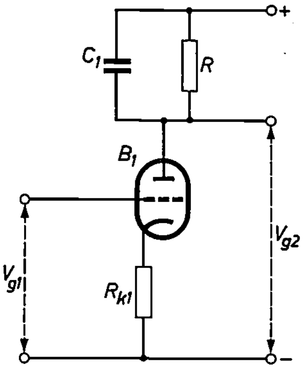


Fig. 10. AC circuit for a low-pass filter.

Introducing feedback here in the same way as represented in fig. 4, we obtain the diagram shown in *fig. 11*. The relation between input and output signal is now:

$$V_{k2} = V_{g2} = -V_{g1}\frac{R_t+R_{k1}}{R_t-R}\,\frac{R}{R_{k1}\left(1+j\omega\dfrac{R_t}{R_t-R}C_1R\right)}. \quad . \quad . \quad . \quad (18)$$

It can be seen from this that the feedback has the same effect on the frequency characteristic as increasing C_1 by a factor $R_t/(R_t-R)$. For limiting the pass-band to very low frequencies it will be easier to use a high value of C_1 than to introduce feedback, and the latter will therefore seldom be used for a simple low-pass filter.

The situation is different, however, if here too use is made of the possibility of connecting two stages in cascade and introducing negative feedback. An

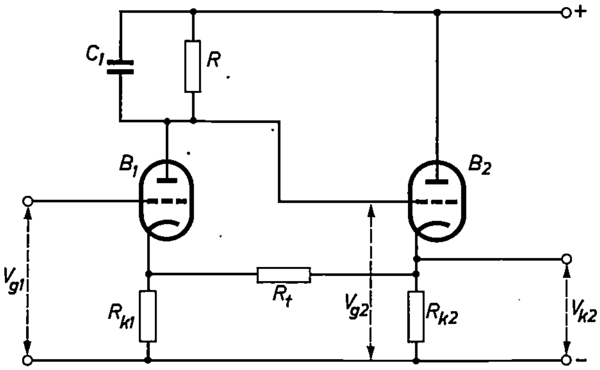


Fig. 11. Low-pass filter with feedback. The effect of the feedback on the frequency characteristic is the same as that of an increase in the capacitance C_1.

arrangement of this kind is shown in *fig. 12*, using two identical filters and negative feedback via the resistor R_{tk} [5]. Assuming again that the resistances in the cathode leads are very high, we can derive the following expression for the relation between the output voltage V_{k3} and the input voltage V_{g1}:

$$V_{\mathrm{k3}} = \frac{A(1+p)V_{\mathrm{g1}}}{(1+\mathrm{j}\omega C_1R)^2 + Ap}. \quad . \quad . \quad (19)$$

Here $A = R^2/(R_{\mathrm{k1}}R_{\mathrm{k2}})$ and $p = R_{\mathrm{k1}}/R_{\mathrm{tk}}$. Since we have assumed that R_{k1}, R_{k2} and R_{k3} are very large,

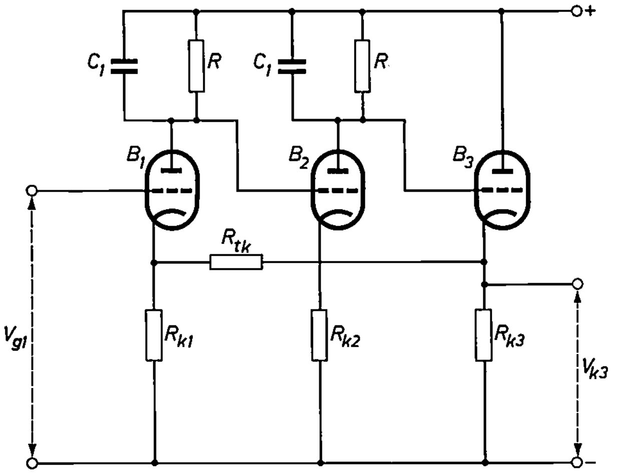


Fig. 12. Low-pass filter consisting of the cascade arrangement of two identical filters as in fig.10, negative feedback being introduced by means of the resistance R_{tk}.

A is the gain of the circuit at zero frequency and with the negative feedback out of operation (i.e. at $R_{\mathrm{tk}} = \infty$). Further, p is a measure of the strength of the negative feedback.

From (19) it follows that for $\omega = 0$ the output voltage is:

$$V_{\mathrm{k3\cdot 0}} = \frac{A(1+p)V_{\mathrm{g1}}}{1+Ap}. \quad . \quad . \quad . \quad (20)$$

In fig. 13 the ratio

$$\frac{|V_{\mathrm{k3}}|}{|V_{\mathrm{k3\cdot 0}}|} = \left| \frac{1+Ap}{(1+\mathrm{j}\omega C_1R)^2 + Ap} \right| \quad . \quad (21)$$

is plotted as a function of ω for various values of $A\,p$,

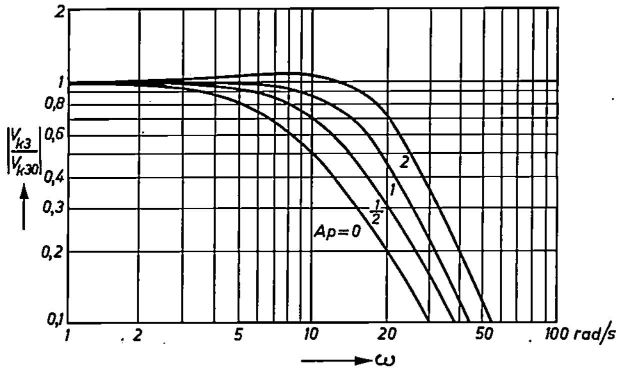


Fig. 13. Frequency characteristics for a low-pass filter as in fig. 12.

keeping the product C_1R constant for all curves (0.1 s). The curve drawn for $Ap = 0$ applies to the case without negative feedback. Comparison with the other curves shows that the use of feedback can lead to a much flatter frequency characteristic, which cannot be obtained with passive *RC* circuits.

With low-pass filters too, it is possible to build more extensive networks by combining filters having different C_1R values. By suitable dimensioning, curves can be obtained which are even flatter at the top and have a steeper descending portion.

Application as a delay network

The analogy existing between *LC* circuits and the active filters described above makes it possible to use these active filters in cases where it is required to delay signals of extremely low frequency without causing any distortion. This can be useful, for example, when building an electrical analogue of a control system involving transmission lags. An electrical analogue of this type can be used for analyzing the stability conditions of the control system.

Where a network is required to transmit each signal undistorted but with a certain retardation, the phase delay must be independent of the frequency, and consequently the phase shift suffered by the various components must be proportional to their frequencies. Furthermore, the amplitude ratio between input and output must be independent of the frequency.

A familiar circuit with which this can be realized in a limited phase-shift region is shown in *fig. 14a*. A less familiar circuit with which the same results can be achieved, and which offers advantages in connection with the following considerations, is given in fig. 14*b*. Assuming for simplicity that R_1 and R_2 are large with respect to R and R_{k}, and choosing the resistances so as to satisfy the equation:

$$\frac{R_1}{R_2} = \frac{2R_{\mathrm{k}}}{R}, \quad . \quad . \quad . \quad . \quad . \quad (22)$$

[5] Since neither of the two individual filters is provided with feedback, this circuit can be designed with only three valves, whereas in fig. 9 four valves were needed. Furthermore, unlike fig. 9, the negative feedback here is obtained with a delta connection of the resistances R_{k1}, R_{k3} and R_{tk}. The reason for this is that, owing to the absence of feedback in the individual filters, the total gain in fig. 12 is much smaller than in fig. 9. To obtain sufficient effect from the feedback, a common resistance in series with R_{k1} and R_{k3} would therefore have to be much larger than R_{kk} in fig. 9. In this case the introduction of a resistance between the cathodes of B_1 and B_3 is a better solution, because such a resistance has less influence on the DC biasing of the two valves.

then the relation between V_g and V_u is given by:

$$\frac{V_g}{V_u} = \frac{R_1 + R_2}{R_2} \frac{1 + j\omega C_1 R}{-1 + j\omega C_1 R}. \quad . \quad (23)$$

The modulus of this expression is indeed independent of ω, and the argument is $2 \tan^{-1}\omega C_1 R \pm 180°$. The voltage V_u thus lags in phase behind V_g by an angle $2 \tan^{-1}\omega C_1 R \pm 180°$. We shall henceforth consider only the frequency-dependent term $\varphi = 2 \tan^{-1} \omega C_1 R$ of this phase shift, since the constant quantity $\pm 180°$ can be compensated, e.g. by using two stages in cascade. In *fig. 15* curve *a* gives the variation of φ as a function of ω. At small values of $\omega C_1 R$, then, $\tan^{-1}\omega C_1 R \approx \omega C_1 R$, and therefore φ is approximately proportional to ω. This applies, however, only up to a limited value of φ (e.g. 60°). The proportionality region can be extended by connecting in cascade with a circuit as in fig. 14 one or more elements whose phase characteristics have linear portions at frequency values differing from zero. This can be done with a circuit as shown in *fig. 16*, for example, which differs from that in fig. 14*b* in that the anode circuit of the valve contains an *LC* circuit instead of the *CR* combination. In this configuration the following relation exists between V_g and V_u:

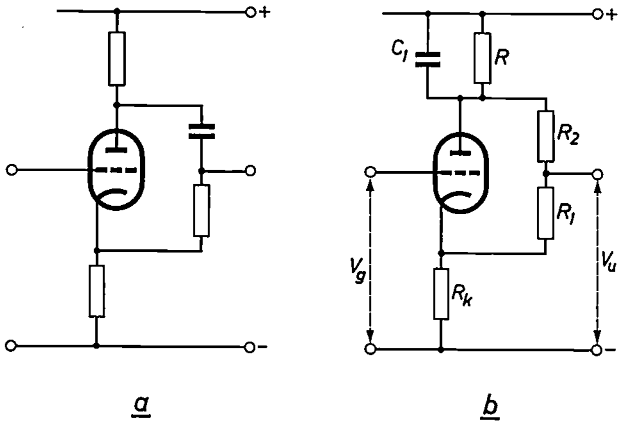


Fig. 14. Circuits in which the signals have a constant phase lag within a limited frequency range, and a constant ratio between the amplitudes of input and output signal.

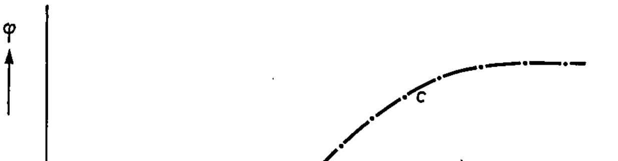


Fig. 15. Phase characteristic of the circuits in fig. 14 (*a*) and of the circuits in figs. 16 and 17 (*b*). Connected in cascade and suitably dimensioned, these circuits give the phase characteristic *c*.

$$\frac{V_g}{V_u} = \frac{R_1 + R_2}{R_2} \frac{1 + j\beta Q}{-1 + j\beta Q}. \quad . \quad (24)$$

(Here β again represents $\omega/\omega_0 - \omega_0/\omega$.) The voltage V_u now lags in phase behind V_g by an angle $2 \tan^{-1} \beta Q \pm 180°$. Fig. 15 gives a plot of $\varphi = 2 \tan^{-1} \beta Q + 180°$ as a function of ω (curve *b*). Using circuits as in fig. 14 and fig. 16 connected in cascade, a phase shift is obtained between input and output voltage which can be found by adding the ordinates of curves *a* and *b* (curve *c*). Choosing suitable circuit values we can now obtain a linear variation over a much wider region than is possible using a circuit as in fig. 14 alone.

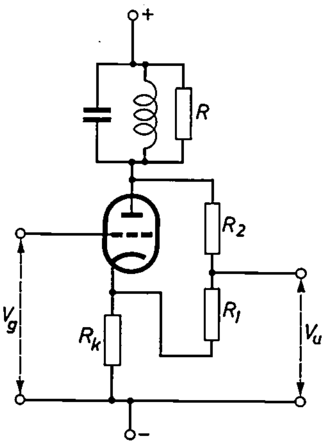


Fig. 16. Circuit giving a phase characteristic that has a linear portion round a frequency differ ng from the value zero.

The principle described can also be applied at very low frequencies when a filter as in fig. 4 is to be used instead of an *LC* circuit. The voltage divider between input and output can be produced by using two resistors R_{t1} and R_{t2} instead of the feedback resistor R_t in the circuit shown in *fig. 17*. The ratio of these resistances should now satisfy the expression:

$$\frac{R_{t1}}{R_{t2}} = \frac{2R_{k1}}{R'}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (25)$$

where R' is given by (5). By means of a cascade arrangement of the circuits of fig. 14 and fig. 17, we can thus obtain a phase characteristic of the form represented by curve *c* in fig. 15.

The "proportionality region" can be still further extended by adding to the circuit one or more filters as in fig. 17, having different resonant frequencies. If these resonant frequencies and the figures of merit are properly chosen, a phase characteristic can be obtained which is almost linear in a wide range of frequencies from zero upwards. This means that signals that possess components lying only in this region have a constant phase delay and therefore undergo a certain delay without any distortion. The figures of merit of the various circuits connected in cascade should, to a good approximation, be virtually proportional to the resonant frequencies. This implies that some of these filters require a fairly high Q, which, within certain limits, can be achieved by means of feedback.

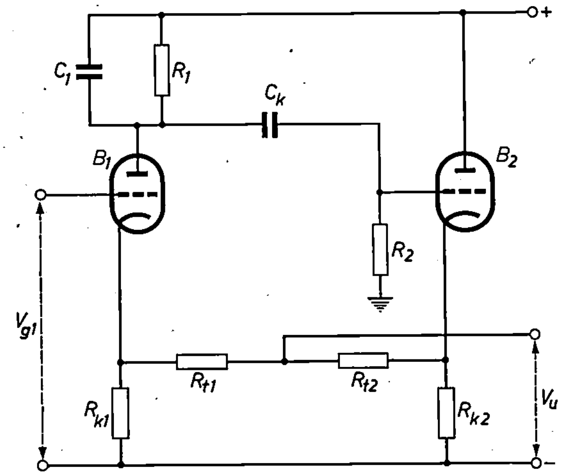


Fig. 17. Circuit which, suitably dimensioned, can be made identical at low frequencies with a circuit as shown in fig. 16.

Final considerations

In the foregoing we have referred to the analogy existing between certain filters composed of resistances and capacitances, and filters using one or more LC circuits. In conclusion we should point to a *difference* between these two circuits that can be of importance in practical applications, namely the fact that the feedback employed to achieve a reasonably high figure of merit causes an increase in the *noise level.* Because of this fact, filters based on the principle described in this article can only be used when the signal level is sufficiently high, higher than that at which filters with LC circuits can be employed.

Finally, it may be mentioned that circuits on the principles described here can also be designed with transistors. In this case, however, the input resistance is much smaller than with valves, which restricts the choice of the circuit elements that can be used.

Summary. Certain active electric networks composed of amplifying valves, resistors and capacitors have characteristics that correspond to those of passive networks containing inductors. At very low frequencies, where coils have certain disadvantages, good use can be made of active networks of this kind. The authors discuss the design of band-pass and low-pass filters, using feedback to achieve a reasonably high figure of merit. They also deal with various circuits for giving these filters exceptionally high stability.

By connecting two filters in cascade and using negative feedback, a frequency characteristic corresponding to that of two coupled LC circuits can be achieved. Finally, reference is made to the possibility of designing networks in this way whose amplitude characteristic is flat within a given frequency range and whose phase characteristic is practically linear, so that signals only containing components in this frequency range can pass through these networks with a certain delay without suffering distortion.

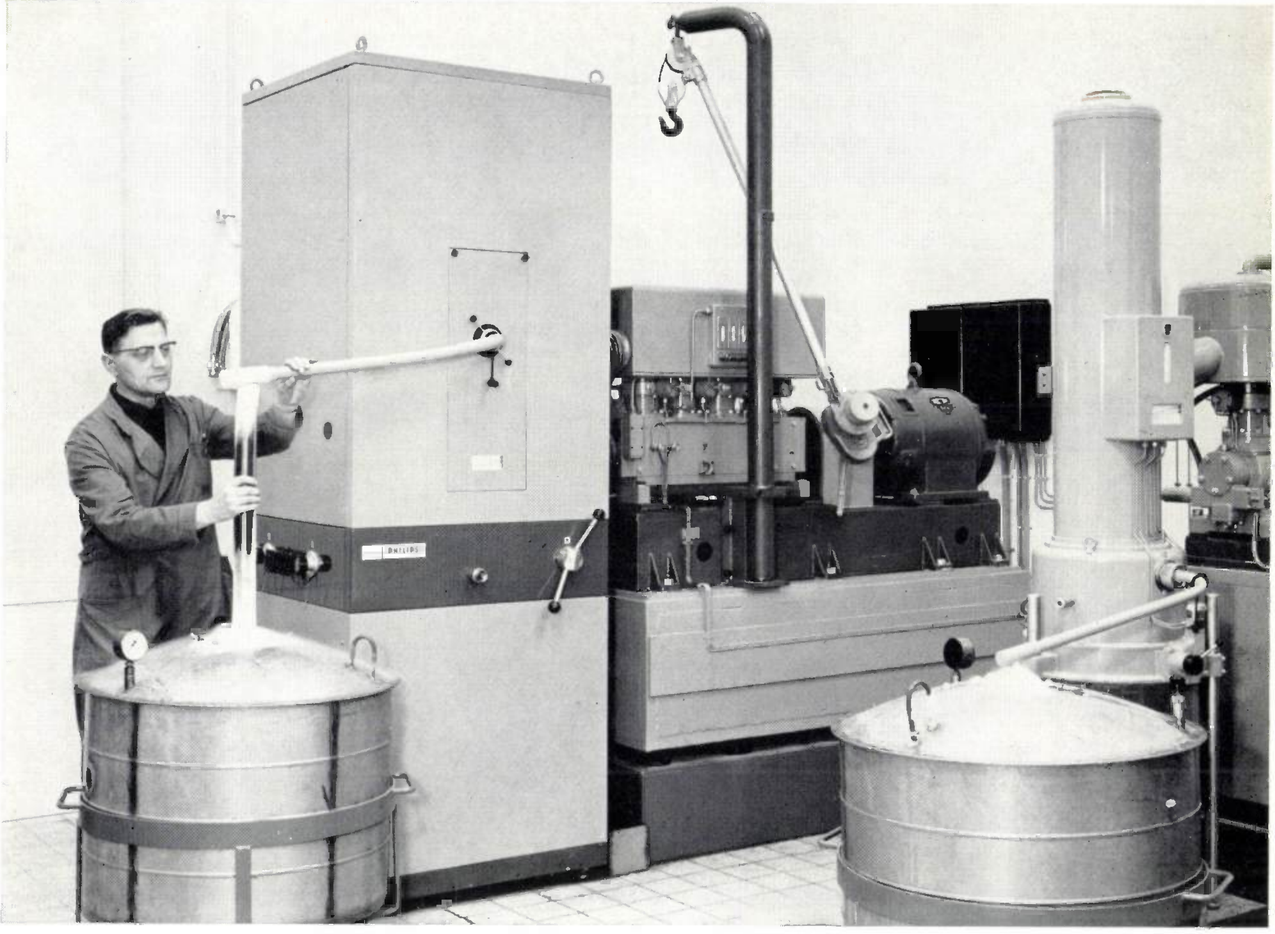
PRODUCTION CENTRE FOR LIQUID NITROGEN

In the last ten years many laboratories have been equipped with a Philips gas refrigerating machine, enabling them to supply their own liquid air requirements. Since in many cases there is a preference for pure liquid nitrogen, many of these machines are used in combination with an air fractionating column, also developed by Philips, which is adapted to the gas refrigerating machine. In our own laboratories — as elsewhere — it is found that the ease with which the liquid air or nitrogen can be obtained greatly stimulates its use. The small installation earlier described [1] [2], which in its present form delivers 6.5 litres of liquid nitrogen an hour, often turns out after some time to be too small. With this in mind, a gas refrigerating machine has been developed which combines four of the original single-cylinder machines. A larger air fractionating column adapted to the large machine is also available.

The photograph on p. 341 was taken in the liquid air

[1] J. W. L. Köhler and C. O. Jonkers, Fundamentals of the gas refrigerating machine, and Construction of a gas refrigerating machine, Philips tech. Rev. **16**, 69-78 and 105-115, 1954/55.

[2] J. van der Ster and J. W. L. Köhler, A small air fractionating column used with a gas refrigerating machine for producing liquid nitrogen, Philips tech. Rev. **20**, 177-187, 1958/59.



production centre in the new research building recently opened by Philips [3]). Partly visible on the far right is a gas refrigerating machine of the single-cylinder type; beside it is the air fractionating column in which the liquid nitrogen is separated. In the centre is the four-cylinder version, with the large air fractionating column in the tall cabinet on the left. This installation delivers liquid nitrogen with a purity better than 99.5%. Production can be regulated to between 15 and 30 litres per hour. The liquid nitrogen is collected in the round vessels in the foreground.

A special feature of the methods of separation and condensation employed is that the air, and the nitrogen to be separated from it and condensed, is nowhere subjected to compression. Apart from the spaces in the gas refrigerating machine where the low-temperature generating Stirling process takes place, the pressure remains atmospheric in all parts of the equipment.

The large installation can operate continuously until it has produced about 7000 litres of liquid nitrogen. The quantity limit is governed by the process of removing the water vapour and carbon dioxide from the air to be condensed. This is done by a "snow separator" [4]). The snow separator is filled to maximum capacity after the production of some 7000 litres of liquid nitrogen. After defrosting — which takes about 5 hours and for which warm air is used — a running-up period of about 1.5 hours is needed before the installation is ready for use again. The warm air is supplied by auxiliary apparatus (not visible in the photograph) which can be connected to the installation. The crank-handle with the two black knobs, which can be seen on the front of the cabinet containing the large air fractionating column, therefore has three positions: running up, operation, defrosting. The small knob to the left of the handle is used for controlling the flow of exhausted oxygen, so as to give the nitrogen the required purity. The panel immediately above this knob contains two liquid pressure gauges, which give readings of the air intake and the oxygen exhaust.

The hoisting tackle on the gas refrigerating machine serves for lifting the head of the machine for periodic maintenance.

[3]) Philips tech. Rev. **24**, 339, 1962/63 (No. 11/12).

[4]) C. J. M. van der Laan and K. Roozendaal, A snow separator for liquid-air installations, Philips tech. Rev. **23**, 48-54, 1961/62.

SUBJECT INDEX, VOLUMES 16-25

Figures in bold type indicate the volume number, and those in ordinary type the page number. For subjects dealt with in volumes **1-15** the reader should refer to the respective decennial indexes at the end of volumes **10**, **15** and **20**.

AUTHOR INDEX, VOLUMES 16-25

Figures in bold type indicate the volume number, and those in ordinary type the page number. Articles marked with an asterisk * are short communications. For articles published in volumes **1-15** the reader should refer to the respective decennial indexes at the end of volumes **10, 15** and **20.**

Duuren, K. van
* A simple and compact arrangement for measuring the β-activity of weak radioactive samples 20,170
Duyfjes, W.
The formulation of pesticides 19,165
Duyster, H. C.
The Philips pavilion at the 1958 Brussels World Fair, IV. Construction of the pavilion in prestressed concrete 20, 27
Edens, A. H.
A method of sealing the window and cone of television picture-tubes 19,318
Einramhof, F.
Visual inspection of moving lamp filaments on a coiling machine 22,239
— and P. Havas
Automatic control of a filament-coiling machine with the aid of preset counters 21,309
Eland, A. J.
* Etch pits on cadmium-sulphide crystals 22,266
* Etch pits on a zinc-sulphide crystal 24, 61
Elenbaas, W.
Fifty years of the high-pressure mercury-vapour lamp 18,167
— and T. Hehenkamp
A new fluorescent lamp in a starterless circuit 17,288
Emmens, H., see Boer, F. de
Enz, U. and H. Zijlstra
Determining magnetic quantities by displacement measurements 25,207
Ermrich, W., see Klopfer, A.
Es, C. W. van, M. Gevers and F. C. de Ronde
Waveguide equipment for 2 mm microwaves,
I. Components 22,113
II. Measuring set-ups 22,181
Espersen, G. A.
Dispenser cathodes for magnetrons 23,316
— and J. W. Rogers
Electron emission of materials for electron tubes 20,269
—, see Plante, R. A. La
Esveldt, C. J., see Wijn, H. P. J.
Ettema, E. H., see Blink, W. P. van den
Exalto, L. J. H.
Temperatures in fittings for incandescent lamps 21,300
Fast, J. D.
Entropy in science and technology,
I. The concept of entropy 16,258
II. Examples and applications 16,298
III. Examples and applications (continued) 16,321
Foreign atoms in metals 16,341
— and D. J. van Ooijen
Hydrogen in iron and steel,
I. Solution and precipitation 24,221
II. Fracturing 24,252
— and F. L. H. M. Stumpers
Entropy in science and technology,
IV. Entropy and information 18,201
Feddema, J.
The application of the X-ray image intensifier,
V. Medical aspects of the image intensifier 17, 88
Fine, S. and C. F. Hendee
Nuclear X-ray sources 18,229
Forbes, R. J.
Glass throughout the Ages 22,282
Fortuin, G. J.
Visual acuity in connection with television 16,172
Foskett, R. D., see Neppiras, E. A.
Francken, J. C.
The resistance network, a simple and accurate aid to the solution of potential problems 21, 10
—, J. de Gier and W. F. Nienhuis
A pentode gun for television picture tubes 18, 73
—, see Waal, J. van der
Fransen, J. J. B. and H. J. R. Perdijk
Barium getter films 19,290
Geels, B. H.
A private automatic branch exchange using high-speed uniselectors 18, 19
Geenen, M. J. W.
* A machine for bend tests 17,246
Geldermans, P., see Wijn, H. P. J.
Gentner, W.
The CERN 600 MeV synchrocyclotron at Geneva, I. Object and design 22,142
Gerthsen, P., J. A. A. Gilsing and M. van Tol
An automatic dew-point hygrometer using Peltier cooling 21,196
Gevers, M., see Es, C. W. van
Gieles, J. P. M.
A 4000 Mc/s wide-band amplifier using a disc-seal triode 19,145
Applications of microwave triodes 22, 16
— and G. Andrieux
A wide-band triode amplifier with an output of 10 W at 4000 Mc/s 21, 41
Gier, J. de, A. C. Kleisma and J. Peper
Secondary emission from the screen of a picture tube 16, 26
—, see Francken, J. C.
Gier, N. A. de, W. van Gool and J. G. van Santen
Photo-resistors made of compressed and sintered cadmium sulphide 20,277
Gilsing, J. A. A., see Gerthsen, P.
Ginkel, H. van
Flying-spot scanners for colour television 21,234
Glässer, W., H. G. Grimmeiss and H. Scholz
* Luminescent P-N junctions in gallium phosphide 25, 20
Goddard, N. E. and H. A. Dell
* A radar sonde system for upper-air measurements 19,258
Goedkoop, J. A.
Neutron diffraction 23, 69
Gool, W. van, see Gier, N. A. de
Goorissen, J.
Segregation and distribution of impurities in the preparation of germanium and silicon 21,185
Gorter, E. W., see Wijn, H. P. J.
Gradstein, S.
* Radiation furnaces of past centuries 23,182
* The calculating machine of Blaise Pascal 24,102
Greefkes, J. A., see Jager, F. de
Grimmeiss, H. G. and H. Koelmans
* P-N luminescence in gallium phosphide 22,360
—, see Glässer, W.
Groendijk, H.
* A noise diode for ultra-high frequencies 20,108
—, see Beer, A. J. F. de
Gude, H. te and E. Schaaff
* A new kind of tuning indicator tube 18,243
Haaf, F. E. L. ten, G. Klein and F. J. Schijff
Monitoring, control and safety equipment for a nuclear reactor of the swimming-pool type,
II. Further description of certain component units 19,273
Haan, E. F. de
* The "Plumbicon", a new television camera tube 24, 57
—, A. van der Drift and P. P. M. Schampers
The "Plumbicon", a new television camera tube 25,133
Haanstra, H. B.
A simple ultramicrotome 17,178
—, J. J. de Jong and J. M. G. Smeets
* The sub-microscopic structure of "Ticonal" G magnet steel 19, 11
—, see Knippenberg, W. F.
Hagedoorn, H. L., see Verster, N. F.
Haisma, J., S. J. van Hoppe, H. de Lang and J. van der Wal
* A small, stable gas laser 24, 95
Haitjema, M. T., W. Kopinga and S. J. Porte
Performance tests on loudspeakers 21,362
Halberstadt, J.
Experiments with radioactive preparations of the acaricide "Tedion V 18" 21,276
—, see Aten, A. H. W.
Halbertsma, N. A.
The birth of a lamp factory in 1891 23,222
Halbmeyer, A. J., see Beek, R. van
Hamacher, E. A. and K. Lowitzsch
The "Norelco" counting-rate computer 17,249
—, see Parrish, W.
Hamaker, H. C.
Applied statistics, an important phase in the development of experimental science 22,105
Hammar, M., A. Rydahl and B. Westerlund
Electronics in the wood industry 23, 29